Retaining the Expert While Reducing Noise in Selection Decisions

Nathan R. Kuncel
Department of Psychology

UNIVERSITY OF MINNESOTA
Driven to Discover®

- Structured Interviews
- Personality Test
- Cognitive Tests
- In-Basket
- Role Play
- Leaderless Group Discussion

→

- Leadership
- Motivation
- Judgment
- Adjustment
- Administration
- Communication
- Interpersonal

→ Person Job Fit Rating

Meehl's Little Book

- In 1954, Paul E. Meehl published a small book entitled "Clinical Versus Statistical Prediction"
- He found overwhelming evidence that decisions based on mechanical combinations of information yield superior decisions.
- This finding was so powerful that it stimulated years of research.
- Many people either attacked or simply ignored the evidence.

Who is successful at Univ. of Minnesota? Admissions Officers Predicting Student GPA

High School Rank + College Admissions Test	r (accuracy) = .45
Holistic Admissions Counselor Judgment based on full file including HS Rank and Admissions Tests	r (accuracy) = .35

Sarbin (1943)

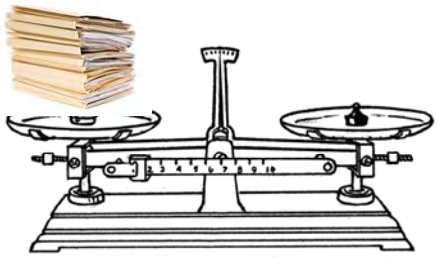
C-files Study

A committee of doctors is reviewing psychiatric files (in folders) to decide which patients to release and who would return.

What other method made decisions that were just as accurate as the ones made by the committee?

Lasky, et al. (1959)

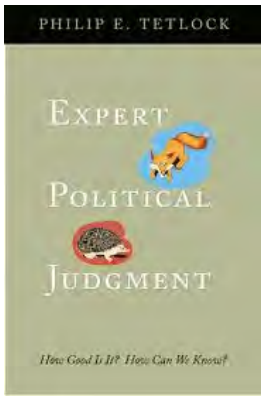
C-files Study



A scale on which the patients' files were weighed!!!

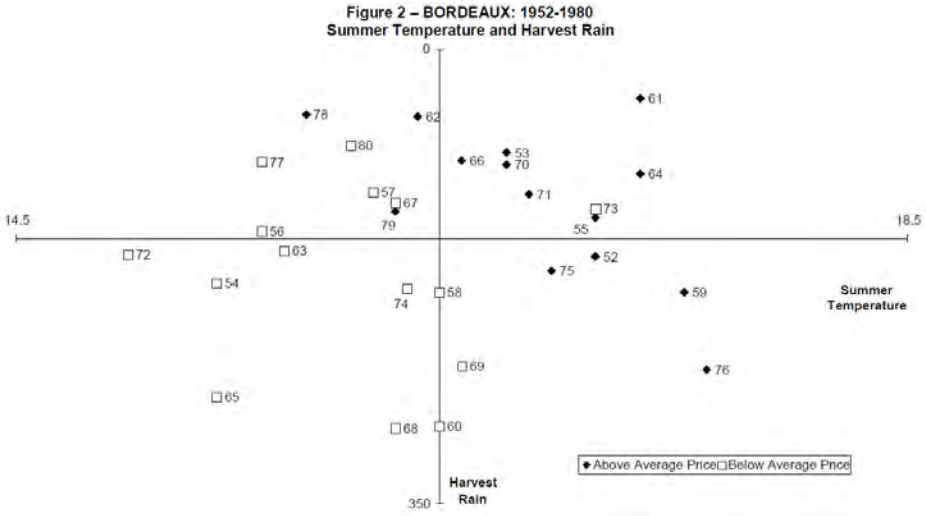
Tetlock (2005)

Examined about 28,000 predictions from experts on politics, economics, & public policy.



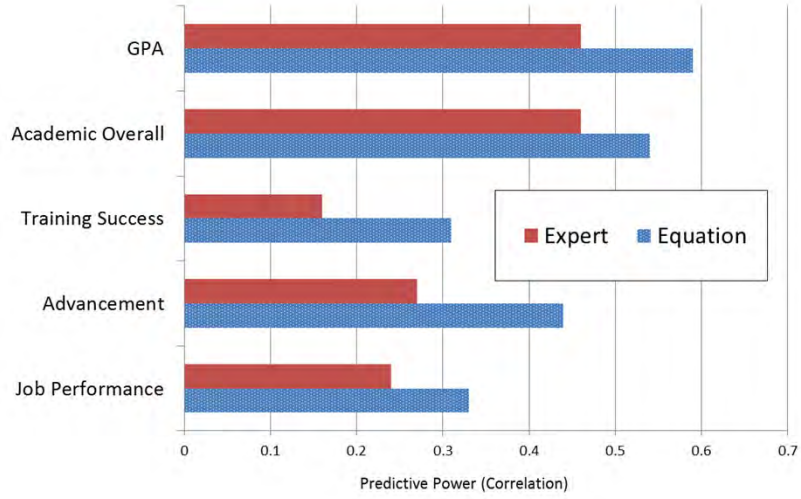
Bordeaux Wine

- Grapes are harvested and made into wine which is not good to drink for years.
- Wine experts taste the raw product and make judgments about how good it will be.
- An economist build a simple equation to predict the value of wine based on rainfall and temperature.
- Wine experts called it all sorts of names including “absurd”.



