Drenth (2008) *Psychology is it applied enough?* revisited

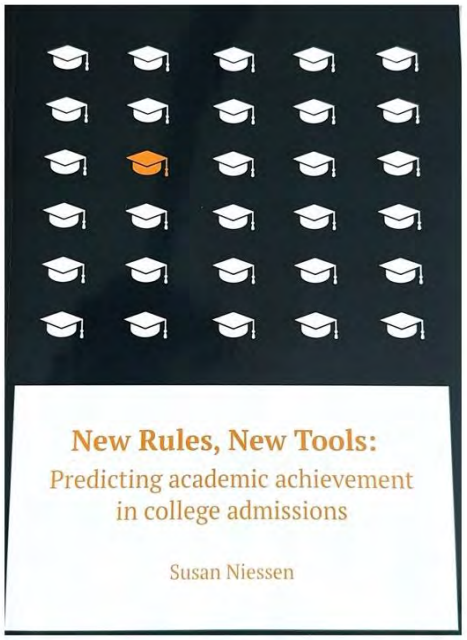
Rob Meijer
University of Groningen

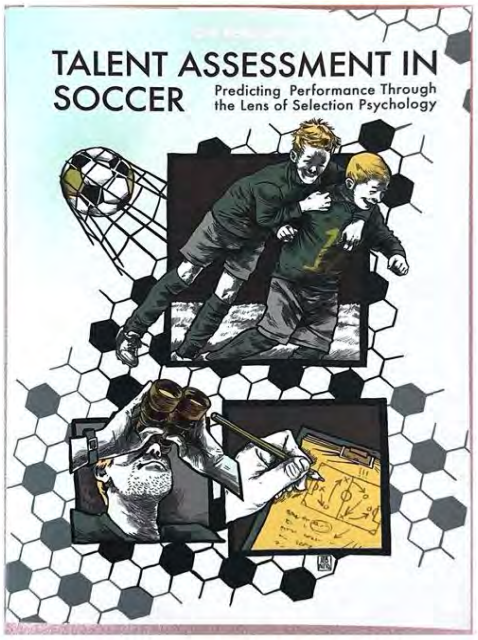
Intro: A personal story

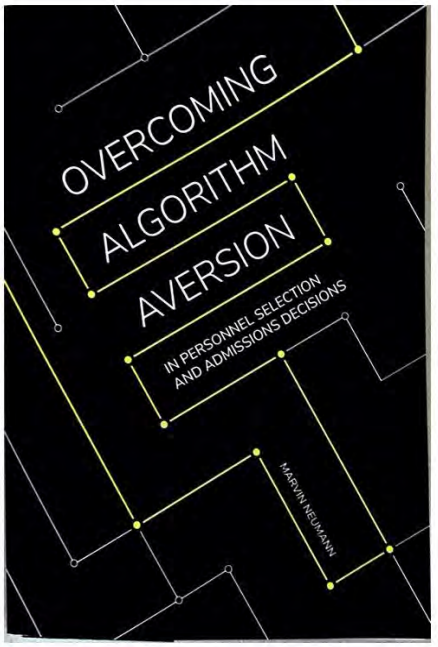
- Psychological testing /psychometrics/personnel selection and admission testing

- Increasing interest in practice, decision making and the ”what people do” and “How that may be improved”

Working together …







I realized …

- .. that a sound test is important, but in applied settings how you use the test to make a decision is what counts

- In practice, valid tests and an optimal regression model are often not used to make decisions

Not new .. but research in this area is not popular

- A special issue on decision making in practice (IJSA) hardly received submissions
- Highhouse, Kuncel
- We ignore an important part of our field: How is our knowledge applied?

Drenth (2008) "Psychology is it Applied Enough"

- Psychology's position om the pure-applied scale
- Is Psychology sufficiently utilized in practice ?
  - Why the underutilization?
  - Recommendations

Why the underutilization by decision makers?

