On the optimal composition of evaluation committees: Evidence from public exams*

Manuel F. Bagues†, Florentino Felgueroso‡ and María J. Pérez-Villadóniga§

March, 2007

Abstract

In this paper we analyse how the field of specialization of evaluation committees may affect the outcome of selection processes. In particular, we show that evaluators tend to assess more severely those dimensions of quality at which they are themselves more knowledgeable. As a result, candidates whose knowledge profile is closer to that of the evaluator will have a higher probability of success. Data from public examinations to the Judicial and Prosecutor Corps in Spain confirm that grades are significantly lower when the members of an evaluation committee are relatively more expert at the field which is evaluated. This confirms the importance of a diverse and balanced composition of selection committees.

1 Introduction

Productivity is hardly ever perfectly observable and evaluators usually base their decisions on some indicators of quality, such as the performance of aptitude tests or interviews. However, it has been widely documented that evaluators’ characteristics may often affect the outcome of selection processes in a way that is not always related to productivity. Specifically, it has been shown that in many situations evaluators tend to give higher ratings to those applicants with similar attitudes, background or observable personal attributes. This tendency, known as the “similar-to-me” effect, has been documented both in psychology and sociology (Byrne and Clore, 1970 and Byrne, 1971) and in organizational processes, such as recruitment (Goldberg, 2005) and supervisors’ assessments of subordinates (Ferris et al., 1994).

In this paper, we focus on another relevant characteristic of evaluators that may influence the assessments given to candidates: the evaluators’ skills profile.

*We are grateful to participants in seminar at the University Carlos III (Madrid).
†Universidad Carlos III, Madrid.
‡Universidad de Oviedo, Asturias and FEDEA.
§Universidad de Oviedo, Asturias.
In fact, most selection processes in the labour market involve measuring candidates’ abilities at several dimensions and the evaluator’s accuracy to assess ability may not be the same in all dimensions.

The model presented is based on two assumptions. First, we consider a context where total productivity is multidimensional, that is it depends on candidates’ ability or knowledge at several dimensions. In particular, we analyse a selection process consisting of a series of qualifying tests that candidates must pass successively and different dimensions of quality are assessed at different stages.

The second assumption is that knowledge is complex: the ability to assess knowledge in a certain dimension depends on the observer’s own knowledge of that dimension. Actually, several studies have empirically shown that individuals who are unskilled at a field are more imprecise in their evaluation of other individuals’ abilities at that field1.

We formally show that when productivity is multidimensional and the evaluator can observe candidates’ quality more accurately at those dimensions at which she is more knowledgeable, she will give more weight to information provided at those dimensions. More precisely, the model suggests that in an optimal evaluation the evaluator will be more demanding at those stages where she can assess quality with more accuracy and, conversely, she will be more benevolent at those fields where her knowledge is relatively more limited. As a consequence, given two equally productive candidates the evaluator will tend to give a higher valuation to the one who excels in the same dimensions as she does. The intuition is simple: candidates whose knowledge profile resembles (more) that of the evaluator will tend to have a higher probability of success, given that they excel precisely at the dimensions that the evaluator can assess with more accuracy and therefore will be taken more into account. On the contrary, candidates whose knowledge profile differs from that of the evaluator will have their quality underestimated, as the dimensions at which they excel will receive less attention.

In order to test the predictions of the model we present empirical evidence coming from a real labour market selection process: the evaluation made to candidates to become judges or prosecutors in Spain. It is a selection process of great importance both because of the number of candidates involved—aabout 5000 each year— and because of the relevance of the jobs at stake: successful candidates become judges or prosecutors2.

The exceptional structure of the public exams to become judges and prosecutors allows us to consistently estimate whether evaluators tend to assess more severely those fields at which their knowledge is greater. First, the allocation of candidates to evaluation committees is decided through a random process so that the distribution of candidates’ quality is, by construction, statistically identical across committees. Second, the composition of the committees that

---

1 See Chi, 1978, Chi et al., 1982 or Kruger and Dunning, 1999.
2 Formally, passing the public examination does not guarantee that candidates automatically become judges or prosecutors, they must take a one-year course. However, there are very few cases, less than 1%, of candidates who have not been successful.
evaluate each set of candidates is not identical: a key difference lies in the speciality of the university professor member of each committee. This allows us to estimate the effect of the committees’ knowledge profile on the evaluations made to candidates.

The results of the empirical analysis reveal the existence of a "similar-to-me effect". The presence, among the evaluators, of a university professor who is an expert in the topic evaluated at a certain test is associated with a more important weight given to this test in the selection process. In particular, we observe that the grades are about 0.65 points lower when the professor member of the committee is an expert in the subject evaluated. Hence, committees seem to be more demanding when evaluating the subject at which they are expert which, in turn, may favour candidates who excel in that dimension.

However, even though this is the best the evaluator can do, if all dimensions of ability are equally important to perform the job it will not be the optimal outcome of the selection process. In this sense, we argue that the legislator or social planner must first define all dimensions of productivity that are required and the weight given to each of them. Then, the evaluation process should be designed in such a way that all dimensions can be evaluated with the same accuracy. To do so, it is important that the composition of evaluation committees is balanced and diverse, so that different evaluators can assess different dimensions according to their evaluation precision offsetting any "similar-to-me effect".