Summer Temp and Harvest Rain for Above and Below Average Priced Bordeaux Wines

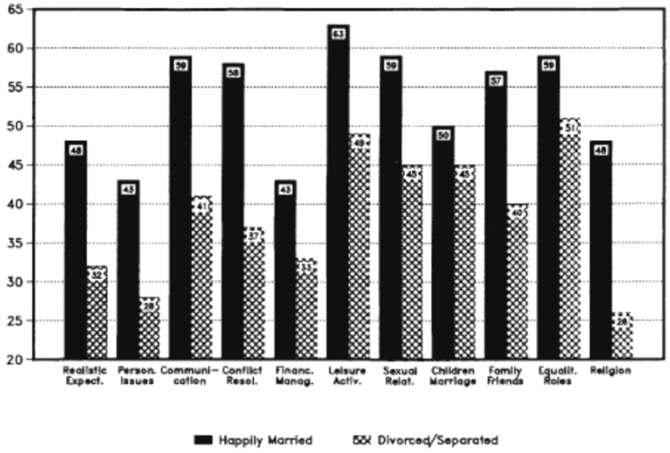
Expert Judgment versus Equations for Predicting Success



Kuncel, Klieger, Connelly, & Ones (2013).

Let's Get Realistic

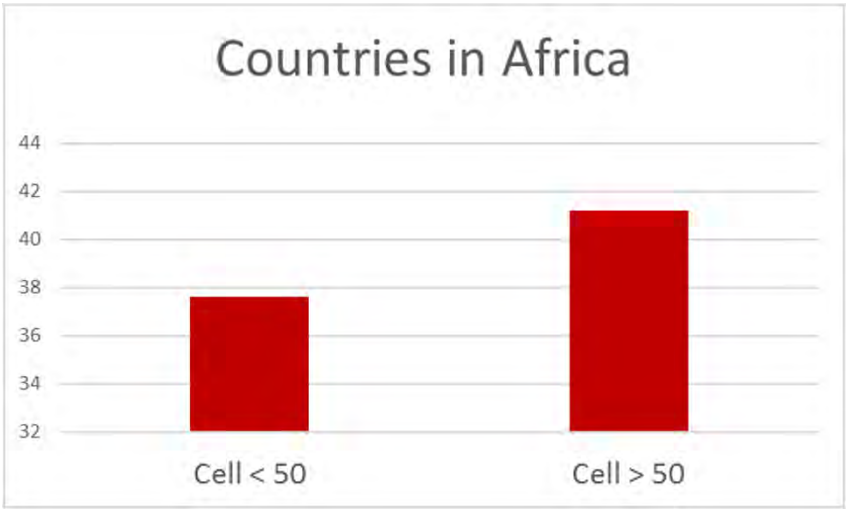
FIGURE 1: Happily Married vs. Divorced/Separated: Positive Couple Agreement (PCA) Scores on PREPARE Categories



My goal is to find interventions that preserve decision maker acceptance while improving decision accuracy

A good start is to understand source of the problem.

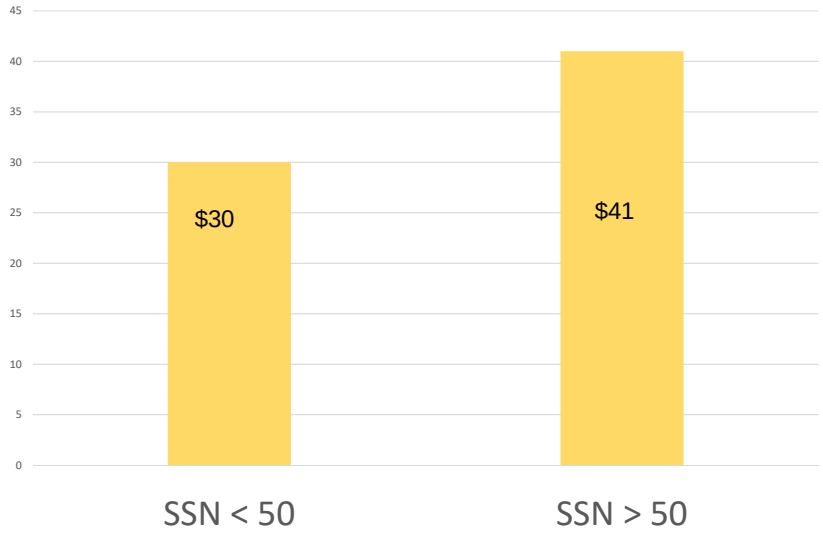
Please write down the last two digits of your phone number.



