

# Subjective and Objective Agency Performance: A Multilevel—Multistage Approach

René Torenvlied<sup>1</sup>

Utrecht University and University of Groningen

Agnes Akkerman

Radboud University and Utrecht University

**Abstract.** Current studies on public sector performance report that indicators of subjective and objective performance are correlated, but not strongly. Although those studies conclude that different indicators measure different aspects of agency performance, no theoretically informed predictions are made about the relative strength of correlations between indicators of public performance. The present paper develops and tests a model of agency performance using a multilevel-multistage approach. The paper distinguishes between five stages (input-, throughput-, output-, and outcome-performance) and four levels (network, agency, time, and client). Special attention is paid to the problem of ‘ecological fallacy’ that lingers behind associations between performance indicators at different levels. From the model testable hypotheses are derived about the effects of different performance indicators on subjective and objective outcome performance. Hypotheses are tested on a large dataset of five educational programs offered by 91 colleges of universities of applied science in The Netherlands between 2002-2005 (n = 305 cohorts). Outcome performance is measured on the basis of data from 5,418 graduates. Multilevel regression analyses show that the multilevel model more than the multistage model is a promising framework for studying the complex correlates of performance in public sectors.

**Key-words:** Public Performance, Performance-Indicators, Multilevel Analysis, Ecological Fallacy

## Introduction

Organizational performance is a multifaceted concept, especially in the public sector, where organizational goals are multidimensional (Boyne, 2003; Provan and Milward, 2001). Public agencies often must intervene in complex situations of individual clients, sometimes with an uncertain political mandate. Whereas profit maximization and shareholder values are relatively straightforward goals of private business, public agencies are often confronted with contradictory demands from heterogeneous interest groups in their environment. Consequently, public sector performance indicators are often equivocal and subject to multiple interpretations, so that all measures of performance are subjective in some sense (Brewer, 2006).

Performance indicators need to meet two criteria to be ‘objective’: (a) these indicators must be an accurate measurement of a dimension of performance, and (b) they must involve some external assessment procedure (Andrews, Boyne and Walker, 2006a: 16).

---

<sup>1</sup> Address all correspondence to: [r.torenvlied@uu.nl](mailto:r.torenvlied@uu.nl). Paper prepared for presentation at the 10<sup>th</sup> Public Management Research Conference (PMRC), Columbus, Ohio, October 1-3, 2009. We thank Jelmer Schalk for his work on data collection. We thank Jim Allen and ROA Maastricht for kindly allowing us to use the data on graduates. The research presented is part of a larger research project of the authors into inter-organizational networks and public performance. Torenvlied acknowledges financial support from the Netherlands Organization for Scientific Research (NWO), Vidi-grant 452-06-001 and from the High Potential program of Utrecht University.

Examples of objective performance data are validated assessments by accountants, or validated test results, for example about school achievements or patient recovery. By contrast, subjective performance measures are often perception data, potentially biased, and certainly not collected using an external procedure. Examples of subjective performance data are perceptions of managers, personnel, or consumers / clients about the quality of services delivered.

The jury is not out about the relative pro's and con's of choosing between subjective and objective performance data. Objective measures of performance are often used to monitor performance and figure predominantly in research focused on organizational effectiveness. However, there is growing consensus that hard indicators by themselves are insufficient for evaluating organizational performance and need to be supplemented by 'soft' indicators, such as perceived program quality (Andrews, Boyne, and Walker, 2006b; Bouckaert and Van de Walle, 2003). But, subjective measures of organizational performance are sometimes criticized because clients may be ill-informed about policies (Brown and Coulter, 1983; Golden, 1992; Kelly and Swindell, 2002).

If the theoretical differences between subjective and objective indicators are substantial, do these indicators also differ much empirically? For the private sector, there is a clear negative answer. Studies on private sector performance reveal that objective and subjective measures of performance all are strongly and positively correlated when applied to comparable constructs (Bommer et al., 1995; Delaney and Huselid, 1996; Dess and Robinson, 1984; Dollinger and Golden, 1992). For the public sector, results are quite different. Andrews et al., (2006a) compared many different indicators of service provision by Welsh local authorities on eight domains (ranging from education to waste and highways) with self-reported evaluations of the managers in charge of these services. They conclude that different indicators measure different aspects of agency performance (Andrews et al., 2006a). The same results are reported by Dücker (2009: 82) who compared the success of a complex health-care transition project as perceived by managers of 24 Dutch hospitals involved, with hard (patient) data on effectiveness.

If we aim to further explore the multifaceted nature of public performance, we should move beyond the observation that different performance indicators measure different aspects of performance. In fact, we *do* expect that indicators for different aspects of performance are associated. The reason is that different agency routines and activities—each for which performance indicators exist—are linked and part of the agency system of operation. The present paper develops and applies a heuristic model for understanding variation in the strength of associations between indicators for public agency performance. The heuristic model builds on a system-theoretical approach to agency operation, which assumes a sequential relation between activities. We argue that performance is a multilevel—multistage phenomenon.

The present paper focuses on subjective and objective indicators for *outcome* performance, such as client satisfaction or effectiveness. From the model, we derive testable predictions about the effects of different types of performance-indicators on outcome performance. Thus, the aim of the paper is not to explain agency performance, but to test hypotheses about associations between performance indicators. The paper studies three related research questions: (1) at what levels of analysis do we observe empirical variation in different performance indicators? (2) What is the association between objective and subjective measures of *outcome* performance at different levels?

(3) What is the association between other performance indicators and the objective and subjective outcome performance? We focus on the performance of colleges in the Dutch system of universities of applied sciences.

The hypotheses are tested on a large dataset of five educational programs offered by 91 colleges of universities of applied science in The Netherlands between 2002-2005 (n = 305 cohorts). One subjective indicator (graduates' satisfaction with the program offered) and one objective indicator (graduates' hourly wage) for 'outcome-performance' are constructed on the basis of data from 5,418 graduates. We apply multilevel analysis to control for the interdependence of cases in the nested design, and to take into account the complex multilevel structure of performance indicators.

## **2. Performance at levels and in stages**

### *2.1 Multilevel nature of agency performance*

Most performance-indicators have a multilevel nature. Figure 1 distinguishes between four levels of agency performance: (1) the individual level; (2) time, (3) the agency; (4) the network. Whether a client is satisfied or dissatisfied with a service delivered is an indicator for agency performance at the individual level. The percentage of clients satisfied with a service delivered is an indicator for agency performance at the agency level. The higher the percentage of clients satisfied, the better the agency performs. The percentage of clients satisfied in a particular year is an indicator for agency performance at the time-level. It may be that client satisfaction shows large fluctuations in time. Finally, if all agencies with high percentages of clients satisfied can be found in one network, client satisfaction is clearly also a network property.

-----  
Insert Figure 1 about here  
-----

Many current studies on the association between subjective and objective indicators for public performance do not take into account the multilevel nature of the performance indicators under study—at least not in one integrated study design. Using bivariate correlations is still a standard (Dess and Robinson, 1984; Dawes, 1999; Kelly and Swindell, 2002). Andrews, Boyne and Walker (2006a: 28) break down bivariate correlations of performance indicators for different years to deal with the multimoment nature of the data.

Multilevel designs have at least two important advantages above bivariate correlations or standard OLS-regression designs for the analysis of subjective and objective performance indicators. In the first place, multilevel designs allow to control client satisfaction or client behavior for various variables at the network-level, agency-level, in time, as well as for characteristics of individual clients (Heinrich and Fournier, 2004; Schalk, Torenvlied and Allen, Forthcoming). Heinrich and Lynn (1999: 133) clearly illustrate the advantages of multilevel research designs for the study of governance. They report that ignoring variation at the individual level, and cross-level interactions between higher-level variables and individual-level variables leads to inaccurate estimates of

effects on performance outcomes. In addition, multilevel analysis has the advantage that it provides information about the level of analysis at which variation is explained: does client satisfaction vary across clients more than between agencies, or does it primarily vary in time? For example, we can analyze how much variance between subjective and objective performance is explained away by agency characteristics, by client characteristics, and by variation in time. To find out, some authors performed separate one-way analyses of variance (Dess and Robinson, 1984).

