Chapter 8
Selection of medical students on the basis of nonacademic skills: Is it worth the trouble?

This chapter was published as:
Abstract

In this chapter we discuss the practical usefulness of selecting future medical students on the basis of increasingly popular nonacademic tests (e.g., multiple mini-interviews, situational judgment tests) in addition to academic tests. Nonacademic tests assess skills such as ethical decision making, communication and collaboration skills, or traits such as conscientiousness. Although several studies showed that performance on nonacademic tests could have a positive relationship with future professional performance, we argue that this relationship should be interpreted in the context of the base rate (the proportion of suitable applicants in the applicant pool) and the selection ratio (the proportion of selected applicants from the applicant pool). We provide some numerical examples in the context of medical student selection. Finally, we suggest that optimizing training in nonacademic skills may be a more successful alternative than selecting students on the basis of these skills.
8.1 Introduction

There is an increasing interest in the selection of future medical students on the basis of nonacademic tests in addition to, or instead of, the traditional cognition-based or knowledge-based academic tests such as the Medical College Admission Test (MCAT) and the UK Clinical Aptitude Test (UKCAT). Nonacademic tests measure skills such as communication skills, professional behavior, and ethical decision making, or traits such as personality characteristics. An example of a nonacademic skills admission test is the multiple mini-interview (MMI; Eva et al., 2009; Eva, Rosenfeld, Reiter, & Norman, 2004). The MMI consists of a series of short, structured interviews and tasks where potential students show their interpersonal skills and ethical standards. Another example is the use of (video-based) situational judgment tests (SJTs) that contain social doctor-patient and doctor-colleague interactions in which applicants have to indicate how they would respond to a particular situation (Lievens, 2013; Patterson, Baron, Carr, Plint, & Lane, 2009). Other admissions tests, often in the form of self-report questionnaires, measure noncognitive traits such as empathy (Adam et al., 2015). In this chapter, we focus on nonacademic skills. However, it is not always clear whether the constructs measured by nonacademic tests should be considered skills, traits, or a combination of both. SJTs, for example, are more related to cognitive abilities when knowledge instructions are used (how should one act), and more related to personality traits when behavioral instructions are used (how would you act; Lievens, Sackett, & Buyse, 2009; McDaniel, Hartman, Whetzel, & Grubb, 2007).

The central idea of using these measures on top of academic measures like high school grade point average (GPA) and standardized test scores (e.g., MCAT scores) is that these measures improve the selection of future medical students. That is, through the use of nonacademic tests such as MMIs or SJTs, the applicants selected will perform better as a doctor than those who were selected on the basis of academic measures only. In 2015, Harris, Walsh, and Lammy were critical about the use of ‘non-knowledge-based tests and situational judgement tests to test desirable professional attitudes, such as empathy and ethical awareness’. Their main argument was that the validity of these tests to predict academic performance was low as compared to knowledge-based tests. Instead, they advocated using knowledge-based tests to select future doctors. Indeed, nonacademic tests should show predictive validity and incremental validity over academic tests to justify their use. However, using academic performance as a criterion may not be very relevant is this case, because the aim is not to predict academic performance, but doctor performance or professional performance. To
show incremental validity, these tests should be reliable instruments (i.e. test results should be consistent across replications), should show positive relationships with relevant criteria (such as professional performance), and should not be strongly related to academic test scores. However, very few studies address incremental validity of nonacademic tests over academic tests using such criteria (Adam et al., 2015 and Lievens, 2013 are exceptions). In addition, statistically significant incremental validity is not necessarily equal to practically relevant incremental validity. In this chapter, we go beyond incremental validity and discuss this topic from a utility approach that originates from the personnel selection literature. The advantage of this approach is that it shows the practical effects of using different selection instruments. Our main message is that in many medical student selection situations, the incremental validity and utility of the use of nonacademic instruments seems to be small and that the recent trend to use these instruments needs a more solid empirical basis.