1. Ignorance
2. Confusion
3. Anti-science attitude
4. Unwillingness
5. Distrust
5. Disappointment
6. Deception

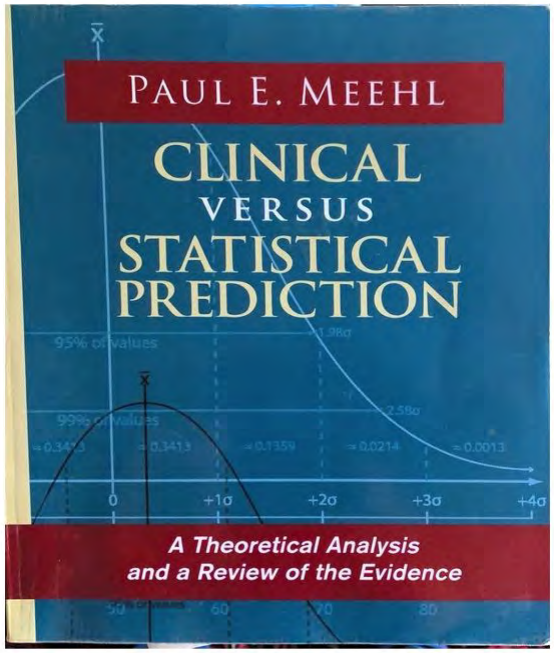
Recommendations: What should psychologists do?

….. ()

1. Behavioral scientist should be more assertive and authorities in presenting their message
2. Communication honest and fair: only empirical results should be the basis
3. Participate in public debate
4. Take a firm stand against pseudo scientific practice and do not be tolerant

The case

› We know since a long time that combining information according to a rule is superior (higher validity) than combining information holistically, intuitively



Important topic

- With increasing algorithm use in society relevant topic because the
- sometimes negative sentiments
- transparency a rule provides in contrast to the black box of a human decision maker' mind

A practical lesson is

- Take a number of valid predictors and combine them according to a rule
- Refrain from (nonsense) “storytelling”

- Problem in practice is: Professionals do not like this and part of the fun of the job is the “storytelling”

"Professional" behavior

- A candidate is suited for a job by interviewing a candidate (often unstructured), look at a profile of NEO scores and role-play scores and then write a report!

- Or interview a candidate look at "the fit" to the organization and then tell yourself a story why this candidate is great !

- This is suboptimal given what we know from the literature

The problem: Inconsistency

	Interview (structured)	Work sample	Conscient.	Something else
Candidate 1	.5	.2	.3	0
Candidate 2	.2	.4	.4	0
Candidate 3	.1	.1	0	.8

So we are inconsistent ! Different candidates get different weights on the same predictor

“I can see that his answers to our questions where ok, but I do not see a *[judge/phd student/doctor etc]* in this guy, difficult to say what it is … “ -----→ exit candidate

Consistent correct weights

	Interview structured	Work sample	Conscient.	Something else
Candidate 1	.5	.3	.2	0
Candidate 2	.5	.3	.2	0
Candidate 3	.5	.3	.2	0

Consistent incorrect weights

	Interview structured	Work sample	Conscient.	Something else
Candidate 1	.2	.3	.5	0
Candidate 2	.2	.3	.5	0
Candidate 3	.2	.3	.5	0

In general: inconsistency is more of a problem than incorrect weights

Do not change the rules during the game

A rule is protecting you

- From weighing all kinds if subtleties in a candidate you think you see but do not improve your prediction because they are error
- From weighting information in a complex way

Overconfidence

- "Overconfidence is powerful source of illusions, primarily determined by the quality and coherence of the story that you can construct, not by its validity.

- If people can construct a simple and coherent story, they will feel confident regardless of how well grounded it is in reality" (Kahneman)

This knowledge is not applied enough

This is bad for practice but also bad for personnel selection as a science, we do not take our own field seriously

› And if we don’t who will?

How can we improve things and help psychologists?