Some More Examples

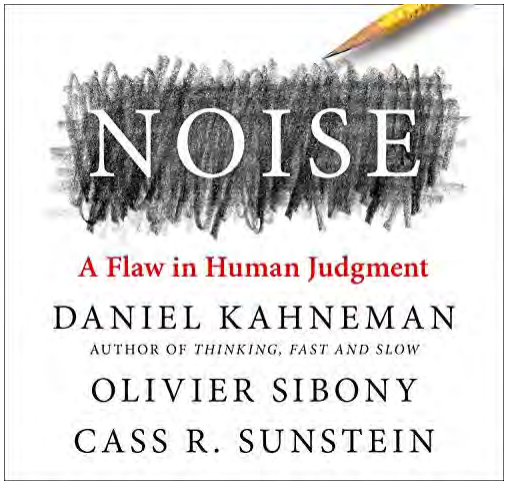
1. Write down the last two digits of your SSN.
2. What would you pay for this lovely toaster?

What Would You Pay?



Simonson & Drolet (2003)

In Addition to Bias Some of the Problem is Decision Making Noise



Two Kinds of Noise

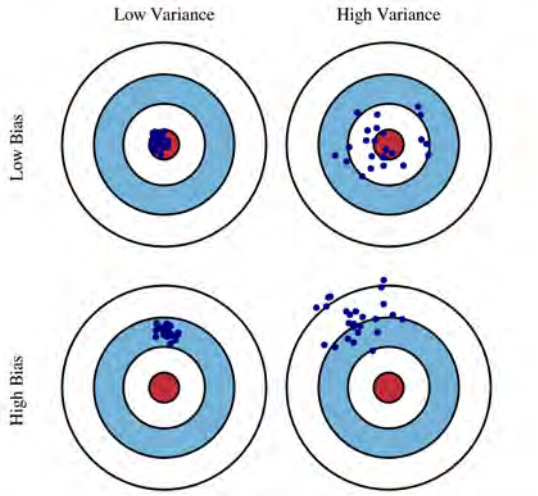
Within Person Noise: Same information, same decision maker, different decisions

Between Person Noise: Same information, different decision makers, different decisions

- Note: Once the decision is assigned to a specific decision maker, decisions may be highly predictable and consistent (i.e., high between person, low within person noise)

Note: We might imagine other levels of analysis and effects, like departmental or organizational noise reflecting difference in local customs or traditions.

Noise and Bias are Not the Same
(but both are a problem)



Fortmann-Roe (2012-2021)

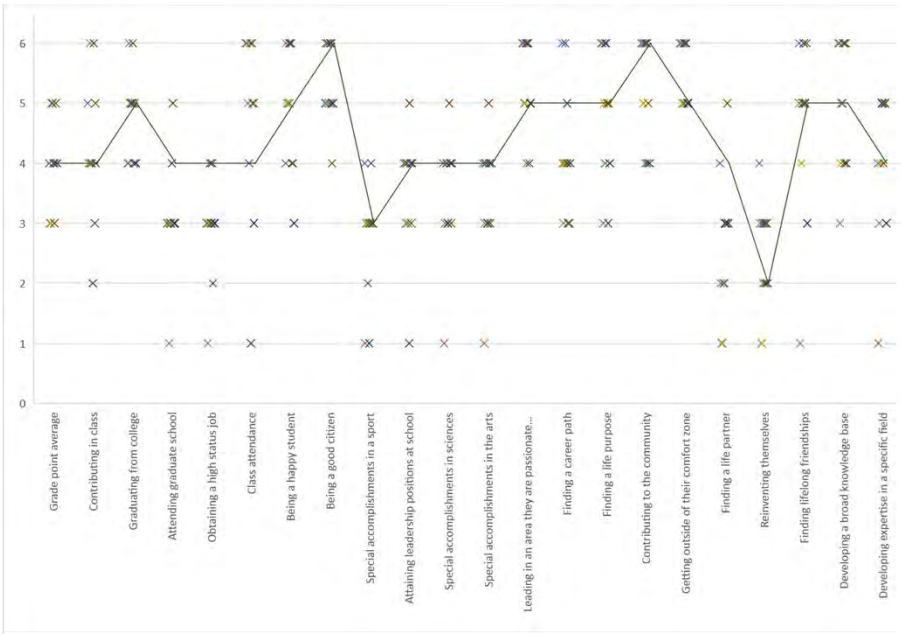
Court Cases in the US

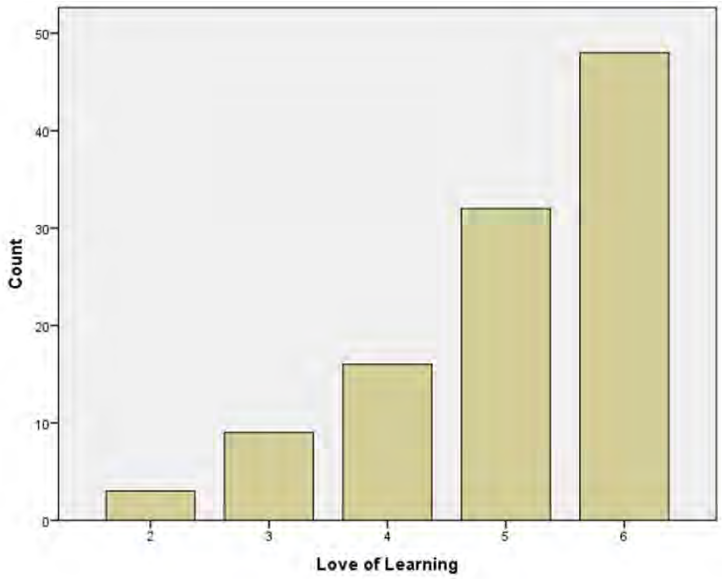
Sentencing Scenario Experiments with Federal Judges

