hudson

Whitepaper:
The Value of Assessment Tools in Personnel Selection

Amelie Vrijdags

February 2022

Table of Contents

1. Introduction

One of the main reasons why HR professionals use assessment tools and tests, is to get an informed idea of how well a candidate will perform in the job they are hiring for.

The **correlation between test scores and work performance** is termed "criterion validity", or known popularly as "predictive validity". The higher this correlation, the better the tool is capable of predicting how well a candidate would perform on the job. Therefore, it makes sense to choose tools with a high predictive validity when designing a selection procedure. Although, predictive validity is not the only important factor as other elements like candidate experience, equal opportunities, and of course cost and practical feasibility are also important.

At the end of 2021, a scientific paper (Sackett et al., 2021) was published that shook up the scientific community, and in extension also many HR professionals involved in personnel selection. This paper by Paul Sackett and colleagues indicates that most of the things we measure or assess during the selection process with the aim of predicting future work performance, do not predict performance as well as we previously thought.

What can we now learn from this? In light of these new findings, which assessment tools are now the most interesting to use, for selection purposes in our own day-to-day practice?

To really understand which lessons to draw from this meta-analysis, Hudson's R&D team has thoroughly analysed Sackett et al.'s new paper, combined its findings with other high-quality research and discussed this in depth with prof. dr. Filip Lievens, world authority in the field of selection and assessment research and one of the co-authors of the paper. We would like to share our insights in the next sections.

2. Summary paper Sacket et al.

2.1 Schmidt & Hunter's classic 1998 study

Let us start from the beginning with the renowned meta-analysis by Schmidt & Hunter (1998). Since its publication, this study has been considered by the HR world as the 'go-to' reference for evidence-based knowledge about the validity of various selection tools.

Schmidt and Hunter (1998) conducted an extensive meta-analysis, aggregating no less than 85 years of research in personnel selection. This resulted in a listing of the average validity of 19 selection procedures from most to least predictive of job performance. The primary conclusion from this meta-analysis was that General Mental Ability (GMA), measured by cognitive tests, was by far the best predictor. And secondly, that most other selection procedures correlated rather strongly with GMA. Which means that only small increments in predictive value were to be expected when combining other procedures with cognitive ability measures.

In the last decade, however, several alternative meta-analyses appeared that questioned and even refuted some of Schmidt and Hunter's (1998) conclusions, (see e.g. Van Iddekinge et al. (2012), and Sackett et al. (2017)).

2.2 The over-correction problem

The current paper by Sackett and colleagues also shows that Schmidt and Hunter's conclusions need to be revisited.

The authors very convincingly prove that the statistical techniques most commonly used in meta-analyses (including the one by Schmidt & Hunter, 1998) often result in (gross) overestimations of 'operational validity'. This operational validity is an estimate of the 'true' correlation between the test scores and work performance, after correcting for artefacts like unreliability of the criterion and range restriction. Usually, researchers tend to overcorrect in meta-analyses, which results in validity estimates that are too high.

Sackett et al (2021), however, use the principle of **conservative estimation**: when there is any uncertainty about how to correct, they rightfully claim it is better to not correct at all than to overcorrect. So the statistical techniques used by Sackett et all. (2021) are more likely to result in a (slight) underestimation of the true validity of selection techniques, rather than an overestimation.

2.3 The new and improved predictive validity estimates

To demonstrate the impact of overcorrecting as it occurred in most meta-analyses, Sacket and colleagues decided to re-analyse Schmidt & Hunter's (1998) landmark study, **using more correct methods** and adding, where available, **more recent research results** to the analyses. Also, **newer selection methods** that were not yet studied at the time of Schmidt & Hunter's paper, were added to the analysis, like, e.g. Emotional Intelligence and Situational Judgment tests. This resulted in 'new' validity estimates that shed a somewhat different light on what was hitherto believed.

The most striking change, compared to the original paper, is that **cognitive ability or GMA measures no longer sit at the top of the list**, but are replaced **by structured interviews** (not unstructured interviews) **as the 'best' predictor of job performance.**

This does not imply that cognitive ability measures have become irrelevant or superfluous, however, we might have overestimated their predictive powers somewhat. Table 1 shows the 'old' and 'new' top five of best selection methods.

Schmidt & Hunter's (1998) top five	Old validity estimate	Sackett et al.'s (2021) top five	New validity estimate
Work samples	.54	Structured interviews	.42
Cognitive ability	.51	Job knowledge	.40
Structured interviews	.51	Empirically-keyed biodata	.38
Job knowledge	.48	Work samples	.33
Integrity	.41	Cognitive ability & integrity	.31

Table 1: Top five strongest predictors of job performance based on operational validity estimates, according to Schmidt & Hunter (1998) and Sackett et al. (2021)

The new top five has a tie in fifth place between cognitive ability and integrity tests. Of the top five predictors from Schmidt and Hunter (1998), all five are again present in the new top five, supplemented with empirically keyed biodata, which was not part of the old top five. So there remains **considerable similarity between prior and current estimates in terms of what rises to the top of a list of the strongest predictors**. However, the magnitudes of the validity estimates differ as the mean across Schmidt and Hunter's top five was .49, while the mean across Sackett et al.'s top five is .37. The **average predictive validity is therefore somewhat lower than previously assumed**.

Figure 1: Operational validity estimates from Sackett, Zhang, Berry & Lievens (2021).

Figure 1 shows Sackett et al.' s entire ranking of selection procedures according to their operational validity estimates.

Notice at the bottom that just counting **the years of relevant work experience has very little, if any, predictive power**. This goes against the popular belief that the more experience you have, the better you will perform on the job.

Interesting newcomers are personality-based emotional intelligence and situational judgment tests, which both perform rather well in terms of predictive power, as they both fall in the top half of the ranking.

