Psychology 405: Psychometric Theory Validity

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

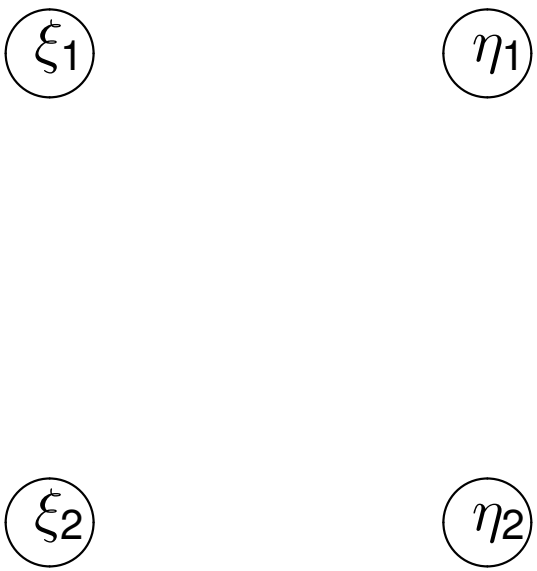
May, 2025

Outline

Observed Variables

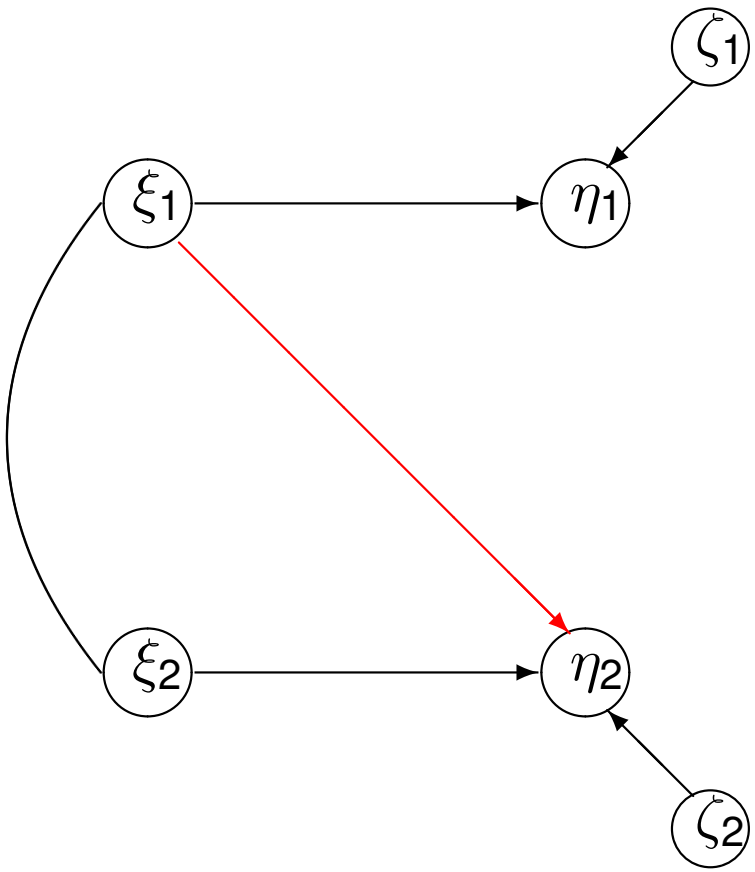
Latent Variables

ξ η



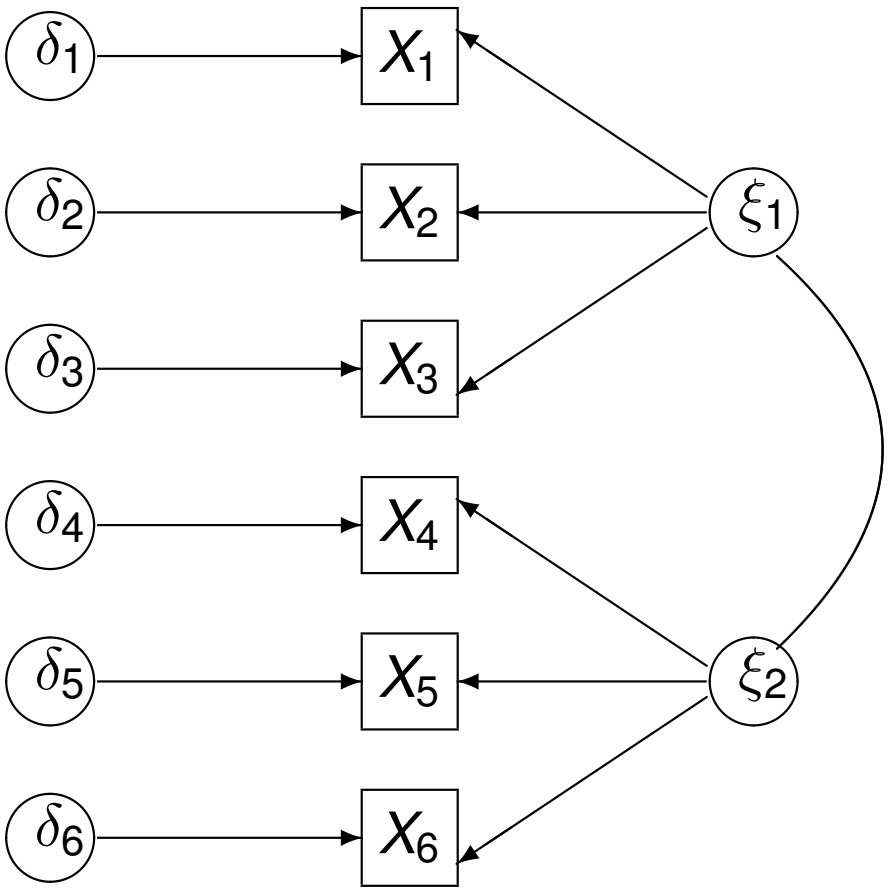
Theory: A regression model of latent variables

ξ η

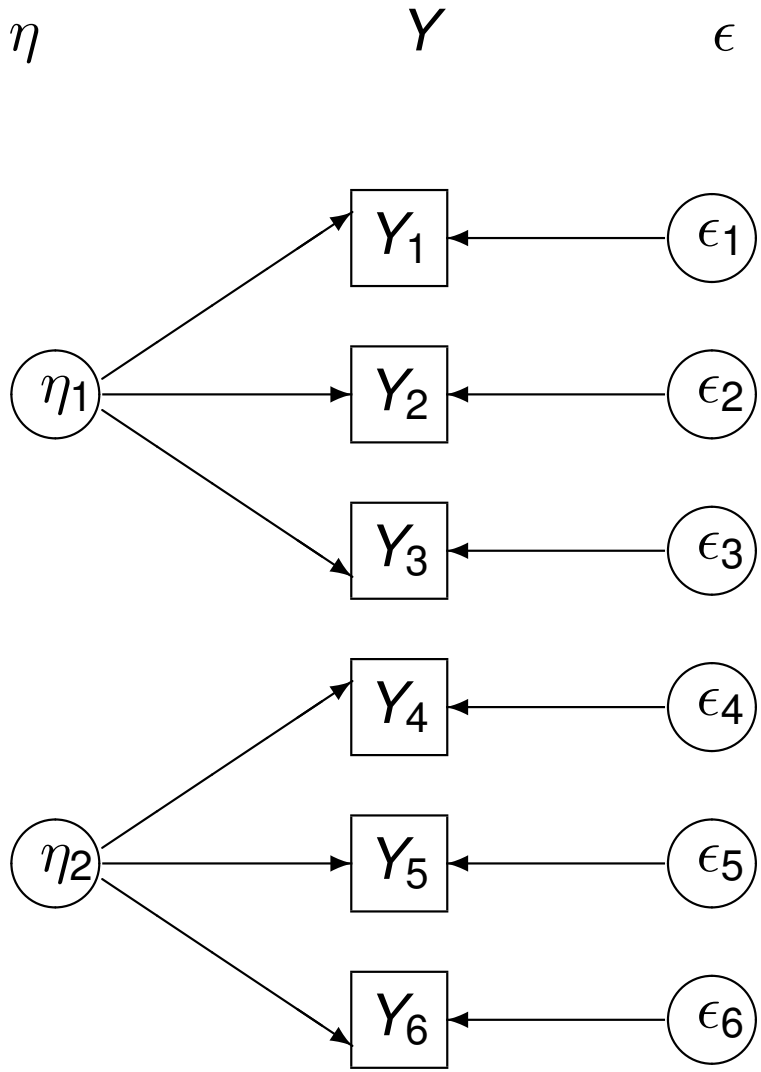


A measurement model for X – Correlated factors

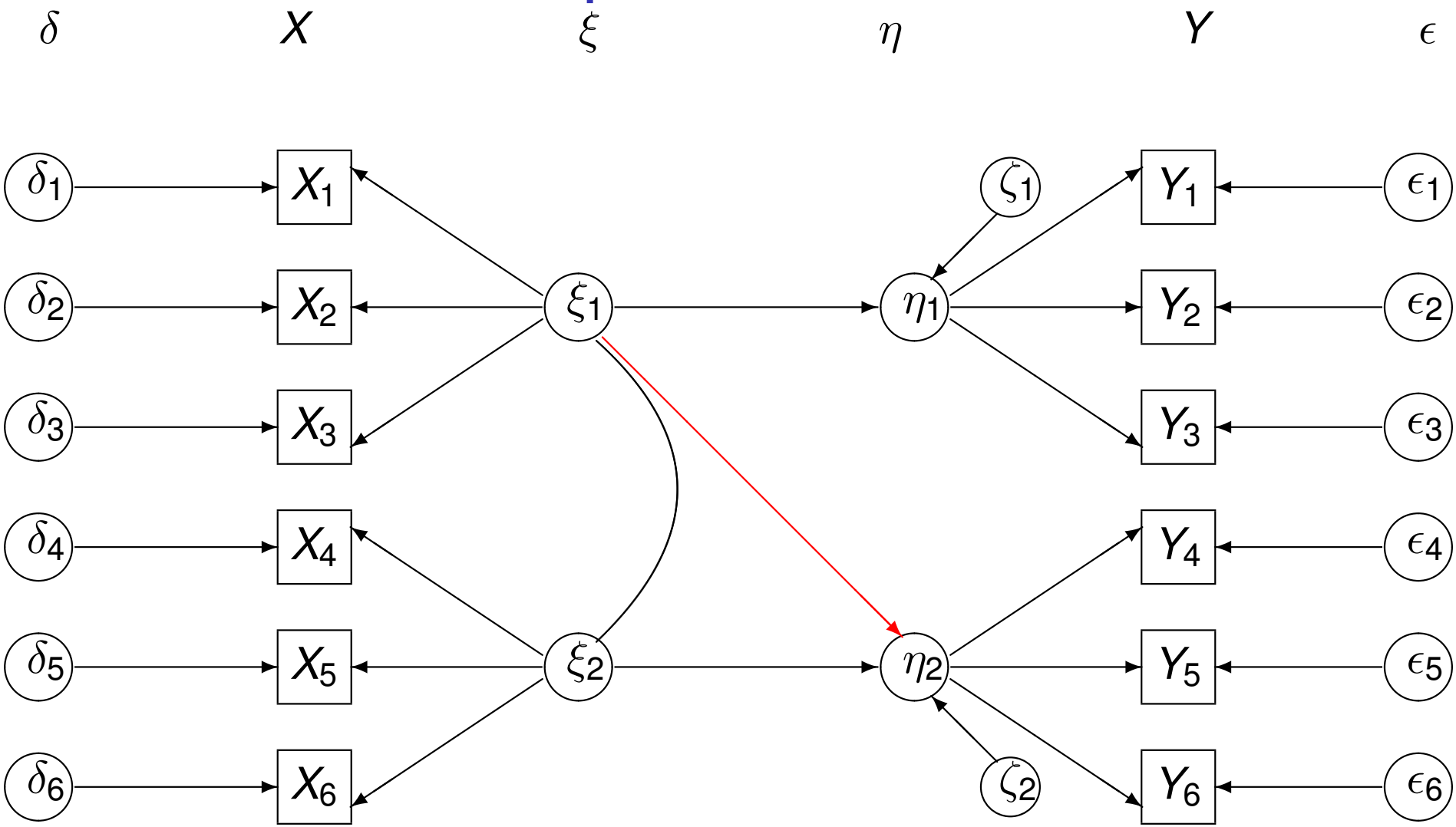
δ X ξ



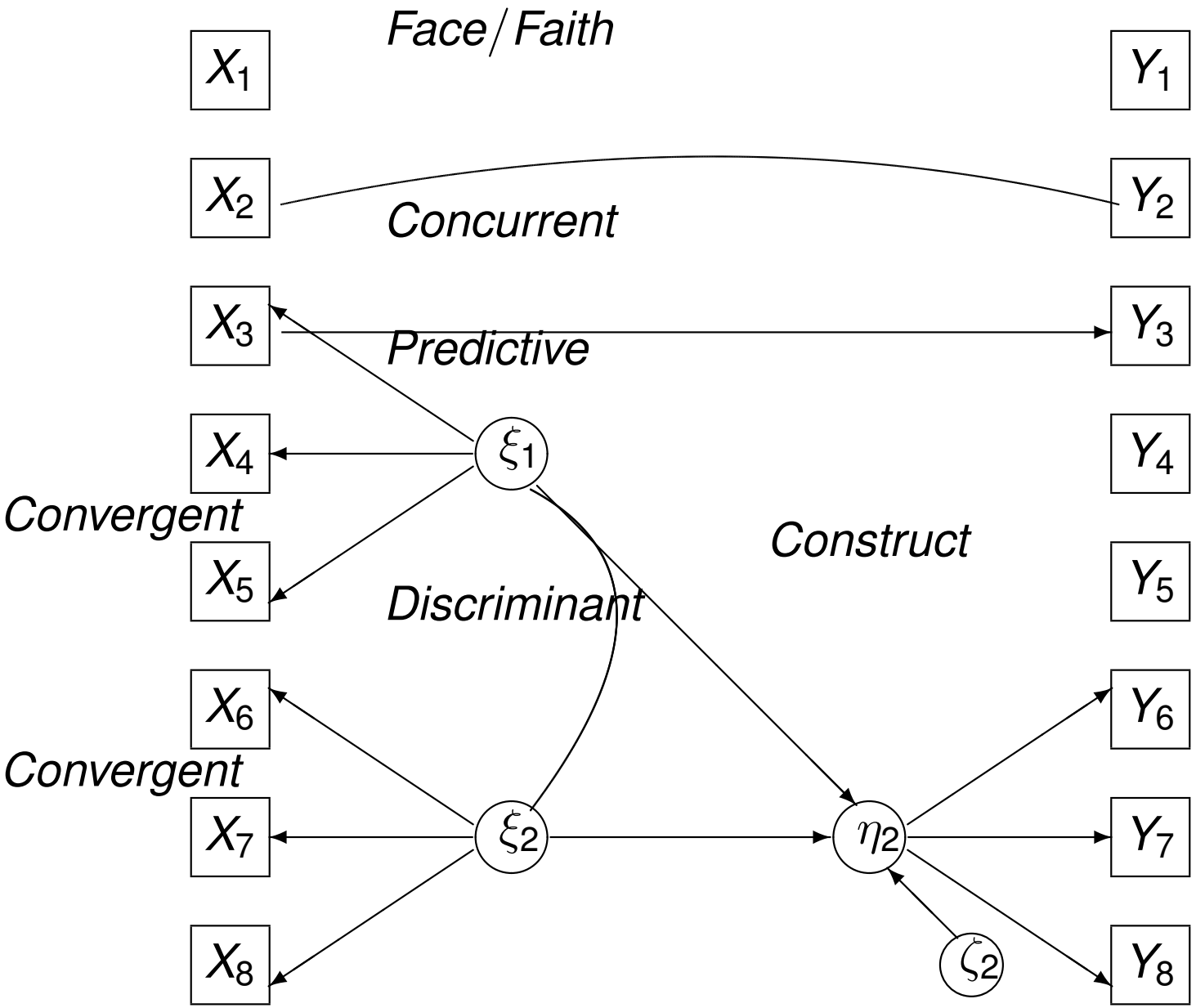
A measurement model for Y - uncorrelated factors



A complete structural model



Types of Validity



Face Validity

Face / Faith

X_1

Representative Content

Seeming relevance

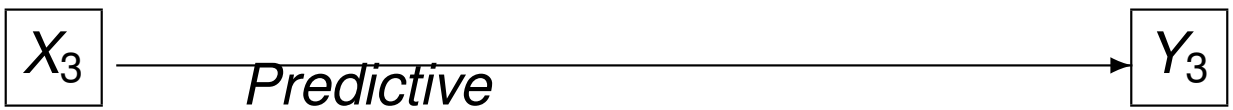
Concurrent Validity

Does a measure correlate with the criterion?

Need to define the criterion.

Assumes that what correlates now will have predictive value.

Predictive Validity



Does a measure correlate with the criterion?

Need to define the criterion.

Allow time to pass

Prediction

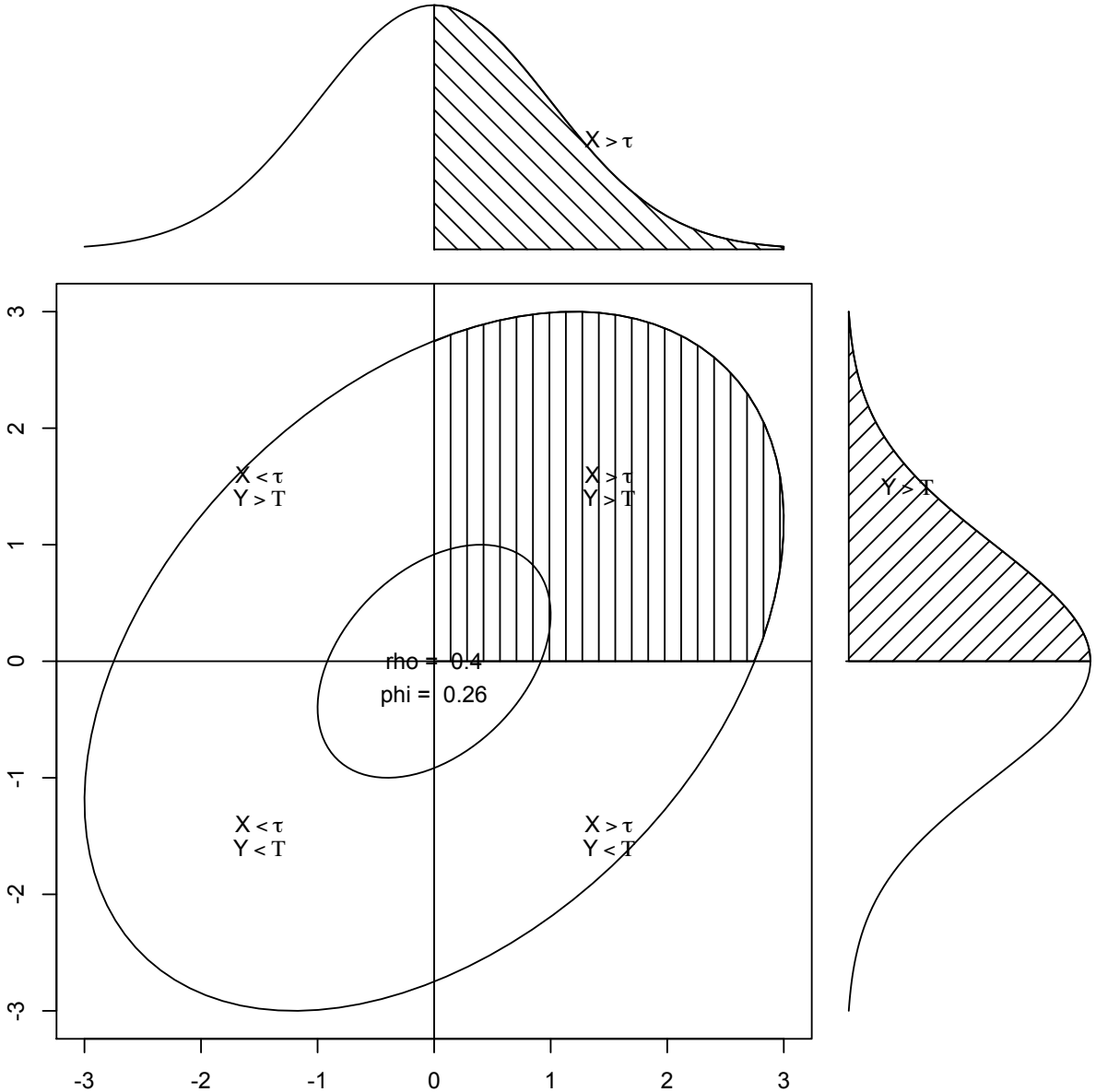
1. Continuous predictor, continuous criterion
   - Regression, multiple regression, correlation
   - Slope of regression implies how much change for unit change in predictor
2. Continuous predictor, dichotomous criterion
   - point bi-serial correlation
3. Dichotomous predictor, dichotomous outcome
   - Phi
   - The Taylor-Russell tables (Taylor and Russell, 1939) and the problem of Selection Ratios and Base Rates

$$\phi = \frac{VP - BR * SR}{\sqrt{(BR)(1 - BR)(SR)(1 - SR)}} \tag{1}$$

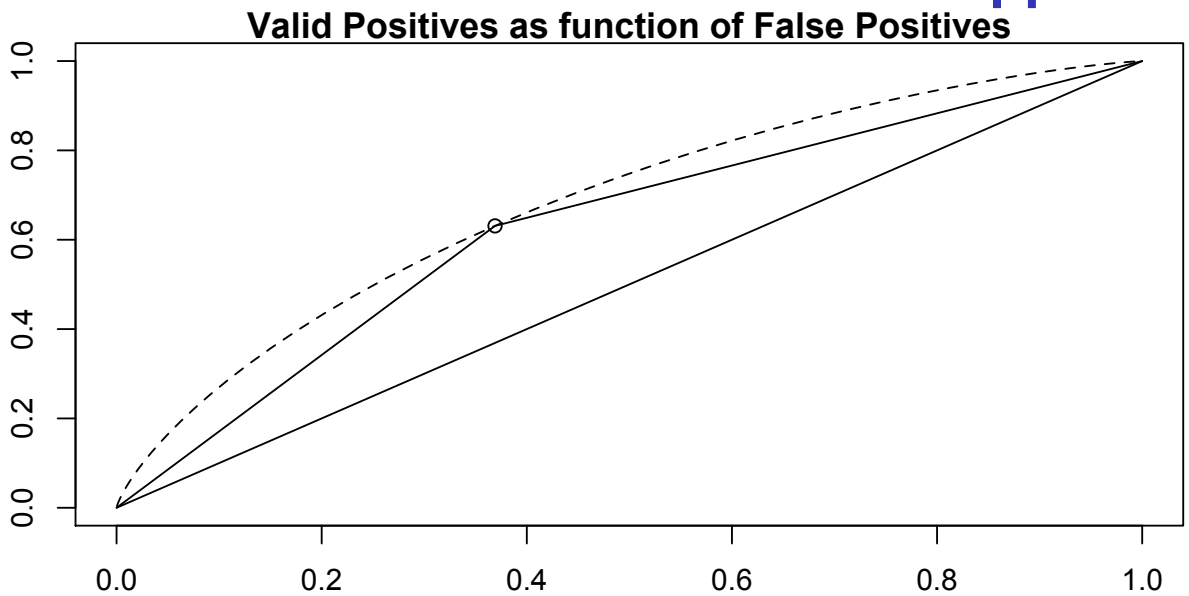
   - Therefore, the number of valid positives is