The second reason for integrating multiple levels of performance into one design is even more important. In multilevel studies, it is a well-known phenomenon that aggregate variables could be associated in quite different ways at different levels of analysis. For example, the satisfaction of patients with their hospital treatment (subjective performance indicator) is likely to be negatively associated with the occurrence of complications (objective performance indicator), controlling for their type of illness and other mediating characteristics. Patients with less or no complications will be more satisfied. We can also study differences between hospitals in their performance on both indicators. The aggregation process changes the substantive meaning of the performance indicator. The percentage of patients satisfied of a hospital is affected by mechanisms different from those that affect patient satisfaction. Hence, it might very well be the case that mean level of satisfaction of patients is not related to the percentage of patients with complications after treatment.

The composition of agency performance indicators from client-level data has profound implications for conclusions drawn on the basis of the performance indicators. Because the substantive meaning of the performance indicator is transformed when aggregating, we cannot make inferences about client opinions and behavior on the basis of agency-level performance indicators (which is called the ‘ecological fallacy’). Only recently a few public management and performance studies pay explicit attention to this phenomenon (Heinrich and Lynn, 1999; Choi, 2008: f.n. 5; Kim, Solomon and Zurlo, 2009: 265). Suppose that in the example of hospitals above, there is no association between the satisfaction of patients of hospitals and their length of stay. A fallacy statement would be the following: patients consider other aspects of hospital treatment to be much more important for their satisfaction. By contrast, an explanation for the absent association between both indicators should be found in characteristics of the hospitals. Any policy advice that follows from (benchmarking) statements about agency performance should be extremely careful in avoiding such statements. A multilevel approach to performance indicators is the only appropriate design for such an analysis—integrating associations at the level of the agency with associations at the level of the client.

## *2.2 A multistage model of agency performance*

Agencies transform the individual demands of clients into outcomes that provide a solution to these demands. For example, hospitals respond to clients with various illnesses and diseases, and aim to cure them by offering a proper treatment. Schools respond to the demands of pupils with various backgrounds and capabilities, and aim to educate them by offering a proper educational program. Social welfare agencies respond to the needs of eligible citizens, and aim to help them getting involved in society again by

offering appropriate support programs. These organizations are in a specific equilibrium (Simon, 1945: 159).

The use of performance indicators for different stages between organizational inputs and outcomes has become quite popular, but has not been studied in a systematic way (Boyne and Law, 1991; Sorber, 1993; Duckett and Swerissen, 1996; Hedley, 1998; Stone and Cutcher-Hershenfeld, 2001). Figure 1 presents a schematic multistage model of agency performance, which connects crucial ‘aspects’ of agency activities from a system-theoretical perspective.<sup>2</sup> Inputs (clients’ demands and needs) are transformed into outputs (agency activities). Throughput activities manage the transformation process from inputs to outputs. Finally, agency outputs are transformed into in outcomes (demands and needs satisfied, problems solved) because agency activities are expected have positive effects. In order to reduce complexity, the stages are assumed to be related in a sequential way: inputs are (by throughput) transformed into outputs, which is transformed into outcomes.

For each of these stages, performance indicators can be singled out. Input-performance can be defined in terms of popularity: the demand for services. For hospitals, an input-performance indicator is the number of referrals by general practitioners. For universities, input-performance is the enrollment of students, or freshmen. Throughput-performance can be defined in terms of efficiency and competence. For hospitals, throughput indicators are waiting lists, nurse-to-bed ratios, average costs per patient, or the level of education of staff. For universities, throughput performance can be defined in terms of student-staff ratios, personnel costs, or solvency. Output-performance can be defined in terms of productivity and quality of services provided. For hospitals, it is the number of patients treated, or the number of patients with complications after treatment. For universities it is the drop-out rate of freshmen, or the diploma-rate. Finally, outcome-performance indicators can be defined in terms of client satisfaction or a more objective measure of value added to clients’ needs and demands.

If the different stages are linked sequentially—and we assume some causal ordering from left to right—the performance indicators must be interrelated to some extent, but not fully. Associations are not perfect because: (a) the stages reflect different aspects of agency performance, and (b) systematic variation will enter the model when moving from inputs to outcomes. First, the multistage model provides a heuristic tool to understand different aspects of performance in their mutual relation. For example, process management—as reflected by throughput performance—reflects a different aspect of agency performance than does agency activity. However, testable hypotheses can be derived from the model about how process-management contributes to output-performance when controlling for input performance. Second, systematic variation from the agency’s environment (for example demographic characteristics) affects its input performance. Changes in the environment will force an agency to adopt more flexible routines and activities (as reflected by its throughput performance) to deliver services in a more variable way (output performance), which may ensure client satisfaction (outcome-performance). Especially the transformation of outputs into outcomes is especially prone

---

<sup>2</sup> In educational research, the use of multiple indicators for school performance is advised (Rumberger and Palardy, 2005) because: (a) some schools may perform better on one type of outcome than another, because specific resources are related to specific outcomes and (b) school performance on one indicator may conflict with performance on another indicator. Meier and O’Toole (2001; 2003) use different objective indicators for the performance of school districts to test effects of network management on performance.

to be co-determined by individual characteristics of clients and their environment, which affects the association between indicators for output-performance and outcome-performance.

### 2.3 Hypotheses

Using a multistage—multilevel approach to performance, we can derive testable hypotheses about: (a) the levels at which we expect that performance indicators vary, (b) the association between subjective and objective performance at the client and agency-levels, and (c) about the associations of input-, throughput-, and output-performance indicators with subjective and objective outcome performance. At which levels do we expect the different performance indicators to vary? We specified four levels of performance: performance can vary across *networks*, between *agencies*, in *time*, and between *clients*. Performance indicators are tied to levels through specific assumptions about the nature of the sequential activities of agencies.

We could safely assume that (networks of) agencies are operating with a relatively stable demand for their services. Notwithstanding that there will always be some variation over time in the demand for services, we expect that these demands primarily vary between types of services (networks of agencies) and between agencies within one network, much more than between moments in time.

*Hypothesis 1a.* Input performance of agencies varies primarily between networks and agencies.

Throughput performance is the extent to which an agency is able to efficiently and competently manage its internal routines and activities. Throughput performance indicators are expected to be stable over time, because changing parameters in agency's internal costs structure or personnel structure is quite difficult. This leads to the following hypothesis:

*Hypothesis 1b.* Throughput performance of agencies is stable over time.

We assumed that the demands for agency services are relatively stable over time. However, with the operational processes of agencies that transform demands (input) into activities (output), more disturbing factors from the agency's environment will come into play. For example, budget cut-downs or a downfall of political support will affect agency activities and output in the short-run, and only later the management of routines and operations is affected. The impact of such environmental factors will vary over time, and thus output performance will start to show variation at this level, where it was not a factor for input-performance or throughput performance. We could label this phenomenon the principle of entropy in multistage performance.

*Hypothesis 1c.* Output performance of agencies varies between networks, agencies, and time.

Outcome performance can be located at least at two levels: the agency level and the client-level. Outcome performance is affected by environmental factors even more than output performance, because it pertains to the transformation of agency outputs into outcomes. This transformation is even more determined by mechanisms at the level of

clients, mostly outside the realms of agency competences. Hence, we expect that such environmental factors at the client level turn performance even more time-dependent at the agency-level, and more dependent on individual client characteristics at the client-level. Thus, we would expect that the principle of entropy in multistage performance is reinforced when moving from outputs to outcomes. This results in the following two hypotheses:

*Hypothesis 1d.* At the time-level the variation in outcome performance is larger than the variation of output performance.