.

3. Implications for our day-to-day selection practice

What do these new findings mean for our day-to-day practice in applicant selection? Should we simply use only the selection tests that come out of this study as highly valid and stop using the tools that are lower in the list?

Well, not exactly, the current meta-analysis covers an extremely wide variety of selection tools and jobs, that are all lumped under the same headings. Therefore, the conclusions of Sackett et al. (2021) are **not a 'one size fits all'**. In a meta-analysis of this size it would be very cumbersome to draw the sort of fine-grained conclusions that are useful for a specific practice. So to really understand which lessons to draw from this meta-analysis, we looked at its results from our perspective, the day-to-day selection practice in Europe/Benelux. This is explained in more detail in the sections below.

3.1 The utility of (modest) predictive validity

In summary, the paper by Sackett et al. (2021) confirms what most seasoned HR professionals have known for a very long time: predicting is hard.

The prediction of work performance with the selection methods under investigation might be less accurate than we used to think, based on Schmidt & Hunter's (1998) paper. However, it is still very much worth the effort to invest in these tools, because **hiring underperforming employees costs organisations a lot of money**.

The utility of a method is the extent to which using that method increases the quality of the selected employees in comparison to a situation where the method is not used. This utility is directly related to predictive validity: the better the predictive validity, the higher the average added value of the selected applicants. Research shows that most employees produce more value than they cost. On average, employees create double the value of their salary. But the difference between high and low performing employees in terms of productivity amounts to 80% of their salary per year (Buyens et al.).

For example, at a salary of €35.000, the difference in productivity between high and low performers amounts to €28.000, per year. By replacing a selection method with no or low validity (like years of experience) with a technique with modest proven validity, the chances of hiring a candidate with sufficient added value will increase substantially.

The new **estimates might seem rather on the low side, but in fact they are not**. To put them into context, we could compare the validity estimates with those of well-known medical procedures that are being administered on a large scale worldwide.

Medical procedure and outcome	Validity	Sample size
Coronary artery bypass surgery and survival at 5 years	0.08	2649
Antihistamine use and reduced runny nose and sneezing	0.11	1023
Effect of NSAID (e.g. Ibuprofen) on pain reduction	0.14	8488
Nicotine patch and smoking abstinence	0.18	5098
Viagra and improved male sexual functioning	0.38	779

 shows the correlation between selected medical procedures and their outcomes. One can see that these validities are a lot lower than most of the estimates in Sackett et al.'s paper. For example, the association between cognitive ability tests and work performance (average validity of .37) is much stronger than the effect of ibuprofen on pain reduction (average validity of .14).

Medical procedure and outcome	Validity	Sample size
Coronary artery bypass surgery and survival at 5 years	0.08	2649
Antihistamine use and reduced runny nose and sneezing	0.11	1023
Effect of NSAID (e.g. Ibuprofen) on pain reduction	0.14	8488
Nicotine patch and smoking abstinence	0.18	5098
Viagra and improved male sexual functioning	0.38	779

Table 2: Examples of the strength of relationship between commonly administered medical procedure and their outcome in terms of the correlation coefficient, which can be considered the validity of that procedure. Taken from Meyer et al. (2001)

Being able to predict human behavior with such high accuracy is rather fantastic if you look at it from the perspective of a medical researcher. But in HR, unfortunately, we got used to the unrealistically inflated estimates that were common in the literature, which makes the new estimates pale in comparison.

3.2 Large variation in predictive value

The operational validity estimates from Sackett et al. (2021) in Figure 1 **cover a large variety of studies, tools and work settings**. It is important to realise that the validity figures reported in meta-analyses are not 'exact' numbers. The values reported are averages, calculated from the validities found in a wide variety of primary studies. In fact, the predictive value of any given selection method varies substantially across studies.

Fortunately, to get an idea of the degree of variability, Sackett and colleagues also reported the **credibility intervals** for each of their estimates. These credibility intervals are visually represented in Figure 22**.** These graphs show that for most selection methods, the variation across primary studies is quite large as the credibility intervals are rather wide. And logically: the more variation, the less reliable the mean estimated validity.

For example, while structured interviews have the highest mean validity in the paper, they also have a very wide credibility interval. This means that an interview's predictive power varies strongly across settings, as it will depend on a myriad of factors: the specific questions asked, the scoring method, question quality, the match (or lack thereof) between questions asked and job characteristics.

And of course a lot will depend on the interviewer: their level of experience, how well were they trained, how effective they are in preventing any personal biases from clouding their judgment and so on. Even among (automated) cognitive ability tests there is significant variability in how well they predict job performance. This also makes sense, as no two cognitive tests are the same, and not all jobs require the same level of cognitive ability.