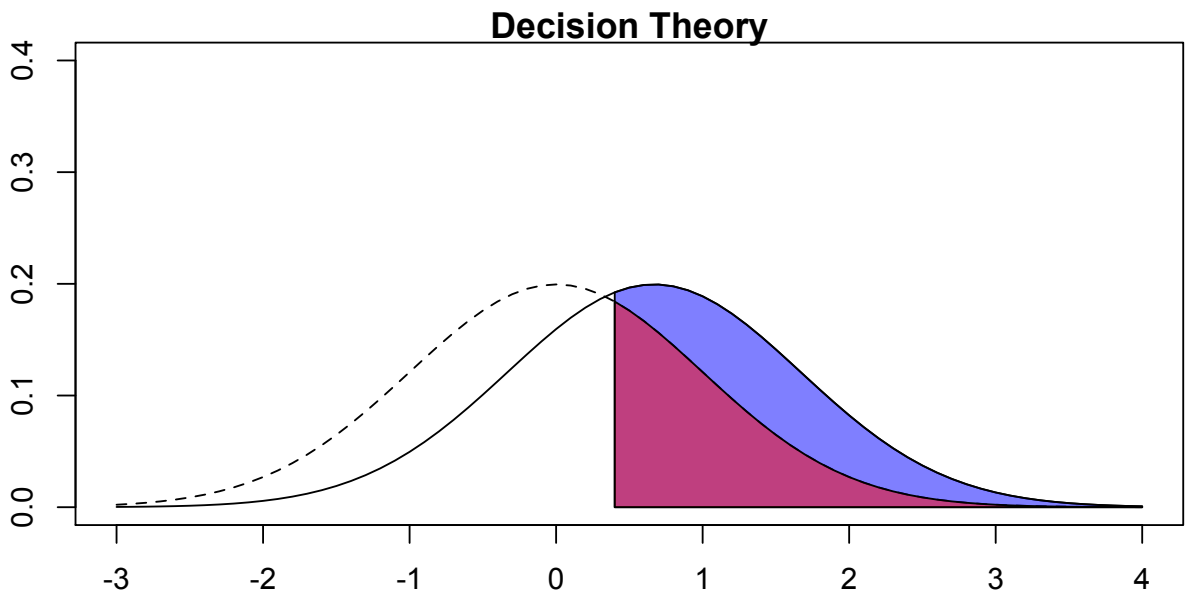
$$VP = BR * SR + \phi\sqrt{(BR)(1 - BR)(SR)(1 - SR)} \tag{2}$$