*Hypothesis 1e.* Outcome performance at the client level varies primarily between clients.

Do we expect differences in the association between subjective and objective performance if we define these indicators at the level of the agency and the level of the individual client? We apply the general thrust of multilevel analysis that variables can display different associations at different levels to outcome performance, and specify the following hypothesis.

*Hypothesis 2.* Subjective and objective indicators for outcome performance are associated differently at the agency level and at the client level.

We assume that the stages in the system of agency operation are linked in a sequential relation, which implies a causal order from left to right. Hence, we may that performance indicators for subsequent stages are more strongly related than performance indicators for stages more distant in the process. Because we focus on outcome-performance, we specify hypotheses for subjective and objective outcome performance. We empirically explore what are the strengths of joint effects of the indicators for the other stages of agency performance on outcome performance.

*Hypothesis 3a.* Indicators for outcome performance are more strongly associated with indicators for input performance than with indicators for throughput performance than with indicators for output performance.

Finally, because subjective measures are assumed to be more biased than objective measures, we can test whether objective outcome performance is more strongly related to performance indicators for the other stages than is subjective performance.

*Hypothesis 3b.* Indicators of agency performance are more strongly related to objective indicators of outcome performance than to subjective indicators of outcome performance.

### **3. Research design and data**

The subjects of the study in the present paper are five different networks of colleges within the Dutch system of universities of applied science (HBOs). This system offers more applied studies compared to the research-oriented system of Dutch universities. The most important formal institution for all colleges is the ‘Netherlands Association of Universities of Applied Sciences’ or *HBO-raad*. The *HBO-raad* has one central board of directors and two different advisory boards, a general advisory board for all universities

of applied sciences and a ‘national program committee’ for each program.<sup>3</sup> The *HBO-raad* is also responsible for special committees, which monitor the implementation of government regulations in the training programs and in which some of the colleges are involved.

The colleges in the studied programs mostly prepare students for a bachelor’s degree, although colleges have recently begun to develop some masters programs. Colleges offer specific, four-year bachelor programs to train students for various professions. Funding is based primarily on total student enrollment but also includes a ‘dynamic demand factor’ that incorporates performance measures, such as dropout rate in the previous year and enrollment in the present year (Kaiser, Vossensteyn, and Koelman, 2001). Thus, these colleges have to compete for both students and resources. However, the colleges also have common interests, such as to effectively lobby, exchange information, and develop joint programs. Colleges are embedded in an inter-college (program) network, but also in local networks, which include the local authority and local regulatory agencies.

From the total of all programs offered by Dutch universities of applied science, five programs were selected for the present study. Two selection criteria were applied. The first criterion was a relatively large program size and enrollment to obtain enough statistical power. For each program, the participating colleges were selected using a prefixed boundary definition, such that only colleges that were members of the ‘national program committee’ of the *HBO-raad* were included. This selection implies that only government-funded colleges are included, which make up the overwhelming majority of all colleges in the Dutch system of education. This translates into 91 colleges with  $n = 18,159$  graduates in the sample. The second criterion aimed to create variation in the level and complexity of professional standards between the programs to which college graduates must comply in their daily work. This variation is associated with the level of (quasi)governmental monitoring and control of the programs. We expect that variation in the level of monitoring and control through professional standards creates variation in the colleges’ performance at the level of program networks. At the same time, our selection of five programs within the same Dutch system of universities of applied science aimed to hold many sector-specific characteristics constant.

The first program selected is the *Primary Education Teacher* program (*PABO*), which involves 28 colleges. These are the largest colleges within the system of universities of applied sciences in the Netherlands (in 2006, total enrollment amounted to 35,000 students). The professional standards for *PABO* graduates are well defined but were not broadly applied. For example, the reputation of the whole sector was recently damaged when a heated political debate and media attention was focused on the poor math and language skills of *PABO*-students. The second program selected is the *Social Work* program (*SPH*), which is a very popular program, with 19 colleges offering many different specializations. For social workers, professional standards are defined but generally have a broader interpretation than the standards for teachers. The third program selected is the program in *Commercial Economics* (*CE*). This program, offered by 15 colleges, educates students for jobs in commerce, (international) business, and marketing. The professional standards are not well defined for this general program. The same

---

<sup>3</sup> Participation of colleges in the general advisory board comprises *direct participation* by directors of ‘monosectoral’ universities of applied science (offering one single program) and *representation* of colleges within ‘multisectoral’ universities of applied science by a university board member.

applies to the fourth program selected, which is *Management, Economics, and Law (MER)*. *MER* is a popular, but very broad, program with subfields of management, organization studies, economics, and law. In total, 13 colleges educate students to become policy-makers, staff advisors, and managers for general functions in businesses. The fifth program selected is the *Nursing Studies* program (*HBO-V*), which is offered by 16 colleges and which trains nurses for various functions in hospitals, with general practitioners, and in the broader health care sector. The graduates of this program must comply with strict, prescribed professional standards to become licensed graduates.

The selection procedure resulted in a nested design, with enough statistical power to test the hypotheses. Ninety-one colleges were nested within five inter-organizational college networks. Data on various measures of college performance were included in the dataset for in four cohorts (2002-2005). We selected two measures of outcome-performance from the sample of graduates to the dataset: their hourly wage as an objective indicator for performance (available  $n = 5,418$  graduates), and their satisfaction with the program offered as a subjective indicator for performance (available  $n = 7,130$  graduates). Below, we further describe the data collection and discuss the main performance indicators and control variables.

### *3.1 Indicators for outcome performance*

The data source for the subjective and objective outcome performance-indicators of is the ‘HBO-monitor’, coordinated by the Research Centre for Education and the Labour Market (ROA). This monitor is a yearly survey among a large sample of all graduates of universities of applied sciences in the Netherlands. Over 85 percent of all colleges participate in the survey. Data are collected between one year and one-and-a-half years after graduation, and the average response rates are about 40-45 percent. We selected the sub-set of all graduates in the sample who graduated at one of the colleges included in the dataset between 2002 and 2005. The HBO-monitor contains several questions about a graduate’s evaluation of PABO performance, their labor market position, as well as individual characteristics. The years 2002-2005 reflect evaluations by separate groups of individual graduates, who are also nested in the colleges. Changes over time thus refer to differences between cohorts, not changes within individuals.

*Graduates’ hourly wage.* We included the hourly wage of a graduate as an objective indicator for college performance. Colleges with graduates who earn relatively more, are assumed to perform better than colleges with graduates who earn small wages. The HBO monitor includes questions about gross monthly income and working time (in full time equivalent). On the basis of these two items, the hourly wage of each graduate was constructed. Obviously, the multilevel research design includes a level specified for programs in order to control for differences between programs in hourly wage. Seven outliers (wages  $> € 100,00$  per hour) were excluded from the analysis. Data were available for 5,418 graduates.

*Graduates’ satisfaction with the program offered.* We included graduates’ satisfaction with the program offered as a subjective indicator for college performance. Although subjective measures of organizational performance sometimes are criticized because

clients may be ill informed about policies (Brown and Coulter, 1983; Kelly and Swindell, 2002), we expect that college graduates possess accurate and detailed knowledge of their past study program. Graduate satisfaction is measured using an item that confronted graduates with the question whether they would choose the same program *at the same institution* again. If not, the graduate could indicate whether: (1) she would choose the same program at a different institution, (2) she would choose a different program altogether, or (3) she would choose not to study again at all. We collapsed the last three categories, creating a dummy variable for graduate satisfaction with the PABO-college.