The structure of the paper is as follows. In the second section we present the theoretical model and the testable implications derived. Then, in section 3 we present the empirical evidence and test the hypothesis. Section four concludes.

2 The Model

Let us assume that an evaluator must carry out a selection process. The evaluator has to make a decision with respect to each candidate: either to select him (a) or to discard him (b). For simplicity we will consider that there are two types of candidates: high productivity (A) and low productivity (B) individuals. The best possible action that the evaluator can take is to select highly productive candidates and reject low productivity ones. Formally, the evaluator’s payoff will be \( \Pi \in (\Pi_H, \Pi_L) \), where \( \Pr(\Pi_H|A,a) = \Pr(\Pi_H|B,b) = 1 \) and \( \Pr(\Pi_H|A,b) = \Pr(\Pi_H|B,a) = 0 \). Without loss of generality, we normalize potential payoffs to \( \Pi_H = 1 \) and \( \Pi_L = 0 \).

The evaluator knows that the proportion of type A individuals in the population is \( \Pr(A) = p \) and the rate of type B candidates is \( \Pr(B) = (1 - p) \). However, there are informational asymmetries that prevent the evaluator from observing ex-ante whether a given candidate belongs to the high or low productivity group. The evaluator can only obtain a signal of the candidate’s quality through the performance of a series of evaluation tests. In particular we will consider a multiple hurdle selection process, that is, the applicant must pass each step in the process. Failure at any step disqualifies him from further consideration. Also, we assume that different dimensions are evaluated at each
stage.

More precisely, the structure of the process is the following. In period 0 nature picks the state of the world, i.e., whether the candidate is highly or low productive, but this information is not observable by the evaluator. In period 1 the first stage of the evaluation takes place and the evaluator sets a threshold \( x_1 \in (0, 1) \) that the candidate must surpass in order to proceed to the second stage of the evaluation. If the candidate obtains a grade below the threshold, he will be definitely excluded from the evaluation process. In period 2, the second stage of the evaluation occurs and the evaluator sets a new threshold \( x_2 \in (0, 1) \).

If the candidate manages to reach it he will have concluded the selection process successfully. Otherwise he will be excluded from the process. Figure 1 describes the sequence of events.

In order to formalize the stochastic relationship between the candidate’s quality and his observed performance at each stage we assume that functions \( g_j(c) \) and \( f_j(c) \) describe, respectively, the probability that a high productivity individual [A] and a low productivity one [B] get a grade \( c \), where \( j = 1, 2 \) indicate the stage that is being evaluated. Therefore, the functions \( g_j(c) \) and \( f_j(c) \) represent the stochastic relationship between a candidate’s quality and the grade he obtains. We assume that the functions \( g_j(c) \) and \( f_j(c) \) satisfy the Monotone Likelihood Ratio Property (MLRP):

\[
\left( \frac{f_j(c)}{g_j(c)} \right)' < 0
\]

The MLRP states that the probability that a candidate who has achieved grade \( c \) in test \( j \) is a high productivity individual (A) is increasing in the grade obtained. That is, given any threshold, the probability that a candidate is type A (B) is higher if he gets a grade above (below) that threshold than if he gets a grade below (above) it.

In order to preserve the symmetry of the analysis, we will also assume that the density functions are symmetric, thus \( g_j(c) = f_j(1 - c) \) and, hence, \( G_j(c) = 1 - F_j(1 - c) \).

The evaluator’s problem lies in setting at each stage a cutoff that the candidate must reach in such a way that the probability of selecting high productivity individuals is maximized and the probability of choosing those of low productivity is minimized. In other words, the evaluator maximizes the probability that high productivity candidates successfully accomplish both stages, subject to the restriction that the proportion of selected candidates must be equal to the proportion of high productivity individuals, \( p \).

Formally, the problem can be represented as follows:

\[
\begin{align*}
\max_{x_1, x_2} V &= \Pr [c_1 > x_1 | A] \Pr [c_2 > x_2 | A] = [1 - G_1(x_1)] [1 - G_2(x_2)] \\
\text{s.t.} p &= \Pr [c_1 > x_1 | B] \Pr [c_2 > x_2 | B] = [1 - G_1(x_1)] [1 - G_2(x_2)] + (1 - p) [1 - F_1(x_1)] [1 - F_2(x_2)] \\
0 &\leq x_1 \leq 1 \quad 0 \leq x_2 \leq 1
\end{align*}
\]
where \( x_1 \) and \( x_2 \) are the thresholds that the candidate must surpass in the first and second stages and the functions \( G_j(c) = \int_0^c g(x)dx \) and \( F_j(c) = \int_0^c f(x)dx \) represent the probability that a type A and a type B candidates do not reach threshold \( c \) in stage \( j \), respectively.