1. Bachelor/Master/Ph.D. education
2. Improve guidelines and standards
3. Research when algorithms are accepted
4. Provide tools (apps) to help decsion making
5. Co-create procedures with the field so that we can improve personnel selection (more valid, but also more fair and transparent)

2. Improve guidelines and standards

- Provide accurate information in guidelines and standards that robust psychological research shows that statistical prediction is “better” than holistic judgment

- Test guidelines in general, and textbooks spread ignorance and confusion when talking about clinical versus actuarial decision making (Meijer et al., 2023)

ITC guidelines do not cover this topic

- *2.1.4 Seek other relevant collateral sources of information.*
- *2.1.6 Ensure that full use is made of all available collateral sources of information.*
- *4. Make clear that the test data represent just one source of information and should always be considered in conjunction with other information.*

This is so vague and ”other information” may also dilute valid information

university of groningen

From Neumann et al (2023)

Tabelle 2

Prädiktive Validität eines strukturierten Interviews in Kombination mit einem zweiten Prädiktor für Arbeitsleistung bei Anwendung der Einheitsgewichtung

Auswahlmethoden (zweiter Prädiktor)	Validität des (zweiten) Prädiktors[a]	Interkorrelation zwischen strukturiertem Interview und zweitem Prädiktor					
		$r = .00$	$r = .10$	$r = .20$	$r = .30$	$r = .40$	$r = .50$
Strukturiertes Interview	.42	-	-	-	-	-	-
Wissenstest	.40	.58	.55	.53	.51	.49	.47
Empirisch entwickelter Biodata-Bogen	.38	.57	.54	.52	.50	.48	.46
Arbeitsprobe	.33	.53	.51	.48	.47	.45	.43
Allgemeine Intelligenz	.31	.52	.49	.47	.45	.44	**.42**
Integritätstest	.31	.52	.49	.47	.45	.44	**.42**
Persönlichkeitsbasierter EI-Test	.30	.51	.49	.46	.45	.43	**.42**
Assessment Center	.29	.50	.48	.46	.44	[illegible]	**.41**
SJT (Wissensinstruktion)	.26	.48	.46	.44	**.42**	**.41**	**.39**
SJT (Verhaltensinstruktion)	.26	.48	.46	.44	**.42**	**.41**	**.39**
Gewissenhaftigkeit – arbeitsbezogen	.25	.47	.45	.43	**.42**	[illegible]	**.39**
Interessentest	.24	.47	.44	.43	**.41**	**.39**	**.38**
Emotionale Stabilität – arbeitsbezogen	.23	.46	.44	**.42**	**.40**	**.39**	**.38**
Fähigkeitsbasierter EI-Test	.22	.45	.43	**.41**	**.40**	**.38**	**.37**
Rational entwickelter Biodata-Bogen	.22	.45	.43	**.41**	**.40**	**.38**	**.37**
Extraversion – arbeitsbezogen	.21	.45	**.42**	**.41**	**.39**	**.38**	**.36**
Gewissenhaftigkeit – allgemein	.19	.43	**.41**	**.39**	**.38**	**.36**	**.35**
Unstrukturiertes Interview	.19	.43	**.41**	**.39**	**.38**	**.36**	**.35**
Verträglichkeit – arbeitsbezogen	.19	.43	**.41**	**.39**	**.38**	**.36**	**.35**
Offenheit – arbeitsbezogen	.12	**.38**	**.36**	**.35**	**.33**	**.32**	**.31**
Extraversion – allgemein	.10	**.37**	**.35**	**.34**	**.32**	**.31**	**.30**
Verträglichkeit – allgemein	.10	**.37**	**.35**	**.34**	**.32**	**.31**	**.30**
Emotionale Stabilität – allgemein	.09	**.36**	**.34**	**.33**	**.32**	**.30**	**.29**
Arbeitserfahrung (in Jahren)	.07	**.35**	**.33**	**.32**	**.30**	**.29**	**.28**
Offenheit – allgemein	.05	**.33**	**.32**	**.30**	**.29**	**.28**	**.27**

Improve things test guidelines/standards

Updating Algemene Standaarden Testgebruik (AST, NIP)

More attention for mechanical decision making

Challenges: How do you combine scores on different instruments to one score

Solutions: provide examples and recipes how to construct rules

5. Co-creation of rule-based decision making

- Research within organizations

- Workshops information through
- Talks
- Experiments with the professionals
- Discuss how they would construct a rule and ask them what they need