Tetrachoric and phi as function of cut points

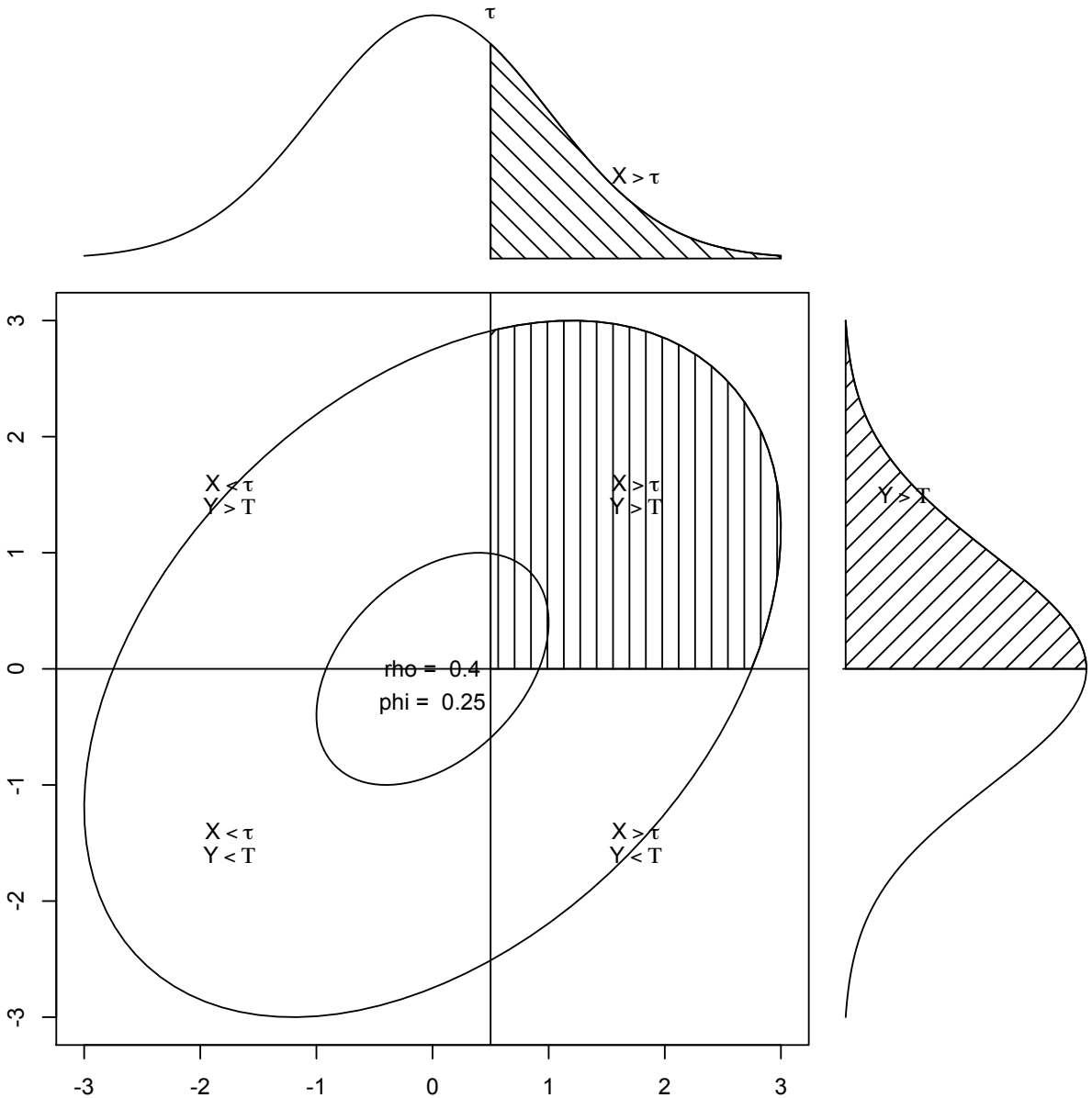


A decision theoretic approach

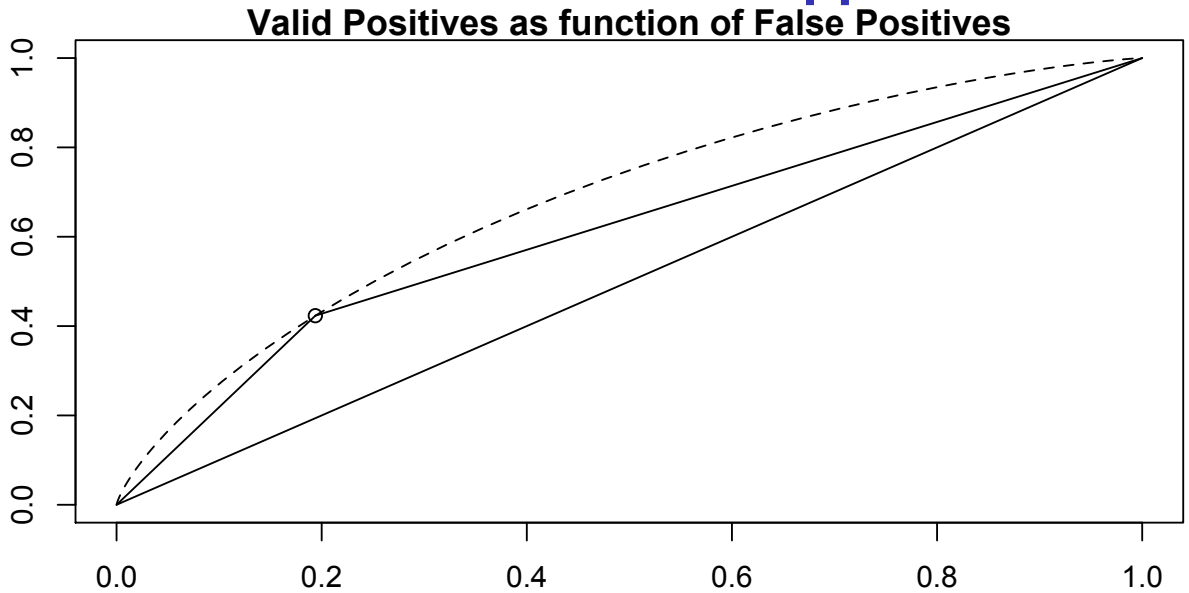


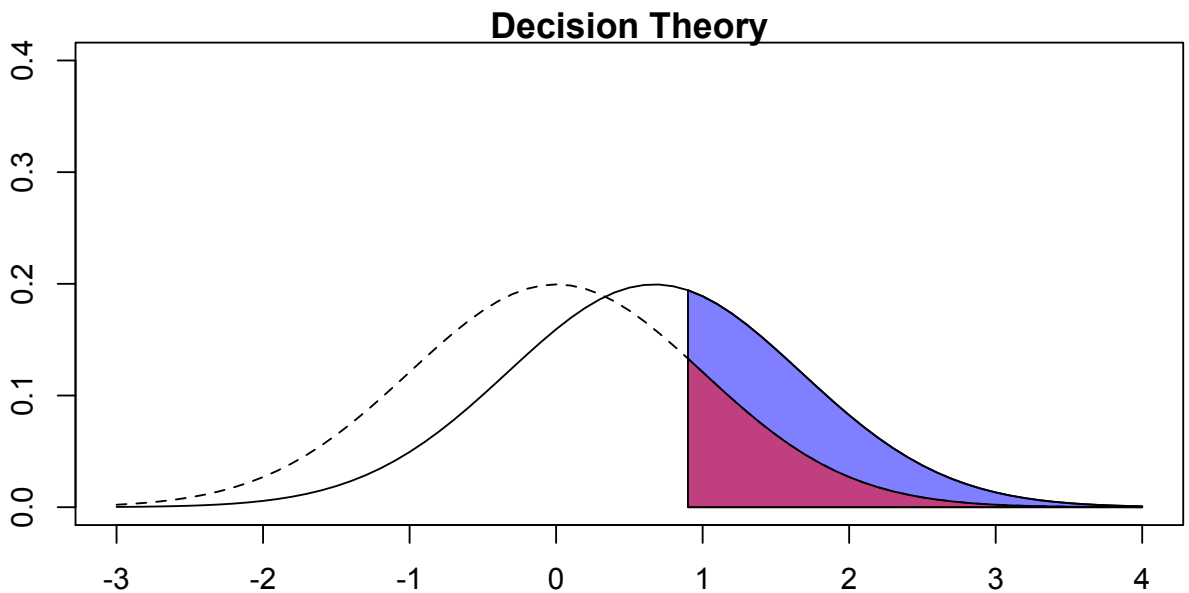


Tetrachoric and phi as function of cut points .5,0

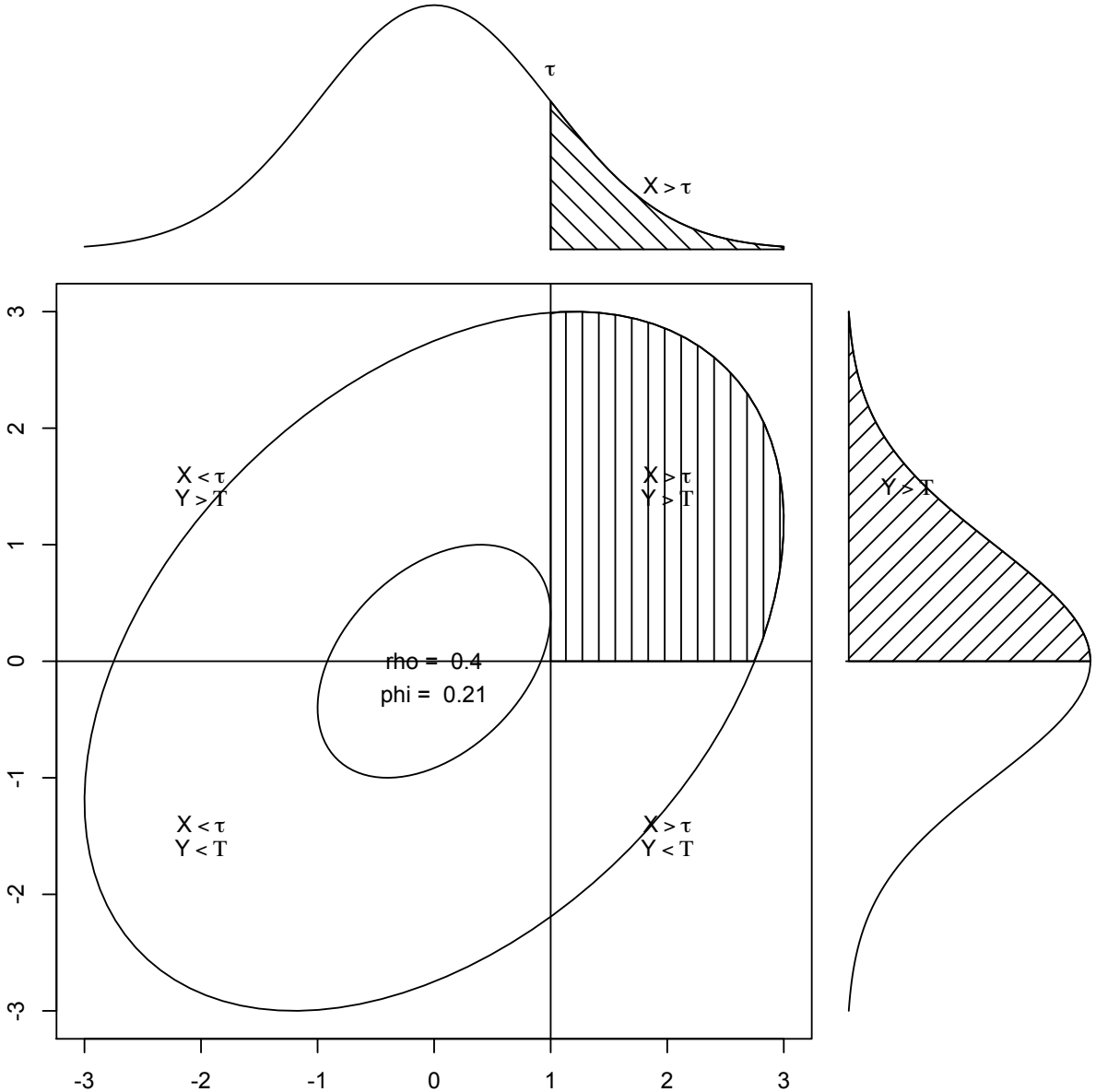


A decision theoretic approach with low beta

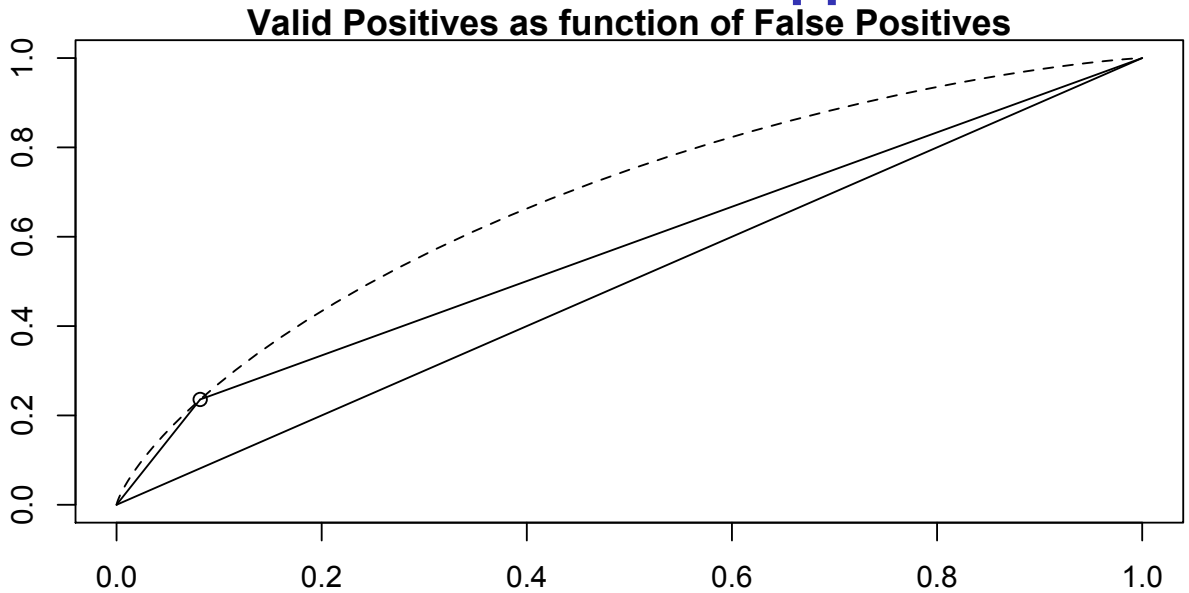


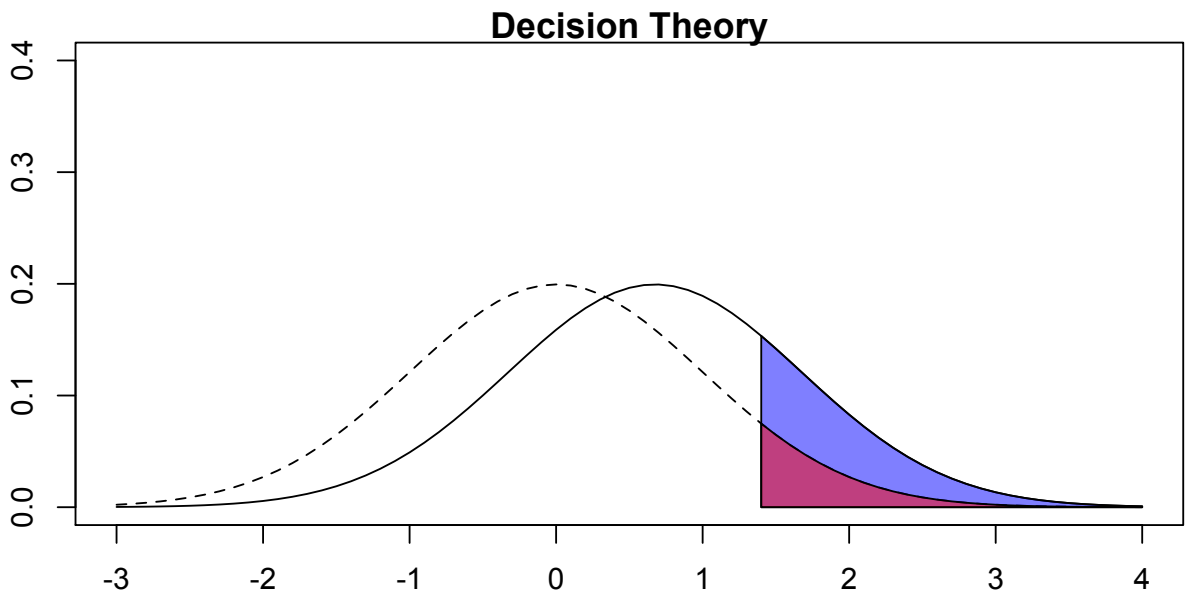


Tetrachoric and phi as function of cut points 1,0

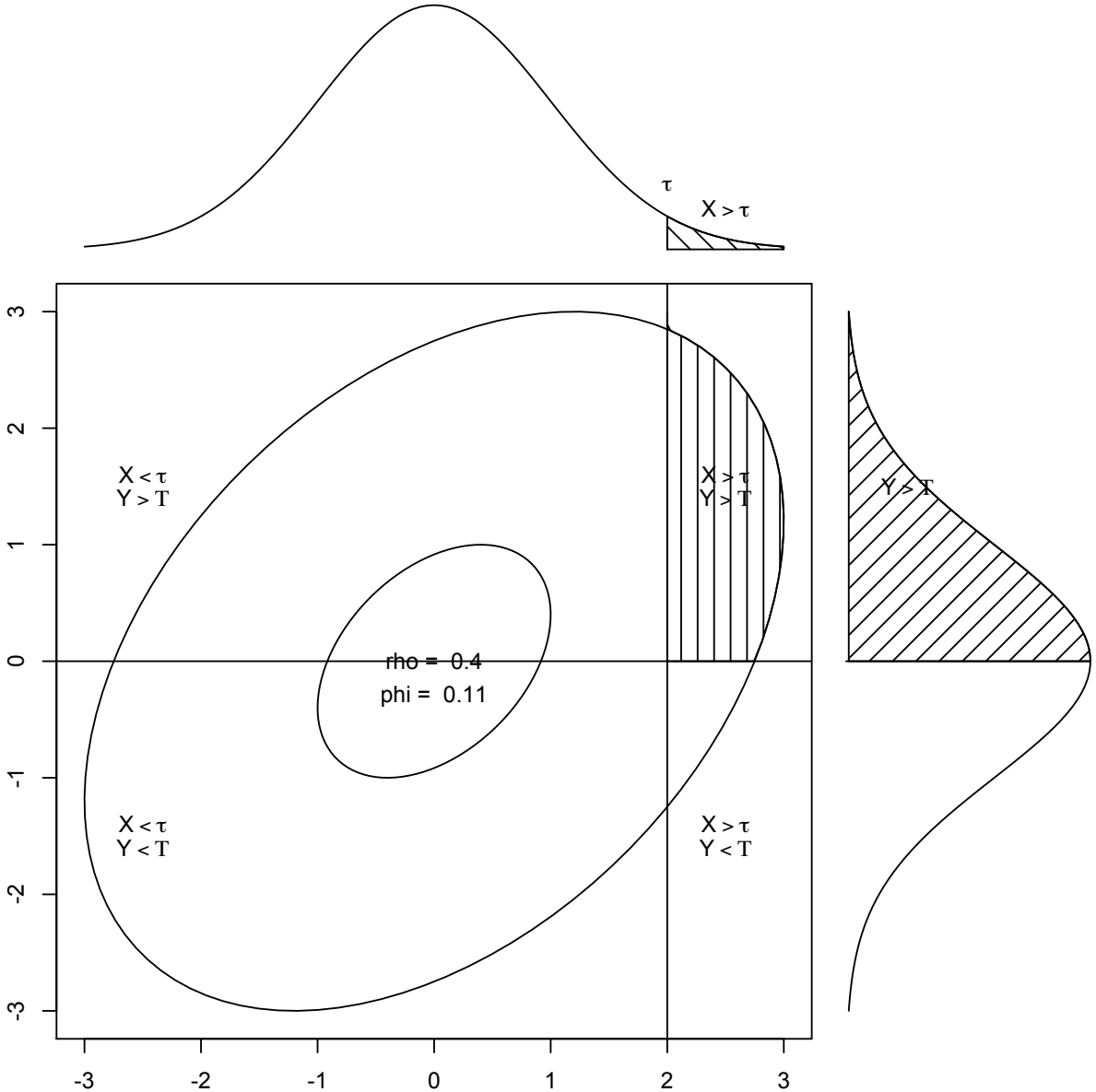


A decision theoretic approach with high beta

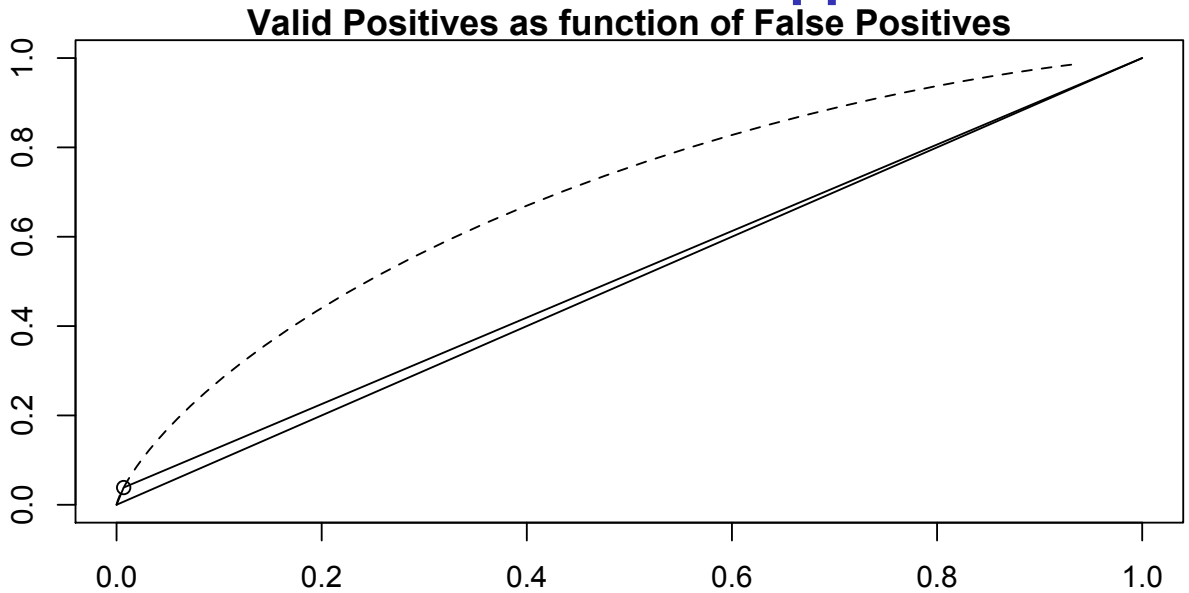


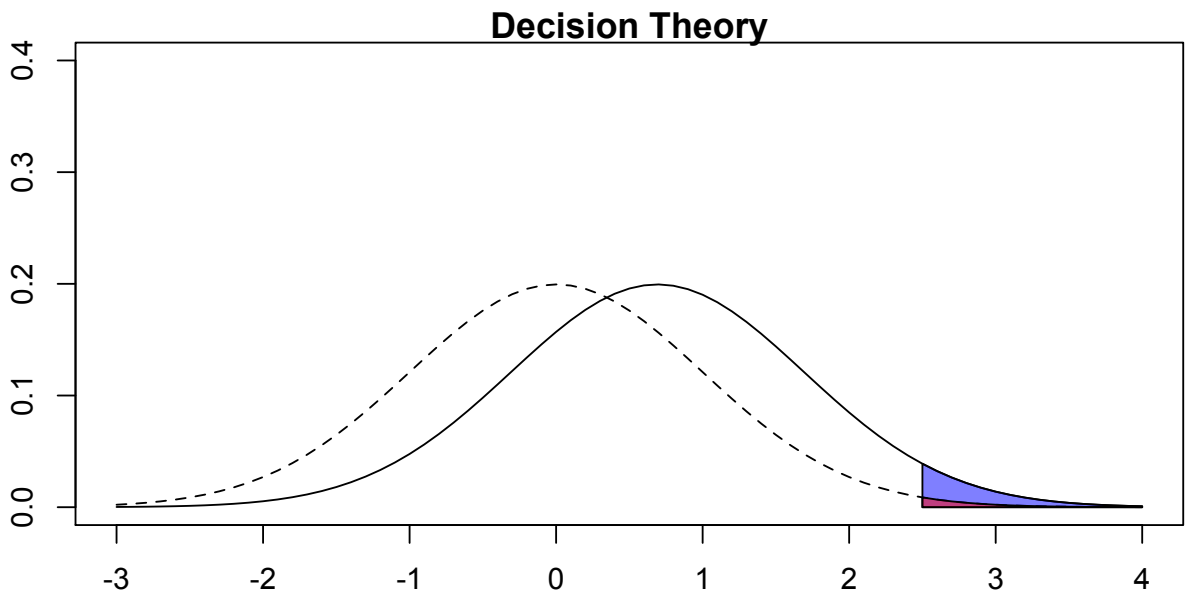


Tetrachoric and phi as function of cut points 2,0

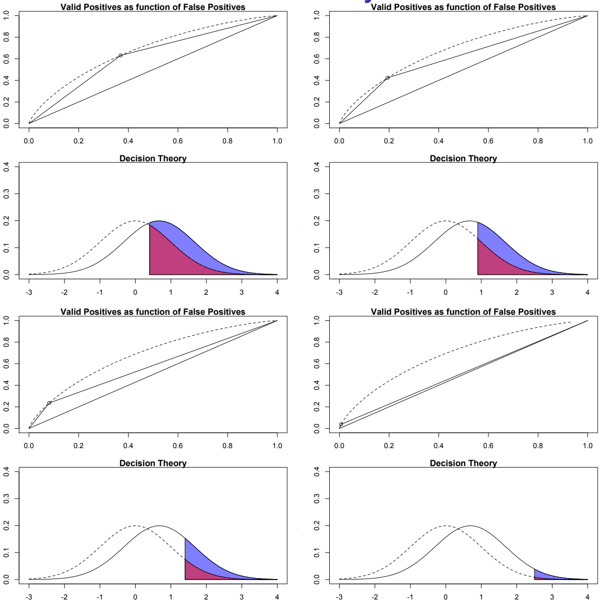


A decision theoretic approach with high beta





A decision theoretic analysis with 4 different cut points



Applying decision theory to a prediction problem: the case of predicting future psychiatric diagnoses from military inductees. (Data from Danielson and Clark (1954) as discussed by Wiggins (1973).

	Predicted Positive	Predicted Negative	Row Totals
True Positive	49	40	99
True Negative	79	336	406
Column Totals	118	376	505

Fraction of Total

	Predicted Positive	Predicted Negative	Row Totals
True Positive	.097	.079	.196
True Negative	.157	.667	.804
Column Totals	.234	.746	1.00

Accuracy = .097 + .667 = .76

Sensitivity = .097/(.097 + .079) = .55

Specificity = .667 / (.667+.157) = .81

Phi = $\frac{.097 - .196 * .234}{\sqrt{.196 * .804 * .234 * .747}} = .32$

What if we went with the base rates and predicted everyone was healthy?

	Predicted Positive	Predicted Negative	Row Totals
True Positive	.0	.196	.196
True Negative	.0	.804	.804
Column Totals	.0	1.000	1.00

Accuracy = 0 +.804 =.80

Sensitivity = 0/.196 = 0

Specificity = .196 / 1.0 = .196

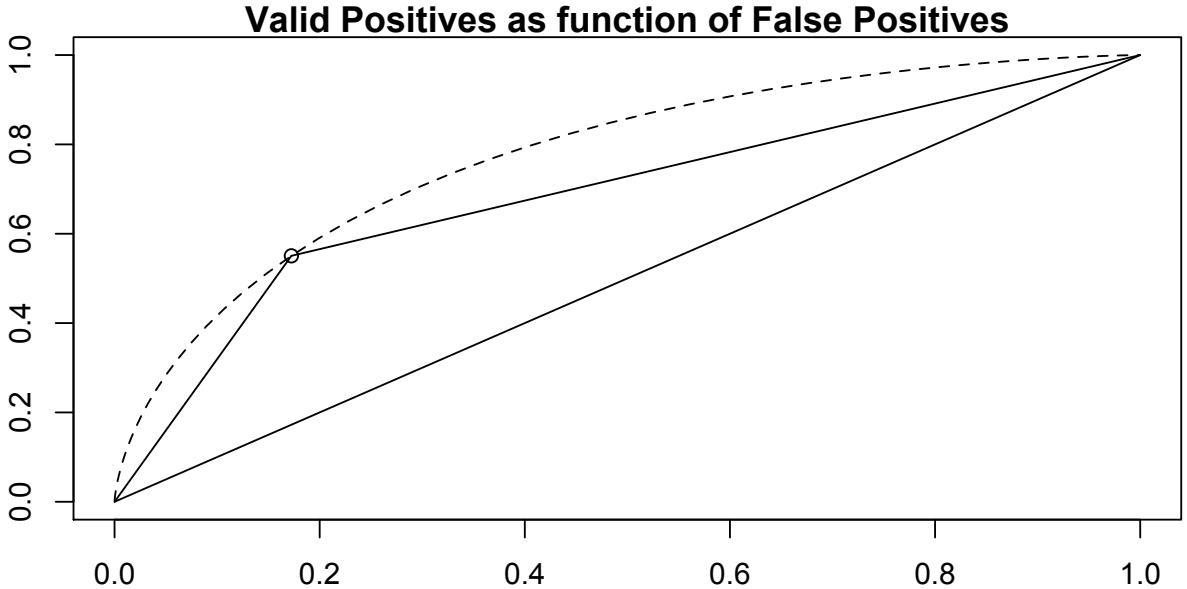
The Danielson and Clark data as a decision problem

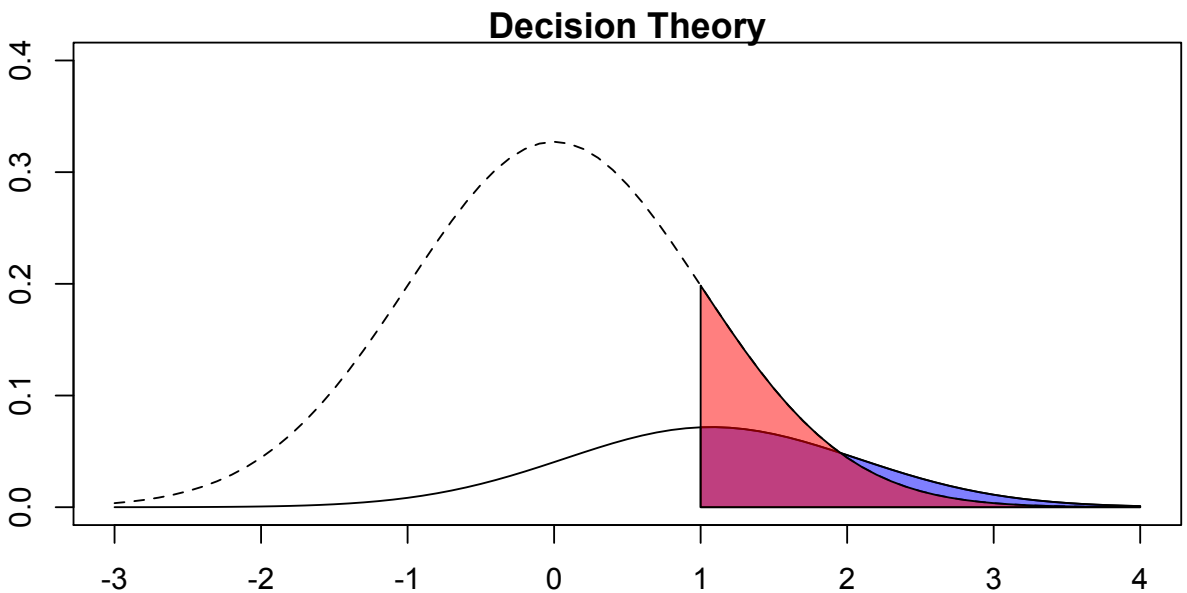
R code

```
dd
         pred.pos pred.neg
obs.pos        49       79
obs.neg        40      336


AUC(dd)
```

```
AUC(dd)

Decision Theory and Area under the Curve

The original data implied the following 2 x 2 table
         Predicted.Pos Predicted.Neg
True.Pos         0.097         0.079
True.Neg         0.157         0.667

Conditional probabilities of
         Predicted.Pos Predicted.Neg
True.Pos          0.55          0.45
True.Neg          0.19          0.81

Accuracy =  0.76  Sensitivity =  0.55   Specificity =  0.81
with Area Under the Curve =  0.76
d.prime =  1  Criterion =  0.88  Beta =  0.15
Observed Phi correlation =  0.32
Inferred latent (tetrachoric) correlation  =  0.53
```

The Danielson and Clark data set as a decision problem





Classics in Prediction and selection

1. Gideon's sequential selection of soldiers (The Hebrew Bible: Judges 6-7, McPherson, 1901)
   - Gideon was ordered to select troops to fight the Midianites,
   - From 32,000 volunteers, selected the 10,000 non-timid but then selected 300 battle ready by observing how they drank water.
2. Selecting spies (OSS Assessment Staff, 1948) and Pilots Army Air Corps selection studies (Dubois, 1947).
3. Kelly and Fiske (1950) (1950) selection of psychology students (Kelly and Fiske, 1951).
4. Astronaut selection, from 10,000 to 7. (selecting for the "Right Stuff" Wolfe, 1970).
5. Peace Corps selection as a process of sequential selection
   - Psych Testing, Peer ratings, Staff ratings after 6 weeks
   - Peer ratings, staff ratings after 12 weeks of training
   - But what was the criterion?

Gideon's assessment (McPherson, 1901)

Predictor set is positively correlated (Dubois, 1947)

60

TABLE 3.8.—*Intercorrelations of classification battery of August 1942, aviation cadets tested at Psychological Research Unit No. 2 and trained in elementary pilot class 43–D*

[N=1,520]

Variable	Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	M	S. D.	Validities	
																								Bombing	Pilot
Printed tests:																									
1. Technical vocabulary-pilot	CE505C		0.36	0.43	0.17	0.15	0.07	0.40	0.29	−0.04	−0.04	0.14	0.24	0.09	0.11	0.09	−0.01	0.11	0.09	0.02	0.15	16.0	7.1	0.10	0.24
2. Technical vocabulary bombardier	CE505C	0.36		.37	.11	.23	.14	.37	.21	.06	.09	.17	.21	.18	.18	.18	.03	.04	.12	.03	.10	4.8	2.7	.00	.10
3. Technical vocabulary-navigator	CE505C	.43	.37		.13	.58	.27	.53	.34	.14	.18	.20	.23	.35	.27	.28	.00	.05	.16	.01	.12	11.3	6.5	.06	.19
4. Speed of identification	CP610A	.17	.11	.13		.11	.07	.09	.11	.13	.10	.47	.35	.03	.16	.22	.06	.09	.18	.10	.16	23.6	6.9	(1)	.24
5. Mathematics	CI702E	.15	.23	.58	.11		.45	.46	.28	.32	.39	.22	.21	.54	.42	.41	−.03	.09	.27	.03	.15	19.9	14.1	.03	.20
6. Numerical approximation	CI706A	.07	.14	.27	.07	.45		.34	.16	.42	.47	.18	.15	.55	.44	.34	.00	.05	.19	.07	.09	10.9	5.6	.02	.09
7. Reading comprehension	CI619D	.40	.37	.53	.09	.46	.34		.36	.20	.25	.23	.26	.45	.40	.35	−.02	.14	.24	.07	.20	33.9	10.4	.15	.26
8. Mechanical comprehension	CI908D	.29	.21	.34	.11	.28	.16	.36		−.02	.04	.15	.22	.30	.19	.17	.00	.27	.17	.07	.24	8.5	5.5	.13	.28
9. Numerical operations I	CI701A	−.04	.08	.14	.13	.32	.42	.20	−.02		.66	.27	.08	.38	.42	.40	.02	.01	.19	.11	.08	16.9	5.8	.15	.06
10. Numerical operations II	CI701A	−.04	.09	.18	.10	.39	.47	.25	.04	.66		.27	.12	.46	.46	.40	−.01	.00	.24	.07	.12	15.7	6.0	.13	.08
11. Spatial orientation part I	CP501B	.14	.17	.29	.47	.22	.18	.23	.16	.27	.27		.46	.21	.35	.39	.04	.12	.27	.14	.24	27.2	5.5	.16	.30
12. Spatial orientation part II	CP503B	.24	.21	.23	.35	.21	.15	.26	.26	.08	.12	.46		.21	.31	.26	.03	.14	.26	.10	.22	21.0	6.8	.15	.23
13. Arithmetic reasoning	CI206B	.09	.18	.35	.03	.54	.55	.45	.30	.38	.46	.21	.21		.48	.35	−.01	.02	.21	.06	.15	12.7	5.5	.13	.17
14. Dial reading	CP622A	.11	.18	.27	.16	.42	.44	.40	.19	.42	.46	.36	.31	.48		.48	−.01	.16	.27	.03	.22	24.0	7.4	.18	.22
15. Table reading	CP621A	.09	.18	.28	.22	.41	.34	.38	.17	.40	.40	.39	.26	.35	.48		.04	.17	.30	.15	.24	39.7	14.1	.25	.22
Apparatus tests:																									
16. SAM steadiness	CM103A	−.01	.03	.00	.06	−.03	.00	−.02	.00	.02	−.01	.04	.03	−.01	−.01	.04		.07	.00	.16	.12	49.5	10.9	.02	.03
17. Two hand coordinator	CM101A	.11	.04	.05	.09	.09	.05	.14	.27	.01	.00	.12	.14	.09	.16	.17	.07		.16	.19	.39	51.3	9.9	.23	.24
18. SAM reaction time	CP611D	.09	.12	.16	.18	.27	.19	.24	.17	.19	.24	.27	.26	.21	.27	.30	.00	.16		.15	.32	51.7	9.9	.26	.21
19. Finger dexterity	CM116A	.02	.03	.01	.10	.03	.07	.07	.07	.11	.07	.14	.10	.06	.09	.15	.16	.19	.15		.22	52.7	10.1	.34	.10
20. Complex coordinator	CM701A	.15	.10	.12	.16	.15	.09	.20	.24	.08	.12	.24	.22	.15	.22	.24	.12	.39	.32	.22		50.4	9.9	.32	.34

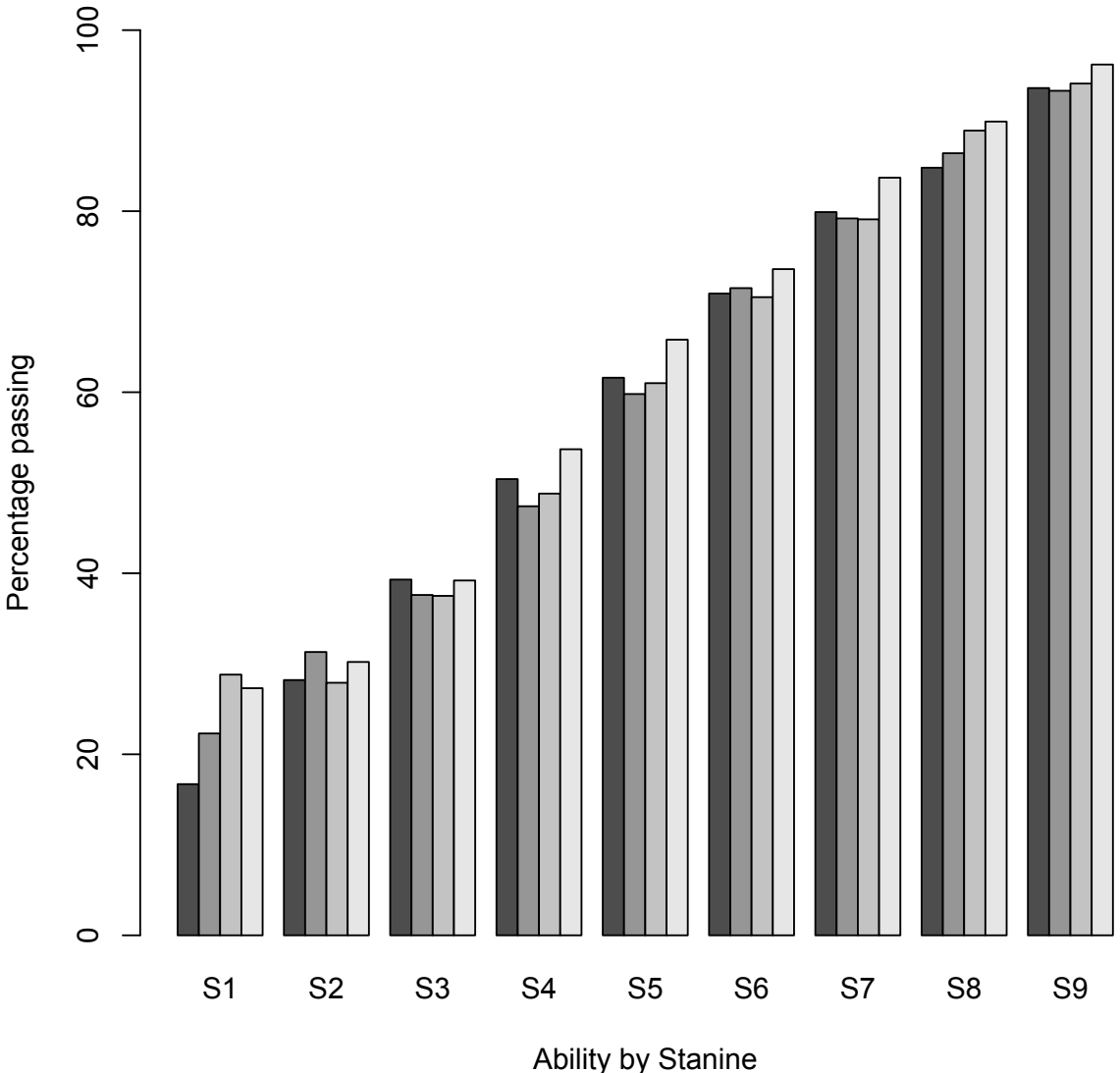
Validities are small but meaningful

tery, together with approximate numbers of cases

Test	Code	Bombardier		Navigator		Pilot	
		r	N	r	N	r	N
Printed tests:							
Reading comprehension [1]	CI614G	0.12	3,200	0.32	700	0.19	7,400
Spatial orientation II	CP703B	.09	3,200	.33	700	.25	9,100
Spatial orientation I	CP701B	.12	3,200	.38	700	.20	9,100
Dial and table reading	CP621A, 622A	.19	3,200	.53	700	.19	3,200
Biographical data—pilot	CE602D					.32	(*)
Biographical data—navigator	CE602D			.23	200	−.03	1,000
Mechanical principles [1]	CI603A	.08	1,800	.13	300	.32	8,100
Technical vocabulary—pilot [1]	CE605C	.01	3,200	.10	700	.30	(*)
Technical vocabulary—navigator	CE605C	.01	3,200	.22	700	.09	13,300
Mathematics [1]	CI702E	.0	3,200	.40	(*)	.08	16,200
Arithmetic reasoning [1]	CI206B	.12	3,200	.45	(*)	.09	10,400
Instrument comprehension I	CI615A					.15	600
Instrument comprehension II	CI616A					.35	600
Numerical operations, front	CI702H	.13	3,200	.28	1,600	.01	9,100
Numerical operations, back	CI702B	.11	3,200	.25	1,600	.02	9,100
Speed of identification	CP610A	.09	3,200	.19	1,500	.18	20,100
Apparatus tests:							
Rotary pursuit [1]	CM806A	.14	1,800	.10	700	.21	8,100
Complex coordination	CM701A	.18	3,200	.21	700	.25	24,100
Finger dexterity	CM116A	.16	3,200	.20	700	.11	18,200
Discrimination reaction time	CP611D	.22	3,200	.36	700	.22	13,700
Two-hand coordination	CM101A	.12	2,200	.26	700	.30	12,600
Rudder control	CM120A					.42	1,000

[1] Pattern of July 1942 included a different form of this test.