From this optimization problem it is possible to determine the optimal levels of difficulty at each stage, i.e. \( x_1^* \) and \( x_2^* \), as a function of the parameters that define the functions \( g_j(c) \) and \( f_j(c) \), and of the unconditional probability \( p \) that a candidate is type A.

**Proposition 1** When an individual’s evaluation ability is the same in both stages, the minimum grade required in each stage will be the same, that is,

\[
\text{If } f_1(.) = f_2(.) \quad \text{and} \quad g_1(.) = g_2(.) \quad \Rightarrow \quad x_1^* = x_2^*
\]

**Proof.** The corresponding lagrangean is given by:

\[
L = [1 - G_1(x_1)][1 - G_2(x_2)] \lambda + \lambda [p - p[1 - G_1(x_1)][1 - G_2(x_2)] - (1 - p) [1 - F_1(x_1)][1 - F_2(x_2)]
\]

In order for \( x_1^* \) and \( x_2^* \) to be the solution to the problem they must satisfy at least one of the following first two conditions and one of the last two ones:

i. \( \frac{\partial L}{\partial x_1} \leq 0 \quad x_1 \geq 0 \quad x_1 \frac{\partial L}{\partial x_1} = 0 \)

ii. \( \frac{\partial L}{\partial x_2} \geq 0 \quad (1 - x_1) \geq 0 \quad (1 - x_1) \frac{\partial L}{\partial x_1} = 0 \)

iii. \( \frac{\partial L}{\partial x_2} \leq 0 \quad x_2 \geq 0 \quad x_2 \frac{\partial L}{\partial x_2} = 0 \)

iv. \( \frac{\partial L}{\partial x_2} \geq 0 \quad (1 - x_2) \geq 0 \quad (1 - x_2) \frac{\partial L}{\partial x_2} = 0 \)

The corner solution \( x_1^* = 0 \) or \( x_2^* = 0 \) means, in practice, that everybody passes the corresponding test. On the other hand, the corner solution \( x_1^* = 1 \) or \( x_2^* = 1 \), i.e., nobody passes one of the tests, does not satisfy the restriction to the maximization problem set up. Any interior solution to the problem must satisfy:

v. \( g_1(x_1^*) f_2(x_2^*) [1 - G_2(x_2^*)][1 - F_1(x_1^*)] = g_2(x_2^*) f_1(x_1^*) [1 - G_1(x_1^*)][1 - F_2(x_2^*)] \)

vi. \( p = p[1 - G_1(x_1^*)][1 - G_2(x_2^*)] + (1 - p) [1 - F_1(x_1^*)][1 - F_2(x_2^*)] \)

If \( f_1(.) = f_2(.) \) and \( g_1(.) = g_2(.) \) it is easy to see that any pair \((x_1^*, x_2^*)\) such that \( x_1^* = x_2^* \) satisfy condition v.

Now, what happens if the evaluator is more accurate when evaluating one of the tests? For simplicity, in this case we will assume that the density functions are linear so that:

\[
\begin{align*}
g_1(x_1) &= 1 - a + 2ax_1 \quad \rightarrow \quad G_1(x_1) = (1 - a)x_1 + ax_1^2 \\
g_2(x_2) &= 1 - b + 2bx_2 \quad \rightarrow \quad G_2(x_2) = (1 - b)x_2 + bx_2^2 \\
f_1(x_1) &= 1 + a - 2ax_1 \quad \rightarrow \quad F_1(x_1) = (1 + a)x_1 - ax_1^2 \\
f_2(x_2) &= 1 + b - 2bx_2 \quad \rightarrow \quad F_2(x_2) = (1 + b)x_2 - bx_2^2
\end{align*}
\]
where $0 \leq a \leq 1$ and $0 \leq b \leq 1$. As shown in Figure 3, the probability of achieving a high value of the signal in both tests is higher for type A candidates.

Parameters $a$ and $b$ represent the evaluator’s accuracy in the first and second tests, respectively. In particular, if $a > b$ the evaluator is more accurate in the evaluation of the first test than of the second. In other words, the probability that the evaluator assigns a high grade to a type A candidate is greater in the first test than in the second one. Similarly, the probability of giving a low evaluation to a type A candidate is lower in the first test. The same reasoning applies for type B candidates.

In a selection process consisting of a series successive and qualifying tests, the greater the accuracy of the evaluator at a certain test, the greater the minimum grade required to pass it. That is,

$$\frac{dx_1^*}{da} > 0; \quad \frac{dx_2^*}{db} > 0$$

At the same time, given that the proportion of candidates to be selected is fixed, rising the threshold at a certain test means lowering the threshold at the other. Hence, the more accurate the evaluator is at the first test with respect to the second, the lower the threshold set at this last one:

$$\frac{dx_1^*}{db} < 0; \quad \frac{dx_2^*}{da} < 0$$

**Proof.** From the first order conditions iii and iv, and substituting into them the expressions for the density and cumulative distribution functions we obtain the equations:

$$(1 - a + 2ax_1^*)(1 + b - 2bx_2^*)(1 - ax_1^*)(1 + bx_2^*) = (1 + a - 2ax_1^*)(1 - b + 2bx_2^*)(1 + ax_1^*)(1 - bx_2^*)$$

$$p = p(1 - x_1^*)(1 + ax_1^*)(1 - x_2^*)(1 + bx_2^*) + (1 - p)(1 - x_1^*)(1 - ax_1^*)(1 - x_2^*)(1 - bx_2^*)$$

These two equations generate a system of equations of a high degree, which makes its analytical resolution for generic values of the parameters enormously difficult. Instead, we have illustrated the optimal solution to the problem by assigning certain values to parameters $p$ and $b$. In Figures 4.1-4.3 we represent, for different proportions of type A candidates, the values that and take as the accuracy of the evaluator in test 1 varies and assuming a fixed accuracy at test 2 ($b = 0.5$).