- Fraud case
  - Mean prison 8.5 with a wide range
  - Including one life sentence!
- Burglary
  - 30 days to 5 years
- In other cases, sentences ranged from probation to years in prison.
- If you know which judge will hear the case, the sentence becomes far more predictable.

Ideas for Solutions

Get Consensus and Then Train Decision Makers





Hypothesis: This leads to between person noise.

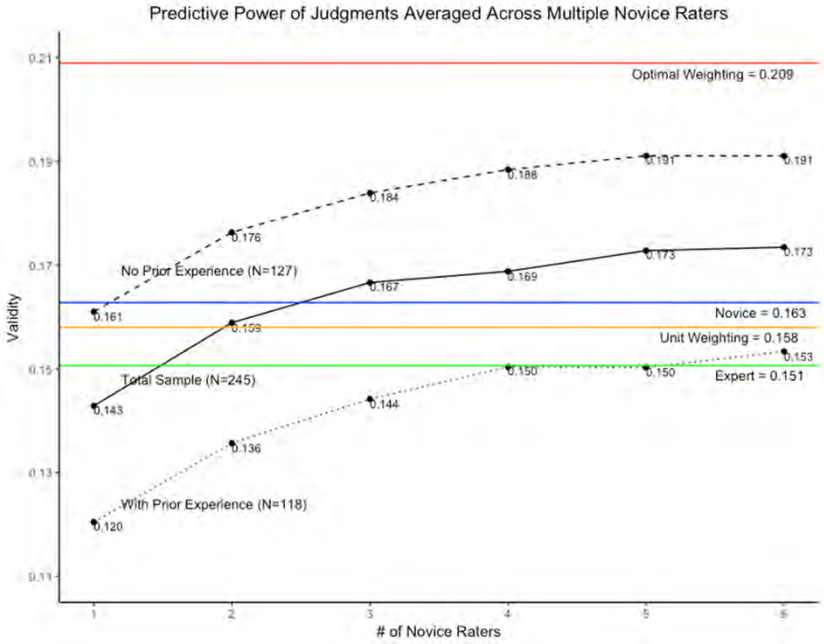
Encourage Aggregating Independent Judgments

Combining Assessment Judgments: Experts vs. Novices

Novices filled out personality and decision style inventories and then evaluated 30 results from real individual assessments for executives at a large Fortune 100 company to provide ratings.

Experts were psychologists who had conducted the individual assessments and each candidate was paired with subsequent supervisory ratings data.

Shu & Kuncel (2016)



What is the Source of the Problem?

Answer:
Human Judgment is Inconsistent

Pushing the Limits of Consistency

Organization	Sample Size
Financial Services Company	231 candidates, 26 assessors
Food Retailer Executive	195 candidates, 23 assessors
Food Retailer Line/Middle Mgmt	421 candidates, 30 assessors

- Structured Interviews
- Personality Test
- Cognitive Tests
- In Basket
- Role Play
- Leaderless Group Discussion

→

- Leadership
- Motivation
- Judgment
- Adjustment
- Administration
- Communication
- Interpersonal

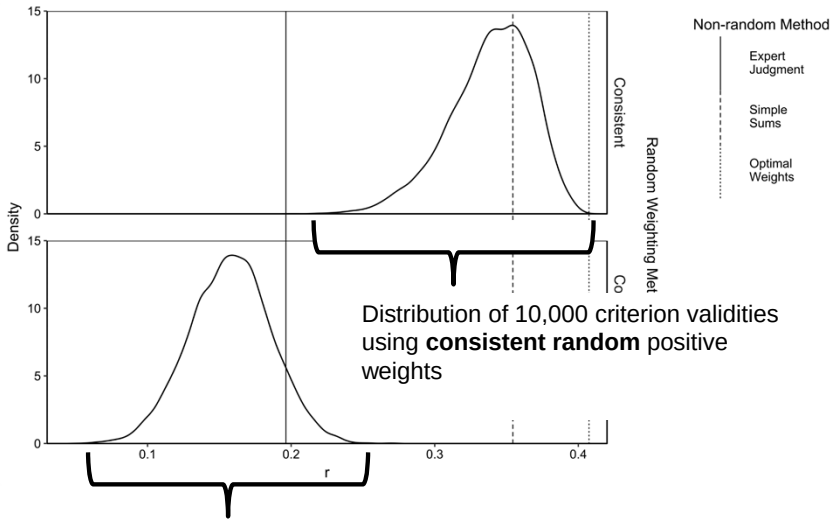
→ Person Job Fit Rating

Yu & Kuncel (2020). *Personnel Assessment and Decisions*

Pushing the Limits of Consistency
Yu & Kuncel (in press)