Implementing advice taken procedures/ statistical decision making

- Discuss findings with professionals
- Make them realize their inconsisteny (“Noise audit’)
- Discuss more optimal procedures
- Help and guide them
- Also clarifies what the problems are
- Input for further research

-

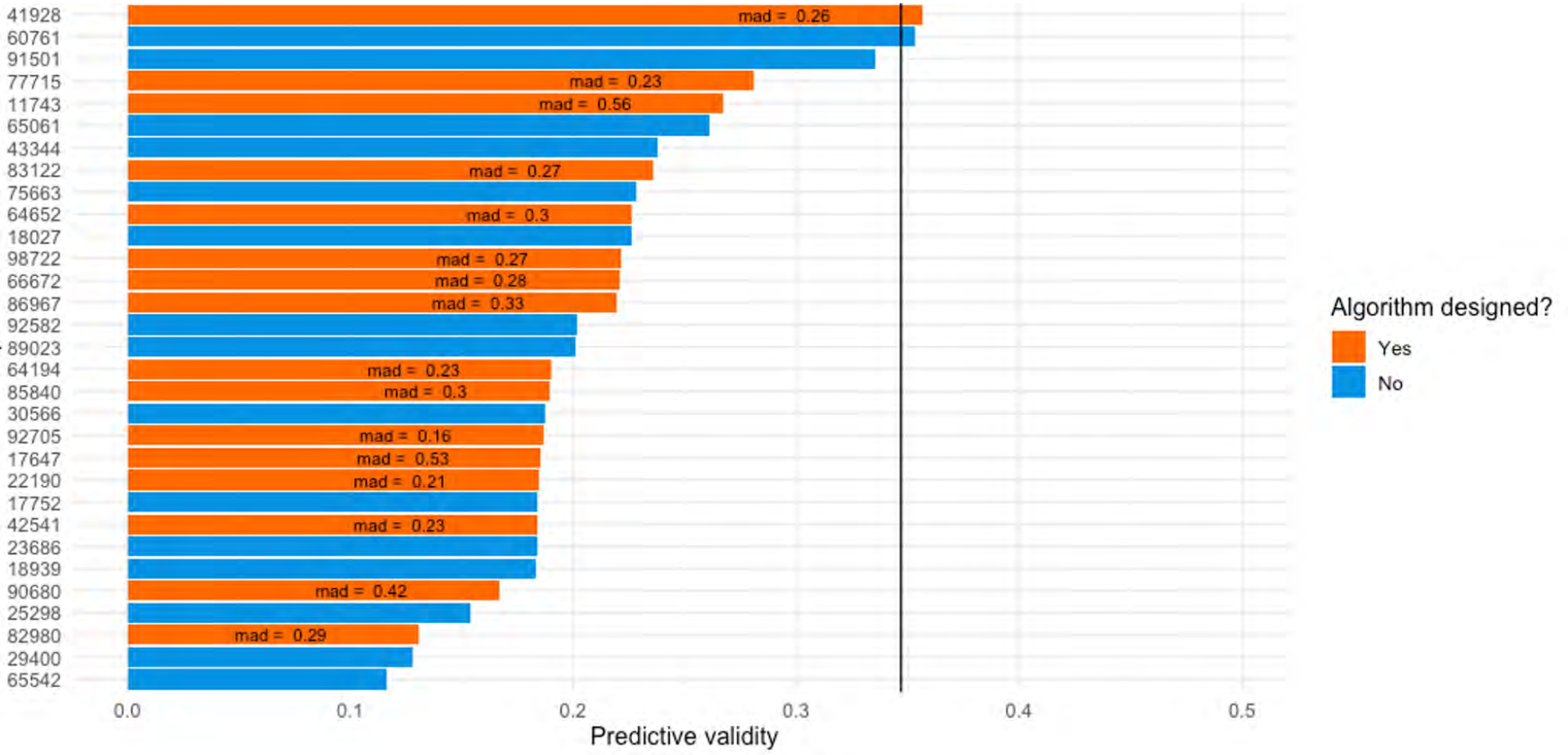
Example: experiment in Workshop (Neumann, Niessen and Meijer)