There may be important differences between the three categories of graduates who would not do the same study program at the same institution again. For the present study, we assume that a graduate who would opt to do the same study program at a different institution, and another graduate who would opt to do a different study program both base their choice upon the basis of the same criterion: their satisfaction with the study program at their institution. Students' assessments of various dimension of college quality, such as practical relevance and coherence of the program and quality of teachers, differs far more strongly between the first and the latter three categories than between each pair of the three contrast categories. Nonetheless, it is possible that the choice for another study program altogether could be based upon a mismatch between the student's preferences and the specific content of the study program, regardless of the institution offering the program.

The variable graduate satisfaction could capture many things, varying from satisfaction with teachers to the evaluation of facilities, or traineeships. However, the measure is a simple and attractive indicator for the evaluation of PABO-performance by graduates. Furthermore, this dummy variable has been used in other studies on program performance as well (Allen and Ramaekers 1999), and it is a core measure used in government study program evaluations and college benchmarking.

### 3.2 Other performance indicators and controls

The data source for the other performance-indicators at the level of the college is the publicly accessible management information system of the universities of applied sciences.<sup>4</sup>

*Input performance.* We use two measures for input performance: student enrollment and number of freshmen. Student enrolment is defined as the number of students that are enrolled in the program for a particular year. Number of freshmen is the number of first-year students who enter the program at a particular year. Obviously, enrolment and number of students are highly correlated and indicate the popularity (or demand) for education at the particular college in a particular year.

*Throughput performance.* We use three measures for throughput performance, as indicators for the competence and efficiency of a college in a particular year. Personnel costs is the total costs for one full time equivalent per student. Student-staff ratio is defined as total student enrollment divided by the total full time equivalent for staff in the year of graduation. Solvency of the college is the ability of the college to pay its debts, as established by the yearly accountant controls and published by the HBO-Council.

---

<sup>4</sup> See: [http://www.hbo-raad.nl/hbo-raad/feiten-en-cijfers/cat\\_view/60-feiten-en-cijfers/63-onderwijs](http://www.hbo-raad.nl/hbo-raad/feiten-en-cijfers/cat_view/60-feiten-en-cijfers/63-onderwijs).

*Output performance.* We use two indicators for output-performance. Diploma rate is defined as the number of graduates in a given year divided by the mean of yearly total enrolment in the PABO over the period 1996-2005. We take the number of graduates relative to the mean enrolment for a long period to rule out the short-term fluctuations in enrolment. The reason is that freshmen enrollment affects diploma rates. When freshmen enrollment increases, diploma rates go down, and college size goes up. High diploma rates are indicators for good performance, because colleges are capable of supporting their students to finish their program in time. Dropout rate is defined as the number of freshmen dropouts as a proportion of the total freshmen student enrolment for a given year. Low dropout rates are indicators of good performance, because colleges are assumed to: (a) raise realistic expectations among potential students about the program that match the actual program, avoiding situations in which students make the wrong choice by enrolling in the program and (b) offer a high-quality program, which motivates the students.<sup>5</sup> Dropout rates are commonly used as an important indicator of school performance (Bryk & Thum, 1989; Lee & Burkam, 2003; McNeal, 1997; Rumberger & Palardy, 2005: 4).

*Individual control variables.* We controlled for individual graduates' gender, ethnicity, and grade point average.

## 4. Results

### 4.1 Variation in performance-indicators at different levels

The first hypotheses make predictions about the levels at which we expect variation in the different performance indicators. For each performance indicator, we estimate the so-called 'intra-class' correlation coefficient  $\rho$ . This coefficient breaks down the total variation in a variable to the different levels specified (the 'empty' model in a multilevel analysis). A large value of the intra-class correlation coefficient implies that most of the variation in performance is observed at that level. Table 1 presents the intra-class correlation coefficients of the college performance indicators for the levels specified in our heuristic model: the network-level (five programs), the agency-level (91 colleges), the time-level (305 cohorts) and for outcome performance the client-level (5,418 / 7,130 graduates).<sup>6</sup> In addition, the table provides information about the significance of the estimates.<sup>7</sup>

-----  
 Insert Table 1 about here  
 -----

<sup>5</sup> Some colleges may apply binding recommendations regarding the continuation of studies in the first year. This could be an alternative explanation for variations in dropout rates.

<sup>6</sup> Restricted maximum likelihood estimation produces more robust estimates when the number of cases at a level is relatively small (Snijders and Bosker 1999). The (restricted) maximum likelihood estimations for variables 'solvency' and 'student-staff ratio' did not converge and are not reported.

<sup>7</sup> Significance of variance estimates is computed using the Wald-test against the standard normal distribution. Because standard-errors are asymptotic, they are valid only for a relatively large  $n$  at the various levels, where large implies that  $n > 100$ . However, when testing fixed effects only in a multilevel analysis, variance estimates appear to be robust for a much smaller  $n$  (Hox 2002: 42).

Table 1 shows that the two indicators for *input*-performance, number of students and number of freshmen, mainly vary between agencies, and partly between programs. There is only little variation between cohorts, which confirms hypothesis 1a: demands for agency services are relatively stable in time. The indicator for *throughput*-performance, personnel costs, does not at all vary between cohorts. Personnel costs vary between agencies and between programs, both sharing 50 percent of total variation. This supports hypothesis 1b.

*Output*-performance was predicted to vary between programs, colleges, and cohorts (hypothesis 1c). Table 1 reveals that, as predicted, the drop-out rate and diploma rate vary at all different levels. Surprisingly, most variation is now observed between cohorts. Neither input-performance, nor throughput performance show much variation between cohorts, but output-performance suddenly varies most between cohorts. Hence, at the cohort-level, mechanisms must be present in the colleges' activities or environment, drive drop-out rates and diploma-rates to heavily fluctuate in time.

Table 1 also shows how much variation exists in subjective and objective *outcome*-performance indicators at different levels. At the *college-level*, the mean percentage of graduates satisfied and the mean hourly wage of a college's graduates show exactly the same pattern in variation across levels. This is a remarkable observation, because mean hourly wage is an objective performance indicator, while mean percentage of graduates satisfied is a subjective performance indicator. As it is the case with output-performance (drop-out rates and diploma-rates), most variation is observed between cohorts, and less variation is observed between colleges. There still is significant, but quite small variation in mean hourly wage and mean graduate satisfaction between programs. Hypothesis 1d predicted that this variation in time is larger than the variation in time of output-indicators. Table 1 shows that this is the case for drop-out rates, but not for diploma-rates. Hence, hypothesis 1d is partly rejected.

At the *graduate level*, a graduate's satisfaction with the program offered and the hourly wage of a graduate, also show highly comparable patterns of variation across levels.<sup>8</sup> Hypothesis 1e, which predicted that for outcome performance at this level almost all variation can be attributed to the individual graduate level, is clearly corroborated. At the higher levels we observe only one to five percent of variation in graduate satisfaction or hourly wage.

#### 4.2 Associations between subjective / objective outcome-performance at different levels

The second hypothesis predicts that the associations between subjective and objective outcome-performance at the individual client-level (graduates) and at the agency level (colleges) are different. To test this hypothesis, we performed a multilevel logistic analysis with individual graduate satisfaction as a discrete dependent variable, and two independent variables: (1) the hourly wage at the graduate-level, and (2) mean hourly wage at the college-level. This analysis effectively provides insight in the joint effects of graduates' hourly wage and colleges' mean hourly wage on graduates' probability to be

---

<sup>8</sup> Graduates' satisfaction is a binary dependent variable, whereas graduates' hourly wage is a continuous dependent variable. For the empty model, intraclass correlations can be computed for binary dependent variables as for continuous dependent variables (Snijders and Bosker, 1999: 209). In logistic multilevel models, the variance of the lowest level (the graduate level) is always fixed at  $\pi^2/3$ .

satisfied with the program offered.<sup>9</sup> Table 2 shows the results of the analysis. First, we estimated the empty model for graduate satisfaction, as a baseline against which to test the associations.<sup>10</sup> The empty model replicates the results presented in the row for ‘graduate satisfaction’ in table 1, but reports the variance estimates and not the fractions of total variation (which are the intra-class correlation coefficients in table 1).