We compared **Expert Judgments** with Unit Weighting, Optimal Weighting, and two Random Weight Conditions

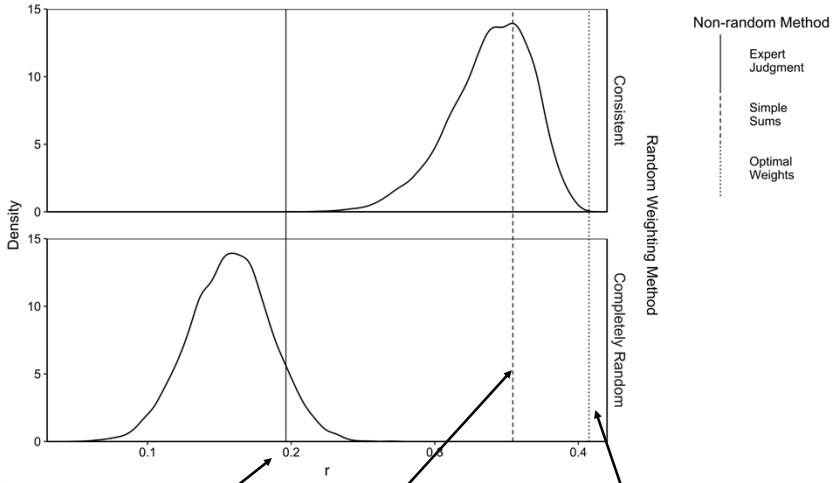
	Consistent Random Weights				Completely Random Weights			
Candidate	Lead	Motive	…	Adj.	Lead	Motive	…	Adj.
1	.02	.36		.19	.35	.45		.12
2	.02	.36		.19	.15	.11		.28
3	.02	.36	…	.19	.43	.06	…	.33
4	.02	.36		.19	.04	.41		.31
5	.02	.36		.19	.22	.17		.09
	10,000X				10,000X			

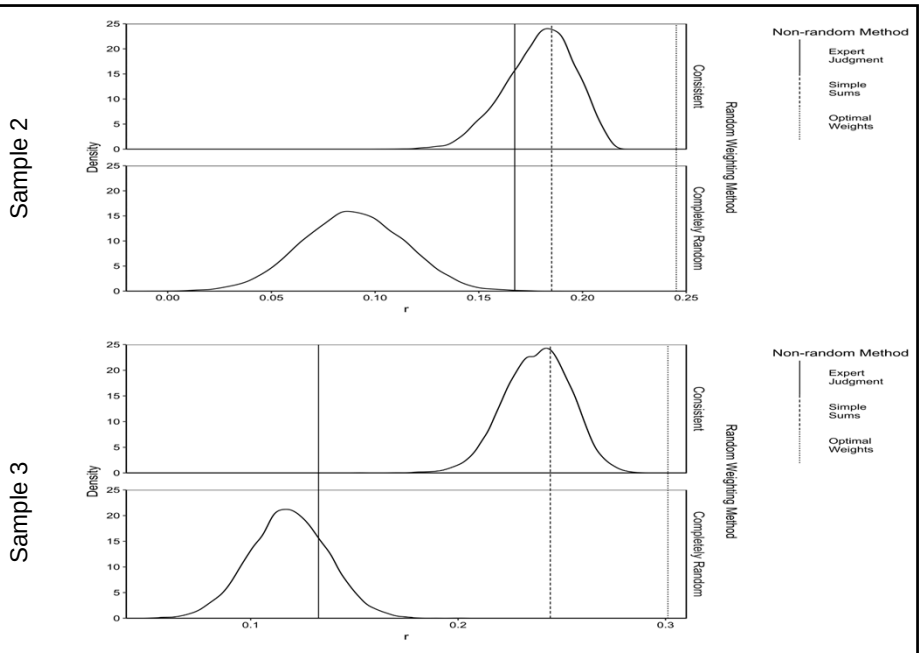


Distribution of 10,000 criterion validities using **consistent random** positive weights

Distribution of 10,000 criterion validities using **completely random** positive weights

Yu & Kuncel (2020)





Yu & Kuncel (2020)

Inconsistency Conclusion

- Assessors are performing only slightly better than a positive random weight generator in predicting job performance.
- Consistent algorithmic data combination, even consistent random weights, are yielding superior prediction of job performance.

But what about unique settings or situations?

Illusion of Uniqueness? Maybe.

Inconsistency isn't necessarily a bad thing. In theory, judgment could improve on simple models.

Decision makers often feel that they are making a unique decision adjusting how they consider the information given the specific setting.