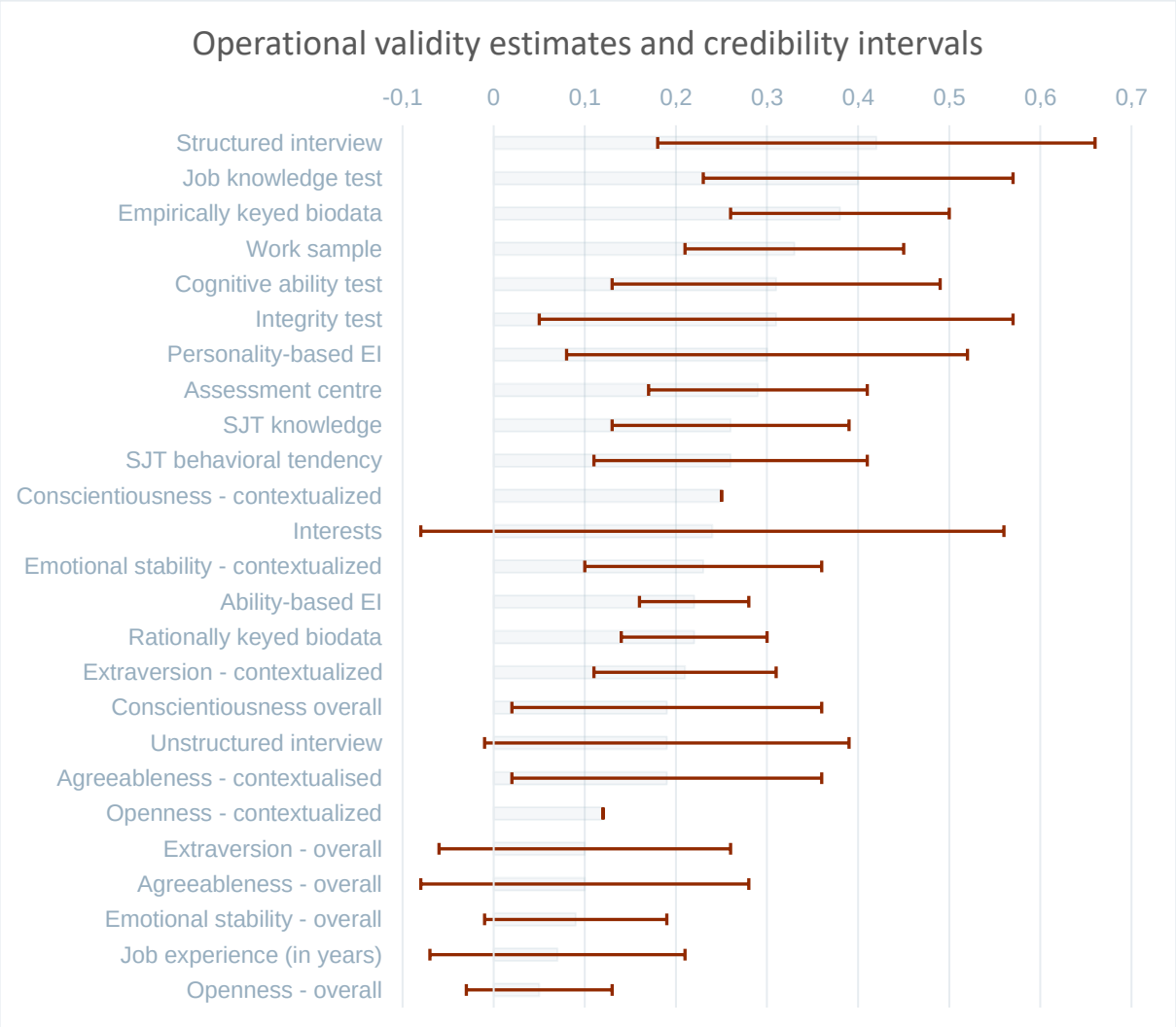

Figure 2: Operational validity estimates and their 80% credibility intervals are taken from Sacket et al. (2021). The bars represent the 80% credibility intervals around the mean operational validities. Two selection techniques (contextualised Conscientiousness and Openness) have no credibility intervals, as their estimated validities are based on one study only.

Figure *2* shows that there is some kind of ranking possible: some tools do seem to predict better than others, but the **ranking is not very precise if you look at the rather high levels of variation in predictive power across settings**. With such a big margin of error it does not seem that useful to focus on exact numbers and rankings. Most of these estimates are not significantly different from one another, there is a big overlap in the credibility intervals around the estimates. So can we really say that, based on these numbers, structured interviews are better predictors than work samples? Or that personality-based Emotional Intelligence is a better predictor than a Situational Judgement Test (SJT)?

Meta-analyses aggregate the conclusions of many separate studies, and conclusions are generalised over all jobs, situations, types of tools within each category, etc. **In order to get a reliable estimate of the predictive validity of a tool in a *specific* context, one should search for more specific studies.** Ideally, test publishers and HR professionals have to continually construct evidence-bases for their own selection activities, in order to evaluate the deployed methods dynamically in their specific contexts. However, high-quality validity research is not straightforward to carry out and requires a lot of resources. Unfortunately, this means that for many organisations, all they can actually use are 'somewhere in the ball-park' meta-analytic validity estimates.

For several frequently used **Hudson tools**, however, we do have multiple *specific* predictive validity studies available. Hudson's validity estimates are either not or conservatively corrected, as recommended by Sacket et al. (2021). Moreover, we can and do carry out extra analyses on specific client populations on request.

3.3 Can we generalise from low stakes research conditions to high-stakes selection conditions?

Another concern that is raised by the authors themselves, is whether applicant-based predictive (high-stakes) studies and incumbent-based concurrent (low-stakes) studies estimate the same thing.

Applicant selection is a high-stakes situation: whether or not the applicant gets the job depends on their performance in the selection procedure. However, the very **large majority** of primary studies reviewed by Sackett et al. (2021) **are incumbent-based concurrent studies, conducted in low stakes situations typical for research settings**: participants are incumbents, their job does not depend on their performance in the study.

We can assume that factors like **faking and social desirability** play much less of a role in low-stakes concurrent studies, and lower validity would be expected in (high-stakes) settings where faking is more prevalent. It is for example known that especially instruments like job interests questionnaires, biodata questionnaires and self-report integrity questionnaires are highly susceptible to faking and social desirable answering when administered in high-stakes selection situations.

Also, **the level of effort** may be lower when incumbents are confronted with difficult-to-solve test items in a low-stake research setting as compared to applicants, who we can assume are much more motivated to give their best effort in pursuit of a job of interest. One can imagine that the lower level of effort in incumbent-based studies has a negative impact on the validity estimates of cognitive tests and other maximal performance tests.