Only the interior solutions to the problem have been represented. In all cases, we observe that as the evaluator’s accuracy in the first test ($a$) increases, the minimum grade required to pass it also increases and, at the same time, the level of demand in the second test decreases.

### 2.1 Testable implications

In this model we have considered a selection process consisting of several qualifying stages. The model predicts that the minimum grade required to pass a
certain stage will be higher the greater the accuracy of the evaluator at that stage. That is, the evaluator will be more demanding at those tests at which he can assess quality more accurately.

What factors does the accuracy with which the evaluator can estimate a candidate’s quality depend on? In the first place, the intrinsic characteristics of the evaluation process may have an influence. For instance, the longer the evaluation process the greater the amount of information will be available to the evaluator to make his assessment. In the second place, the evaluator’s own characteristics may also be relevant. An individual’s evaluation ability at a certain dimension may depend critically on his own knowledge at that dimension. Several studies have found that in chess (Chi, 1978), physics (Chi et al., 1982) or grammar (Kruger and Dunning, 1999) the lower the evaluator’s quality the more imprecise her evaluation on other individuals’ quality was.

An implication of the above model will be that in a selection process consisting of several qualifying stages, evaluators will tend to be more demanding at those stages at which his knowledge is greater.

Corollary 2 In a selection process consisting of successive and qualifying tests, the greater the evaluator’s knowledge at the dimension assessed in a certain test, the greater the grade demanded to pass that test.

3 Empirical analysis

A great amount of information is required in order to estimate empirically how the evaluator’s different levels of knowledge at each dimension may affect her assessments candidates’ quality. It is necessary to collect, at each dimension, information on the candidates’ and evaluators’ true quality. As well, it is also necessary to observe the assessment that each candidate has received from the evaluator. Such detailed information is hardly ever available in labour market selection processes so economists often appeal to laboratory experiments.

In this paper, however, we present evidence from a real selection process in the labour market: the evaluation made to candidates to become judges or prosecutors in Spain in the years 2003, 2004 and 2005. In Spain, in order to become judge or prosecutor, as to access any position in the public sector, individuals are required to pass a selection process at the national level3. More than 5000 candidates take these tests, called yearly, to fill between 100 and 200 posts.

Two intrinsic characteristics of this selection process make the empirical analysis feasible. One the one hand, the first stage of the process consists in taking a multiple choice test. The existence of this preliminary test allows us to obtain an unbiased proxy of the candidates’ real quality. On the other hand, once the number of candidates who have achieved a minimum grade at the first test is determined, the number of committees that are to evaluate the following stages is decided. Then candidates are allocated to evaluation committees.

---

3 See Bagues (2005).
through a random mechanism, which guarantees that the average quality of candidates evaluated by each committee is statistically similar. This allows us to test consistently whether the “treatment”, in this setup the knowledge profile of evaluators, has any effect on the candidates’ grades.

3.1 The public examinations to Spanish Judicial and Fiscal Corps

The examinations to become judges or prosecutors were carried out separately until year 2001. From then on they are carried out jointly following certain common rules that include, since 2003 a preliminary multiple choice test. We now describe the main features of the exams in detail.

3.1.1 Structure of the exams

The public examination consists of three theoretical tests, all of them qualifying: the first one is a multiple choice test and the other two are oral tests.

![Figure 4 here]

The topics that candidates are evaluated on are drawn from the set of all possible topics, which are listed in the State Bulletin (BOE) beforehand. This set of topics is divided into two main blocks, each of them including different subjects. In the first test, candidates must answer, in written form, one hundred multiple choice questions. Around 80% of these questions are drawn from the first block of topics and the remaining 20% refer to the second block. Candidates who manage to attain a certain minimum grade in this preliminary test are called to take the first oral test. In this test each candidate has to develop, within 75 minutes, five topics drawn randomly from the first block of topics. Two of the topics are in civil law, two in criminal law and one in general theory of law or constitutional law. At the end of each session candidates are evaluated. The committee first votes in order to decide whether the candidate passes or fails, the President deciding in case of a tie. For every candidate that passes each member of the committee gives a grade ranging from 0 to 5 per topic. The final grade is obtained by adding all grades, excluding the maximum and the minimum, and dividing the total sum by the number of computed gradings. Finally, those candidates who manage to pass the first oral test are called to take the second oral test. The structure and grading of this test is same as

---

4 The evaluation process is described in each call. See, for example, "Agreement of 8th May 2006 of the Selection Commission", BOE n°3, 13th May 2006, p. 18330.