Let's Test It: Local Versus General Models in Hiring

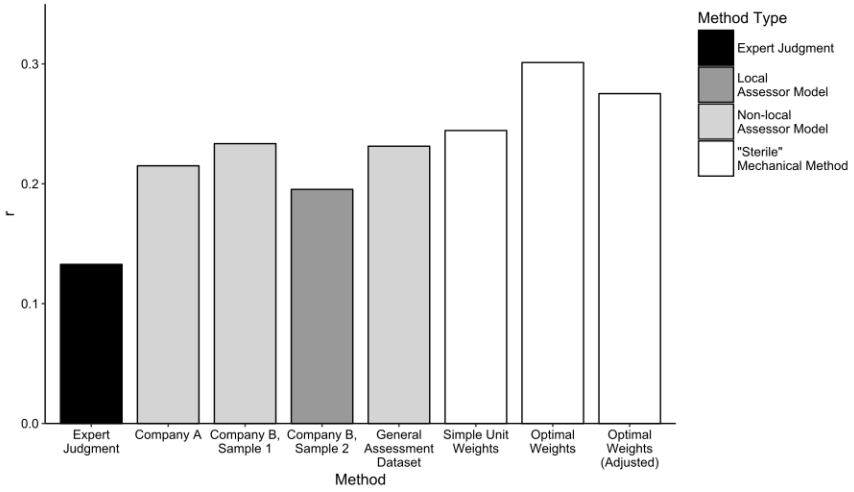
- 3 Validation Data Sets
- General Dataset
  - 16,143 candidates
  - 176 assessors
  - 683 organizations
  - 1971-2000
- If we model decision making in the general dataset will those models perform better or worse than decisions from the local dataset?

Predicting Expert Overall Person-Job Fit Ratings from Assessment Data

Dataset	Dimension	B	SE	R^2
Company A	Motivation	.01	.05	.69
	Judgment	.21	.03	
	Administrative	.04	.03	
	Communication	.10	.06	
	Interpersonal	.20	.06	
	Leadership	.34	.06	
	Adjustment	.13	.06	
Company B, Sample 1	Motivation	.27	.11	.54
	Judgment	.26	.09	
	Administrative	.15	.12	
	Communication	.08	.11	
	Interpersonal	.21	.10	
	Leadership	.10	.09	
	Adjustment	.29	.11	
Company B, Sample 2	Motivation	.04	.07	.57
	Judgment	.28	.03	
	Administrative	-.05	.07	
	Communication	.05	.08	
	Interpersonal	.21	.05	
	Leadership	.27	.05	
	Adjustment	.14	.06	
General Assessment Dataset	Motivation	.09	.01	.69
	Judgment	.22	.01	
	Administrative	.13	.01	
	Communication	.11	.01	
	Interpersonal	.20	.01	
	Leadership	.24	.01	
	Adjustment	.21	.01	

Note: Judgment and Leadership were given considerable weight in all models and, it turns out, these dimensions are some of the most predictive.

The Illusion of Uniqueness



Yu & Kuncel (2022).

Anchoring as a Partial Solution

Using the Anchoring Bias to Debias Decisions: Shu & Kuncel Experiment

We used a large validation data set to present new applicant data to decision makers and then evaluated their accuracy.

"Please use the following applicant information to make 40 hiring decisions. For this job, measures of conscientiousness are generally the best predictors followed by neuroticism and agreeableness."

N = 1,234

"If you can make better decisions than your peers you will get more $$$."

Applicant: "James"

Neuroticism:	43%
Extroversion:	72%
Openness:	54%
Agreeableness:	64%
Conscientiousness:	68%

"You will also be given a algorithmic combination of the applicant data. This index works quite well but is not perfect. If you can make better decisions than your peers you will get more $$$."

Applicant: "Alice"
Hireability Index: 60%

Neuroticism:	43%
Extroversion:	72%
Openness:	54%
Agreeableness:	64%
Conscientiousness:	68%

Using a Bias to Debias Decisions: Shu & Kuncel Experiment

Please use the following applicant information to make 40 hiring decisions. For this job, measures of conscientiousness are generally the best predictors followed by neuroticism and agreeableness.

This is an Anchor

nic combination
works quite
make better
ill get more $$$

Simple average of predictor scores (not optimal).

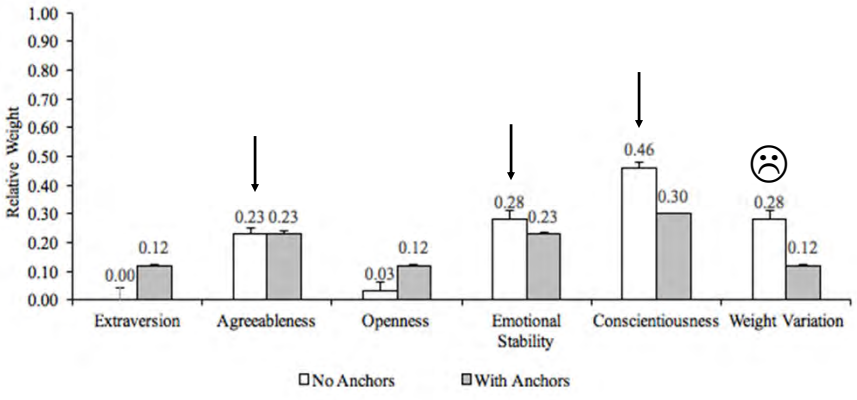
Applicant: James

Neuroticism:	43%
Extroversion:	72%
Openness:	54%
Agreeableness:	64%
Conscientiousness:	68%