3.4 Can we generalise validity estimates to today's world of work?

Many of the primary studies reviewed in Sackett et al. (2021) are old to very old, many are pre-1970s, and the **majority is from before the year 2000**. With the introduction of the internet and many other new technologies, the work floor has changed a lot in the last few decades. Nowadays, employment opportunities increasingly lie in jobs requiring higher-level social or analytical skills, or both. Moreover, selection practices have changed tremendously over the last decade by incorporating more technology with a **heightened focus on competencies of the 21ste century**. Can we confidently generalise findings from the previous century to the current one?

One the other hand, we should not conclude that older research results like these have no value at all in today's world. The validity estimates quantify the *relationship* (correlation) between two variables (selection tool results and work performance), and not the separate variables themselves. So it is not because our idea of what constitutes good work performance may have changed tremendously over the last decades, that the relationship between work performance measures and selection test results has changed accordingly. Nevertheless, we can doubt whether it is still relevant to know that work sample tests of the previous century predicted the way of working at that time.

3.5 Variations in the predicted "Job performance"

Besides large variations in the specific tools studied and work settings covered, there are also **large variations in what is being predicted** in the primary studies.

"Job performance" is the criterion that is predicted in all studies, but studies vary substantially in the way in which job performance measured. It is not so straightforward to verify whether an employee, after being hired, performs well or not. In some jobs (especially high-complexity jobs) performance can be very tricky to measure. Logically, the **validity of a selection method will depend a lot on the way in which the criterion (i.e. job performance) is measured**.

Job performance can be measured with so-called objective criteria (such as production or sales numbers, number of client complaints, absenteeism, et cetera) or using subjective criteria, i.e. performance as evaluated by the employee's supervisor. Ideally, both objective and subjective criteria are used to assess performance, because they complement one another. This becomes clear when we look at the **low correlations (from 0.20 to 0.39)** that are usually observed **between objective job performance criteria and subjective supervisor ratings** (Lievens, 2020).

The vast majority of the primary studies reviewed by Sackett et al. (2021) uses (subjective) supervisory ratings as criteria. However, supervisory ratings of job performance do not reflect the same thing across studies, which makes it difficult to directly compare validity estimates across predictors, as also stated by the authors of the paper. The authors explain that if raters focus primarily on **task performance**, they will give a different evaluation than if asked to evaluate **broader performance**, including citizenship or counterproductivity components of performance.

A famous example in this respect is an earlier study by Sackett et al. (2017), who found that validity studies for cognitive ability tests tend to focus on task performance only, while a much broader definition of performance is used when the validity of assessment centres is studied. Therefore, in their meta-analysis they included only studies in which a cognitive ability test and an AC were administered to the *same* individuals, and both measures are correlated with the *same* (broad) job performance criterion. The results were surprising, as assessment centres (mean validity = .44) produced a much higher predictive validity than cognitive ability tests (mean validity = 0.22). This goes to say that **when selection methods are compared on the basis of exactly the same job performance criterion, the findings can be completely different**. Therefore, Sackett and colleagues warn us that we should be very cautious when comparing validity estimates across meta-analyses, as we do not have clear understanding of the specific components underlying performance ratings.

3.6 The European view

The study of Sackett et al. (2021) is an international meta-analysis, which draws on a lot of primary studies that were conducted in the US. **But not all of these studies are relevant in a European/Benelux selection context.** To be able to deploy Sackett et al.'s results for optimising selection and assessment processes in our context, it is useful to have a look at the findings from a European angle and 'translate' them to the European context. In this section, we explain in detail how to interpret some of the tools and accompanying results included in Figure 1 if your organisation operates in a European context.

The study includes **a number of tools that are rarely used in Europe** for selection purposes. These tools could be left out of the overview for the European context. For other tools, the **terminology used is not commonly known in Europe**, **or could be interpreted differently**. For the European context it therefore makes sense to rename those tools, so it becomes more clear what type of test it is and how it is used in Europe.

- **Work sample**. There are two kinds of work samples: (a) *psychomotor tests*, which are used to assess physical actions in lower complexity jobs. Performance is then scored on the basis of e.g. the time needed, or the number of mistakes made, and (b) work samples where candidates do not have to perform any physical actions, but they need to take decisions, negotiate or debate. These work samples can be called "*simulation exercises*" and are often used in an Assessment Center. Role plays, oral presentations, case studies and group discussions are all examples of the second type of work samples. (Lievens, 2020).

    ➔ *European view:* replace the name 'work samples' by 'simulation exercise', as these are the most commonly used and known works samples in our selection context.

- **Biodata questionnaires** probe behaviours and events that occurred early in one's life and that are proven to correlate with work performance. Typical biodata questionnaire items ask about things like the amount of "parental warmth" received when growing up, or the number of hours spent studying as a student. Although these aspects may be predictive of work performance, such questions are highly fakeable and most candidates in Europe will experience them as a serious breach of their privacy (Lievens, 2020). Therefore, these types of tests are therefore hardly ever used over here. We usually focus on the job-relevant tools only, even if by doing so we might miss out on some highly predictive information.

    ➔ European view: leave Biodata questionnaires out of the overview, as these are rarely used in Europe, for privacy reasons

- **Integrity tests**. Self-report integrity questionnaires are so transparent that they are found to barely work in high-stakes selection contexts, when you can win or lose a job based on your answers (e.g. Van Iddekinge et al., 2012). Similar to biodata questionnaires, integrity tests are also highly fakeable. Socially desirable responding is more of a problem here in Europe than in the US: Europeans will e.g. not easily admit that they have once stolen from an employer (a question typically asked in an 'overt' integrity questionnaire). In the US, however, before taking part in a selection process, candidates are often required to sign an 'honesty contract' which formally states that any answers given are truthful and honest. And, the legal consequences for breaching such a contract can be quite severe. The consequences of faking or dishonesty are much more serious in the highly legalised US context than in Europe, which results in a higher predictive power for easily fakeable tests. But unfortunately, these types of tests simply do not work over here.