### 8.2 Incremental Validity of Additional Selection Instruments

The effects of additional predictors above academic tests to select students will be largest when the correlation between the predictors is low. However, several studies have shown that nonacademic skills and academic, cognition-based measures are not independent. For example, in a recent study on medical student selection from the Netherlands (Schripsema, van Trigt, Borleffs, & Cohen-Schotanus, 2014) it was concluded that:

Top pre-university GPA students also achieved the highest possible score in the professionalism course most often. In this course, nonacademic variables such as interpersonal and communication skills, ethical decision-making, reflection and professional behaviour are assessed. The overall high performance of the top pre-university GPA group suggests that applicants who perform well academically might also have an advantage in the so-called nonacademic domain. (p. 1207)

Thus, in this course, students were judged on showing professional behaviour and their reflections on professional behaviour.

These findings are also in line with a study from the USA in which it was concluded that ‘the relationships between cognitive and noncognitive subdomains of the licensing examinations reported here ranged between $r = .17$ and $r = .43$ and correlations generally increased with trainees’ seniority’ (Eva et al., 2009). This increase with trainees’ seniority also reflected that nonacademic skills are, to a large extent, trainable. Note that the noncognitive skills mentioned in this study
are similar to what we consider nonacademic skills. In a more general (nonmedical) context, a meta-analysis found moderate relationships between SJT scores and cognitive ability, with a stronger relationship for tests that were based on job analysis and tests that provided knowledge instructions (McDaniel et al., 2007). Similar findings were reported in a medical context (Lievens et al., 2009).

Thus, the studies cited above showed the same pattern; namely that academic and nonacademic skills are positively correlated. However, nonacademic skills may still improve predictions based on only academic scores. For example, in a Belgian study, Lievens (2013) investigated the additional explained variance of a video-based SJT that measured interpersonal skills above high school GPA and knowledge test scores in a hierarchical regression analysis. The SJT had significant added value for predicting four outcomes: interpersonal GPA, an interpersonal skills assessment, doctor performance, and performance on a case-based panel interview, with additional portions of explained variance of 4.4%, 1.4%, 2.2%, and 3.4%, respectively. On the basis of these results it was concluded that ‘video-based SJTs as measures of procedural knowledge about interpersonal behaviour show promise as complements to cognitive examination components’. These percentages are in agreement with a meta-analysis (McDaniel et al., 2007) that found incremental validities of SJTs over cognitive ability of 3–5% with job performance as a criterion. However, what do these numbers mean in terms of practical relevance?

### 8.3 Practical Usefulness of Additional Selection Instruments

To evaluate the practical usefulness of a selection procedure we should not only take the correlations between the predictors and the criterion scores (predictive validity) into account, but also the selection ratio (the proportion of potential students who are selected), and the base rate (the proportion of potential students who would be successful without using tests; Taylor & Russel, 1939). The base rate is particularly interesting: Imagine that we would not select applicants, what proportion of the applicants would then show sufficient skills when they get their licensing degree?

There are models that show the interplay between selection ratio, base rate, and the relationship between predictor and criterion scores. The most popular model is the Taylor-Russell model (1939), which provides the success ratio – the proportion of admitted medical students who will perform well as doctors, for a given base rate, selection ratio, and predictive validity. This model allows us to compare the result of a selection procedure to the base rate (i.e. when there is no selection), or to compare the success ratios under different selection procedures.
The general message of this model is that when the base rate is high or the selection ratio is high, the effect of selection is low. This is easy to understand: if most applicants will be successful, or if almost all applicants will be selected anyway, selection will only have a small effect on the quality of the admitted students, regardless of the predictive value of the selection instruments.

8.4 An Illustration

How does this translate to the selection of future medical students? To answer this question, we need information about the selection ratio and the base rate. In general, medical schools are selective, and thus the selection ratio will be moderate to low. In the USA, for example, selection ratios range from .02 for some very selective medical schools to approximately .60. In the Netherlands, the selection ratio is around .60 (Niessen & Meijer, 2015). No data seems to be available for the exact base rate of successful doctor performance, but say we estimate the base rate at around .80. The estimated base rate is high for two reasons: (1) Students who apply to study medicine are not a random sample from the general population but are already strongly preselected on the basis of academic and nonacademic skills as a result of high school selection and training. (2) Students are trained in academic and nonacademic skills during medical education.