The assessment of pilots – how to show a .45 correlation makes a difference



Predicting clinical psychologists – Kelly and Fiske

1. Multiple predictors of graduate school performance: Kelly and Fiske (1950), Multiple predictors
2. Ability, Interests, temperament (each with r $\approx$ .2 -.25) have multiple R of .4-.5
3. Are they able, interested and stable?

VA study: overview

- Researchers
  - nearly 40 cooperating clinical training programs
  - $\approx$ 75 psychologists on research staff
- Participants
  - 3/4 of those entering graduate training in 1946, 1947, 1948
  - N = 160, 128, 545 (selected down to 98)
- Measures
  - Objective tests
  - Clinical assessments

Objective instruments

- Ability
  - Millers Analogy Test
  - Thurstone Tests of Primary Mental Abilities
- Temperament and Character
  - Minnesota Multiphasic Personality Inventory
  - Guildord Martin Battery of Personality Inventories
- Interests, Values
  - Allport-Vernon Scale of Values
  - Strong Vocational Interest Blank
  - Kuder Preference Record

Assessment ratings

- Seven days of tests, interviews and “other” procedures
  - Three raters spent a week studying 4 trainees
  - Staff time devoted to each candidate was at least 7 man-days
- Ratings based on interviews, projective tests, role playing
  - Ratings on:
    - 22 descriptive variables (e.g., cooperativeness, talkativeness)
    - 10 evaluative variabels (e.g., social adjustment, emotional expression)
    - 11 predictive variables (e.g. academic, diagnostician, overall suitability)

Criterion variables after 2 years

- Training status (Failure, still in Training, Ph.D. obtained)
- 2nd year evaluations
  - Skill in clinical diagnosis
  - Skill in individual psychotherapy
  - Skill in Research
  - Preference for hiring
- Generally high correlations among all the criteria

High correlations among the criteria

Intercorrelations among selected criterion evaluations

N = 130 P-3 trainees evaluated in the spring of 1949, for whom all evaluation measures were available.

		Clinical Diagnosis		Individual Therapy		Research		Preference for Hiring	
		Univ.[1]	Instal.[2]	Univ.	Instal.	Univ.	Instal.	Univ.	Instal.
Clinical Diagnosis:	Univ.	*72*	47	81		55		88	
	Instal.		*79*		60		54		65
Individual Therapy:	Univ.		38	*73*	63	48		78	
	Instal.	62			*86*		37		74
Research:	Univ.		28		31	*65*	54	76	
	Instal.	11		30			*85*		56
Preference for Hiring:	Univ.		34		70		31	*71*	56
	Instal.	44		38		40			*85*

[1] University staff evaluations.
[2] Installation staff evaluations.

The more they know about you, the more they will judge you

Assessor	Information on Which Predictions Were Based										Criterion Evaluations							
	Rorschach	TAT	Sentence Compl.	Bender-Gestalt	Credentials	Objective Tests	Autobiographical Materials	Interview	Situations	Other	Clinical Diagnosis		Individual Therapy		Research		Preference for Hiring	
											Univ.	Instal.	Univ.	Instal.	Univ.	Instal.	Univ.	Instal.
A. Assessment Ratings																		
Projectivist	X										02	08	07	01	17	22	24*	13
		X									17	18	17	17	−06	−07	17	16
			X								18	32**	04	15	25*	25*	12	20
				X							−01	−03	−05	00	00	06	−01	11
Proj. Integration	X	X	X	X							−05	−02	−11	01	12	03	08	18
Initial					X						13	10	22	25*	14	21	09	25*
Interviewer					X			X			21	04	22	16	13	20	09	14
Intensive					X						16	07	16	14	24*	15	20	20
Interviewer					X	X					36**	22	26*	19	32**	22	34**	28*
					X	X	X				19	08	20	16	35**	24*	30**	23*
	X	X	X	X	X	X	X				24*	21	27*	22	40**	33**	33**	33**
	X	X	X	X	X	X	X	X			24*	24*	30**	22	28*	22	31**	33**
Pre-Conference	X	X	X	X	X	X	X				38*	29	32*	19	44**	27	48**	37*
Situationists (Pooled Rating)									X		30**	28*	23*	25*	31**	22	18	36**
Prelim. Pooled	X	X	X	X	X	X	X	X			22	13	26*	22	21	24*	29*	35**
Final Pooled	X	X	X	X	X	X	X	X	X	X	23*	11	30**	18	12	14	21	21

Objectives are just as good

B. Objective Test Scores

Miller Analogies	24*	15	06	05	24*	23*	23*	18
Strong Test								
Psychologist—1938	29*	14	20	21	35**	31**	27*	19
Psychologist—1948 (Kriedt)	36**	20	28*	21	41**	32**	31**	20
Psychologist, Clinical, 1948 (Kriedt)	22	30**	18	16	07	07	17	09
Psychologist, VA Clinical (This Project)	36**	33**	35**	32**	33**	27*	36**	25*
Allport-Vernon Theoretical	23*	−03	16	04	25*	15	23*	−02
Guilford-Martin								
C—Lack of Cycloid Disposition	29*	24*	29*	17	19	11	28*	16
N—Lack of Nervous Tenseness and Irritability	28*	23*	21	24*	21	06	22	20

Interviews might actually hurt!

The finding that the interview did not add to, but actually tended to decrease, the validity of clinical judgments made in the 1947 assessment program was confirmed by submitting the paper and-pencil materials on these same candidates to a later assessment staff which made predictions without any face-to-face contact with the assessee. Under these conditions, the new staff made predictions with slightly higher validities than those made by the staff in 1947, who had the additional data from the interview, situation tests, etc.

Interests matter

The VA Clinical Psychologist key, developed by this project on the basis of the responses of full time VA psychologists, regularly yields relatively high correlations with all criterion evaluations, and compares favorably with the best predictions based on assessment ratings. Other psychologist keys, including the original (1938) general psychologist key and two developed by Kriedt (2), do fairly well. Not shown in the table is a correlation of .61 (N = 44) between scores based on the psychologist key (1938) and the scores made on the objective test of Knowledge of Clinical Psychology three years later. Thus, scores from a single objective test obtainable by mail, at little cost, predicted each of several criteria as well as any of the clinical judgments made in the entire assessment program

Motivation

Our findings suggest that, in selection for professional training, more attention might well be given to the role of motivation, Perhaps at the level of graduate training, we need establish only a minimal cutting score on tests of intellectual aptitudes; beyond that point, the strength of motivation and the absence of conflicting drives may be the determining factors in success in professional training, and even in the conduct of professional duties.

Faith validity of interviews

Many who have seen our results have been disturbed by the findings regarding the validity for this selection problem of specific techniques which are felt by many professional psychologists to have a high degree of face-validity (or is it faith validity?). Thus, it was the firm conviction of the staff of the OSS assessment program that the global evaluation of a person permits much more accurate predictions of his future performance than can possibly be achieved by a more segmental approach. Unfortunately, the OSS data did not provide a conclusive answer to this question. Our own findings to date serve to raise doubts concerning the validity of this general proposition.

We must evaluate our judgments

Evidence such as that accumulating in this project serves to remind us of the fallibility of the human being both as a measuring device and as an integrator of data. In laboratories, in factories, and in accounting offices, it has been found necessary to supplement his sensory and perceptual capacities with an elaborate array of measuring instruments and computing devices. Pending the gradual development of better measures of psychological variables and comparable aids for combining them, we must continue to rely heavily on human judgment. In so doing, however, we must be continually aware of the magnitude of the errors of such judgments. These errors can be minimized by placing greatest reliance on measures of demonstrated reliability and validity.

Putting it together

We are, in fact, rather encouraged at the probability of being able to predict such criteria with a multiple R of around .50 on the basis of an inexpensive test battery which may be administered without requiring the applicant to present himself at the university of his choice.

More recent prediction studies

1. Longitudinal study of life time accomplishments (Terman and Oden, 1947, 1959; Oden, 1968) and mortality (Friedman et al., 1995).
2. Meta analyses of graduate school prediction (Kuncel et al., 1998, 2001; Kuncel and Hezlett, 2007)
3. Predicting from early adolescence Benbow et al. (1996); Lubinski and Benbow (2000); Lubinski et al. (2001); Lubinski and Benbow (2006); Lubinski (2016)
4. Predicting mortality from measures at age 11 (Deary et al., 2004; Deary and Batty, 2007; Deary et al., 2007, 2013)

Kuncel et al. meta analysis predicting graduate school performance

Table 2
Meta-Analysis of GRE and UGPA Validities: Total Sample

Predictor	N	k	r_{obs}	SD_{obs}	SD_{res}	ρ	SD_ρ	90% credibility interval
GGPA								
Verbal	14,156	103	.23	.14	.10	**.34**	.15	.09 to .59
Quantitative	14,425	103	.21	.11	.06	**.32**	.08	.19 to .45
Analytical	1,928	20	.24	.12	.04	**.36**	.06	.26 to .46
Subject	2,413	22	.31	.12	.05	**.41**	.07	.30 to .52
UGPA[a]	9,748	58	.28	.13	.10	**.30**	.11	.12 to .48
1st-year GGPA								
Verbal	45,615	1,231	.24	.19	.09	**.34**	.12	.14 to .54
Quantitative	45,618	1,231	.24	.19	.08	**.38**	.12	.18 to .58
Analytical	36,325	1,080	.24	.19	.06	**.36**	.09	.21 to .51
Subject	10,225	98	.34	.11	.03	**.45**	.04	.38 to .52
UGPA[a]	42,193	1,178	.30	.18	.10	**.33**	.10	.17 to .49
Comprehensive exam scores[b]								
Verbal[c]	1,198	11	.34	.16	.12	**.44**	.15	.19 to .69
Quantitative[c]	1,194	11	.19	.11	.04	**.26**	.06	.16 to .36
Subject[d]	534	4	.43	.07	.00	**.51**	.00	.51 to .51
UGPA[a]	592	6	.12	.05	.00	**.12**	.00	.12 to .12
Faculty ratings								
Verbal	4,766	35	.23	.12	.08	**.42**	.14	.19 to .65
Quantitative	5,112	34	.25	.10	.02	**.47**	.04	.40 to .54
Analytical	1,982	9	.23	.05	.00	**.35**	.00	.35 to .35
Subject	879	12	.30	.16	.11	**.50**	.18	.20 to .80
UGPA[a]	3,695	22	.25	.12	.10	**.35**	.14	.12 to .58
Degree attainment[a]								
Verbal	6,304	32	.14	.14	.12	**.18**	.16	−.08 to .44
Quantitative	6,304	32	.14	.17	.15	**.20**	.20	−.13 to .53
Analytical	1,233	16	.08	.25	.22	**.11**	.30	−.38 to .60
Subject	2,575	11	.32	.16	.14	**.39**	.17	.11 to .67
UGPA[a]	6,315	33	.12	.17	.16	**.12**	.16	−.14 to .38

Kuncel et al. meta analysis predicting graduate school performance

Table 9
GRE and UGPA Unit-Weighted Composite Predicting GGPA and Faculty Ratings

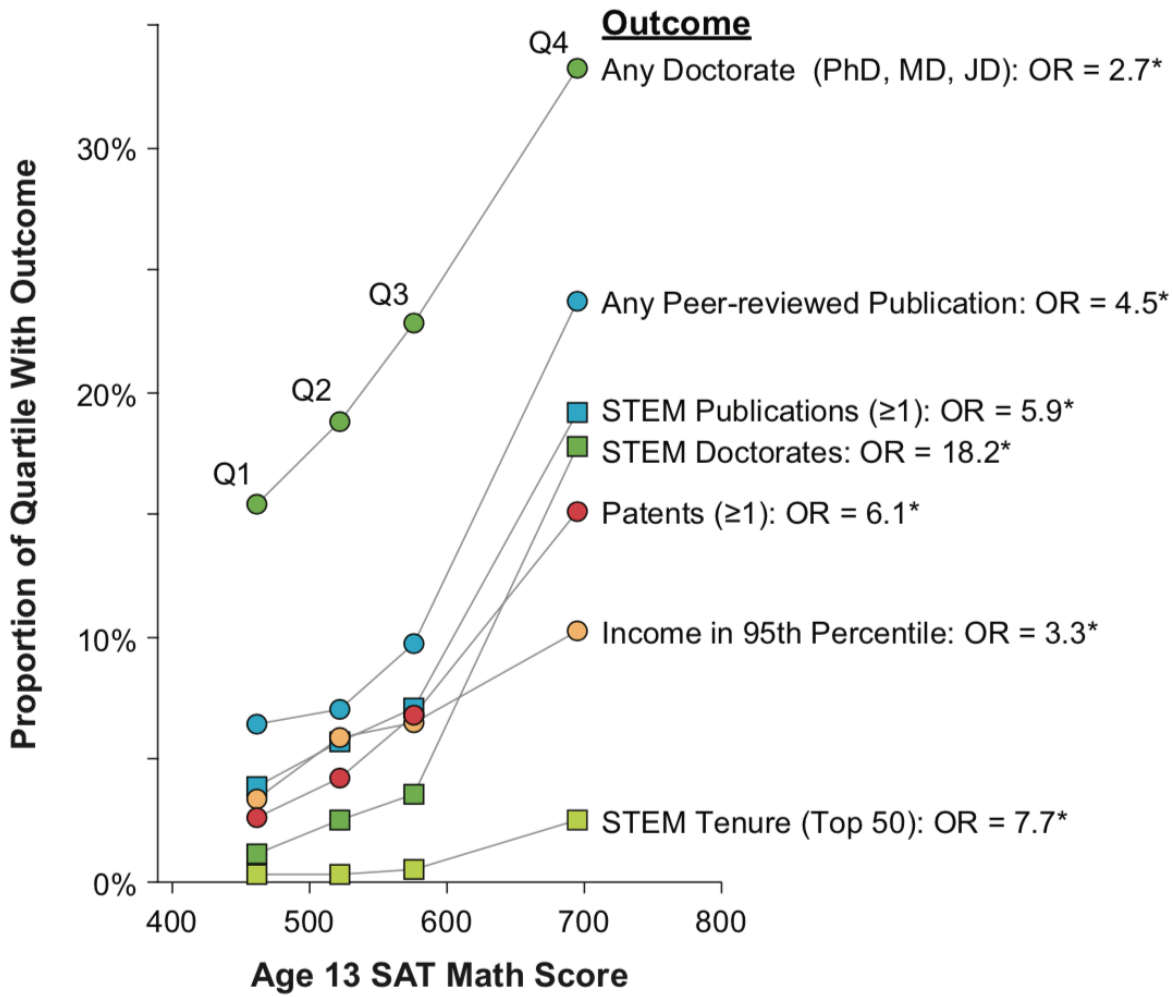
Predictor set	Predictive validity of unit-weighted composite	Predictive validity of composite plus UGPA (unit weighted)
Verbal	.41	.48
Quantitative	.42	.50
Analytical	.38	.46
Subject	.49	.54
Verbal + Quantitative	.46	.53
Verbal + Quantitative + Analytical	.45	.50
Verbal + Quantitative + Subject	.52	.56
Verbal + Quantitative + Analytical + Subject	.50	.54

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; GGPA = graduate grade point average.

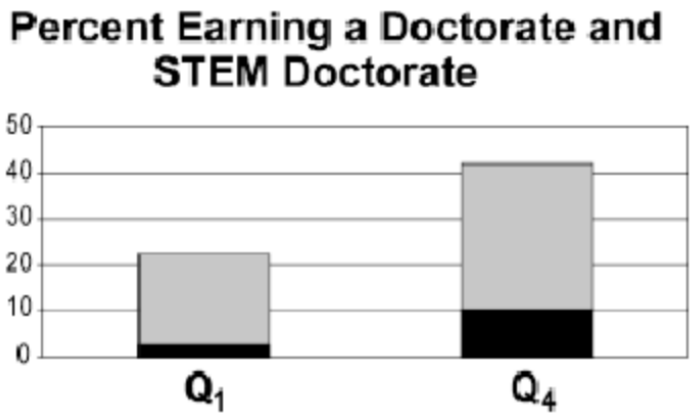
Restrictions of range

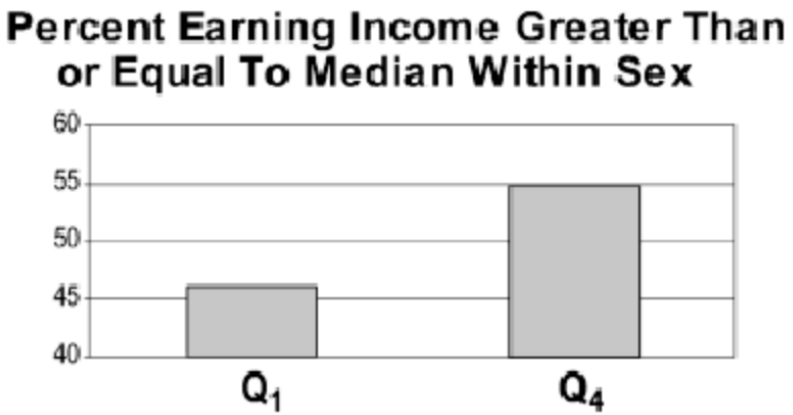
1. Many claim that ability does not predict above about 2 sd.
2. But David Lubinski points out that there are 6 sd above the mean.
3. Validity of SAT is partially limited by range restriction. see (Lubinski and Benbow, 2000, 2006)
4. Consider giving SATs to 12-13 year olds
   - SAT M $\geq$390 or SAT V $\geq$370 (top 1 in 100)
   - SAT M $\geq$ 500 or SAT V $\geq$ 430 (top 1 in 200)
   - SAT M $\geq$ 700 or SAT V $\geq$ 430 (top 1 in 10,000)
5. Longitudinal study started by Camille Benbow while at Johns Hopkins Benbow et al. (1996)

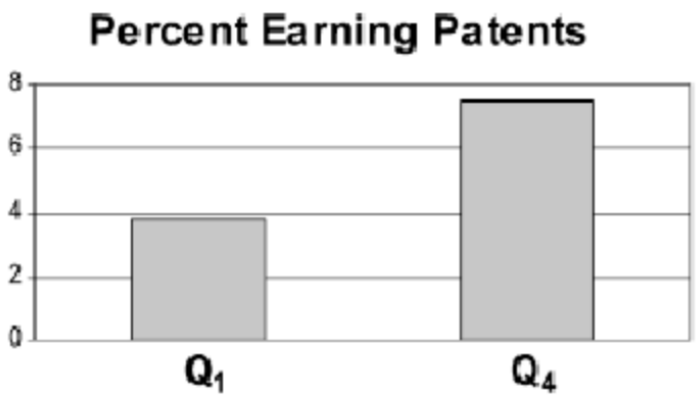
Benbow and Lubinski: Beyond the threshold