    ➔ *European view:* leave self-report integrity tests out of the overview, as these do not work in high-stakes selection contexts

- **Non-contextualised personality questionnaires.** These are generic personality questionnaires that do not specify that the items refer to the work context only. Most people experience their own personality as somewhat different at work as compared to their private life. Therefore, candidates could get confused and have difficulties answering the questions (should they be thinking about their work context or their personal life when answering). Non-contextualised personality questionnaires consequently perform worse in terms of predictive power, than clearly work-related questionnaires, as is also shown in Figure 1. Additionally, non-contextualised personality questionnaires might seem to ask for rather private information with no clear link to the work context, creating also a feeling that this breaches the candidate's privacy. Therefore, when personality questionnaires are used for selection in Europe they are mostly contextualised, as it is specified that the frame of reference is the work context.

    ➔ *European view*: leave out the non-contextualised personality questionnaire dimensions, as these are generally not used for selection purposes.

- **Job interests questionnaires**: Are almost never used in high-stakes selection as they are very transparent and easily fakeable. Not many candidates would admit that they are not really interested in the job they are applying for. With the same remarks as above, people in Europe tend to fake more easily, when this impacts their chances of getting hired, as getting caught faking has less severe consequences than in the US.

    ➔ *European view:* leave out the job interest questionnaires, as these do not work over here in a high-stakes selection context

A few **other selection tools** reported in Figure 1 might **need some more background information** on the specifics to be able to understand the results:

- **Personality-based Emotional Intelligence** or "mixed Emotional Intelligence" (e.g. Joseph et al. 2015) should not be confused with much better known ability-based Emotional Intelligence tests that consider emotional intelligence to be an actual ability or facet of intelligence. Personality-based emotional intelligence is in fact a **personality-based compound** that encompasses a constellation of personality traits, affect and self-perceived abilities rather than aptitude. It is indeed in the combination of different relevant personality traits (in this case labelled 'Emotional Intelligence') that we can often see whether someone is fit for a job or not.

  ➔ European view: rename Personality-based Emotional Intelligence (EI) to "Personality-based compound (EI)", to avoid confusion with other ability-based EI tests, and to stress the fact that this is a compound of personality-type questionnaire results.

- **The "Big Five" personality factors Conscientiousness, Emotional Stability, Extraversion, Agreeableness and Openness** are separate dimensions of personality. The validity estimates reported in Figure 1 indicate the predictive power of these dimensions for work performance in general, i.e. for all types of jobs. However, while the role of intelligence is comparable for many jobs, the role of personality traits varies a lot from job to job. This means that while the validity of any dimension may be very low for some jobs, it is much higher for other jobs. Openness, for example, was shown to be a good predictor for jobs requiring a high level of creativity and adaptability, while Agreeableness is a good predictor of success in customer service jobs (Lievens 2020). It is thus rather striking that the validities of the Big Five personality factors for all jobs are still rather high in this meta analysis.

➔ *European view:* add “all jobs” to the personality factor results, to clarify that this is the predictive power of that one factor for any job. The validity of individual personality dimensions can be higher if we match them with specific jobs for which that personality aspect is highly important.

— **Assessment Centre:** Assessment Centre (AC) results are not often used for large groups of candidates. On the contrary; ACs are mostly only used for ‘selecting in’: meaning that usually only the final candidate takes part in an AC to decide whether they will get the job or not. ACs are hardly ever used for ‘selecting out’: where the entire applicant group takes an AC to filter out (select-out) the weakest candidates. This means that there is a larger restriction of range than for ACs than for other selection tools, like for example knowledge tests which are often used as a pre-selection tool for selecting out purposes. The weaker candidates will hardly ever be in the pool of candidates taking an assessment centre. So the fact that the validity estimates of Sackett et al. (2021) show that even within that pool of top candidates, ACs allow us to differentiate between high performers and less strong performers, makes these results particularly favourable for ACs as a tool for selecting in. If ACs were to be used for selecting out purposes and larger groups of candidates would take an AC, the validity estimates would be even higher. Unfortunately, the number of such studies is rather small, and therefore they don’t carry much weight in meta-analyses.

➔ *European view:* add “(select-in)” to the label Assessment Centre, to stress the fact that the validity estimate was calculated on a very restricted pool of candidates.

Integrating all this information into Figure 1 resulted in Figure 3, which is an adaptation of Sackett et al.’s ranking of operational validity estimates to the European/Benelux context.