As a result of the interplay between base rate, selection ratio, and (multiple) correlation between predictor scores and criterion scores, we can determine the success ratio. The question is what the difference in the success ratio is when we use an academic test and a nonacademic test, in contrast with using an academic test alone. Instead of consulting the printed Taylor-Russell tables to investigate the added value of a given admission procedure, calculations can also be made using a web-application (McClellan, 1996). In Table 8.1, we show these percentages for three hypothetical base rates: .70, .80, and .90, and four hypothetical selection ratios: .60, .30, .10, and .05. These selection ratios represent moderate selectivity, like the Dutch context, strict selection, like in France, where approximately 20% of the students are selected for the second year of medical school after one year of studying, and very strict selection as often found in the USA and the UK. For this illustration we used the results described in Lievens (2013) since results about the incremental validity of a nonacademic test over academic tests were reported, and suitable criterion measures were used. In this study, knowledge-based tests showed a correlation of \( r = .07 \) with doctor performance. Adding an SJT resulted in an increase in explained variance of \( \Delta R^2 = .022 \), yielding a multiple validity coefficient of \( R = .16 \). Note that this correlation is much lower than the often reported correlations between academic test scores and academic performance in medical school. For example, in a study from the USA (Julian, 2005) median
correlations of $r = .46$ were reported between MCAT scores and mean grades over the first 3 years. The low correlation is probably due to the more heterogeneous criterion of doctor performance as compared with the criterion of school grades, although the definition of ‘doctor performance’ was unclear in Lievens’s (2013) study. To investigate the utility of these findings, we determined the success ratios when the (multiple) validity coefficient increased from .07 to .16.

Table 8.1
**Success ratio as a function of the base rate (BR), selection ratio (SR) and predictor-criterion correlation.**

<table>
<thead>
<tr>
<th>SR</th>
<th>$R = .07$</th>
<th>$R = .16$</th>
<th>$\Delta%$</th>
<th>$R = .07$</th>
<th>$R = .16$</th>
<th>$\Delta%$</th>
<th>$R = .07$</th>
<th>$R = .16$</th>
<th>$\Delta%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.60</td>
<td>71.6</td>
<td>73.6</td>
<td>2.0</td>
<td>81.3</td>
<td>82.9</td>
<td>1.6</td>
<td>90.8</td>
<td>91.9</td>
<td>1.1</td>
</tr>
<tr>
<td>.30</td>
<td>72.8</td>
<td>76.3</td>
<td>3.5</td>
<td>82.2</td>
<td>85.0</td>
<td>2.8</td>
<td>91.4</td>
<td>93.1</td>
<td>1.7</td>
</tr>
<tr>
<td>.10</td>
<td>74.2</td>
<td>79.2</td>
<td>5.0</td>
<td>83.3</td>
<td>87.2</td>
<td>3.9</td>
<td>92.0</td>
<td>94.3</td>
<td>2.3</td>
</tr>
<tr>
<td>.05</td>
<td>74.9</td>
<td>80.6</td>
<td>5.7</td>
<td>83.8</td>
<td>88.2</td>
<td>4.4</td>
<td>92.3</td>
<td>94.8</td>
<td>2.5</td>
</tr>
</tbody>
</table>

*Note. Combined $R^2 = .07^2 + .022 = 0.027$, combined $R = .16$; $\Delta\%$ denotes the differences in more successful decisions when the multiple correlation increase from $R = .07$ to $R = .16$ (or equivalently $\Delta R^2 = .022$).*

In Table 8.1, it can be seen that for a selection ratio of .60 and a base rate of .80, the selection procedure with only an academic test will result in an additional 1.3% successful students compared with the base rate. Now, suppose that through the use of an additional nonacademic test the predictive validity (multiple correlation) increases from .07 to .16. The reader can see in Table 8.1 that the percentage of successful students increases by 2.9% compared to the base rate. However, note that the difference compared to using just the academic test is only 1.6% (1.3% versus 2.9%). For a base rate of .90 this percentage is around 1.1% and for a base rate of .70 it is approximately 2.0%. Even for a selection ratio of .30, an increase in $R^2$ of .022 only leads to a difference of 2.8% using a base rate of .80, 1.7% for a base rate of .90 and 3.5% for a base rate of .70. Thus, when we assume that 80% of medical school applicants would be successful in their medical job, using this SJT only provides a modest increase in the percentage of successful doctors. Only when selection is strict and base rates are not very high, using this additional
nonacademic test yields a larger, but still modest, increase in successful doctors (4.4% for a base rate of .80 and a strict selection ratio of .05).