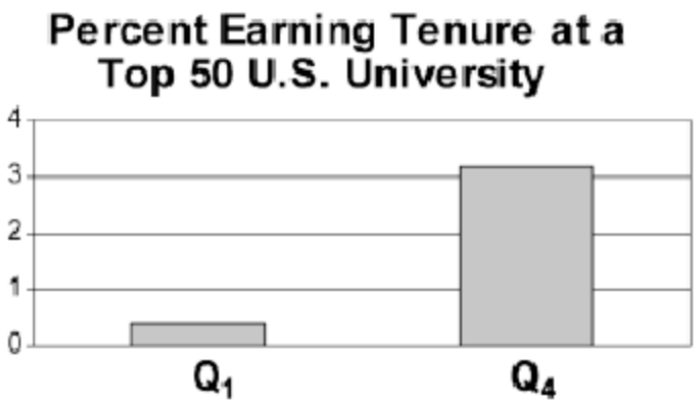


Tenure differs even among the top 1%









Income of top 1 in 10,000

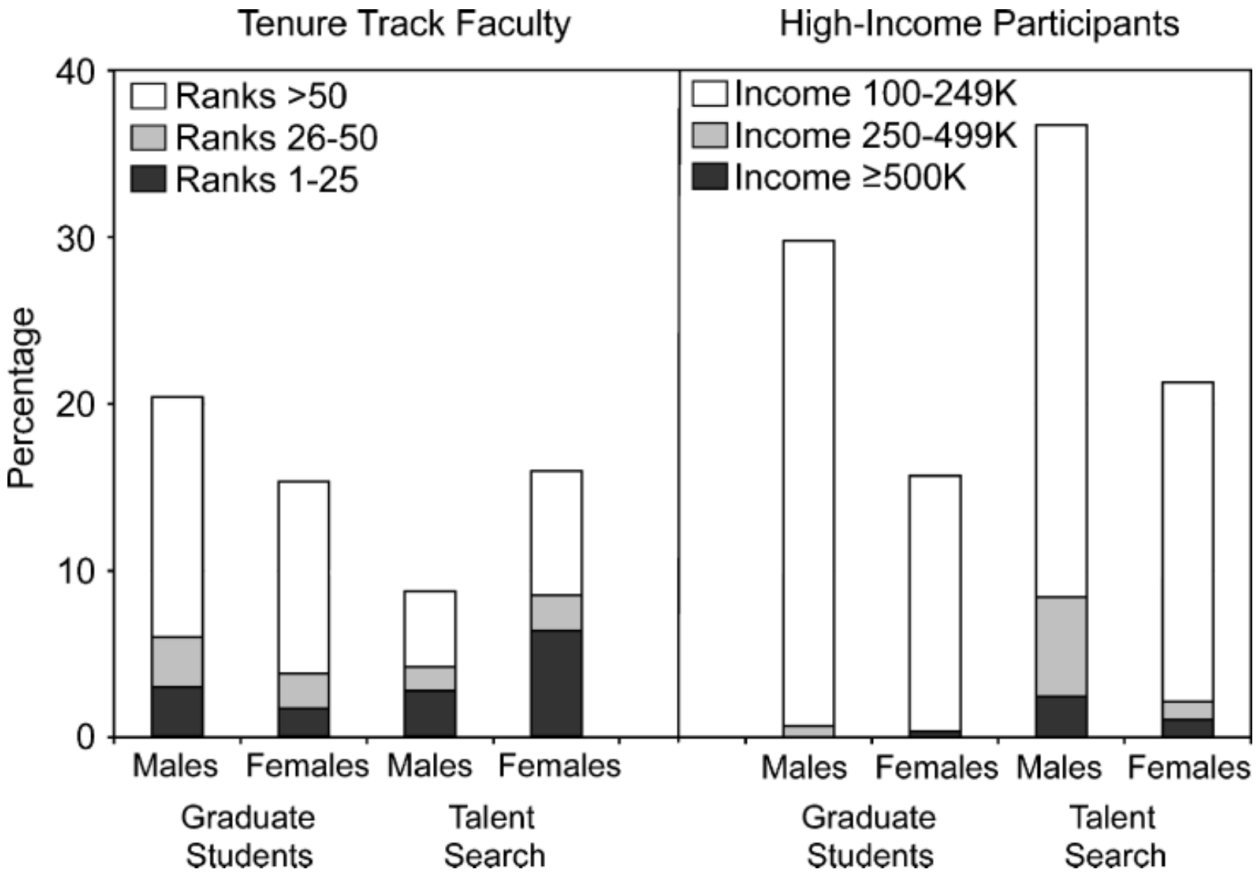
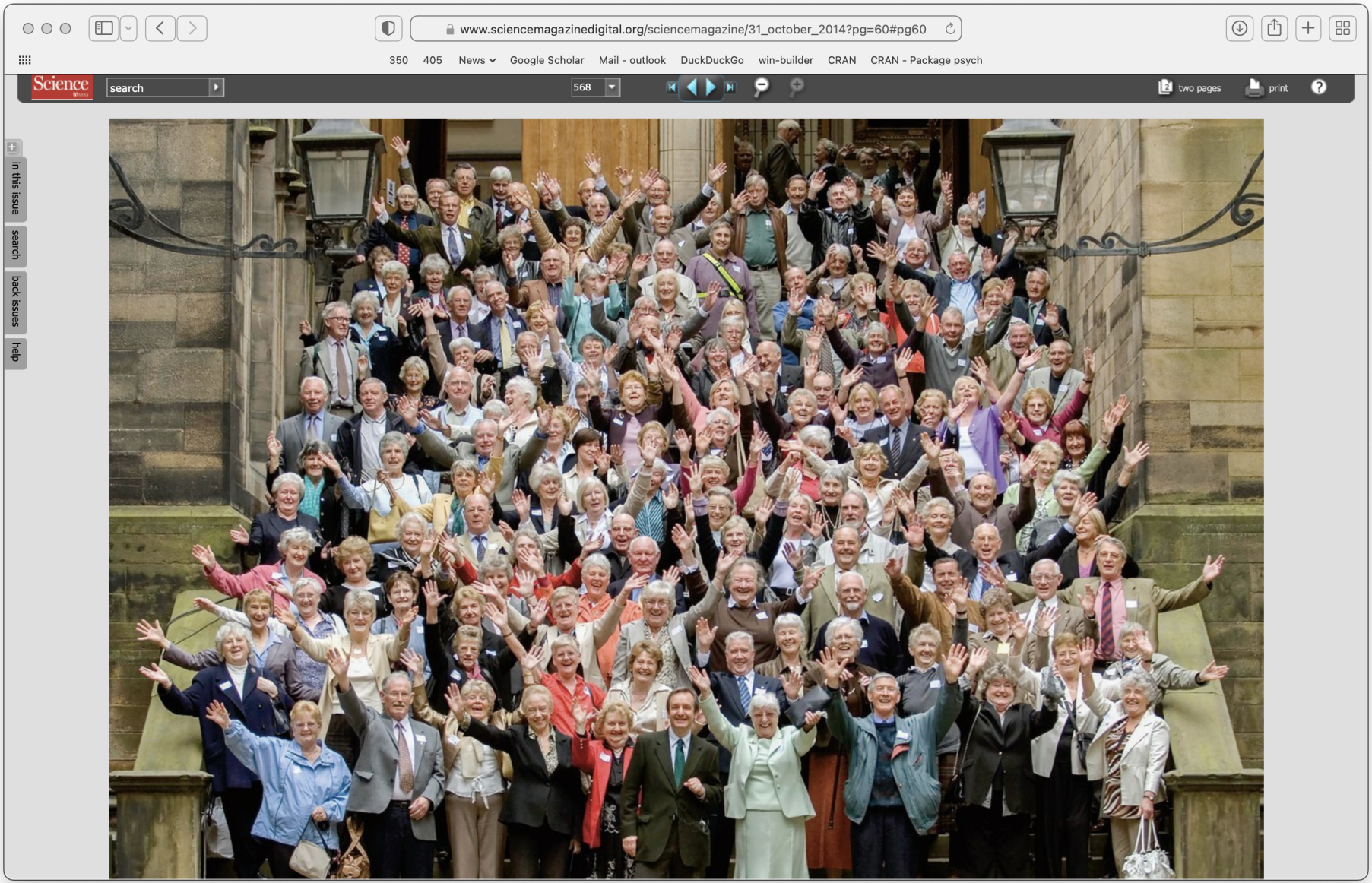


Fig. 3. Percentage of graduate students (Cohort 5) and talent-search participants (Cohort 3, top 1 in 10,000) with tenure-track or tenured positions (left) and annual incomes of $100,000 or more (right) at follow-up. The data shown here are based on samples of 299 and 287 male and female graduate students, respectively, and 286 and 94 male and female talent-search participants, respectively. From Lubinski, Benbow, Webb, and Bleske-Rechek (2006).

Predicting the ultimate criterion: mortality

1. Scottish Mental Survey of 1932 tested all 11 year olds in Scotland.
2. Repeated the data collection in 1947 for another cohort
3. In 1997 Ian Deary and his colleagues found the original testing records and started a longitudinal study (the Midlothian Birth Cohort) (Deary et al., 2004; Deary, 2008, 2009) and started a new field: cognitive epidemiology.
4. They recruited participants from the 1932 and 1947 cohorts to do retesting, health workups, MRI scans, and examined mortality statistics.
5. See Underwood (2014) for a account of the work.

The Midlothian sample



Deary: the Scottish sample and test retest reliability

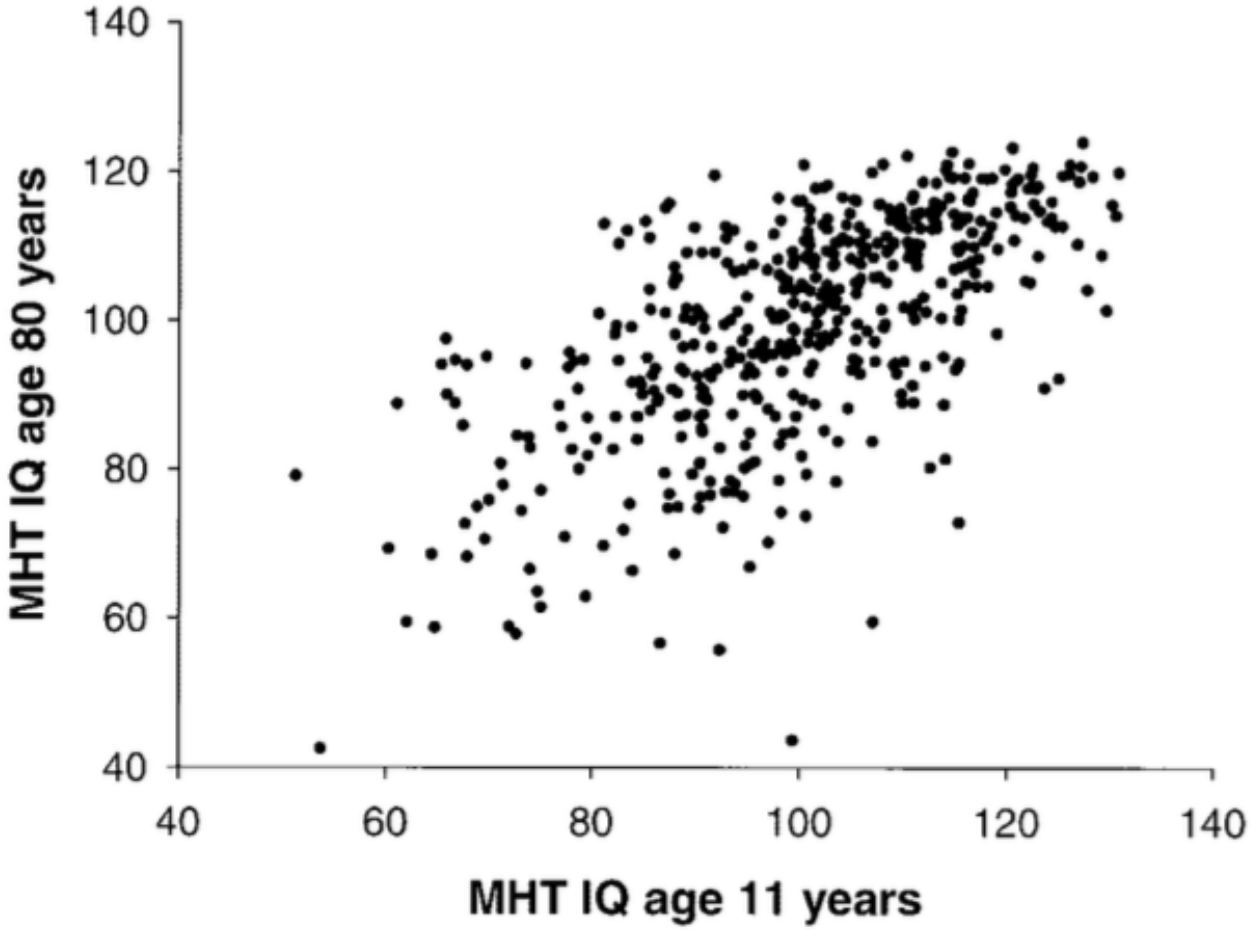
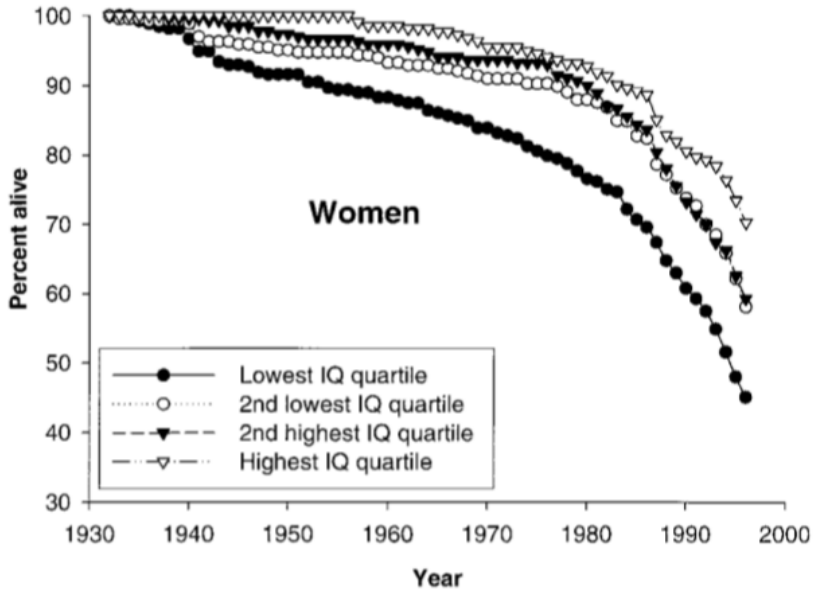
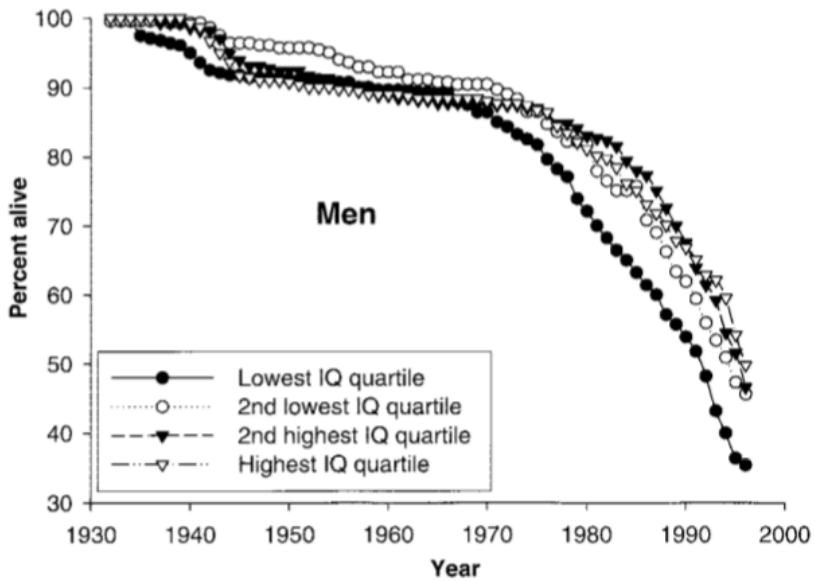


Figure 3. Scattergram of age-corrected Moray House Test (MHT) scores at age 11 and age 80 for participants in the Lothian Birth Cohort 1921 of the Scottish Mental Survey 1932.

Deary: the Scottish sample and mortality





Deary: the Scottish sample and mortality: a model

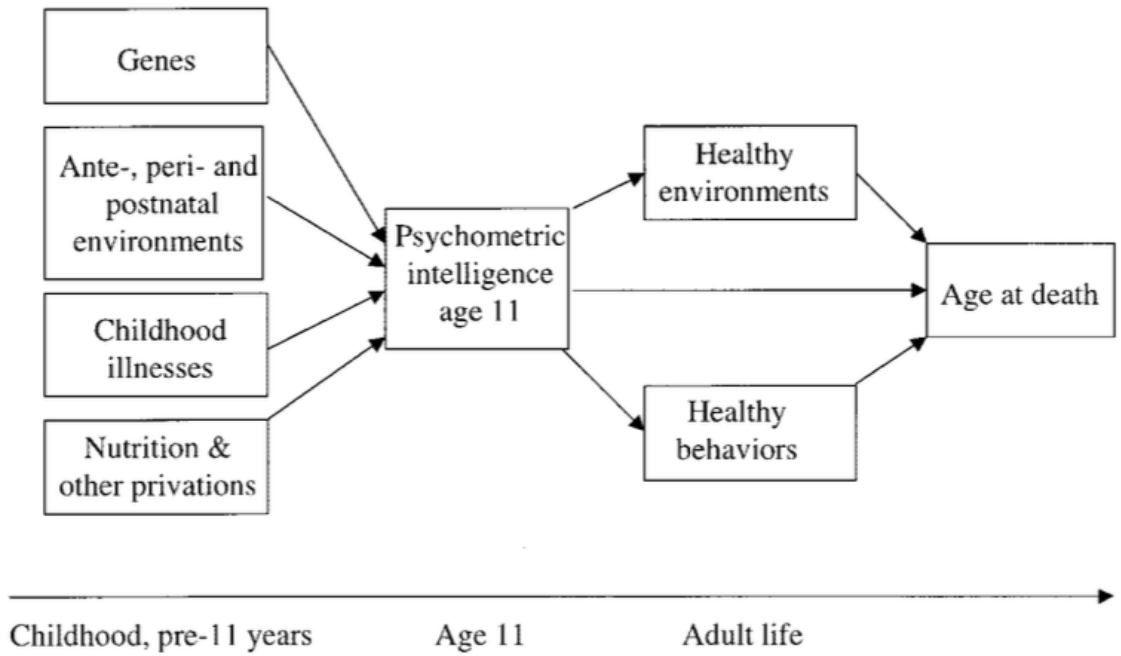
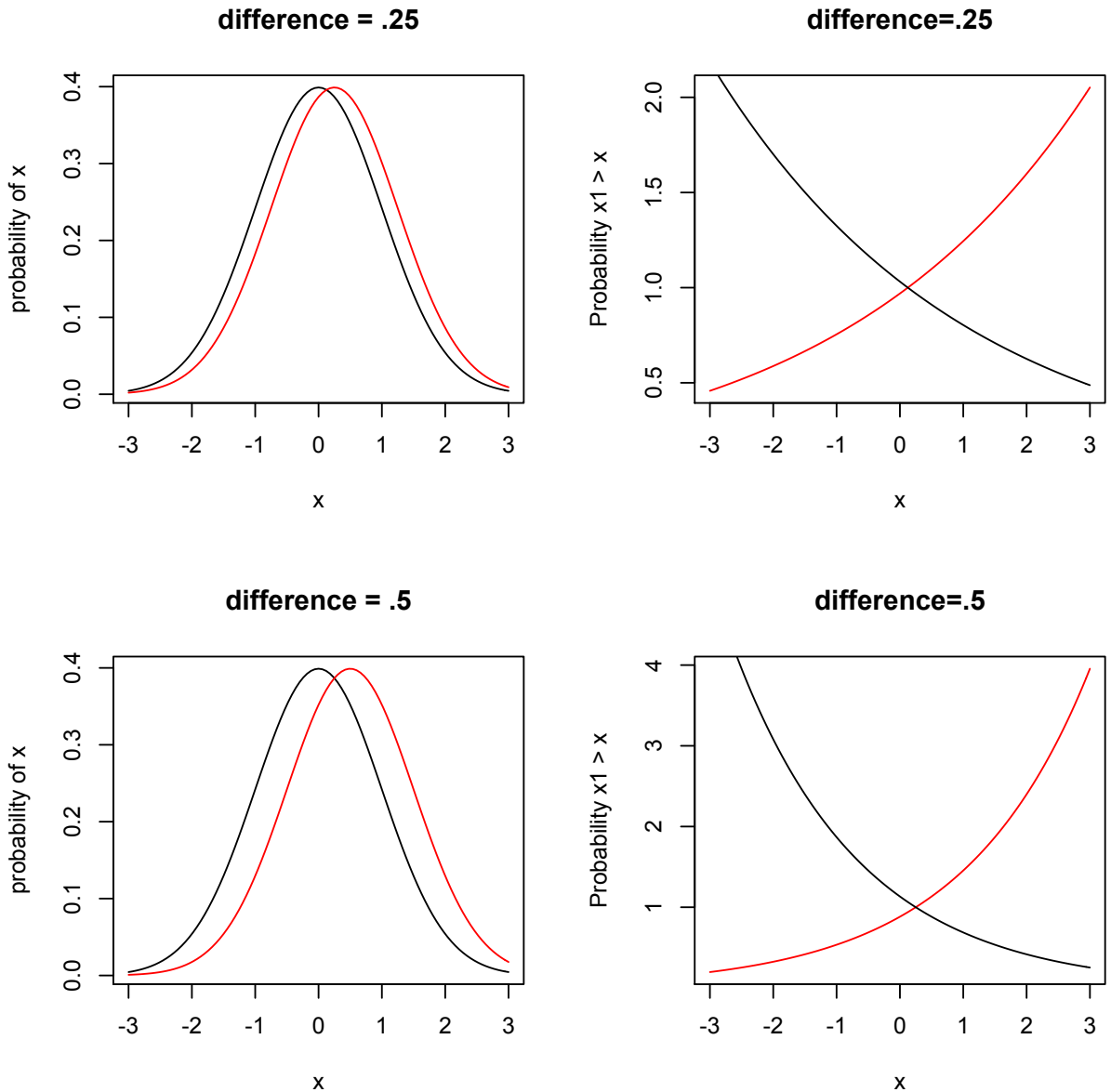
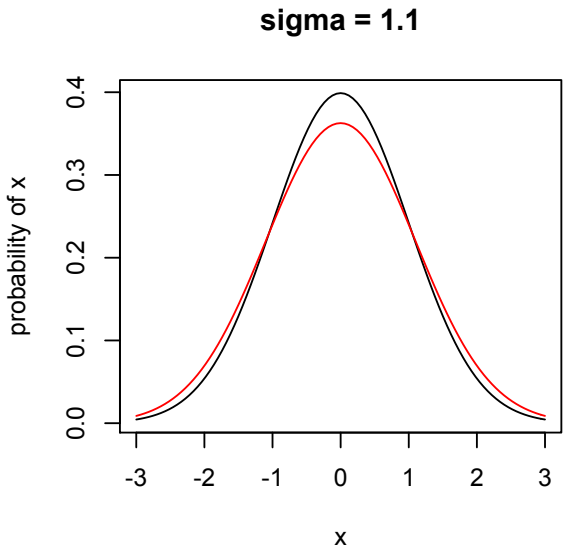


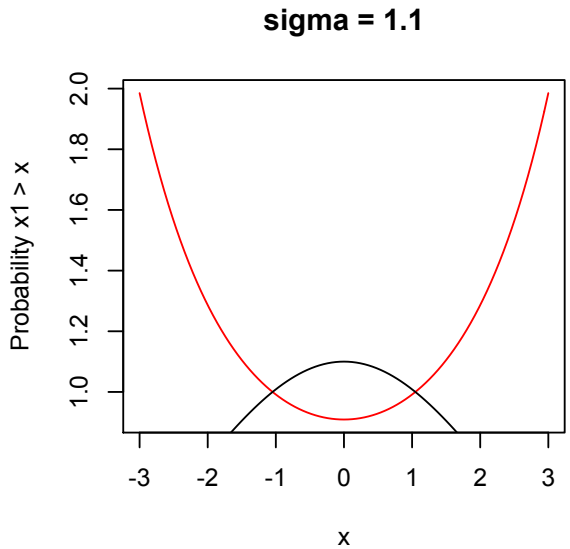
Figure 6. Some possible influences and pathways linking mental ability in childhood and survival. From *Brain and Longevity: Perspectives in Longevity* (p. 162, Figure 3), by C. Finch, J.-M. Robine, & Y. Christen (Eds.), 2003, Berlin: Springer. Copyright 2003 by Springer. Adapted with permission.