-----  
Insert Table 2 about here  
-----

Model 2 estimates the effect of individual graduates’ hourly wage on their probability to be satisfied with the program offered. Table 2 shows that this effect is highly significant and positive ( $\beta = 0.049$ ,  $p > .001$ ). Hence, at the individual level of the graduates, the objective and subjective indicators for performance are positively associated. We observe that the variation at the higher levels (cohorts, colleges, programs) barely changes compared to the empty model, which is due to the fact that all covariance between graduates’ hourly wage and their satisfaction can be found at the individual level. The model improves significantly compared with model 1 ( $LR = 20.48$ ,  $df = 1$ ,  $p < .001$ ).

In Model 3 we add the effect of colleges’ mean graduates’ satisfaction between 2002 and 2005 to the multilevel analysis. Surprisingly, mean graduates’ satisfaction now has a strong *negative* and highly significant effect on graduates’ satisfaction ( $\beta = -0.206$ ,  $p > .001$ ).<sup>11</sup> This negative effect implies that colleges delivering graduates with higher wages, tend to have less satisfied graduates. (More precisely stated: the mean hourly wage of the graduates of a college reduces the probability that its graduates are satisfied about the program offered—even though individual hourly wage positively affects the probability of being satisfied.) If we rely on the two college-level outcome-performance indicators, we conclude that hourly wage and graduate satisfaction are negatively correlated for  $n = 91$  colleges. If we rely on the two graduate-level outcome-performance indicators, we conclude that hourly wage and graduate satisfaction are positively correlated for  $n = 5,418$  graduates. This is a classical, textbook example of paradoxes that may occur when aggregating data—indeed, hypothesis 2 is firmly corroborated. We find opposite associations when we analyze subjective and objective indicators for performance at different levels.

-----  
Insert Figure 2 about here  
-----

To obtain a better understanding of the phenomenon under study, we constructed two bivariate plots of hourly wage against the probability of being satisfied. In figure 2, we plotted the 91 colleges’ mean wage of graduates against the colleges’ probability that graduates are satisfied (an estimate for the percentage students satisfied). The figure clearly shows that both measures are associated, with the linear predictor having a negative slope. For three, relatively sizeable colleges in Figure 1, we constructed three

---

<sup>9</sup> An analysis with hourly wage as a dependent variable makes little sense because we assume that objective performance affects subjective performance and that the reverse does not hold. It is highly unlikely that a graduate’s satisfaction with the program offered has an effect on his wage.

<sup>10</sup> In the multilevel logistic analysis we use the ‘Laplace’ method for maximum likelihood estimation. This method allows us to compare the deviances as indicators for model improvement in a logistic model specification (Snijders & Bosker, 1999).

<sup>11</sup> Although model 3 does not improve significantly compared with model 2 ( $LR = 3.14$ ,  $df = 1$ ,  $p < .10$ ).

separate bivariate plots of graduates hourly wage against their probability of being satisfied. These plots are presented in Figure 3 for four cohorts. The first plot is for a college with a low mean hourly wage, and we clearly observe a positive association. Note that the mean hourly wage of all subjects is indeed relatively low, and that their mean probability of being satisfied is relatively high. For the two colleges with an intermediate mean hourly wage and a high mean hourly wage of graduates, the linear association is much stronger. But note that mean satisfaction goes down, as mean hourly wage goes up. The 91 positive trend-lines within colleges move from up-right to down-left in the wage—satisfaction space.

#### *4.3 Associations between multistage performance and outcome-performance*

Hypotheses 3a and 3b make predictions about the associations between input-performance, throughput-performance, output performance, and subjective/objective outcome performance. Table 3 provides the results of bivariate associations between all these indicators for performance. The observed associations for the colleges nicely replicate the results reported by Boyne et al. (2006a) for Welsh municipal services: associations are small but significant.

-----  
Insert Table 3 about here  
-----

We specified a number of multilevel models with graduate satisfaction and hourly wage as dependent variables, and the other performance indicators as independent variables to test hypotheses 3a and 3b. Thus, these model show how much outcome performance in terms of a subjective (graduate satisfaction) and an objective (hourly wage) indicator is associated with a combination of the other performance-indicators. In addition, a multilevel analysis allows us to observe at which levels most of the variance is explained. Note that the aim of these analyses is not to provide substantive explanations of performance, but to test how strongly the performance indicators for each stage affect outcome performance.

To test hypothesis 3a, with regard to the relative strength of the joint effects of performance indicators for different stages on outcome performance, we estimated three models for graduates' satisfaction and three models for graduates' hourly wage. The first model is the empty model (which provides information about the intra-class correlation coefficient). The second model includes the performance indicators for output-performance (drop-out rate and diploma-rate), for throughput-performance (student-staff ratio and solvency), and for input-performance (number of freshmen). As a control we included mean hourly wage in the model for graduates' satisfaction, as we did in table 2. The third model controls these effects for various variables at the individual level of the graduates. Results are presented in table 4.

-----  
Insert Table 4 about here  
-----

With respect to graduates' satisfaction, only the throughput-performance indicator 'solvency' positively and significantly affects graduates' satisfaction, and remains to do so after controlling for individual characteristics of graduates. For the other performance

indicators, we find no effects. Inspection of the random parts of the models, the variance estimates, shows that at the program level most variance is explained (a 50 percent reduction), but only after controlling for individual characteristics of graduates. This reduction in (unexplained) variance is evidence for the existence of selection effects at the program level. The performance model does not improve significantly, compared to the empty model ( $LR = 3.66$ ,  $df = 5$ , n.s.). Hence, there is little association between the different performance indicators and graduate satisfaction.

With respect to graduates' hourly wage, the picture is quite different. In the performance model, drop-out rate, student-staff ratio, and the number of freshmen have positive and significant effects on the hourly wage of graduates. Student-staff-ratio has a surprising *positive* effect on hourly wage. Apparently, those colleges with least staff per student, deliver graduates with relatively high mean wages. The effect of number of freshmen disappears after controlling for individual characteristics.<sup>12</sup> Hence, we have strong effects of indicators for output-performance, throughput-performance, and input-performance on objective outcome performance. The relative strength of these effects, however, do not match the predictions of hypothesis 3a.

Inspection of the random parts of the models for graduates' hourly wages, show some interesting results. Compared to the empty model, the performance model explains quite some variance at the levels of the colleges and the programs. The estimated variance at the program level drops with 78 percent, while the estimated variance at the college level drops with 36 percent. Compared to the empty model, the performance model improves, but this improvement is barely significant ( $LR = 9.38$ ,  $df = 5$ ,  $p < .10$ ). Only after introducing individual controls, the model improves highly significantly ( $LR = 69.12$ ,  $df = 8$ ,  $p < .001$ ). Remarkably, the reduction in unexplained variance is very large again at the program level, indicating again that selection effects are at work here.

Finally, the differences we found between the effects for graduates' satisfaction (subjective outcome performance) and the effects for hourly wage (objective outcome performance) are exactly what was predicted by hypothesis 3a. Objective outcome performance is more strongly related to other indicators than is subjective outcome performance.

## 5. Conclusion

In the present paper we aimed to test hypotheses about associations between performance indicators at multiple levels of analysis and for multiple stages of the agency 'production process'. Quite surprisingly, the multilevel nature of performance indicators has not triggered much scholarly attention yet, while the covariance between performance indicators highly depends on the specific levels to which these indicators are linked: the network-level, the agency-level, the time-level, or the client-level. For 91 colleges within the Dutch system of universities of applied science, with 7,130 graduates, we showed that

---

<sup>12</sup>A brief inspection of the effects of the control variables in the individual controls models in table 3 adds to our confidence in the analyses. Female graduates are not significantly more satisfied than male graduates, but have significantly lower wages. Non-native Dutch are significantly more satisfied than native Dutch graduates, but also have significantly lower wages. Graduates' grade point-average significantly contributes to both their satisfaction and their hourly wage.

different performance indicators show considerably different patterns in variation at different levels of agency performance.