5 In 2003 100% of the questions referred to the first block of topics. Since 2004 about 80% of the multiple choice questions are drawn from the first block and 20% from the second. The first block of topics includes civil law, penal law, general theory of law and constitutional law. The second block includes civil and criminal procedural law, administrative law, commercial law and labour law.

6 The minimum grade required is not set ex-ante, but it is decided ex-post by the Selection Commission once all multiple choice tests have been corrected.
the first one, although the topics are drawn from the second block of the programme: two questions on procedural law, two on administrative law and one on commercial or labour law.

3.1.2 Composition of the evaluation committees

Given the high number of candidates that take the public exam each year, once the number of candidates that have passed the preliminary test is determined, the number of evaluation committees is decided and candidates are allocated to these committees. In general, a committee is formed for every 500 candidates. The allocation of candidates to committees is random: candidates are ranked in alphabetical order and then distributed to committees, the distribution starting from a letter drawn from a lottery.

According to articles 304 and 305 of the Organic Law on the Judicial Power (Ley Orgánica del Poder Judicial) evaluation committees are composed by nine members. The committee is designated by the Selection Commission and is presided by a magistrate of the Supreme Court of Justice or chief prosecutor. The other eight members are two magistrates, two prosecutors, one public defender, one lawyer with over ten years of professional experience, a court secretary and a law university professor. The professors are usually expert in civil, criminal or procedural law.

3.2 The data

To carry out the empirical analysis we have compiled data from public examinations to become judge or prosecutor in 2003, 2004 and 2005. About 4500 candidates were evaluated each year. Our database contains information on the candidate’s gender and geographical origin and on the candidate’s experience, that is the number of times he has taken the public exam. In table 1 we show information from candidates. Most of the candidates are females (only about 32% are males) and about 22% of all candidates reside in Madrid. Also, nearly 80% of candidates had taken the exam at least once before, about 61% had taken it at least twice before and 43% had taken it three or more times before.

With respect to the evaluations received, we can observe which candidates have passed each exam and, for those that have succeeded, we know the grade they were given. On average, about 38% passed the multiple choice test. Out of the candidates who took the first oral exam, around 23% passed it and about 38% of those who took the second oral exam succeeded became judges or prosecutors.

The theoretical analysis developed in the previous section predicts that in a selection process consisting of several successive and qualifying stages, the difficulty of exams will be higher the more knowledgeable the evaluator is at the subject that is being evaluated. Although the composition of the evaluation committees is very similar a very useful feature for our analysis is the presence,
among the members of each committee, of a Professor who is specialized in some of the topics subject to evaluation. Over the period analysed 15 committees were formed. As shown in Table 2, in some cases the Professor was specialized in civil or criminal law, which are evaluated in the first oral exam and, in other cases, the professor was an expert in one of the fields evaluated in the second test, procedural law. Each of these areas makes up 40% of the total content of the exam that is being evaluated. In order to test whether differences in the evaluators’ knowledge profiles generate affect their evaluations we will exploit the variability in the knowledge profile of the professor member of each committee.

### 3.3 Results

There are several factors that may affect the grade $y_{ijc}$ given to a candidate $i$ who has taken test $j$ and has been evaluated by committee $c$. The following equation models such factors:

$$y_{ijc} = \alpha + \gamma C_i + \sum_{t=2003}^{2005} \delta_t D_t + \varepsilon_{ijc}$$

On the one hand, the difficulty of each test may depend on the number of job openings and on the number of candidates that take the exam. The inclusion of the vector of year dummies $D_t$ captures this effect. On the other hand, the vector $C_i$ includes a series of personal characteristics of the candidate: gender, experience and geographic origin, which may also be correlated with the candidate’s quality. Finally, $\varepsilon_{ijc}$ captures those unobservable factors that affect the grade.

[Table 3 here]

In column 1 of table 3 we show the relationship between these factors and the grades obtained by candidates in the preliminary multiple choice test. The grades are significantly increasing with experience: for instance, a candidate who had taken the exam three times before gets, on average, about 15 points (out of 100) more than a novel candidate. Moreover, males tend to get about 2.5 more points than females and those individuals living in Madrid get about 3 points less. Both differences are statistically significant.

In columns 2 and 3 we present the determinants of the grades obtained in the first and second oral exams, respectively, including the grade obtained in the multiple choice test $[y_{i,test}]$ and its square $[y_{i,test}^2]$ as control variables. Given that the multiple choice test is corrected anonymously before the allocation to

---

1 The interest of this paper focuses on the effect of the evaluator’s relative knowledge of different topics on the assessments given. It is important to remark that these public exams are also subject to several evaluation biases related to the conditions of the evaluation or to the presence of sexism or nepotism (see Bagues, 2005).
committees is decided, this grade can be interpreted as a proxy of candidates’ quality which is unbiased with respect to the committee characteristics. In both oral exams we only have information on the grade obtained by those candidates that have passed the exam. Observed grades range from a minimum of 12.5 to a maximum of 25. Employing the tobit estimation method which allows to account for left censoring we can observe that in the first oral exam, conditional on the grade obtained in the multiple choice test, males tend to get lower grades. This result suggests either that males have worse verbal abilities or that there is some kind of discrimination against them on the part of evaluators\(^8\). The rest of the personal characteristics do not seem to have any effect on the grades, conditional on the multiple choice test mark. In the second oral test we observe that candidates who had taken the exam at least three times before tend to obtain relatively lower grades.