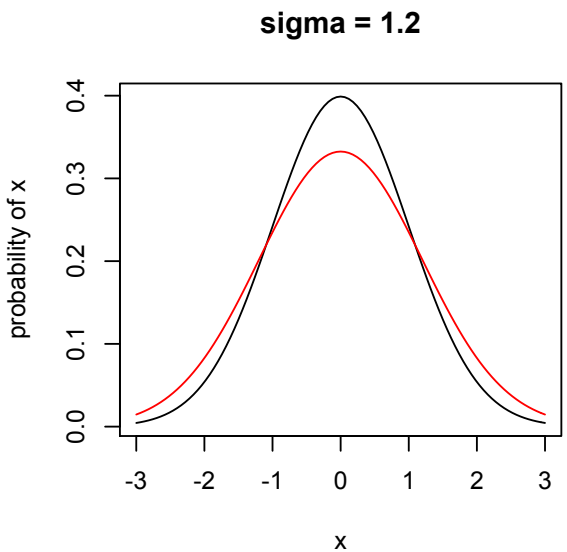
Mean differences and extreme scores

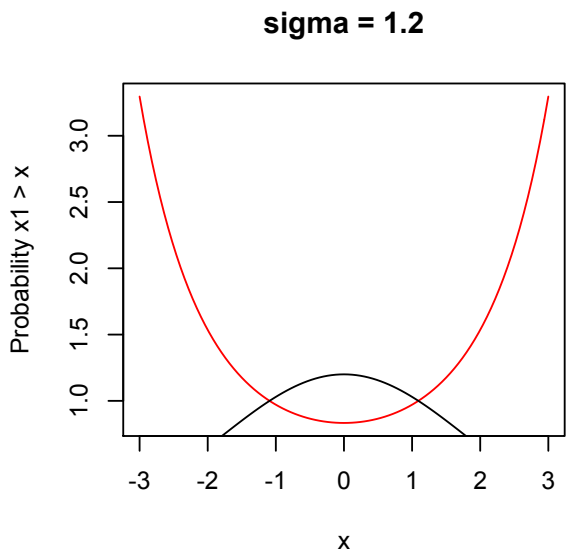


Variances differfences and extreme scores

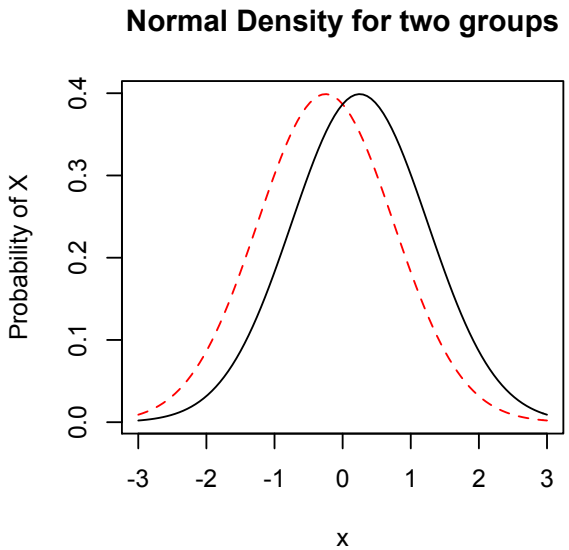


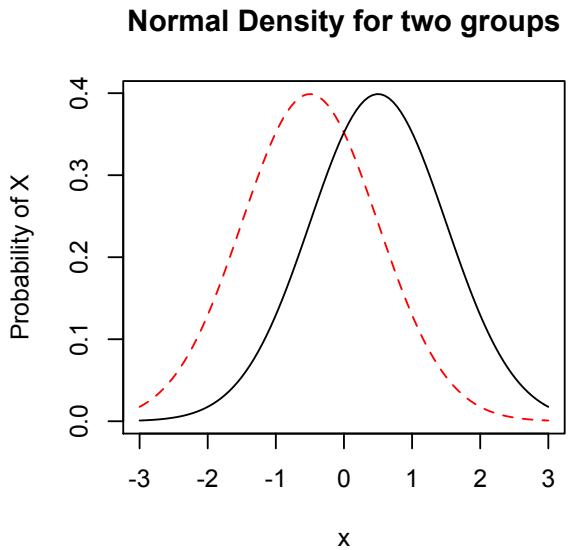


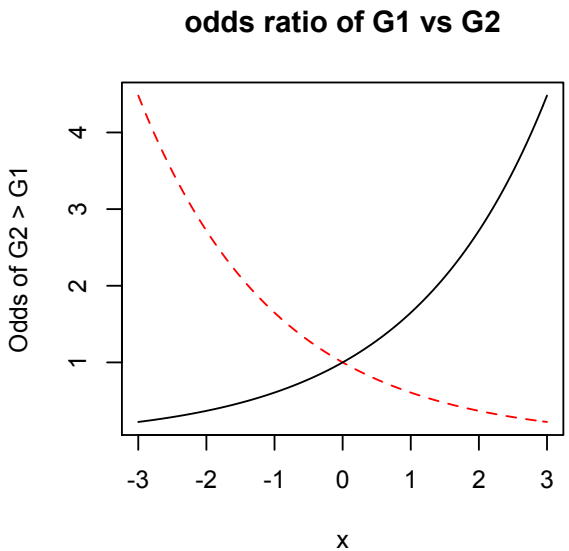


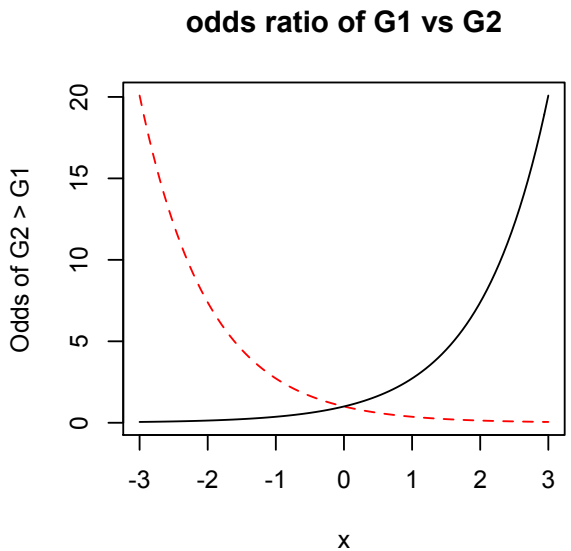


Mean differences and extreme scores: odds ratio







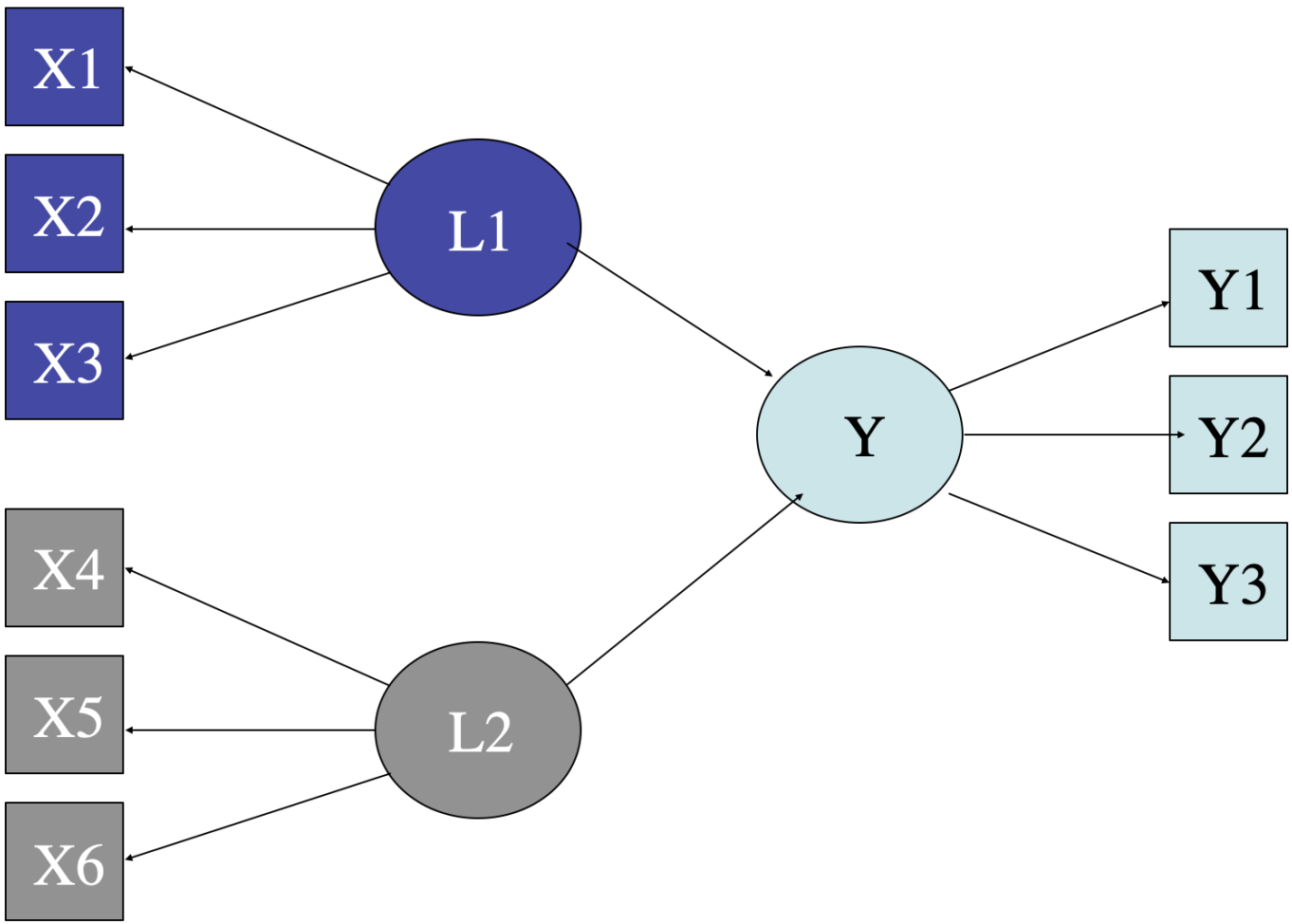


Validation as a process

1. The many kinds of validity (Loevinger, 1957)
2. Construct validation (Cronbach and Meehl, 1955)
3. The multi-Trait-Multi-Method Matrix (Campbell and Fiske, 1959)

The need for other variables

Construct Validity: Convergent, Discriminant, Incremental



The Multi-Trait-Multi-Method Correlation matrix

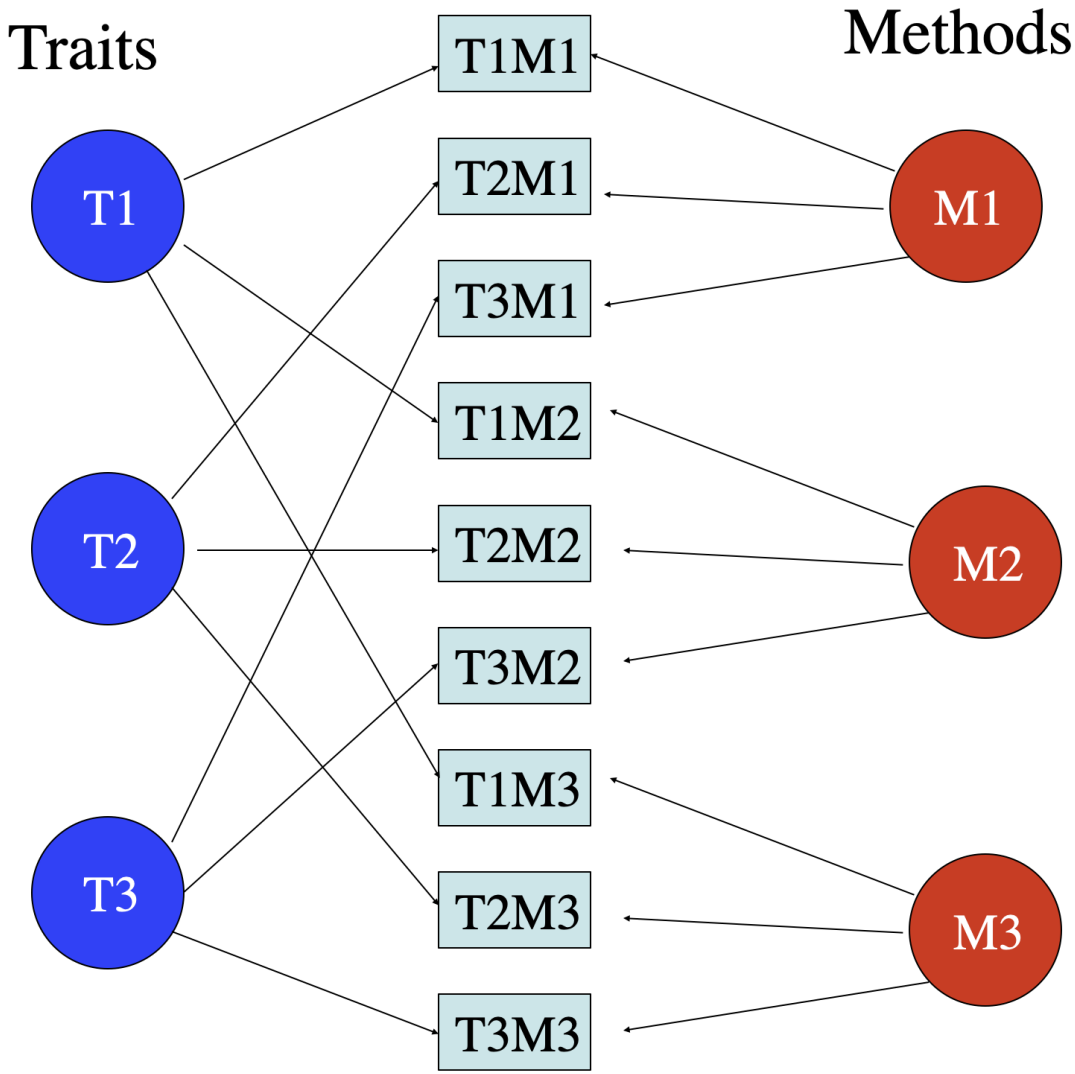
	T1M1	T2M1	T3M1	T1M2	T2M2	T3M2	T1M3	T2M3	T3M3
T1M1	T1M1								
T2M1	M1	T2M1							
T3M1	M1	M1	T3M1						
T1M2	T1			T1M2					
T2M2		T2		M2	T2M2				
T3M2			T3	M2	M2	T3M2			
T1M3	T1			T1			T1M3		
T2M3		T2			T2		M3	T2M3	
T3M3			T3			T3	M3	M3	T3M3

Mono-Method, Mono trait = reliability

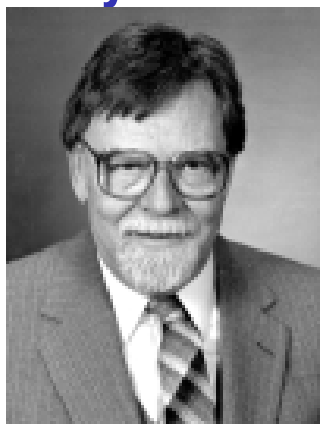
Hetero Method, Mono Trait = convergent validity

Hetero Method, Hetero Trait = discriminant validity

The Multi-Trait-Multi-Method as a factor model



Construct Validity as an extension of True Score Theory

1. Construct validity in terms of the structure of latent variables was introduced by Cronbach and Meehl (1955). This was probably partly as a counter to behaviorism.
2. Elaborated by Loevinger (1957) who dismissed the idea of mere “practical" validity.
3. Construct validity could be conceptualized
   - Convergent: different measures of the same construct should go together
   - Divergent: measures of different constructs should not go together
   - Incremental: a measure should add something .

   A test should be defined by what it measures and what it does not measure.

Construct validity and the “Nomological Net"

1. Tests did not have validity, they were part of a network of validity.
2. Best exemplified in the Multi-Trait – Multi-Method Matrix of Campbell and Fiske (1959).


Agreement between Self Report and Peer Ratings
An example of a Multi-Trait–Multi-Method Matrix

Table: Self report and peer report from the SAPA-project. Correlations reported by Zola et al. (2021). Reliabilities on the main diagonal. Raw correlations below the diagonal. Correlations corrected for reliability above the diagonal. Upper left quadrant reflects SAPA Personality Inventory scores (Condon, 2018) for 158,631 participants, mean n/item = 18,180. Other quadrants reflect 908 peer rated participants. Data from the `zola` dataset in the *psychTools* package.

	Self Report					Peer Ratings				
Variable	Agrbl	Cnscn	Nrtcs	Extrv	Opnnn	Agrbl	Cnscn	Stblt	Extrv	IntlO
Agreeableness	**0.87**	0.32	-0.14	0.28	0.09	**0.75**	0.21	0.18	0.34	0.22
Conscientiousness	0.28	**0.87**	-0.20	0.13	0.06	0.16	**0.78**	0.22	**0.42**	0.13
Neuroticism	-0.12	-0.18	**0.90**	-0.28	-0.10	-0.01	-0.16	**-0.78**	-0.40	-0.25
Extraversion	0.25	0.12	-0.25	**0.90**	0.14	0.01	-0.01	0.07	**0.71**	0.14
Opennness	0.08	0.05	-0.09	0.13	**0.86**	-0.14	-0.06	0.10	0.17	**0.49**
Agreeableness	**0.47**	0.10	-0.01	0.00	-0.09	**0.45**	0.36	**0.47**	0.15	**0.44**
Conscientiousness	0.15	**0.55**	-0.12	-0.01	-0.04	0.18	**0.58**	**0.42**	**0.41**	**0.47**
Stability	0.13	0.16	**-0.58**	0.05	0.07	0.25	0.25	**0.60**	0.38	**0.52**
Extraversion	0.23	0.28	-0.27	**0.49**	0.11	0.07	0.23	0.22	**0.52**	0.32
IntellectOpenness	0.14	0.08	-0.15	0.09	0.30	0.19	0.24	0.27	0.15	**0.44**

The unfortunate emphasis upon construct validity reduced the emphasis upon the practical use of tests

1. In a response to operationalism, construct validity was in strong contrast to three other approaches.
2. Constructs, as embedded in nomological networks, were seen as theoretical concepts and could only be evaluated in terms of the pattern of correlations.
3. Criterion-oriented validation procedures, on the other hand, harkened back to the operational definitions of behaviorism.
   - Concurrent validity is the correlation with a current criterion.
   - Predictive validity is the correlation with a future criterion.
4. Content validity was established by showing that the test items were a sample of a universe in which the investigator is interested.

Loevinger and the boiling of eggs

1. Favorably quoting Marschak, Loevinger said: (p 641) “A theory provides us with solutions which are potentially useful for a large class of decisions.
2. Hence, the more we know about its properties the better. If we merely want to know how long it takes to boil an egg, the best is to boil one or two without going into the chemistry of protein molecules. The need for chemistry is due to our want to do other and new things "
3. She goes on to say “The argument against classical criterion-oriented psychometrics is thus two-fold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry.
4. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful."

In case we did not understand

“the most fruitful direction for the development of psychometric devices, and hence of psychometric theory, is toward measurement of traits which have real existence in some sense; that this orientation is antithetical to one which places first emphasis on prediction, decisions, or "utility;" that most decision-oriented psychometric studies would be more fruitfully formulated as trait-oriented studies;
and that such legitimately pressing decisions as must inevitably be made will also best be served by a predominantly trait-oriented psychometrics." Loevinger (1957)

But see “The seductive beauty of latent variable models: or why I don’t believe in the Easter Bunny" (Revelle, 2024) for another perspective.

The persistent myth of the validity of the interview

1. It has been known since 1950 that interviews are appealing but do not work.
2. Everyone relies on their feeling that they work, remembering successes, forgetting failures.
3. Clinical versus actuarial prediction Dawes et al. (1989)
4. Experience is not a good teacher when the feedback is slow Dawes (1989)
5. Belief in the unstructured interview Dana et al. (2013)
6. Summarized very well in Dawes (2009)

Medical School Admissions

1. As discussed by Dawes (2009), DeVaul et al. (1987) examined the effect of interview ratings on later success in medical school.
2. Of 2200 applicants to medical school, 800 were invited to interview and were interviewed
3. Of these, 150 of the top 350 were offered positions
4. The state then provided funding for an additional 50 students
5. only the 700-800 ranked students were still available
6. After four years: “Even when the top 50 students in committee preference were compared with the 50 applicants, there were no differences. Thus, the least desirable candidates performed as well as the most desirable.

How useful are items?

1. The common observation/belief is that items have low correlations with other items.
2. From a classical reliability perspective:
   Item variance = general + group + specific + error.
3. The “gospel" is that items are mainly error variance.
4. This is true from a latent variable perspective, but less true if we actually examine item variance.
5. Perhaps 20% of an item is general factor variance, another 10-20% group variance but about 40% is specific and reliable variance.
6. We can see this by doing a variance decomposition of items that are repeated across time (Revelle and Condon, 2019)
7. So what?
8. Lets look at the correlates of items.

Items as analogous to SNPs in GWAS studies

1. In Genome Wide Association Studies one examines phenotypic variation as it correlates with differences in SNP frequencies across the genome.
2. We can do the same by examining phenotypic variation and correlation across the persome (Möttus et al., 2019)
3. A typical approach is to show the correlations and their probability values (corrected for multiple tests)
   - Typically displayed in “Manhattan Plots" across the genome. We do this across the “Persome".
4. First show plots for an open source data set (spi) available in the *psych* package (Revelle, 2025).
   - This is a set of 135 temperament items (Condon, 2018), with 10 criteria for 4,000 subjects.
5. Then do the same for items from the Big 5, then an extended set (the little 27) then for a bigger data set with even more items.
6. Finally, we show the profile correlations of college majors and occupations for 255K participants across 900 items

Sapa Personality Inventory (`spi` data set)

Table: Descriptive statistics for the 10 criteria variable taken from the `spi` data set.

Variable	vars	n	mean	sd	median	trmmd	mad	min	max	range	skew	krtss
age	1	4000	26.90	11.49	23	25.02	7.41	11	90	79	1.45	1.80
sex	2	3946	1.60	0.49	2	1.62	0.00	1	2	1	-0.39	-1.85
health	3	3536	3.51	0.98	4	3.54	1.48	1	5	4	-0.25	-0.42
p1edu	4	3051	4.72	2.39	5	4.77	4.45	1	8	7	-0.11	-1.33
p2edu	5	2896	4.33	2.32	5	4.28	4.45	1	8	7	0.09	-1.33
education	6	3330	4.10	2.21	3	4.00	1.48	1	8	7	0.41	-1.04
wellness	7	3311	1.54	0.50	2	1.55	0.00	1	2	1	-0.17	-1.97
exer	8	3310	3.57	1.60	4	3.60	1.48	1	6	5	-0.35	-1.06
smoke	9	3348	2.19	2.04	1	1.70	0.00	1	9	8	1.83	2.19
ER	10	3347	1.16	0.48	1	1.03	0.00	1	4	3	3.42	12.74

spi items are taken from Condon (2018)

Scale development and cross validation

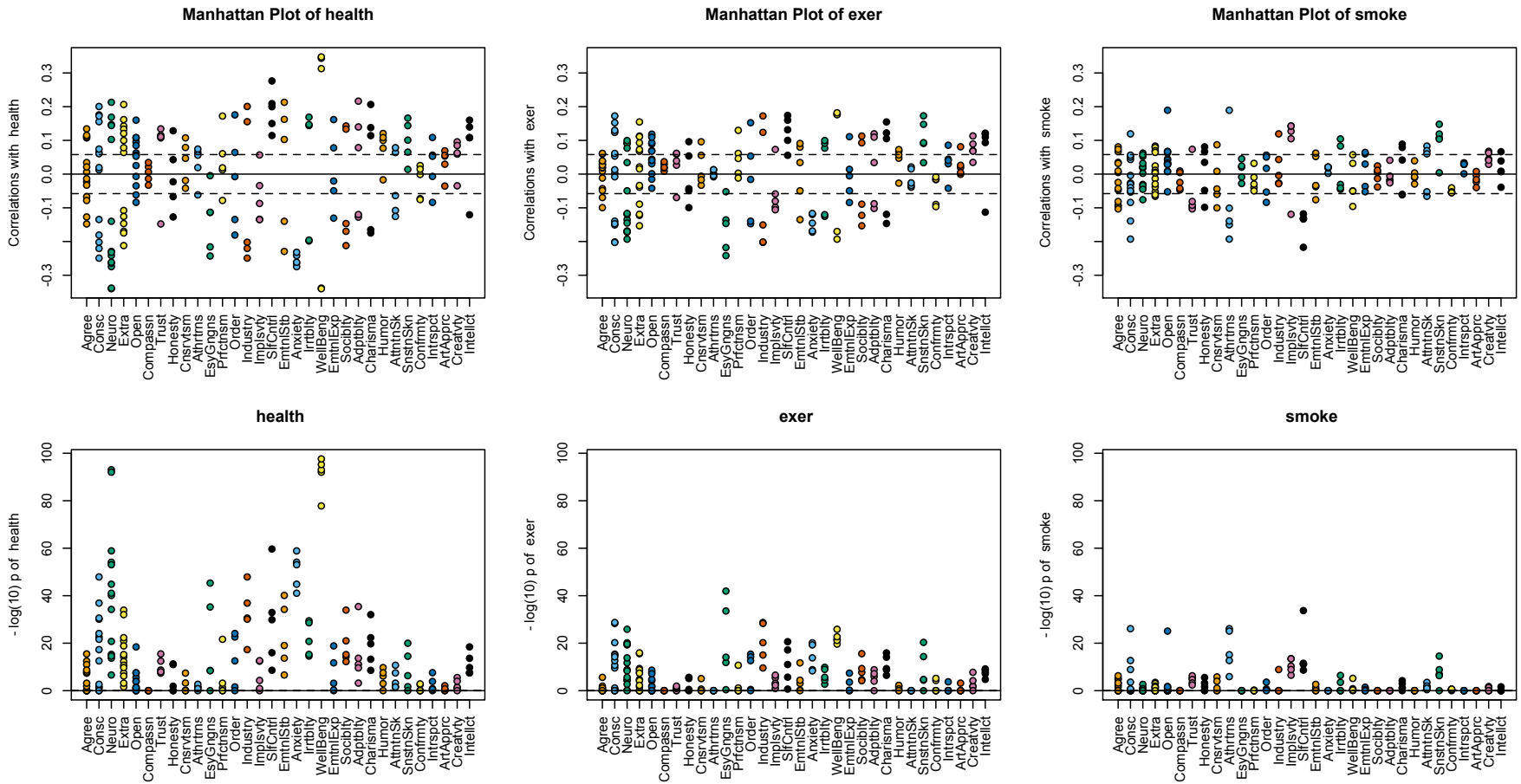
1. Weights based upon data are best fits for those data
2. Need to “Cross Validate" on a different set
3. Original cross validation technique was to split the sample into 2, derive on first half, report the validities on the second half
4. KFold cross validation splits the data into K parts, derives the model on K-1 parts and then validate it on the remaining part. Repeat this K times (folds) and then average across folds.
5. Boot Strap Aggregation (“bagging") takes many (100 - 1000) bootstrap samples and then aggregates across the hold out sample. Bootstrap automatically produces a hold out since 62.3% of subjects are in the derivation sample and 37.7% are in the holdout for each iteration.
6. The `bestScales` function does either K-fold or bagging and produces the Best Items Scale that is Cross-validated Unit-weighted, Informative and Transparent (Elleman et al., 2020) .