- Predict job success on the basis

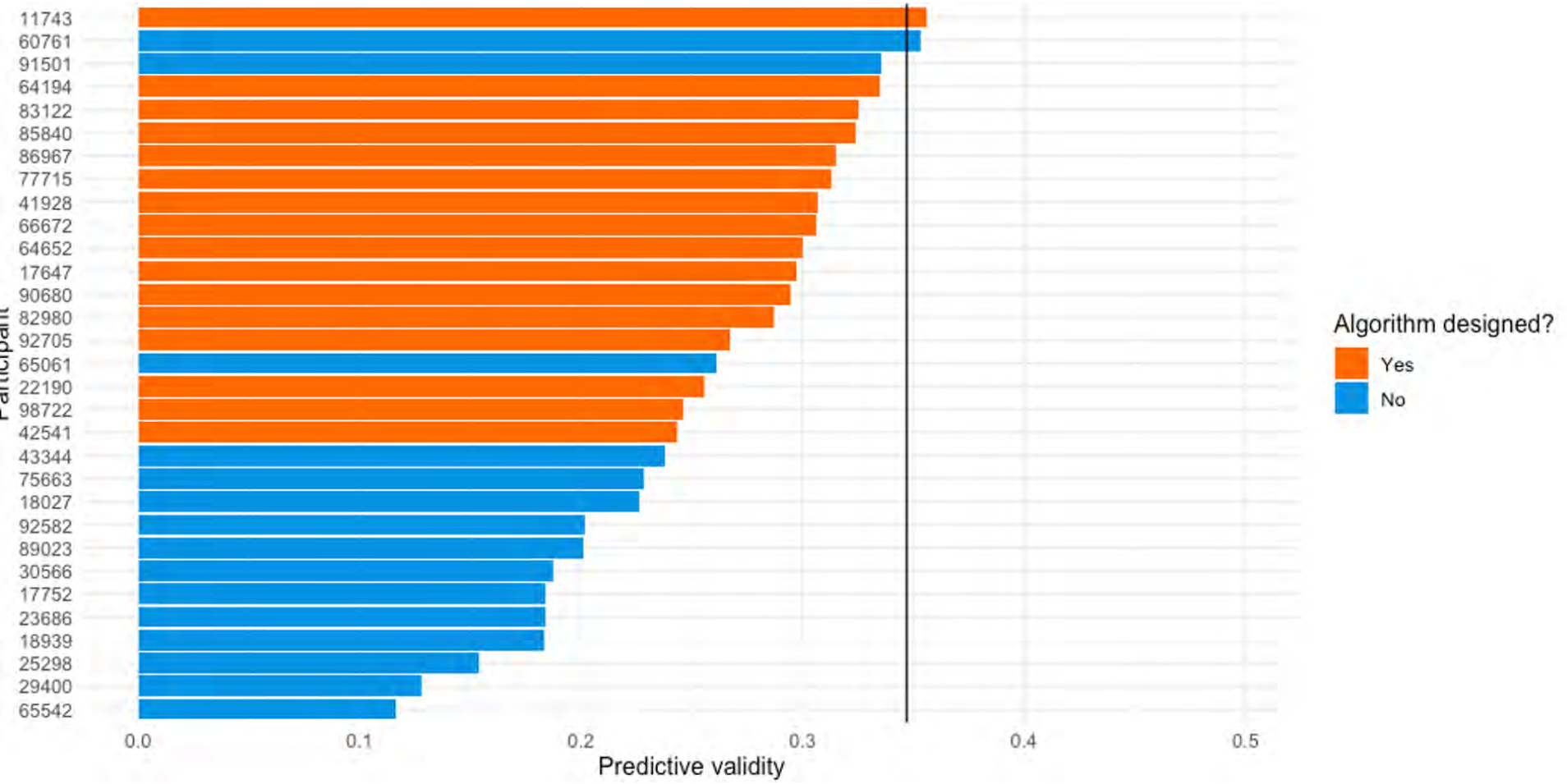
- Cognitive ability scores
- Unstructurered interview scores
- Conscientiousness scores

- Rwo conditions: (1) holistic and (2) design your own rule and you get the result from this rule

Example of experiment result



if you would have followed the rule …



Noise audit (Kahneman et al, 2021)

- Using predictors that are used to make holistic descions in the company
- When constructing a rule: How would you weight ?
- What kind of rules would you like to use?