In order to assess the effect of the committee knowledge profile on candidates’ assessments we include the dummy variable \([S_{jc}]\) which captures whether the Professor member of committee \(c\) that evaluated the oral test \(j\) was an expert in the subjects evaluated:

\[
y_{ijc} = \alpha + \beta_1 y_{i,test} + \beta_2 y_{i, test}^2 + \lambda S_{jc} + \gamma C_i + \sum_{t=2003}^{2005} \delta_t D_t + \varepsilon_{ijc}
\]

There are two features of the selection process that allow us to estimate consistently the effect of the committee’s knowledge profile \([S_{jc}]\) on the grades achieved in the oral exams. With respect to the first oral test, given that both the assignment of candidates to committees and the appointment of the university professor are randomly determined, it is guaranteed that the professor’s knowledge is orthogonal to the quality of the candidates that she has to evaluate and to any other factors that may affect the grade, both observable and unobservable. In statistical terms,

\[
E(S_{1c} \cdot \varepsilon_{i1c} | D_t) = 0
\]

Exploiting this orthogonality condition it is possible to estimate the effect of \([S_{1c}]\) on the first oral exam grades.

The other feature of the evaluation process that allows us to estimate consistently the effect of the evaluator’s knowledge profile on the grades of the second oral exam is the existence of the preliminary multiple choice test. Although the average quality of candidates that take this exam may not be the same in all committees—it will depend on how benevolent the evaluation of the first oral exam has been—knowledge of the grade that each candidate has achieved in the multiple choice test allows us to correct the potential sample bias:

\[
E(S_{2c} \cdot \varepsilon_{i2c} | D_t, y_{i,test}) = 0
\]

Before interpreting the empirical results we must make the following remark. While in the model presented in section 2 the evaluator sets different thresholds

\(^8\)Bagues and Esteve-Volart find that male evaluators tend to favor female candidates and discriminate against male ones (2007).
in different stages depending on her knowledge of the topics evaluated, in the evaluation of candidates to become judges or prosecutors the threshold is the same in both oral exams (12.5) regardless of the committee’s field of expertise. The effect of the committee’s knowledge profile would then be reflected by making it more difficult for candidates to reach the minimum threshold in the stage where the professor is an expert. This would result in lower average grades at that stage where her quality can be assessed more accurately relative to committees where the professor is not an expert in that stage.

Columns 1 and 2 in Table 4 show the determinants of the grades in the first and second oral exams, respectively.

[Table 4 here]

The speciality of the university professor has a significant effect on the grades, both statistically and economically. Grades were 0.65 points lower in those committees where the professor was an expert in criminal or civil law—both topics are evaluated in the first oral test—relative to those committees where the professor was an expert in procedural law, which is evaluated in the second oral exam (column 1). That is, consistently with the predictions of the theoretical model, evaluators are relatively more demanding when they are more knowledgeable on the subject that is being evaluated. Regarding the determinants of grades in the second oral test, results are qualitatively similar to those obtained for the first oral test. When the Professor is an expert in one of the topics evaluated in this test—procedural law—grades have been 0.65 points lower although in this case, due to the small sample size, the result is only significant at the 15 percent level.

Therefore, analysis of the grades obtained by candidates on both oral exams suggests that evaluators’ knowledge profile may affect the minimum threshold of knowledge that is required at each test.

Although the random assignment of candidates to evaluation committees guarantees that, ex-ante, candidates who are evaluated at each committee have statistically similar quality, we cannot discard the fact that some candidates may behave strategically and, aware of the importance of the evaluators’ knowledge profile, have devoted more time to study those topics at which the committee is an expert. In this case, the estimated coefficient must be considered as a lower bound of the total effect of the knowledge profile on the evaluation standards.

4 Conclusion

It has been argued that the ratings given to candidates could be affected by evaluators characteristics. In particular, evaluators may tend to give higher ratings to those candidates with whom they share similar attitudes or other personal attributes. In this paper we argue that this similarity effect may also be relevant in terms of skills. Specifically, we consider a set up where productivity is multidimensional and the accuracy to assess knowledge in a certain dimension depends on the evaluator’s own knowledge of that dimension. Moreover, we
present a multiple hurdle selection process where different fields are evaluated at different stages. The model suggests that an optimal evaluation implies that the evaluator will be more demanding at those dimensions where her knowledge is greater and, therefore, can evaluate with more accuracy and will be more benevolent at those fields where her knowledge is relatively more limited. As a consequence, candidates who excel in the same dimensions as the evaluator will have a greater probability of being selected as they will tend to do relatively better in those dimensions that are evaluated more harshly.