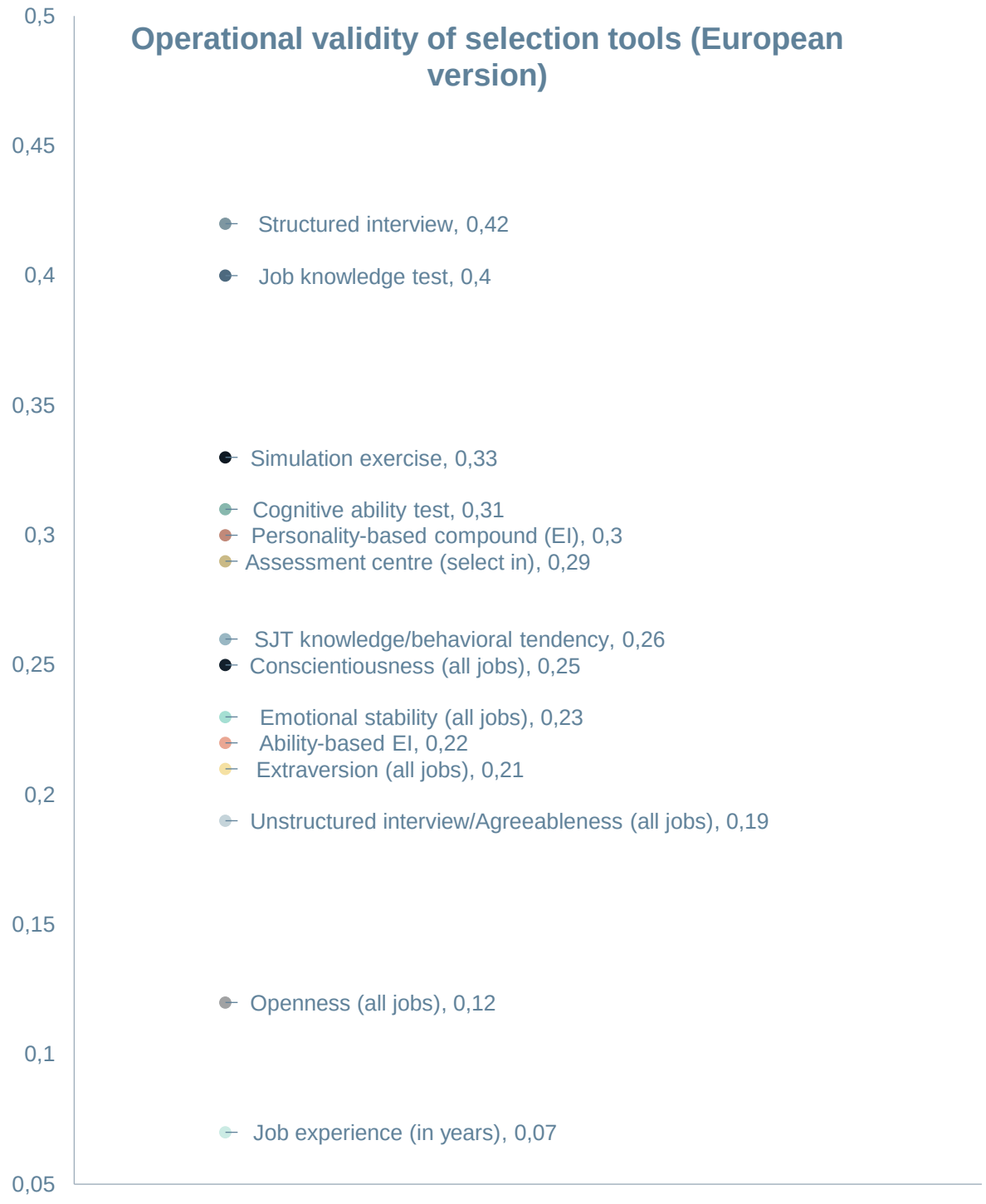

Figure 3 Operational validity estimates from Sackett, Zhang, Berry & Lievens (2021), updated to the European context

When we also apply this to Figure 1, showing also the variability in predictive value, this results in Figure 4Figure 3, which is an adaptation of Sackett et al.'s ranking (including the credibility intervals) to the European/Benelux context.