Internal consistency and correlations of the Big 5

Table: Reliabilities and correlations of the Big 5 scales from the `spi` data set. α reliability is on the diagonal of the correlations.

Variable	ω_{total}	ω_h	Agree	Consc	Neuro	Extra	Open
Agreeableness	0.91	0.61	0.87				
Conscientiousness	0.89	0.61	0.24	0.86			
Neuroticism	0.93	0.71	-0.12	-0.19	0.90		
Extraversion	0.92	0.70	0.23	0.07	-0.20	0.89	
Openness	0.88	0.72	0.00	0.01	-0.12	0.13	0.84

Item Validities: Manhattan plots



Cross validation of predictions

1. 10 criteria from the `spi` data set.
2. Linear regression using the spi-Big 5 (14 item scales for each of 5 dimensions)
3. `bestScales` solution choosing items from the spi to predict criteria
4. The “little 27" lower level factors of the `spi`
5. Multiple regression using all 135 items
6. All models developed on random subset of 2,000 subjects, cross validated on the remaining 2,000.

Why and what is cross validation?

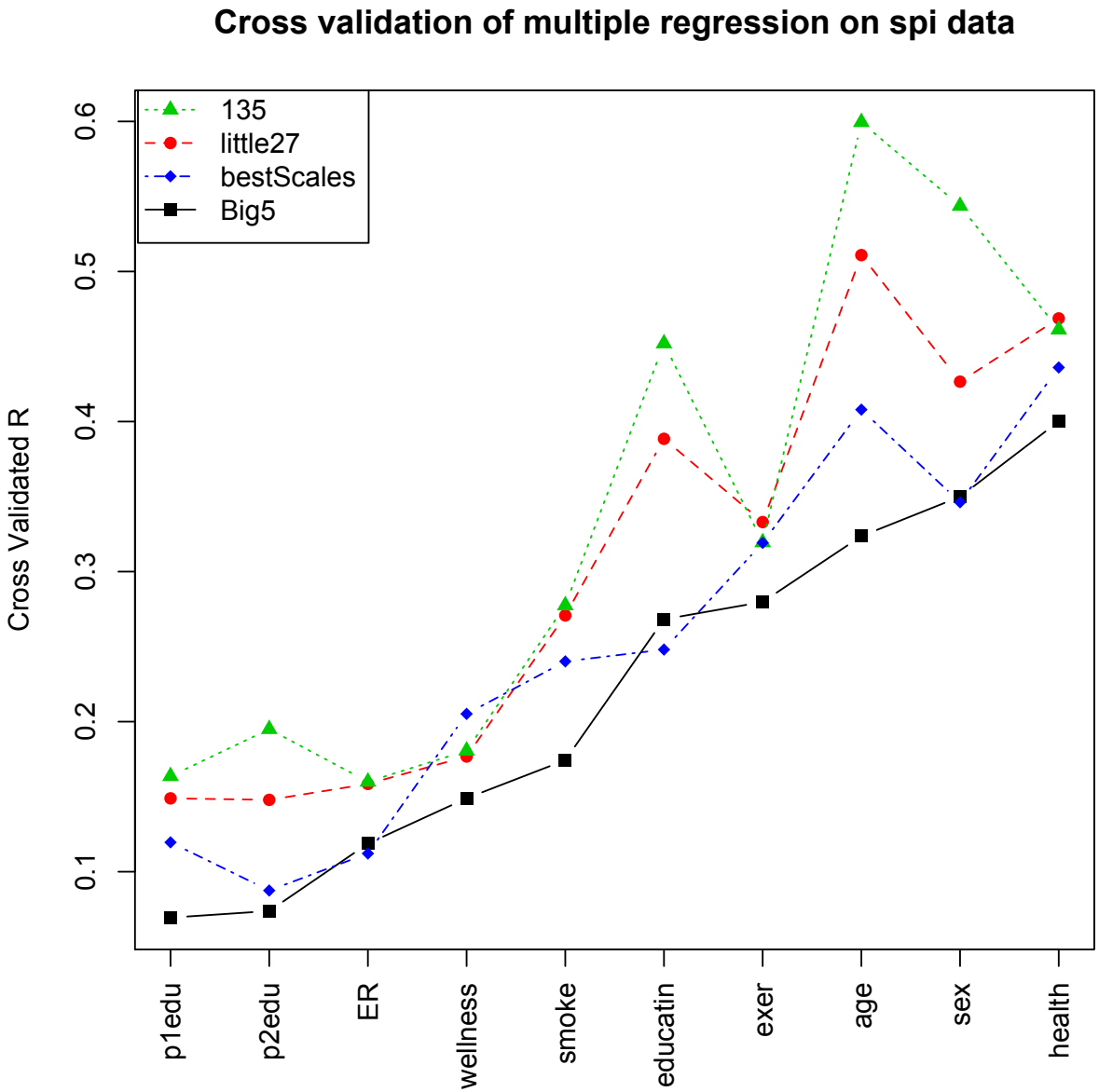
1. Items or scales selected to predict will best fit derivation sample.
2. Regressions will “shrink" in other samples.
3. Particularly a problem with small samples and a large number of predictors.
4. Perhaps the best discussion is a delightful paper by Cureton (1950): “Validity, reliability, and baloney".

Profile Correlations

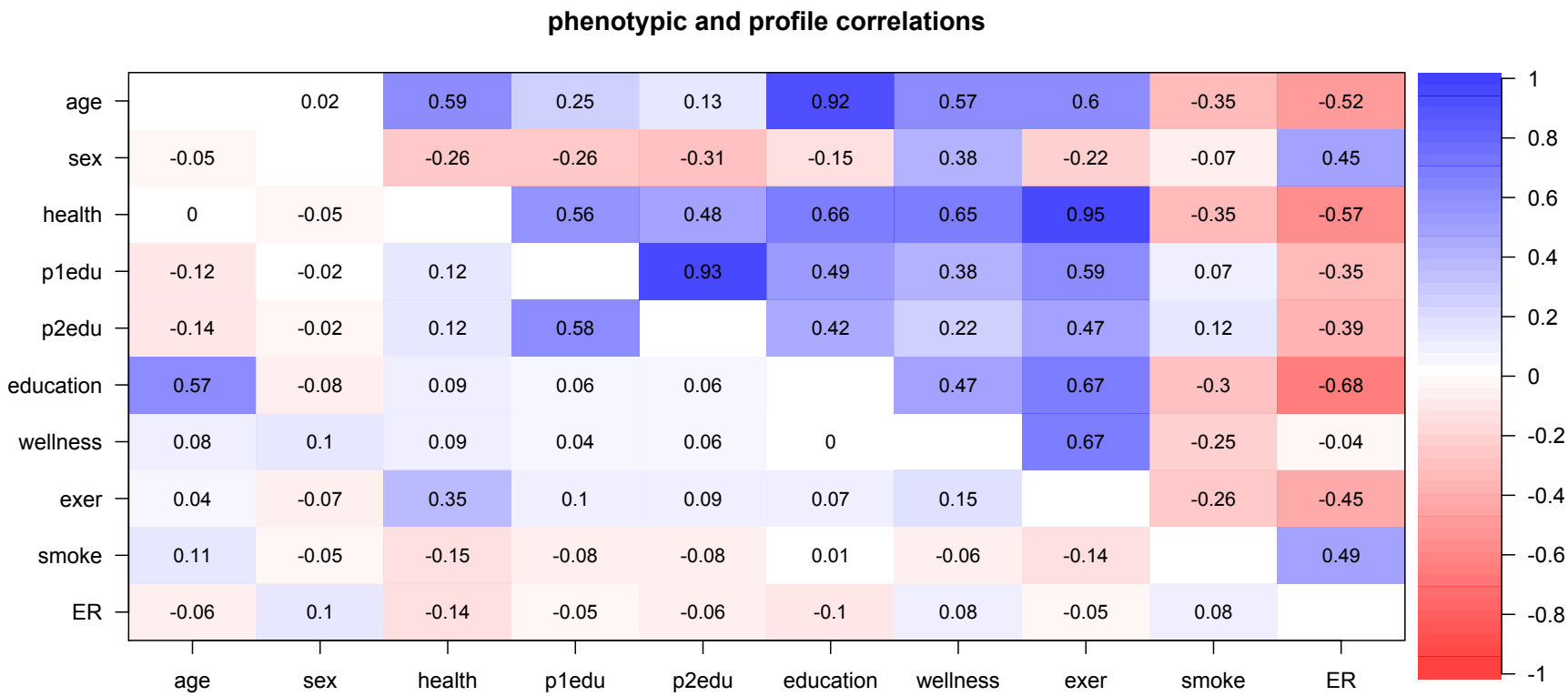
1. Normally, we examine the correlations of scales with criteria and criteria with criteria
2. Analogous to the “genetic correlation" which is the correlation of phenotypes across the genome, we can find the “persome correlation" which is the correlation of the criteria across the persome.
3. These reflect the amount that the predictable variance in one outcome is the same as the predictable variance in another outcome.
4. The next slide compares phenotypic correlations with persomic corrrelations.

See Revelle et al. (2021) for more examples and the data and code for doing these examles.

Cross validated correlations predictions 10 different criteria.



Comparing phenotypic to profile correlations



The SAPA project emphasizes item validities

1. Random sampling of many items
2. Basic psychometrics of composites
3. Best Scales approach to scale construction
4. Profile analysis of groups
5. For an example: Open science, open items PWAS or Persome Wide Association Studies.

SAPA measures self report

1. How to validate the self reports of the SAPA project
2. SAPA participants were asked to nominate anonymous friends
3. These friends then gave peer ratings
4. Zola et al. (2021) reported the validity of self report personality items from the SAPA personality inventory (SPI) (Condon, 2018) in terms of 30 peer reports on 8 dimensions. Here are the polychoric correlations of these items. spi items were collected using SAPA procedures for 158,631 participants (mean n/item = 18,180), 908 of whom received peer ratings..
5. The Multitrait-multimethod correlations were found from the correlations.

Zola validitiies

R code

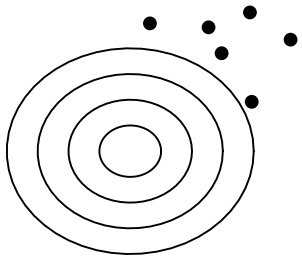
```
scores <- psych::scoreOverlap(zola.keys[c(1:5,33:37)],zola) #MTMM of I
```

```
lowerMat(scores$cor)
                  Agrbl Cnscn Nrtcs Extrv Opnnn Agrbl Cnscn Stblt Extrv IntlO
Agreeableness      1.00
Conscientiousness  0.28  1.00
Neuroticism       -0.12 -0.18  1.00
Extraversion       0.25  0.12 -0.25  1.00
Openness           0.08  0.05 -0.09  0.13  1.00
Agreeableness      0.47  0.10 -0.01  0.00 -0.09  1.00
Conscientiousness  0.15  0.55 -0.12 -0.01 -0.04  0.18  1.00
Stability          0.13  0.16 -0.58  0.05  0.07  0.25  0.25  1.00
Extraversion       0.23  0.28 -0.27  0.49  0.11  0.07  0.23  0.22  1.00
IntellectOpenness  0.14  0.08 -0.15  0.09  0.30  0.19  0.24  0.27  0.15  1.00

lowerMat(scores$MIMS)  #average item correlations within and between domains
                  Agrbl Cnscn Nrtcs Extrv Opnnn Agrbl Cnscn Stblt Extrv IntlO
Agreeableness      0.33
Conscientiousness  0.10  0.32
Neuroticism       -0.05 -0.07  0.38
Extraversion       0.10  0.05 -0.11  0.39
Openness           0.03  0.02 -0.03  0.05  0.30
Agreeableness      0.18  0.04  0.00  0.00 -0.03  0.17
Conscientiousness  0.06  0.23 -0.05  0.00 -0.02  0.07  0.26
Stability          0.05  0.07 -0.25  0.02  0.03  0.10  0.11  0.28
Extraversion       0.09  0.11 -0.11  0.21  0.04  0.03  0.10  0.09  0.21
IntellectOpenness  0.05  0.03 -0.06  0.03  0.11  0.07  0.10  0.11  0.06  0.16
```

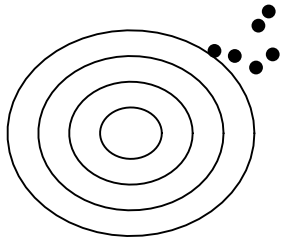
Introduction to the question

1. In a brilliant manuscript which I had the good fortune to review, Mijke Rhemtulla developed the “Dart Board" validity/reliability metaphor.
   - This was based on a strong assumption that validity can be defined as what a factor measures.
   - That is, validity is factorial validity.
   - Reliability is just how well we measure the construct.
   - Validity is the ratio of internal consistency to test-retest reliability.
2. Dartboard validity wants scales to be internally consistent measures of single constructs.
3. Dartboard validity equates validity with how well the test measures a construct.

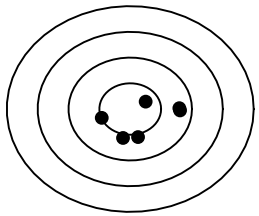
Reliability and Validity as dart throwing



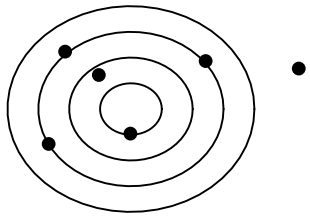
Unreliable and Invalid



Reliable and Invalid



Reliable and Valid



Unreliable but Valid

1. Unfortunately for Mijke, I had just given a keynote address at ISSID entitled “The seductive beauty of latent variables" (Revelle, 2023)
    - That paper was an attack on our beloved application of latent variable models and argued that we should worry more about prediction than factorial homegeneity.
    - I even suggested that to believe in latent variables was akin to believing in the Easter Bunny or the Tooth Fairy.
2. In addition, I had recently published an article with Alice Eagly “Understanding the Magnitude of Psychological Differences Between Women and Men Requires Seeing the Forest and the Trees" (Eagly and Revelle, 2022) which examined the effect of aggregation on reliability and validity.
    - That paper showed that while aggregation could increase reliability, aggregating unrelated concepts could increase validity.
    - It rediscovered Gulliksen (1950).

Which set of items (X1..X4) have the highest validity when predicting Y?

A) $\alpha = .73$ $R_y = ?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.4	1.0			
X3	0.4	0.4	1.0		
X4	0.4	0.4	0.4	1.0	
Y	0.2	0.2	0.2	0.2	1.0

B) $\alpha = .63$ $R_y = ?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.3	1.0			
X3	0.3	0.3	1.0		
X4	0.3	0.3	0.3	1.0	
Y	0.2	0.2	0.2	0.2	1.0

C) $\alpha = .5$ $R_y = .?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.2	1.0			
X3	0.2	0.2	1.0		
X4	0.2	0.2	0.2	1.0	
Y	0.2	0.2	0.2	0.2	1.0

D) $\alpha = .31$ $R_y = ?$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.1	1.0			
X3	0.1	0.1	1.0		
X4	0.1	0.1	0.1	1.0	
Y	0.2	0.2	0.2	0.2	1.0

Please rank order these four cells in terms of validity.

Which set of items (X1..X4) have the highest validity when predicting Y?

A) $\alpha = .73$ $R_y = .27$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.4	1.0			
X3	0.4	0.4	1.0		
X4	0.4	0.4	0.4	1.0	
Y	0.2	0.2	0.2	0.2	1.0

B) $\alpha = .63$ $R_y = .29$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.3	1.0			
X3	0.3	0.3	1.0		
X4	0.3	0.3	0.3	1.0	
Y	0.2	0.2	0.2	0.2	1.0

C) $\alpha = .5$ $R_y = .32$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.2	1.0			
X3	0.2	0.2	1.0		
X4	0.2	0.2	0.2	1.0	
Y	0.2	0.2	0.2	0.2	1.0

D) $\alpha = .31$ $R_y = .35$

Variable	X1	X2	X3	X4	Y
X1	1.0				
X2	0.1	1.0			
X3	0.1	0.1	1.0		
X4	0.1	0.1	0.1	1.0	
Y	0.2	0.2	0.2	0.2	1.0

Validity is higher the lower the internal consistency.

Validity and reliability: a short digression

1. Although we know from Spearman that we can correct for reliability to find the “True" relationship between two variables, this does not help us in the real world.
2. Reliability is incorrectly associated with internal consistency which leads to such derivations as coefficients KR20 (Kuder and Richardson, 1937), λ_3 (Guttman, 1945) or α (Cronbach, 1951).
3. Expressed terms of inter-item correlations, this is just $\frac{k\bar{r}}{1+(k-1)\bar{r}}$ and increases with test length (k) and the average interitem correlation ($\bar{r}$)
4. However, validity of a k item test (r_{y_k}) or the correlation with an external criterion, Y, also increases with test length, and the average item validity ($\bar{r_y}$) but decreases as the inter-item correlation increases $r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k+k*(k-1)\bar{r}}}$.

Reliability and Validity

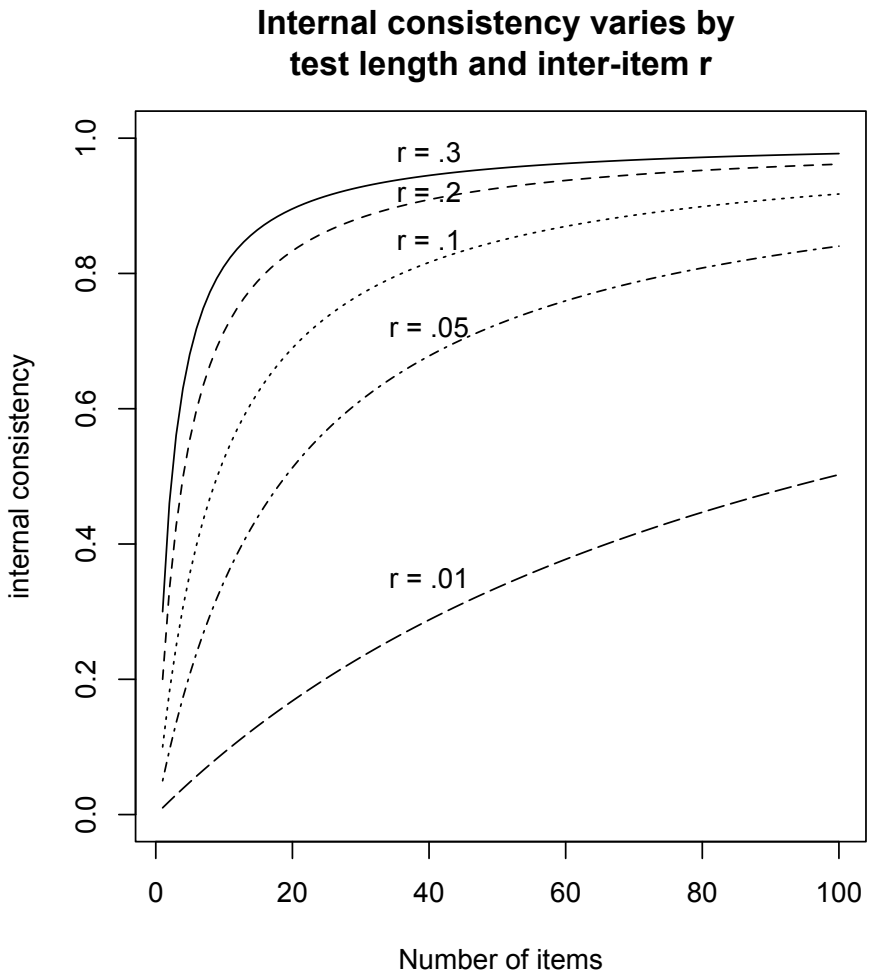
1. Lets unpack these two equations.
   Internal consistency