The exceptional structure of the public exams to become judges and prosecutors in Spain allows us to consistently estimate the predictions of the above model. Given the random allocation of candidates to evaluation committees and the different speciality of the university professors across committees we can estimate the effect of the committees’ knowledge profile on the evaluations received by candidates.

The empirical analysis reveal the existence of a “similar-to-me” effect in profile of skills. The presence, among evaluators, of a university professor who is an expert in the topic evaluated at a certain test is associated with this test being relatively more difficult. In particular, the grades are about 0.65 points lower when the professor member of the committee is an expert in the subject evaluated.

These results confirm the importance of selection committees being diverse and balanced in terms of the skills possessed by their members. Otherwise candidates whose knowledge profile is closer to the evaluators’ one will be favoured.
References


Figure 1. Sequence of the evaluation process
Note: we are assuming that the evaluator is more accurate when assessing stage 1 than stage 2

Figure 1: Figure 2. Probability of achieving grade c in each stage

Figure 3.a. Threshold required in each stage when b=0.5 and p=10%
Figure 3.b. Threshold required in each stage when b=0.5 and p=25%

Figure 3.c. Threshold required in each stage when b=0.5 and p=50%
Candidates take multiple choice test

Stage 1

Candidates take 1st oral test

Stage 2

Candidates take 2nd oral test

Stage 3

Nº of successful candidates known and committees are formed

If $c_1 > 12.5$

takes next exam. Else, rejected

If $c_1 > 12.5$

gets the job. If not, rejected

Figure 4. The timing of the public examination
<table>
<thead>
<tr>
<th></th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registered</td>
<td>4974</td>
<td>4732</td>
<td>4084</td>
<td>13790</td>
</tr>
<tr>
<td>Male</td>
<td>31,4%</td>
<td>31,8%</td>
<td>31,3%</td>
<td>31,5%</td>
</tr>
<tr>
<td>Lives in Madrid</td>
<td>21,8%</td>
<td>21,8%</td>
<td>21,1%</td>
<td>21,6%</td>
</tr>
<tr>
<td>Experience=1</td>
<td>77,5%</td>
<td>78,0%</td>
<td>82,3%</td>
<td>79,1%</td>
</tr>
<tr>
<td>Experience&gt;=2</td>
<td>60,0%</td>
<td>59,8%</td>
<td>65,7%</td>
<td>61,6%</td>
</tr>
<tr>
<td>Experience&gt;=3</td>
<td>48,0%</td>
<td>44,8%</td>
<td>48,7%</td>
<td>47,1%</td>
</tr>
<tr>
<td>Nº of candidates who took the test</td>
<td>85,6%</td>
<td>81,7%</td>
<td>83,4%</td>
<td>83,6%</td>
</tr>
<tr>
<td>Multiple choice test successful candidates</td>
<td>28,7%</td>
<td>43,4%</td>
<td>43,7%</td>
<td>38,2%</td>
</tr>
<tr>
<td>1st oral exam successful candidates</td>
<td>13,6%</td>
<td>27,3%</td>
<td>19,6%</td>
<td>20,0%</td>
</tr>
<tr>
<td>2nd oral exam successful candidates</td>
<td>29,9%</td>
<td>41,9%</td>
<td>39,1%</td>
<td>38,3%</td>
</tr>
</tbody>
</table>

Table 1. Descriptive data – candidates’ characteristics

<table>
<thead>
<tr>
<th>Committee</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Committee 1</td>
<td>Procedural</td>
<td>Procedural</td>
<td>Procedural</td>
</tr>
<tr>
<td>Committee 2</td>
<td>Procedural</td>
<td>Procedural</td>
<td>Procedural</td>
</tr>
<tr>
<td>Committee 3</td>
<td>Civil</td>
<td>Criminal</td>
<td>Procedural</td>
</tr>
<tr>
<td>Committee 4</td>
<td>Criminal</td>
<td>Criminal</td>
<td>Procedural</td>
</tr>
<tr>
<td>Committee 5</td>
<td>Criminal</td>
<td>Criminal</td>
<td>Procedural</td>
</tr>
<tr>
<td>Committee 6</td>
<td>Procedural</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Civil and Criminal Law are included in the first block of topics. Procedural Law is included in the second block.