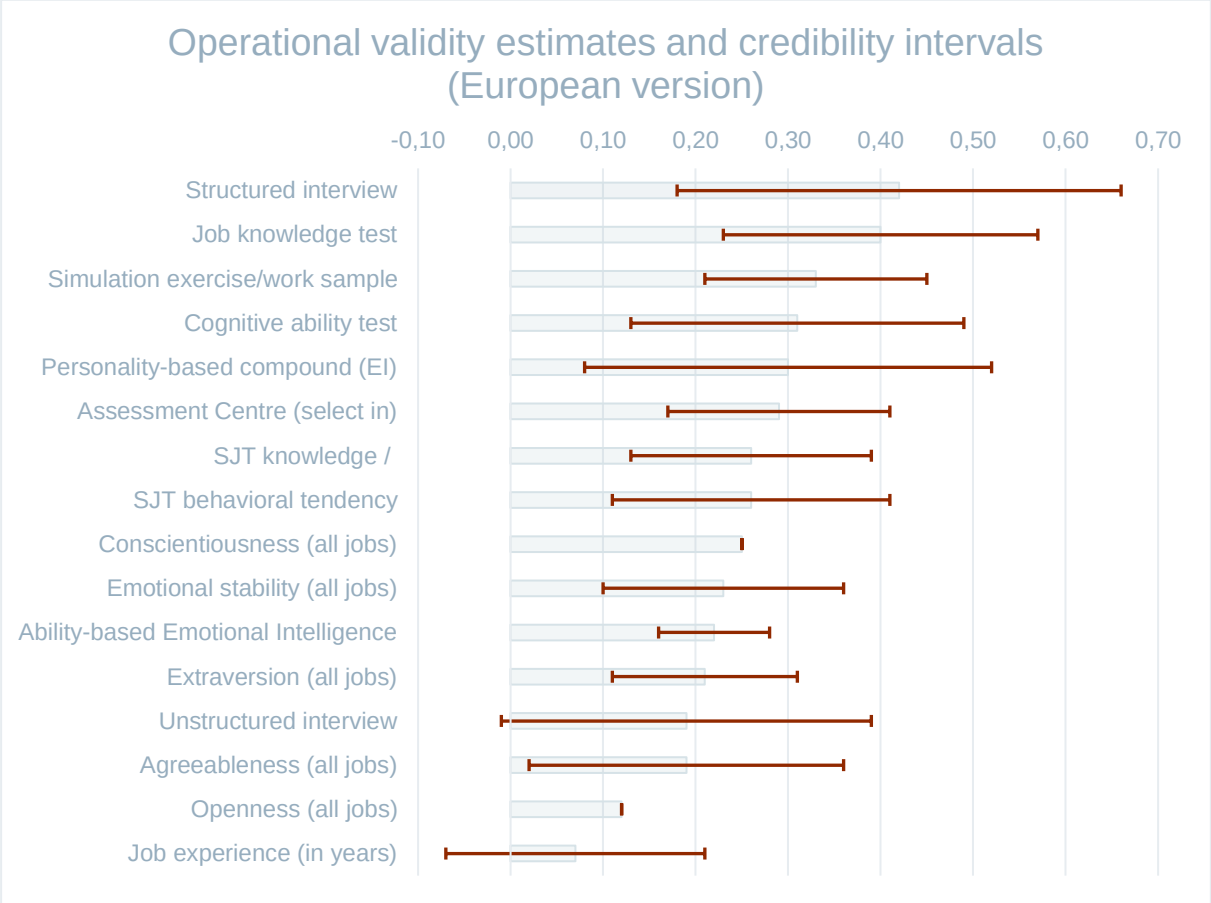

Figure 4: Operational validity estimates and their 80% credibility intervals, adapted to the European/Benelux context

4. Validity is not all that matters

For many organisations, other aspects—such as providing equal opportunities, procedure cost and candidate experience—might be at least as important as predictive validity when setting up a selection procedure.

4.1 Equal opportunities: lack of adverse impact

A tool is said to have **adverse impact when people from majority and minority groups (gender, cultural background, age, ..) do not have an equal chance of passing the test; of being selected on the basis of their test score.**

There is adverse impact if the proportion of candidates from one specific group (usually minority group) who pass the test is significantly smaller than the proportion of candidates from the other groups passing the test. In such a case candidates from one group would have a lower chance of being selected on the basis of their test scores than candidates of other groups. Many government institutions risk facing litigation if adverse impact can be demonstrated in their selection procedure. Some selection procedures yield higher subgroup score differences than others. The latter tests should be chosen if an organisation seeks to minimise adverse impact in their selection procedure.

The current paper by Sackett et al; (2021) also reports the Black-White subgroup mean difference for each selection predictor. But of course, in the European context, Black and White differences are less important than other ethnic differences, gender differences and age differences.

For example, it has been consistently shown in research (also in the current paper) that most ethnic minority groups (except for Asians) score lower (on average) on **cognitive ability tests** than the ethnic majority. This is usually not due to the fact that cognitive tests are ethnically biased. The lower average scores can in most cases be attributed to a true difference in the ability that is being measured.

Possible causes for this robust observation are mainly societal: less mastery of the dominant language, lower educational level, lower socio-economic status, etc. are more often observed in the ethnic minority groups and could lead to lower (average) cognitive ability.

Some other examples of true differences: men tend to score higher on average on numerical reasoning tests than women, while women score higher on average on personality aspects like Altruism. Older people score higher on average on verbal reasoning tests, while younger people score higher on abstract reasoning (Hough et al., 2001).

As various tools and instruments impact different minority groups in different ways, **combining these tools** in a selection procedure (as is usually done in an Assessment Center) is usually **a good way to level out the adverse impact that a single tool may have.** Specifically when opting for cognitive ability measures, one could consider focusing on sufficiently specific and relevant cognitive ability subtests (e.g. language-free) to minimise the adverse effects (Wee et al., 2014).

4.2 Cost and practical feasibility

Another aspect that is of primordial importance in real-life settings is the cost in terms of budget and resources of any selection procedure. For example, while some **face-to-face procedures** like interviews or work samples have a high predictive validity in some contexts, they also **require a lot of resources**, especially when dealing with a very large applicant base. A computerised questionnaire or test, that can be administered remotely, might be much more practical and cost-effective in those situations.

4.3 Candidate experience

Another factor that is becoming more and more important is 'candidate experience': how do candidates experience the selection procedure?

With the ongoing 'War for talent' and the rise of social media, employer branding is becoming more and more important. Organisations now need a positive reputation to be able attract the top candidates. And an organisation's reputation is to a large extent influenced by how candidates feel about that organisation's recruitment and selection procedure.

Research has shown that when people have a **positive experience**, they not only rate the **organisation as more attractive**, and are more likely to **accept a job offer**, but they will also **perform better** during the selection procedure.

There are two important conditions for someone to have a positive assessment experience (Hausknecht et al., 2004):

1. **How relevant do they perceive the tests** to be for the job they are applying for? People experience a high **face validity** when they feel that the test content, the questions asked are really related to the job they are applying for, to what they would do on the job.
   Tests like reasoning ability tests and personality questionnaires generally have a low face-validity, people do not see the link with work performance, while in role plays and SJTs this link is clearly present (although such tests have a much lower predictive validity)

2. **How fair they perceive the** application and **testing procedure** to be? They want to feel that that they have the same resources and opportunities to perform as anyone else, that the rules to select the best candidate are logical, just and transparent. And they must feel respected by recruiters throughout the procedure.
   Candidates will perceive the procedure as fair if they are provided with sufficient information and feedback, and if they are treated respectfully by the assessors and/or test guidance.

This goes to say that different assessment selection procedures impact candidate experience in different ways, which should also be taken into account when choosing a procedure. Very intrusive methods, that ask for private information sometimes with no clear link to the job in question (e.g. biodata), or appear to treat candidates with suspicion (e.g. integrity questionnaires), might not be a good choice when candidate experience is considered important by an organisation.

4.4 Incremental validity

Another important aspect to consider is what happens when different techniques are applied together. This is related to incremental validity: what is the **added value** in terms of predictive validity of one technique **when you are already using another technique**. Cognitive ability tests are then often used as a baseline to compare with. Structured interviews, personality questionnaires and work samples offer a significant increase in predictive validity when combined with a cognitive ability test (Lievens, 2020). Moreover, Assessment Centres are known to explain a sizeable proportion of variance in job performance beyond cognitive ability and personality (Meriac et al., 2008).

Most other techniques tend to correlate more strongly, i.e. have more overlap, with cognitive ability, so they offer less additional predictive power value when combined with a cognitive ability test.