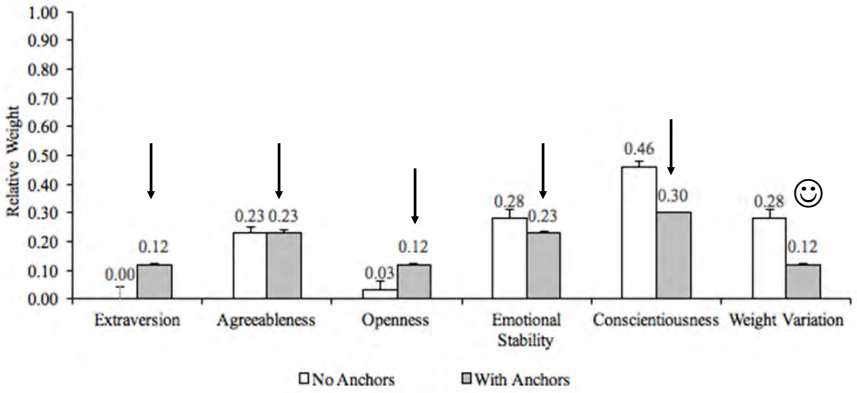
Applicant: Alice
Hireability Index: 60%

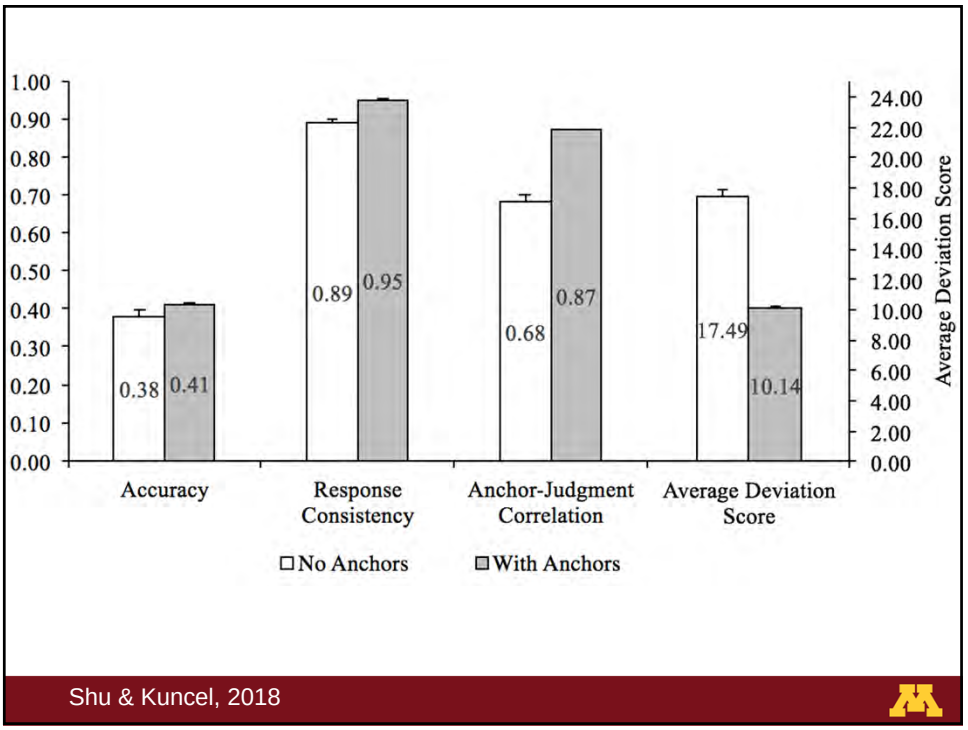
Neuroticism:	43%
Extroversion:	72%
Openness:	54%
Agreeableness:	64%
Conscientiousness:	68%

Effects of Anchors on Weighting Policy: How are they using the information?



Effects of Anchors on Weighting Policy: How are they using the information?





Shu & Kuncel, 2018

Validating Hiring Systems: Focus on the Average Performance

Decision Making of both the hiring manager and the applicant can undermine the most beautiful hiring system.

Kuncel (2018). *JDM in Staffing Research and Practice*.

You Develop a New Hiring System

Good job analysis

Careful predictor build

Quality criterion measure

r = .45

All is right in the world, yes?

Kuncel (2018). *JDM in Staffing Research and Practice*.

No

- In this case, the correlation between hiring decisions and your assessment scores are ZERO.
- Judgments were unrelated to your system.
- The organization has realized **negative** utility from your work.

The Effect of Our Work on Decision Making is the Key

The Cassandra Problem

Our validation methods assume that decision makers listen to us.

Often they don't.

R = 1.0 doesn't mean anything if decision makers chose employees based on their firm handshakes.

Kuncel (2018) *JDM in Staffing Research and Practice*.