- Sometimes large discrepancies between professionals in the same company

Some general reactions

- In general positive, should we also not make our tools more mechanical ?
- Psychologist should have the last say and I should be able to make exceptions
- ----------
- How should we make rules ?
- Is this allowed wrt ethical guidelines

Finally

- Consider the validity of our decisions, not only those of (optimal combinations of) test scores

- This would would greatly increase the "reality value" of personnel selection research and make it more applied and concrete

Finally

- It also increases the relevance of our research
- For example, how to construct rules is often not obvious and not that simple for professionals
- Researching and writing tutorials how to do that is important !

- Thanks for your attention!

Some future dreams

- In science we now specify our research plan and submit it to a public registry (preregistration)

- (Selection) companies may specify which instruments they use and how they will combine the data according to a rule, they will make this public on a public registry for top quality assessment companies

Instead of

- .. Is the specialist in providing insight into talent and the developmental potential to people. Teams and organizations. We do this by optimally combining our advisers' knowledge and experience with the digital tools we've developed

- We call this method [fancy name]. Digital tools provide insights, our people give meaning to those insights, thus enabling clear and informed decision making

A final observation

- We can discuss whether intelligence correlates .3 or .4 with later job performance
- But in practice we found a correlation of -.15 because “I am not interested in another smart guy” it is all about if he fits the team !

Grove and Meehl (1996)

- Of the 136 studies, 64 favored the actuary by this criterion, 64 showed approximately equivalent accuracy, and 8 favored the clinician.
- explanation of these deviant studies is that they arose by a combination of random sampling errors (8 deviant out of 136) and the clinicians' informational advantage in being provided with more data than the actuarial formula.
- One who is strongly predisposed toward informal judgment might prefer to interpret this lopsided box score as in the following way: "There are a small minority of prediction contexts where an informal procedure does better than a formal one." Alternatively, if mathematical considerations, judgment research, and cognitive science have led us to assign a strong prior probability that a formal procedure should be expected to excel, we may properly say, "Empirical rese provides no clear, replicated, robust examples of the informal method's superiority."

- IRS – INSPIRE resislience scale (e.g.,optimism, self-efficacy etc), stanine scores 5 deterrminants and 7 coping scales provided to the psychologsists
- NEO 5 traits, 30 subscales, stanine scores, with explanation what they mean
- Roleplay: scored on 4 diemsnions by actor and assessor after consenses, psychologist gets 4 scores
- Intervuew semi structured input input from major event list, some competencies are discusse others are observed, some from the role
- Empathie, besluitvaardigheid en authoriy from role play

How do textbooks on psychological testing discuss statistical versus statistical decision making? (Meijer et al. 2023)

Table 3.1 Scores that reflect the way textbooks discuss different criteria

	Conclusion in line with literature	Data collection/combination	Robustness weights	Exceptions to the rule	Transparency
Anastasi and Urbina (1997)	1	1	0	1	0
Aiken (2003)	2	2	0	0	0
Murphy and Davidshofer (2005)	2	2	3	2	0
Kline (2005)	0	0	0	0	0
Domino and Domino (2006)	2	2	0	0	0
Reynolds and Livingston (2012)	2	0	0	2	0
Kaplan and Saccuzzo (2013)	2	0	0	0	0
Gregory (2013)	2	0	0	0	0
Hogan (2015)	1	0	1	0	0
Miller et al. (2015)	2	1	0	0	0
Cohen and Swerdijk (2015)	1	2	0	0	0
Furr (2018)	0	0	0	0	0
Cooper (2019)	0	0	0	0	0

Note: 0 = not discussed, 1 = incorrect description; 2 = description lacks important points; 3 = fair description