Table 2. Field of expertise of the University Professor by year and committee
<table>
<thead>
<tr>
<th>Method of estimation</th>
<th>Grade of multiple choice test</th>
<th>Grade of 1st oral exam</th>
<th>Grade of 2nd oral exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>2.507262***</td>
<td>-.4322136**</td>
<td>.2120666</td>
</tr>
<tr>
<td></td>
<td>.3120593</td>
<td>.1980134</td>
<td>.4569611</td>
</tr>
<tr>
<td>Experience&gt;=1</td>
<td>9.840791***</td>
<td>.3601315</td>
<td>-.2332698</td>
</tr>
<tr>
<td></td>
<td>.4592692</td>
<td>.4146223</td>
<td>1.008028</td>
</tr>
<tr>
<td>Experience&gt;=2</td>
<td>4.287419***</td>
<td>.0259175</td>
<td>-.8155122</td>
</tr>
<tr>
<td></td>
<td>.4835482</td>
<td>.307512</td>
<td>.6917929</td>
</tr>
<tr>
<td>Experience&gt;=3</td>
<td>2.448574***</td>
<td>-.2476011</td>
<td>-1.073427**</td>
</tr>
<tr>
<td></td>
<td>.4097732</td>
<td>.2376584</td>
<td>.5324699</td>
</tr>
<tr>
<td>Lives in Madrid</td>
<td>-2.912037***</td>
<td>.0932183</td>
<td>.3715962</td>
</tr>
<tr>
<td></td>
<td>.3486699</td>
<td>.2326358</td>
<td>.528272</td>
</tr>
<tr>
<td>Multiple choice mark</td>
<td>-.1061171</td>
<td>1.047889**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.1444718</td>
<td>.5135562</td>
<td></td>
</tr>
<tr>
<td>Multiple choice mark^2</td>
<td>.0028869***</td>
<td>-.0052421*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.0009297</td>
<td>.0031833</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>58.14503***</td>
<td>-3.544344</td>
<td>-40.9405**</td>
</tr>
<tr>
<td></td>
<td>.3855815</td>
<td>5.608419</td>
<td>20.60468</td>
</tr>
<tr>
<td>Year Dummies</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Adjusted R-2/Pseudo R-2</td>
<td>0.2088</td>
<td>0.0546</td>
<td>0.0178</td>
</tr>
<tr>
<td>N</td>
<td>11535</td>
<td>5268</td>
<td>1334</td>
</tr>
<tr>
<td>Uncensored</td>
<td>1338</td>
<td>508</td>
<td></td>
</tr>
</tbody>
</table>

Note: The public exam consists of three qualifying tests: a multiple choice test and two oral tests. The grade of the first test ranges between 0 and 100. The grades of the oral tests are only publicly noticed to candidates who have passed the test and the grade ranges between 12.5 and 25. ***, **, * indicate significance at level 1, 5 and 10%, respectively.

Table 3. Determinants of grades in public examinations to become judges or prosecutors, years 2003-2005.
<table>
<thead>
<tr>
<th>Method of estimation</th>
<th>Grade of 1st oral exam</th>
<th>Grade of 2nd oral exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>-.4216027*** .188752</td>
<td>.1973175 .4563662</td>
</tr>
<tr>
<td></td>
<td>.4130829 1.006337</td>
<td></td>
</tr>
<tr>
<td>Experience&gt;=1</td>
<td>.3323369 -.191188</td>
<td>.4130829 1.006337</td>
</tr>
<tr>
<td></td>
<td>.3064834 .6904476</td>
<td></td>
</tr>
<tr>
<td>Experience&gt;=2</td>
<td>.0175474 -.8052902</td>
<td>.1973175 .4563662</td>
</tr>
<tr>
<td></td>
<td>.3064834 .6904476</td>
<td></td>
</tr>
<tr>
<td>Experience&gt;=3</td>
<td>-.2368191 -1.090424***</td>
<td>.2369021 .531686</td>
</tr>
<tr>
<td></td>
<td>.2319684 .5275331</td>
<td></td>
</tr>
<tr>
<td>Lives in Madrid</td>
<td>.1184353 .3553961</td>
<td>.1184353 .3553961</td>
</tr>
<tr>
<td></td>
<td>.2319684 .5275331</td>
<td></td>
</tr>
<tr>
<td>Multiple choice mark</td>
<td>-.0978699 1.052627***</td>
<td>.1440431 .5130884</td>
</tr>
<tr>
<td></td>
<td>.1440431 .5130884</td>
<td></td>
</tr>
<tr>
<td>Multiple choice mark^2</td>
<td>.0028312*** -.0052743*</td>
<td>.0028312*** -.0052743*</td>
</tr>
<tr>
<td></td>
<td>.1985976 .4535374</td>
<td></td>
</tr>
<tr>
<td>Expert in field evaluated</td>
<td>-.6439556*** -.6448508</td>
<td>-.6439556*** -.6448508</td>
</tr>
<tr>
<td></td>
<td>.1985976 .4535374</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-3.486254 -40.78771</td>
<td>5.591295 20.58595</td>
</tr>
<tr>
<td></td>
<td>5.591295 20.58595</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year Dummies</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted R-2/Pseudo R-2</td>
<td>0.0555</td>
<td>0.0182</td>
</tr>
<tr>
<td>N</td>
<td>5268</td>
<td>1334</td>
</tr>
<tr>
<td>Uncensored</td>
<td>1338</td>
<td>508</td>
</tr>
</tbody>
</table>

Note: The grades of the oral tests are only available for candidates who passed the test and range between 12.5 and 25. The committee is expert in the 1st oral exam if the professor is specialized in civil or criminal law. If the professor’s field of expertise is procedural law, the committee is expert in oral exam 2.

***, **, * indicate significance at level 1, 5 and 10%, respectively.

Table 3. The effect of committees’ field of knowledge on grades in public examinations to become judges or prosecutors, years 2003-2005.