4.5 Trade-offs: the perfect tool?

And now? Which selection method is the best one?

Up to date, no single selection tool exists with perfect scores on all important aspects: demonstrating high predictive and incremental validity as well as low adverse impact, at a low cost, while being experienced positively by most applicants. All tools have certain advantages and disadvantages and trade-offs exist between the important aspects. To compose and devise a selection procedure, **HR professionals need to carefully and critically consider all these aspects**, taking into account the client's specific context and desires and the requirements of the job they are hiring for.

	Predictive validity	Candidate experience	Equal opportunities (Low adverse impact)	Cost
Unstructured (biographic) interview	+	++(+)	+	€€
Structured interview (CBI)	++(+)	++	++	€€
SJT/E-tray	+(+)	++	+++	€€
Work sample/ Simulation exercise	++(+)	++(+)	++	€€(€)
Personality questionnaire	++	+	++(+)	€
Reasoning ability tests	++(+)	+	+	€
Assessment Centre	++(+)	++(+)	++	€€€

Table 3: main tools that are often used for selection, with their simplified 'ratings' (based on scientific findings in the papers of Sackett et al. 2020, Hough et al. 2001 and Hudson's own research)

To assist in making an informed decision on which tools to use, we have created an **overview table** (Table 3), **listing the main tools** that are often used by Hudson and its clients for selection. This table lists the tools' simplified 'ratings' (based on scientific findings in the papers of Sackett et al. 2020, Hough et al. 2001, and Hudson's own research, i.e. analyses conducted on our tools) for the important aspects: predictive validity, equal opportunities (low adverse impact), candidate experience and cost.

Sometimes plusses/euros are shown between brackets, because findings vary largely for that aspect. A lot will depend for example on assessor or interviewer characteristics, how well they have been trained (e.g. more training leads to an increase in predictive validity) and how objective/unbiased they are. Also the choice and weighting of tools in an assessment centre, and the information provided to candidates will have a strong impact on all the factors. For custom-made exercises a lot will depend on the quality and standardisation of the development procedure.

5. Conclusion

At the end of 2021, Paul Sackett and colleagues published a new landmark meta-analysis that indicates that most of the aspects we measure or assess during the selection process do not predict performance as well as we previously thought. However, the **validity estimates are still considered to be very high**. Being able to predict behavior with such high accuracy is rather fantastic if you consider the complexity of human nature.

The authors re-analysed the classic Schmidt & Hunter (1998) meta-analysis with more correct methods which resulted in a few striking changes compared to the original paper. The most striking one being that cognitive ability or GMA measures are no longer the best predictors of job performance, but are replaced by structured interviews at the top of the list.

When interpreting the results reported in the new paper and drawing lessons for our day-to-day practice, we have to take into account that (like all research) the new study comes with a few limitations:

- First of all, validity estimates in meta-analyses typically cover a large variety of studies, tools and work settings, they are not 'exact' numbers. In fact**,** the predictive value of any given selection method varies substantially across studies. This implies it is less useful to focus on exact numbers and rankings.

- The large majority of primary studies reviewed are incumbent-based and conducted in low-stakes situations typical for research settings. We can assume that factors like faking and social desirability play much less of a role in such

conditions compared to real high-stakes selection contexts. Moreover, the level of effort may be lower with incumbents which may affect the validity estimates.

- The way in which the criterion (i.e. job performance) is measured varies a lot across primary studies. Therefore, we should be very cautious when comparing validity estimates with one another.
- There are quite a number of tools reported in the list (such as biodata, work samples, job interests questionnaires, et cetera) that are less relevant or need to be labelled differently in the European/Benelux context, for various reasons.
- And, last but not least, validity is not all that matters. For many organisations, other aspects—such as providing equal opportunities, procedure cost and candidate experience—might be at least as important as predictive validity when setting up a selection procedure.

The main takeaway is that there is no "one size fits all" solution and that most tools on the market remain useful and relevant, provided that they are deployed under the right circumstances. No single tool exists with perfect scores on all important parameters, which implies that devising a selection procedure will always involve some sort of trade-off.

Therefore, it is always important for HR professionals to keep to three simple rules:

1. align selection methods as much as possible with the job requirements for which one is selecting.
2. bring in experienced test administrators, assessors and interviewers, who have been thoroughly and properly trained.
3. opt for tools that were developed according to the highest quality standards.

6. References

Buyens, D., De Schampelaere, V., Verbrigghe, J., & Verhaeghe, S. *Rekrutering en Selectie in kaart: Huidige Tendensen en Uitdagingen voor Morgen.* Federgon / Vlerick Business School.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639-683.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1-2), 152–194.

Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology, 100*(2), 298–342. https://doi.org/10.1037/a0037681

Lievens, F. (2020). *Human Resource Management: Back to basics*. LannooCampus, Leuven.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*(5), 1042–1052. https://doi.org/10.1037/0021-9010.93.5.1042

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*(2), 128–165. https://doi.org/10.1037/0003-066X.56.2.128

Sackett, P. R., Shewach, O. R., & Keiser, H.N. (2017). Assessment Centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102(10)*, 1435—1477.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology.* Advance online publication. https://doi.org/10.1037/apl0000994

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124 (2)*, 262—274.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology, 97*(3), 499–530. https://doi.org/10.1037/a0021196

Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than g: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology, 99*(4), 547–563. https://doi.org/10.1037/a0035183

Contact

t. +32 497 52 63 86
amelie.vrijdags@hudsonsolutions.com
www.hudsonsolutions.com

Hudson Belgium N.V./S.A.
Moutstraat 56 9000 Ghent, Belgium
t. +32 (0)9 222 26 95
info@hudsonsolutions.com
www.hudsonsolutions.com