We Are Correlation Collectors and Predictor Optimizers

Table 1
Predictive Validity for Overall Job Performance of General Mental Ability (GMA) Scores Combined With a Second Predictor Using (Standardized) Multiple Regression

Personnel measures	Validity (r)	Multiple R	Gain in validity from adding supplement	% increase in validity	Standardized regression weights	
					GMA	Supplement
GMA tests[a]	.51					
Work sample tests[b]	.54	.63	.12	24%	.36	.41
Integrity tests[c]	.41	.65	.14	27%	.51	.41
Conscientiousness tests[d]	.31	.60	.09	18%	.51	.31
Employment interviews (structured)[e]	.51	.63	.12	24%	.39	.39
Employment interviews (unstructured)[f]	.38	.55	.04	8%	.43	.22
Job knowledge tests[g]	.48	.58	.07	14%	.36	.31
Job tryout procedure[h]	.44	.58	.07	14%	.40	.20
Peer ratings[i]	.49	.58	.07	14%	.35	.31
T & E behavioral consistency method[j]	.45	.58	.07	14%	.39	.31
Reference checks[k]	.26	.57	.06	12%	.51	.26
Job experience (years)[l]	.18	.54	.03	6%	.51	.18
Biographical data measures[m]	.35	.52	.01	2%	.45	.13
Assessment centers[n]	.37	.53	.02	4%	.43	.15
T & E point method[o]	.11	.52	.01	2%	.39	.29
Years of education[p]	.10	.52	.01	2%	.51	.10
Interests[q]	.10	.52	.01	2%	.51	.10
Graphology[r]	.02	.51	.00	0%	.51	.02
Age[s]	−.01	.51	.00	0%	.51	−.01

Schmidt Hunter (1998) cited over 6,000 times

"The primary inference of concern in an employment context is that test scores predict subsequent work behavior."
– Van Iddekinge & Ployhart (2008)

(I think Chad and Rob are exceptional scholars but I disagree about this point).

"The primary inference of concern in an employment context is that hiring systems cause better people to be hired."
– Kuncel (right now)

Incremental Validity Part 1:
Adding a New Predictor to the Hiring System

Predictor A, r = .30, ΔR^2 = .00

Predictor A = Failure, Drop

Right?

Kuncel (2018) *JDM in Staffing Research and Practice*.

No

Predictor A is

- Highly face valid
- Sexy (for an assessment)
- Gets a heavy weight in decision making…forcing out irrelevant factors.
- Decision quality improves.

Incremental Validity Part 2:
This Time It's Personal

Predictor C, r = .15, ΔR^2 = .01
Fast, Very low cost, People like it

Predictor C = It helps, Keep

Right?

No

- **Predictor C** has strong *narrative* qualities that cause it to get excessive judgment weight dragging down overall decision quality.
- In other words, people now ignore the other higher quality information in the system.

Kuncel (2018) *JDM in Staffing Research and Practice*.

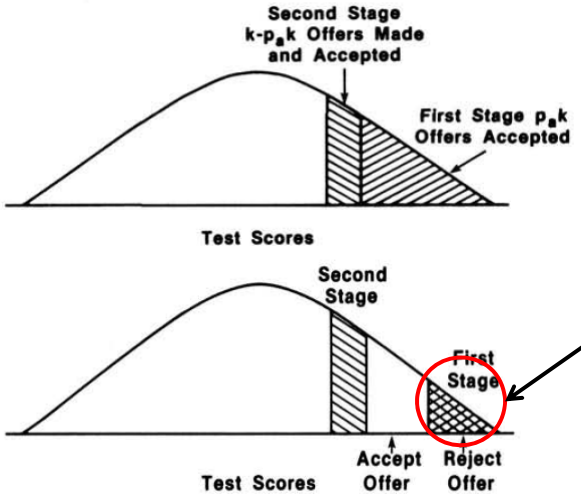
Thank You!

Nathan Kuncel

kunce001@umn.edu

UNIVERSITY OF MINNESOTA
Driven to Discover®

The Best May Not Accept



They don't love you

Murphy (1986)

The Best May Not Apply

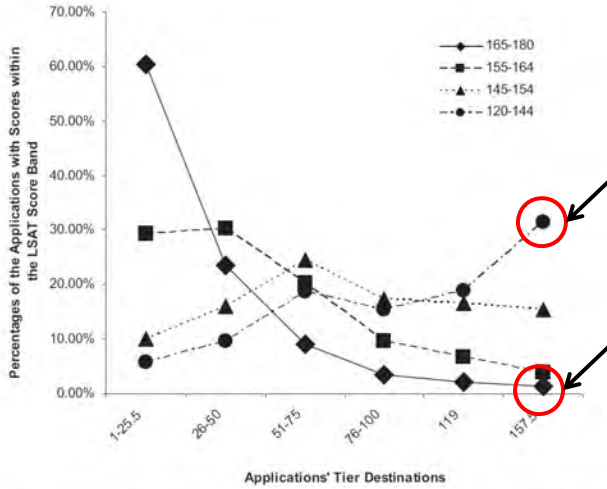


Figure 1. Percentage of applications from four Law School Admission Test (LSAT) score bands to six law school tiers.

Weakest LSAT, Lowest Ranked Schools

Strongest LSAT, Lowest Ranked Schools

Kuncel & Klieger (2007)