We illustrated that markedly different—and sometimes paradoxical—associations can exist between performance indicators when moving from one level of performance to another level. Subjective performance of colleges at the level of graduates—their satisfaction with the program offered—is positively associated with objective performance at the level of graduates—their hourly wage. But when we move one level of aggregation, to that of the college, the mean percentage of graduates satisfied by a college between 2002 and 2005 is *negatively* associated with the mean hourly wage of a college's graduates. Thus, quite different mechanisms may affect performance at different levels, and we should keep this in mind when studying the determinants of performance, formulating policy advices, or benchmarking the performance of agencies against some baseline. Yet, we found strong associations between subjective and objective indicators for outcome-performance: both at the level of the college as well as at the level of the graduates.

The second aim of the present paper was to build on current studies, which conclude that different indicators capture different aspects of agency performance. We integrated our multilevel analyses with the popular multi-stage (system) model of performance indicators, with the idea that the causal and sequential process of agency operations offers an overarching explanation for associations between indicators. The analyses in the present paper are both hopeful and somewhat disappointing. Results are hopeful in the sense that different indicators display patterns in variation between levels as expected. Entropy entered into college performance in terms of strong fluctuations in time when moving from input to output and from output to outcomes. The environment of the colleges did not so much affect their performance at the entrance of the college system, but does so at exit. Stable enrollment numbers changed into fluctuating drop-out rates and diploma rates of students, and fluctuating satisfaction and wages of graduates.

Results are somewhat disappointing with respect to the strength of associations between input-performance, throughput-performance, and output-performance on the one hand, and outcome performance on the other hand. We replicated the findings of Boyne et al. (2006a). We found a few strong effects, but many indicators were not related to outcome performance. Hence, despite its popularity, the multistage model of performance might not be a promising avenue for providing an overarching explanation for associations between performance indicators—even when combined with a multilevel approach which takes into account co-variation at different levels. On the other hand, the aim of the paper was not to explain agency performance, but to make informed predictions on associations between indicators. On the basis of the (co-)variation we found in and between different indicators at different levels, we have become much better informed where in the colleges' activities, and at what levels of performance, we can further search for mechanisms that explain performance.

## References

- Allen, Jim and Ger Ramaekers. 1999. *De arbeidsmarktpositie van afgestudeerden in het hoger onderwijs: HBO-monitor 1998. The Hague: HBO-Raad.*
- Andrews, Rhys, George A. Boyne, and Richard M. Walker. 2006a. Subjective and objective measures of organizational performance: An empirical exploration. In *Public service performance: perspectives*

- on measurement and management, eds George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole Jr., and Richard M. Walker, 14-34. Cambridge, NY: Cambridge University Press.
- Andrews, Rhys, George A. Boyne, and Richard M. Walker. 2006b. Strategy Content and Organizational Performance: An Empirical Analysis. *Public Administration Review* 66:1, 52-63.
- Bommer, William H., Jonathan Johnson, Gregory A. Rich, and Philip M. Podsakoff. 1995. On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology* 48:3, 587-605.
- Boyne, George A. 2003. What is public service improvement? *Public Administration*, 81: 211-27.
- Boyne, George and Jennifer Law. 1991. Accountability and local authority annual reports: The case of Welsh district councils. *Financial Accountability & Management* 7:3, 179-194.
- Bouckaert, Geert, and Steven van de Walle. 2003. Comparing measures of citizen trust and user satisfaction as indicators of 'Good Governance': Difficulties in linking trust and satisfaction indicators. *International Review of Administrative Sciences* 69(3):329-343.
- Brewer, Gene A. 2006. All measures of performance are subjective. In *Public service performance: perspectives on measurement and management*, eds George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole Jr., and Richard M. Walker, 35-54. Cambridge, NY: Cambridge University Press.
- Brown, Karin and Philip B. Coulter. 1983. Subjective and objective measures of police service delivery, *Public Administration Review*, 43: 50-58.
- Bryk, Anthony S. and Yeow M. Thum, 1989. The effects of high school organization on dropping out: An exploratory investigation. *American Educational Research Journal*, 26: 353-383.
- Choi, Sungjoo. 2008. Diversity in the US federal government: Diversity management and employee turnover in federal agencies. *Journal of Public Administration and Theory* 19: 603-630.
- Dawes, John. 1999. The relationship between subjective and objective company performance measures in market orientation research: further empirical evidence. *Marketing Bulletin* 10: 65-75.
- Delaney, J. T. and M.A. Huselid. 1996. The impact of human resource management practices on perceptions of organizational performance. *Academy of Management Journal* 39: 949-969.
- Dess, Gregory G. and Richard B. Robinson. 1984. Measuring organizational performance in the absence of objective measures: The case of the privately-held firm and conglomerate business units. *Strategic Management Journal* 5: 265-273.
- Dollinger, M.J. and P.A. Golden. 1992. Interorganizational and collective strategies in small firms: Environmental effects and performance. *Journal of Management* 18: 695-715.
- Dückers, M.L.A. 2009. *Changing hospital care: Evaluation of a multilayered organisational development and quality programme*. Utrecht: NIVEL.
- Duckett, Stephen and Hal Swerissen. 1996. Specific Purpose Programs in Human Services and Health: Moving from an Input to an Output and Outcome Focus. *Australian Journal of Public Administration* 55:3, 7-17.
- Golden, B.R. 1992. Is the past the past – or is it? The use of retrospective accounts as indicators of past strategies. *Academy of Management Journal* 35: 848-860.
- Hedley, Timothy P. 1998. Measuring Public Sector Effectiveness Using Private Sector Methods. *Public Productivity & Management Review* 21:3, 251-258.
- Heinrich, Carolyn J. and Larry E. Lynn, Jr. 1999. Means and ends: A comparative study of empirical methods for investigating governance and performance. *Journal of Public Administration and Theory* 11: 109-138.
- Heinrich, Carolyn J. and Elizabeth Fournier. 2004. Dimensions of publicness and performance in substance abuse treatment organizations. *Journal of Policy Analysis and Management* 23:1, 49-70.
- Hox, J. 2002. *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Kaiser, Frans, Hans Vossensteyn, and Jos Koelman 2001. *Public funding of higher education: A comparative study of funding mechanisms in ten countries*. CHEPS: Enschede.
- Kelly, Janet M., and David Swindell. 2002. A multiple-indicator approach to municipal service evaluation: Correlating performance measurement and citizen satisfaction across jurisdictions. *Public Administration Review* 62:610-620.
- Kim, Tae Kuen, Phyllis Solomon, and Karen A. Zurlo. 2009. Applying hierarchical linear modelling (HLM) to social work administration. *Administration in Social Work* 33:3, 262-227.
- Lee, Valerie E. and Burkam, David T. 2003. Dropping out of high school: The role of school organization and structure. *American Education Research Journal*, 40: 553-593.