Given that tests like MMIs and SJTs are complex instruments to develop and do require extensive resources, one may ask whether these extra resources pay off in practical student selection. For example, let us consider the selection of future medical students in the Netherlands. In 2015 there were 2,785 places available in eight different Dutch medical schools. With a base rate of .80 and a selection ratio of .60, this would yield 2,264 successful doctors with only academic selection, and 2,309 successful doctors with academic and nonacademic selection. The difference between selection with an academic test only and selection using both tests is 45 more successful doctors when both tests are used. In the Netherlands, these 45 extra successes are divided between eight medical schools, so on average this would result in approximately six extra successes per school. This number is small, whereas the costs of selection are often not trivial. Each medical school may take these numbers and the costs into account when deciding to use nonacademic tests.

A counterargument is that, although there may be few future medical students who are unsuited for the medical profession, it is still very important to detect these few students because they may bring serious harm to patients and to medical institutions. This is true, but we should show how well instruments like nonacademic tests meet this goal based on analyses as shown earlier. Following our example, let us assume that 20% of the medical school applicants are not suitable to be doctors. Using an SJT in addition to an academic test would reduce the proportion of unsuitable doctors within the population of doctors by between 1.6% and 4.4%, depending on the selection ratio.

There are two other arguments why colleges, in our view, should take calculations like the ones provided above into account in selecting students on the basis of nonacademic skills. The first is that the environment also determines professional behaviour (e.g., Johns, 2006). The second argument is that many (or at least some) nonacademic skills are trainable (Hook & Pfeiffer, 2007) and that practice and experience is very important. In fact, a reasonable question is: What is more effective, selecting students based on nonacademic skills, or improving the curriculum and working conditions with respect to teaching and supporting these skills?

8.5 Discussion
Medical school admissions need not be based on predicted academic performance or even in improved medical job performance. Medical schools can opt to select
future medical students on the basis of particular talents (e.g., leadership, wanting to work in certain domains, etc.) or on the basis of school diversity. For example, some schools use standardized test scores as diagnostic tools to predict which students need extra help in their studies (Julian, 2005). However, it is important to realise that when predictors are chosen on the basis of the strength of the relationship with a criterion, the base rate and the selection ratio play a crucial role in the practical outcomes. Because instruments like SJTs and MMIs are psychometrically more complex, more expensive to develop, and sometimes less reliable than academic instruments (McDaniel et al., 2007), the added value may be modest compared to the costs. We illustrated this in the context of medical school selection, but our illustration does not necessarily show a general result in terms of the utility of nonacademic tests. Selection ratio and base rate may differ across medical schools and across countries. So, there may be situations where it pays off to add nonacademic skills tests to academic tests.

Besides utility models that show increments in success ratios, there are utility models that can be used to estimate the economic consequences of implementing selection instruments (Brogden, 1949; Cascio, 1980; Cronbach & Gleser, 1965). However, it is not easy to illustrate these models given the uncertainty of how to estimate some of the parameters, such as the standard deviation of economic gain from increased performance (Holling, 1989 provides some solutions). In addition, the costs of educating a student who is unsuitable for medical practice is also difficult to determine and will vary considerably across countries and universities (e.g., Mellenbergh, 1995). An economic utility analysis is complex and beyond the scope of this chapter, but we encourage decision makers to explore the applications of these models to their admission procedures.

Another argument for selecting future students based on nonacademic skills is that it may result in self-selection, leading to applicants who perhaps fit the profile of a future successful doctor better. If this would be the case, then the base rate would change as a result of the selection procedure. An implicit assumption under the Taylor-Russell model is that the base rate remains constant regardless of the selection procedure in place. However, there are some indications that this assumption may be false (Wouters, 2017). In that case, applying the Taylor-Russell model would yield an underestimation of the increase in the success ratio.

We would like to stress that we are not against selecting future medical students on the basis of nonacademic skills, but we would argue for future research that is aimed at the incremental validity and utility of these instruments in the context of base rates and selection ratio. As we showed, statistically significant incremental
validity is not necessarily practically relevant. In addition, in future research, much more attention should be paid to the criterion variables. Doctor performance is a complex variable that is not taken into account in many studies or is not operationalized clearly. For example, in Lievens's (2013) study, students following a career in general practice were studied. It may be the case that nonacademic skills like social skills are more important for this specialty than for other medical specialties.