- McNeal, Ralph B. 1997. High school dropouts: A closer examination of school effects. *Social Science Quarterly*, 78: 209-222.
- Meier, Kenneth J. and Laurence J. O'Toole Jr. 2001. Managerial strategies and behavior in networks: A model with evidence from U.S. public education. *Journal of Public Administration Research and Theory*, 11: 271-93.
- Meier, Kenneth J., and Laurence J. O'Toole Jr. 2003. Public management and educational performance: The impact of managerial networking. *Public Administration Review* 63(6):689-699.
- Provan, Keith G. and H. Brinton Milward. 2001. Do networks really work? A framework for evaluating public-sector organizational networks. *Public Administration Review* 61:414-23.
- Rumberger, Russell and Gregory J. Palardy 2005. Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American Educational Research Journal* 42: 3-42.
- Schalk, Jelmer, René Torenvlied and Jim Allen. Forthcoming. Network embeddedness and public agency performance: The strength of strong ties in Dutch higher education. *Journal of Public Administration Research and Theory*, online published July 31, 2009.
- Simon, Herbert A. 1945. *Administrative behavior*. (4<sup>th</sup> edition). New York, NY: Free Press.
- Snijders, Tom A.B., and Roel J. Bosker. 1999. *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London: Thousand Oaks.
- Sorber, Bram. 1993. Performance Measurement in the Central Government Departments of the Netherlands. *Public Productivity & Management Review* 17:1, 59-68
- Stone, Melissa M. and Susan Cutcher-Gershenfeld. 2001. Challenges of measuring performance in non-profit organizations. In Patrice Flynn and Virginia A. Hodgkinson (eds.), *Measuring the impact of the non-profit sector* (pp. 33-54). New York, NY: Kluwer Academic Publishers.

**Table 1**

Intra-class correlation coefficients for seven performance measures and four levels (z-values of variance estimates between parentheses).

	Network-level (5 programs)	Agency-level (91 colleges)	Time-level (305 cohorts)	Client-level
<i>Input-performance</i>				
Number of students	0.266 (2.137)*	0.623 (12.987)	0.111 (20.608)	
Number of freshmen	0.239 (2.139)*	0.565 (12.617)	0.195 (20.707)	
<i>Throughput-performance</i>				
Personnel costs	0.467 (2.688)**	0.478 (13.015)	0.055 (20.845)	
<i>Output-performance</i>				
Dropout rate	0.401 (2.715)**	0.261 (8.554)	0.339 (20.966)	
Diploma rate	0.225 (2.323)**	0.315 (7.551)	0.459 (20.696)	
<i>Outcome-performance</i>				
Mean % graduates satisfied	0.186 (2.180)*	0.397 (9.618)	0.417 (20.799)	
Mean graduate hourly wage	0.180 (2.010)*	0.367 (8.527)	0.453 (20.809)	
Graduate satisfaction N = 7,130	0.010 (1.235) <sup>ns</sup>	0.046 (4.960)	0.012 (3.564)	0.933 (n.a.)
Graduate hourly wage N = 5,418	0.008 (1.149) <sup>ns</sup>	0.026 (3.869)	0.024 (5.570)	0.942 (80.254)

All  $p$ -values < .0001, except for <sup>ns</sup>  $p$  > .05; \*  $p$  < .05; \*\*  $p$  < .01. Computation of variance estimates and standard errors based on restricted maximum likelihood estimation.

**Table 2**  
Multilevel logistic regression of (mean) hourly wage on graduates' satisfaction (N=5,418).

	Model 1	Model 2	Model 3
<i>Fixed part</i>			
Mean hourly wage (2002-2005)			-0.206 (0.057) <sup>***</sup>
Hourly wage		0.049 (0.006) <sup>***</sup>	0.051 (0.006) <sup>***</sup>
Constant	0.865 (0.108) <sup>**</sup>	0.250 (0.133) <sup>*</sup>	2.835 (0.725) <sup>***</sup>
<i>Random part</i>			
Program level (n=5)	0.044 (0.035)	0.049 (0.039)	0.033 (0.028)
College level (n=91)	0.174 (0.037)	0.188 (0.040)	0.163 (0.035)
Cohort level (n=305)	0.041 (0.015)	0.039 (0.015)	0.038 (0.015)
Graduate level	3.26	3.29	3.29
<i>Deviance</i> <sup>‡</sup>	3826.267	3805.791	3802.656

\*  $p < .10$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ . † Multilevel logistic regression using (maximum likelihood) Laplace estimation; ‡ Deviance =  $-2\log(\text{likelihood})$ .

**Table 3**

Correlations of eight performance indicators for 91 colleges between 2001 and 2005 (reported significance-levels based on clustered<sup>†</sup> standard errors of estimates in bivariate regression).

Number of students	(a1)	(a1)						
Number of freshmen	(a2)	.96 <sup>***</sup>	(a2)					
Student-staff ratio	(b1)	.22 <sup>*</sup>	.26 <sup>*</sup>	(b1)				
Solvency	(b2)	.18 <sup>*</sup>	-.14	-.05	(b2)			
Dropout rate	(c1)	.14 <sup>*</sup>	.15 <sup>*</sup>	.17 <sup>*</sup>	-.12	(c1)		
Diploma rate	(c2)	-.25 <sup>***</sup>	-.27 <sup>**</sup>	-.18 <sup>*</sup>	.02	-.31 <sup>*</sup>	(c2)	
Graduate satisfaction	(d1)	-.05	-.02	-.09	.02	-.25 <sup>**</sup>	.11 <sup>*</sup>	(d1)
Graduate hour wage	(d2)	.26 <sup>**</sup>	.26 <sup>*</sup>	.33 <sup>***</sup>	.02	.32 <sup>***</sup>	-.18 <sup>***</sup>	-.17 <sup>*</sup>

<sup>†</sup> Clustering variable = college; <sup>\*</sup>  $p < .10$ ; <sup>\*</sup>  $p < .05$ ; <sup>\*\*</sup>  $p < .01$  <sup>\*\*\*</sup>  $p < .001$  (two-tailed).

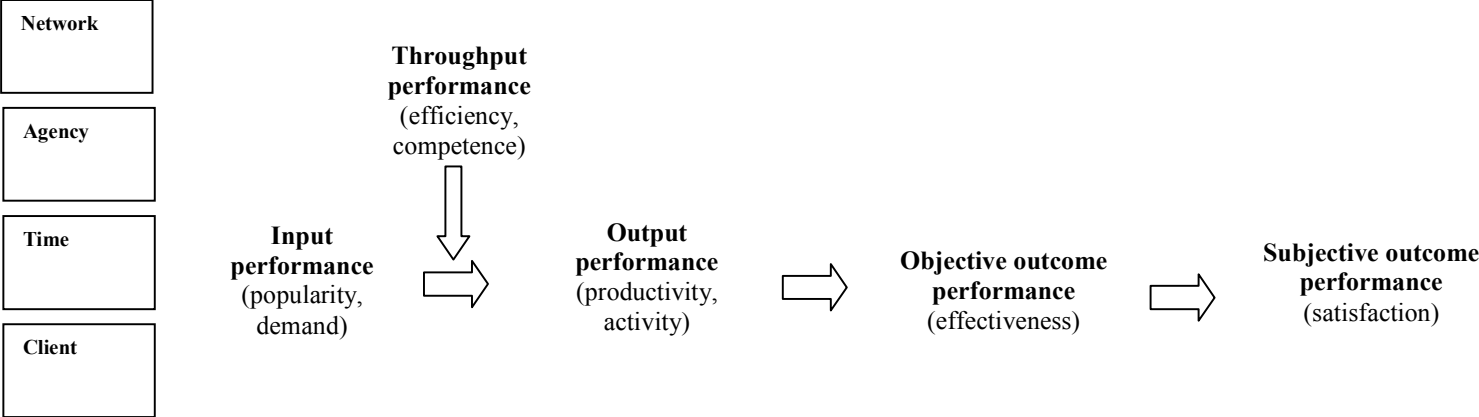
**Table 4**  
Multilevel regressions of graduates' satisfaction<sup>†</sup> and graduates' hourly wage for four cohorts (2002-2005); N= 5,418 graduates.

	Graduates' satisfaction			Graduates' hourly wage		
	Empty model	Performance model	Individual controls	Empty model	Performance model	Individual controls
<i>College level</i>						
Drop-out rate		-0.003 (0.005)	-0.00 (0.005)		0.041 (0.010) <sup>***</sup>	0.035 (0.009) <sup>***</sup>
Diploma rate		-0.098 (0.765)	0.037 (0.778)		-0.252 (1.601)	0.526 (1.568)
Student-staff ratio		0.025 (0.033)	0.214 (0.762)		0.211 (0.060) <sup>***</sup>	0.209 (0.059) <sup>***</sup>
Solvency		0.782 (0.350) <sup>*</sup>	0.732 (0.348) <sup>*</sup>		0.916 (0.625)	0.977 (0.616)
Number of freshmen		0.001 (0.001)	0.001 (0.001)		0.075 (0.044) <sup>*</sup>	0.061 (0.043)
Mean hourly wage		-0.200 (0.062) <sup>***</sup>	-0.235 (0.061) <sup>***</sup>			
<i>Graduate level</i>						
Hourly wage of graduates			0.047 (0.005) <sup>***</sup>			
Gender (1= female)			0.003 (0.054)			-1.379 (0.101) <sup>***</sup>
Ethnicity (1 = non native)			0.244 (0.069) <sup>***</sup>			-0.141 (0.139) <sup>***</sup>
Grade Point Average			0.138 (0.019) <sup>***</sup>			0.285 (0.034) <sup>***</sup>
Constant	0.865 (0.108) <sup>***</sup>	2.680 (0.800) <sup>***</sup>	1.964 (0.802)	12.698 (0.206) <sup>***</sup>	8.053 (0.988) <sup>***</sup>	9.574 (1.004)
$\sigma^2$ program level	0.044 (0.035) <sup>***</sup>	0.023 (0.022)	0.011 (0.014)	0.161 (0.140)	0.036 (0.053)	0.013 (0.039)
$\sigma^2$ college level	0.174 (0.037)	0.164 (0.035)	0.163 (0.035)	0.525 (0.136)	0.338 (0.103)	0.338 (0.101)
$\sigma^2$ cohort level	0.041 (0.015)	0.035 (0.015)	0.033 (0.015)	0.475 (0.085)	0.429 (0.081)	0.397 (0.076)
$\sigma^2$ graduate level	3.26	3.26	3.26	18.964 (0.236)	18.964 (0.236)	18.632 (0.231)
<i>Deviance</i> = -2log(likelihood)	3826.267	3822.608	3785.249	19042.999	19033.624	18973.844

<sup>†</sup> Multilevel logistic regression using (maximum likelihood) Laplace estimation; <sup>\*</sup> $p < .10$ ; <sup>\*</sup> $p < .05$ ; <sup>\*\*</sup> $p < .01$ ; <sup>\*\*\*</sup> $p < .001$ .

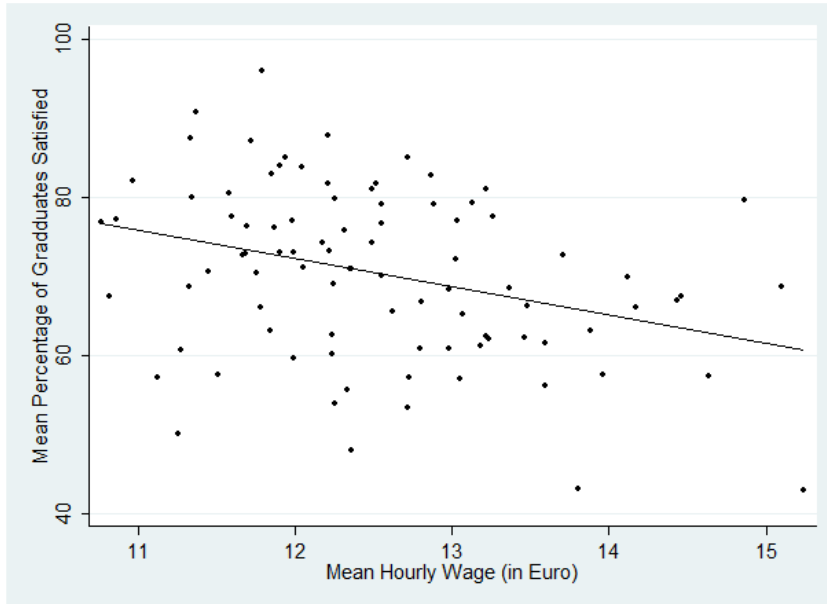
**Figure 1**  
Schematic overview of four types of agency performance at four levels of analysis.

---



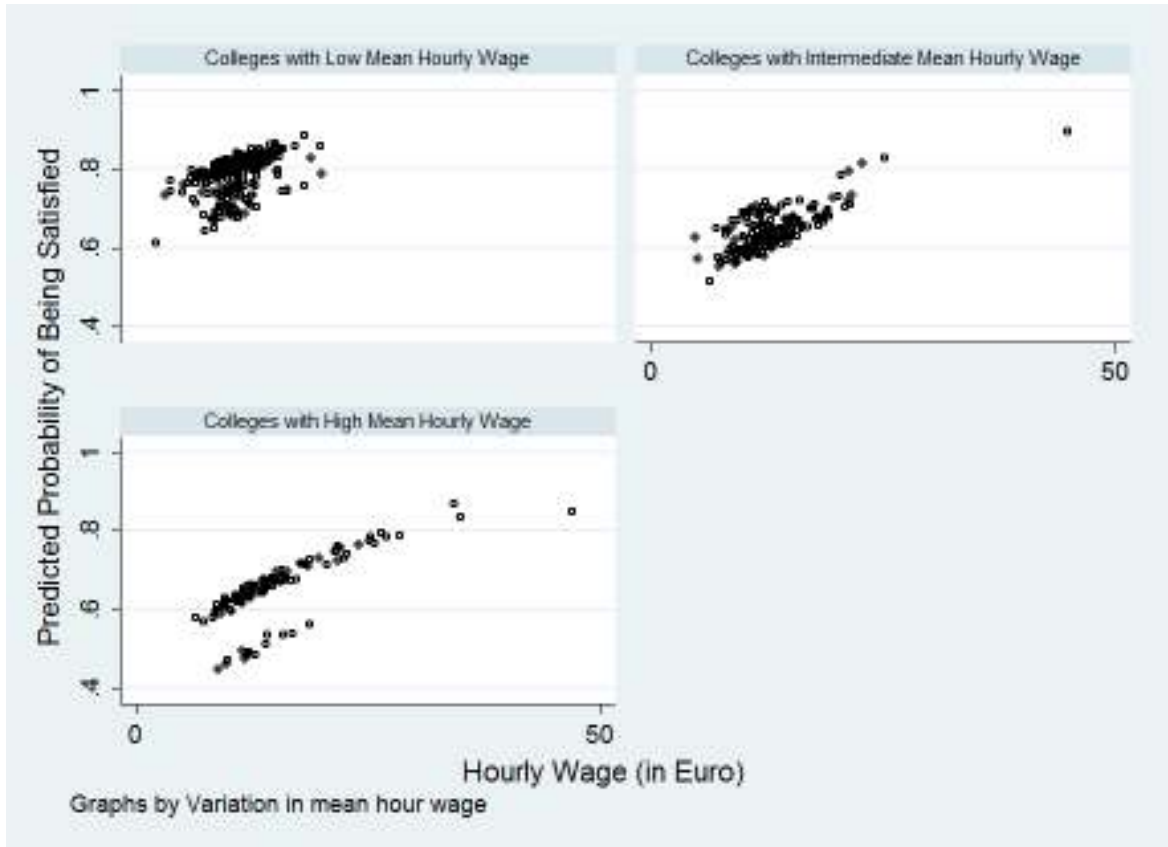
**Figure 2**

Negative association between subjective and objective performance at the level of the college: 'mean hourly wage' and 'mean percentage of graduates satisfied'.



**Figure 3**

Positive associations between subjective and objective performance at the level of the graduate: 'hourly wage' and 'probability of being satisfied'<sup>†</sup> of graduates for some colleges with low, intermediate, and high mean hourly wage.



<sup>†</sup> Probabilities computed for cases on the basis of model 3 in Table 2.