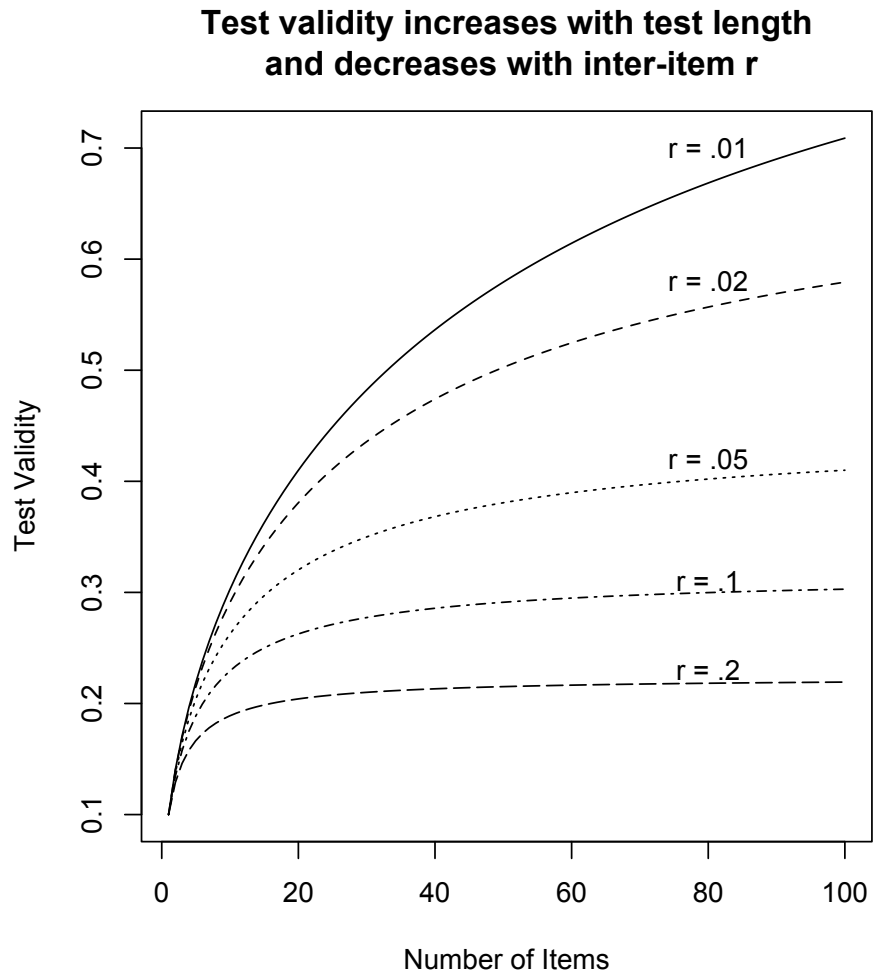
$$\lambda_3 = \alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \tag{3}$$

2. but validity

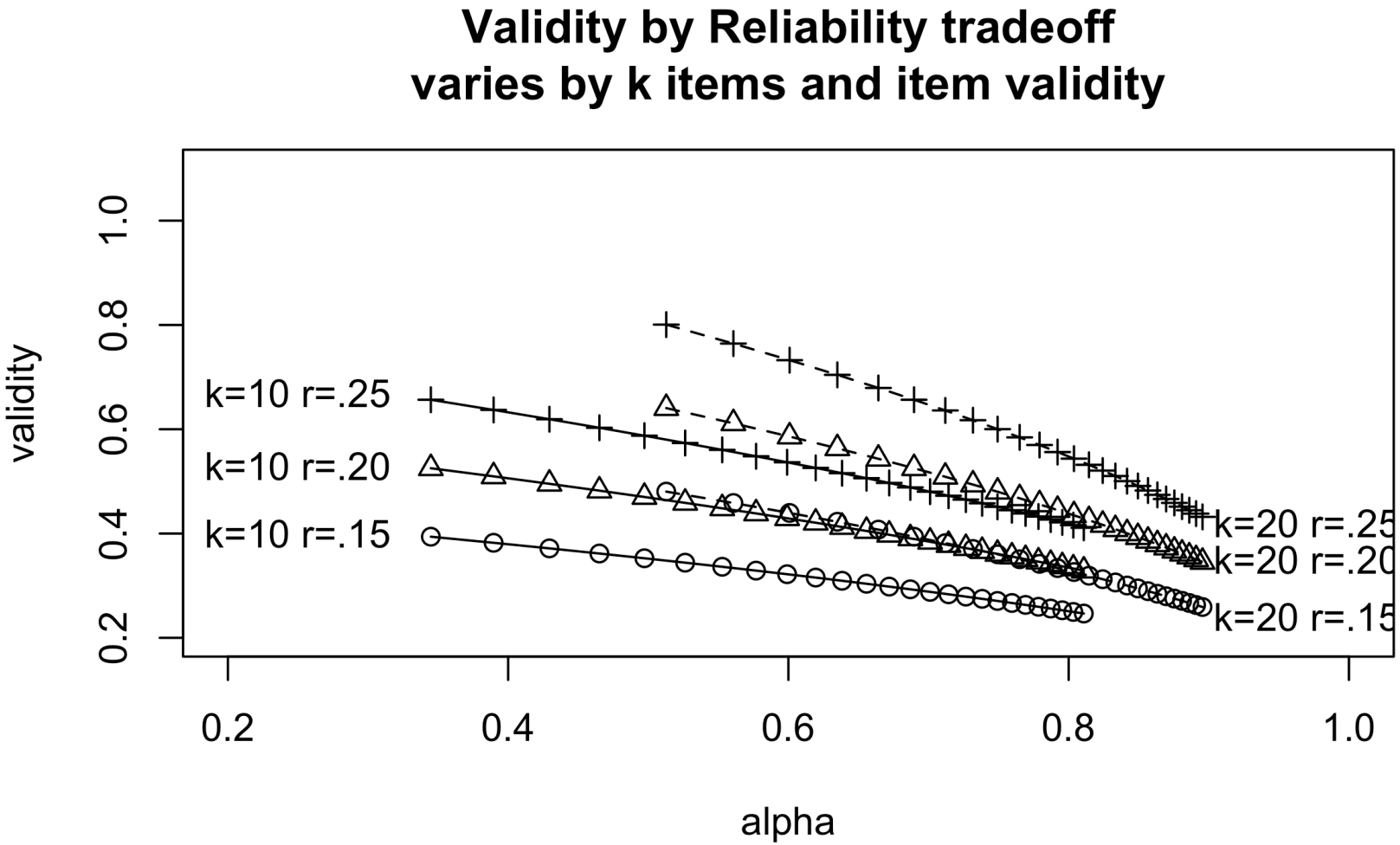
$$r_{y_k} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k + k * (k-1)\bar{r}}}. \tag{4}$$

The trade off between test consistency and test validity





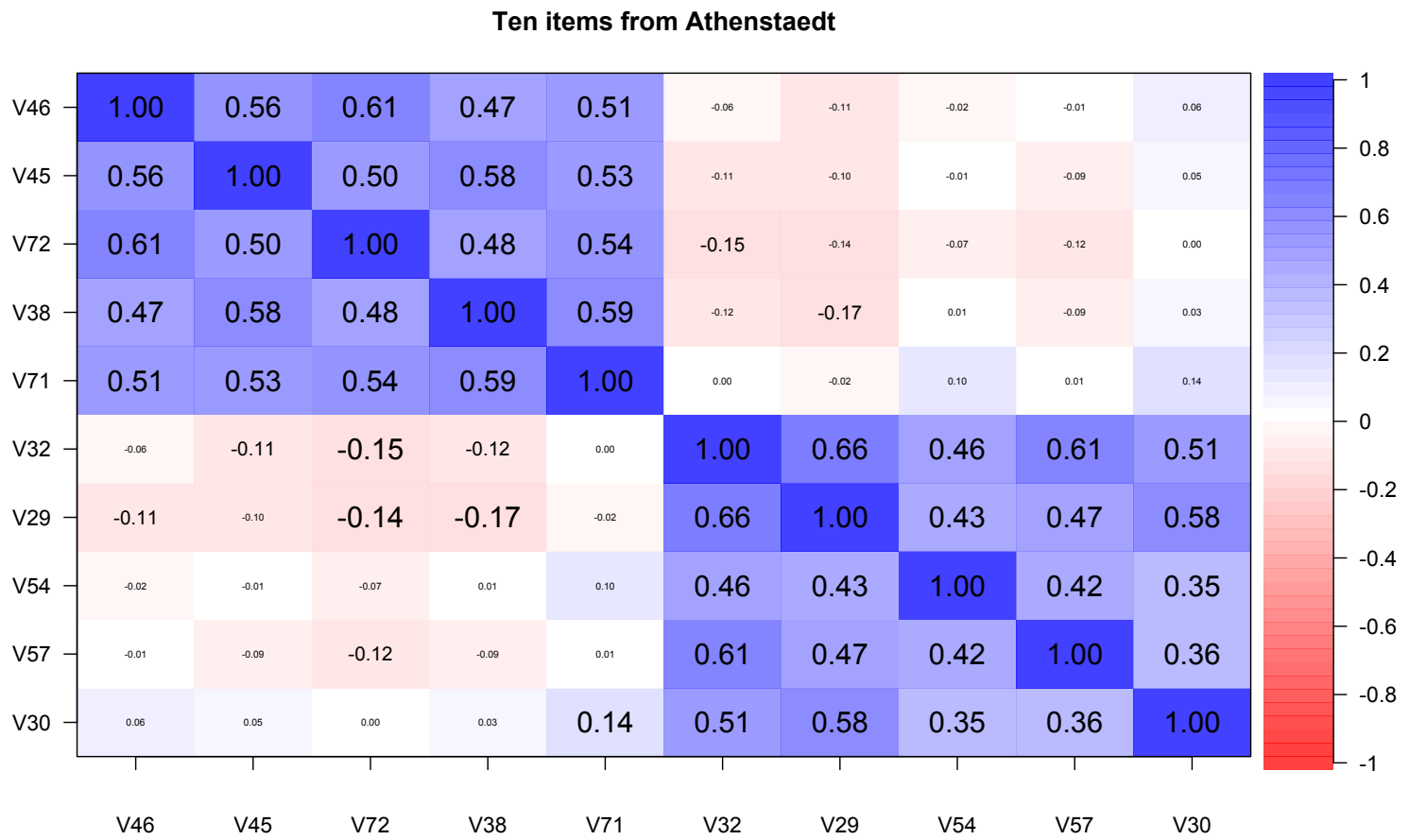
The trade off between test consistency and test validity



Increasing validity implies increasing the diversity of the item content

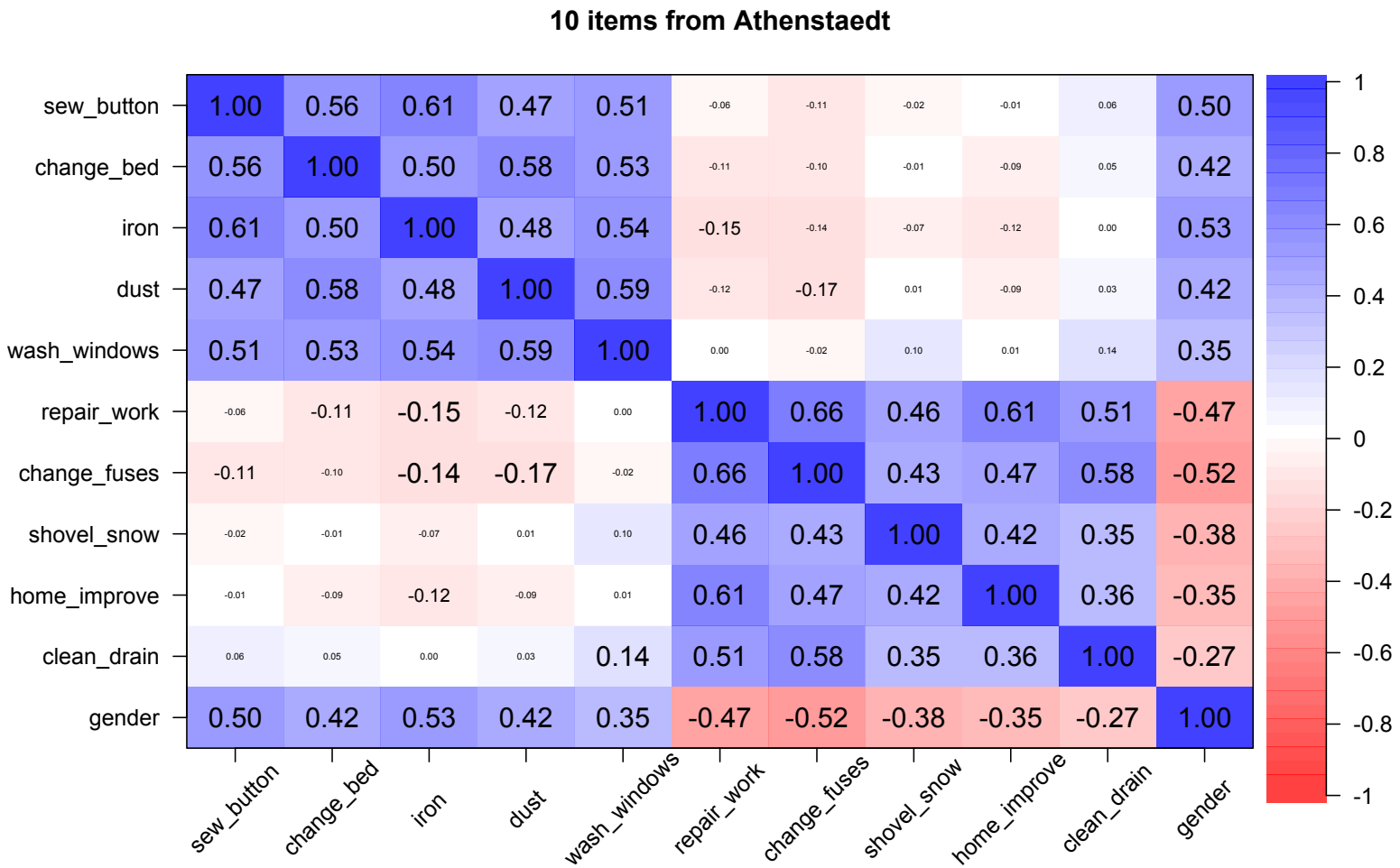
1. The goal of construct validity is have pure measures with high internal consistency. (Measure one thing well).
2. And highly correlated measures of the same constructs.
3. But if the goal is predictive validity, we should minimize internal consistency and have independent predictors.
4. By emphasizing practical validity, we are ignoring most of what we have been taught (and teach) about reliability (Revelle and Condon, 2018, 2019) and scale construction (Revelle and Garner, 2023).
5. Variations on this theme have been discussed before by (Condon et al., 2021; Möttus et al., 2020).

10 items from Athenstaedt (2003)



Clearly a two factor solution (using the inter-ocular trauma test).

10 items from Athenstaedt (2003) predict gender



Clearly a two factor solution but with some interesting correlations with gender.

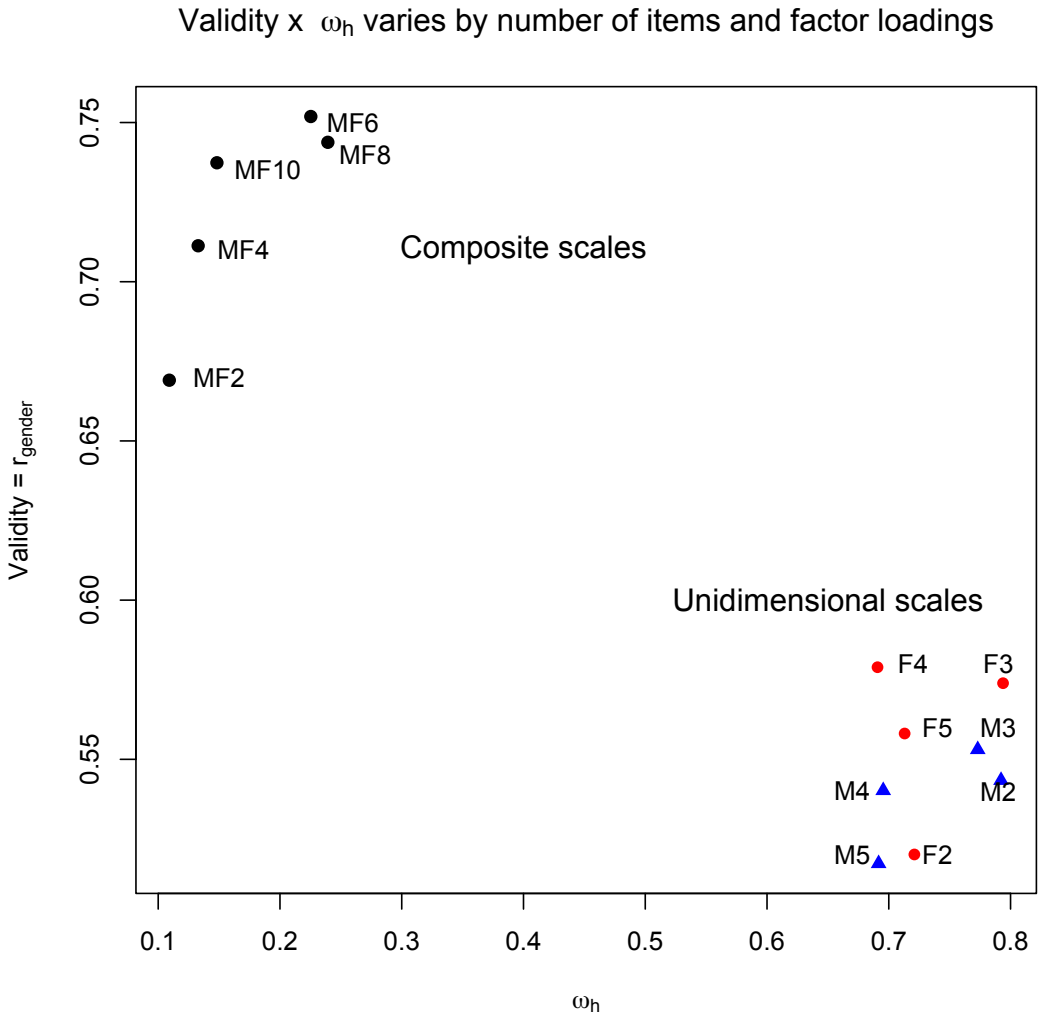
Form various short scales

1. It is easy to form 2 ... 5 item short and factorially pure scales from these items. (F2 ... F5, or M2 ... M5)
2. Equally easy to form 2 .. 10 item composite scales mixing M and F content (MF2 ... MF10)
3. Just M or just F scales are very internally consistent ($\omega_h = .72$ ... .85) and reasonably valid ($r_{gender} = .52$ ... .58)
4. But the composite (MF) scales are much less internally consistent ($\omega_h = .11$ ... .23, $\alpha = .11$ ... .77) and more valid ($r_{gender} = .67$ ... .75)

Reliability and Validity for Short M, F, and MF scales

Relability and Validity

Scale	ω_h	α	r_{gender}
F2	0.72	0.72	0.52
F3	0.79	0.79	0.57
F4	0.69	0.82	0.58
F5	0.71	0.85	0.56
M2	0.79	0.79	0.54
M3	0.77	0.76	0.55
M4	0.70	0.81	0.54
M5	0.69	0.82	0.52
MF2	0.11	0.11	0.67
MF4	0.13	0.59	0.71
MF6	0.23	0.69	0.75
MF8	0.24	0.75	0.74
MF10	0.15	0.77	0.74



Darts or Fishing Spears versus Fishing Nets

1. The M and F scales are sharper spears (more internally consistent) and have a clear one factor solution.
2. And the mixed composite scales are looser (less internally consistent), less clear construct (multifactorial) and more net like.
3. But Fishing Nets catch more fish (have higher validities) than do Spears.
4. Perhaps it is time to not focus on construct validity or factorial purity but rather on predictive validity.

Athenstaedt, U. (2003). On the content and structure of the gender role self-concept: Including gender-stereotypical behaviors in addition to traits. *Psychology of Women Quarterly*, 27(4):309–318.

Benbow, C. P., Lubinski, D. J., and Stanley, J. C. (1996). *Intellectual talent: psychometric and social issues*. Johns Hopkins University Press, Baltimore.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(8):81–105.

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*.

Condon, D. M., Wood, D., Möttus, R., Booth, T., Costani, G., Greiff, S., Johnson, W., Lukaszesksi, A., Murray, A., Revelle, W., Wright, A. G., Ziegler, M., and Zimmerman, J. (2021). Bottom Up Construction of a Personality Taxonomy. *European Journal of Psychological Assessment*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement*, 10(1):94–96.

Dana, J., Dawes, R., and Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment & Decision Making*, 8(5):512–520.

Danielson, J. R. and Clark, J. H. (1954). A personality inventory for induction screening. *Journal of Clinical Psychology*, 10(2):137 – 143.

Dawes, R. (2009). *House of cards*. Simon and Schuster.

Dawes, R. M. (1989). Experience and validity of clinical judgment: The illusory correlation. *Behavioral Sciences & the Law*, 7(4):457–467.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.

Deary, I. J. (2008). Why do intelligent people live longer? *Nature*, 456(7219):175–176.

Deary, I. J. (2009). Introduction to the special issue on cognitive epidemiology. *Intelligence*, 37:517–519.

Deary, I. J. and Batty, G. D. (2007). Cognitive epidemiology. *British Medical Journal*, 61(5):378–384.

Deary, I. J., Pattie, A., and Starr, J. M. (2013). The stability of intelligence from age 11 to age 90 years: The Lothian Birth Cohort of 1921. *Psychological Science*, 24(12):2361–2368.

Deary, I. J., Strand, S., Smith, P., and Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1):13–21.

Deary, I. J., Whiteman, M., Starr, J., Whalley, L., and Fox, H. (2004). The impact of childhood intelligence on later life: Following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, 86:130–147.

DeVaul, R. A., Jervey, F., Chappell, J. A., Caver, P., Short, B., and O'Keefe, S. (1987). Medical school performance of initially rejected students. *JAMA*, 257(1):47–51.

Dubois, P. H. (1947). The classification program report no. 2. Army air forces aviation psychology program research reports, Army Air Forces, Defense Documentation Center Defense Supply Agency.

Eagly, A. H. and Revelle, W. (2022). Understanding the Magnitude of Psychological Differences Between Women and Men Requires Seeing the Forest and the Trees. *Perspectives on Psychological Science*, 17(5):1339–1358.

Elleman, L. G., McDougald, S., Revelle, W., and Condon, D. (2020). That takes the BISCUIT: a comparative study of predictive accuracy and parsimony of four statistical learning techniques in personality data, with data missingness conditions. *European Journal of Psychological Assessment*, 36(6):948–958.

Friedman, H. S., Tucker, J. S., Schwartz, J. E., Tomlinson-Keasey, C., Martin, L. R., Wingard, D. L., and Criqui, M. H. (1995).

Psychosocial and behavioral predictors of longevity: The aging and death of the “termites". *American Psychologist*, 50(2):69 – 78.

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley & Sons, Inc.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282.

Kelly, E. L. and Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist*, 5(8):395 – 406.

Kelly, E. L. and Fiske, D. W. (1951). *The prediction of performance in clinical psychology*. University of Michigan Press, Ann Arbor, Michigan.

Kuder, G. and Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

Kuncel, N. R., Campbell, J. P., and Ones, D. S. (1998). Validity of the graduate record examination: Estimated or tacitly known? *American Psychologist*, 53(5):567–568.

Kuncel, N. R. and Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, 315(5815):1080–1081.

Kuncel, N. R., Hezlett, S. A., and Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127(1):162 – 181.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports Monograph Supplement 9*, 3:635–694.

Lubinski, D. (2016). From Terman to today: A century of findings on intellectual precocity. *Review of Educational Research*.

Lubinski, D. and Benbow, C. P. (2000). States of excellence. *American Psychologist*, 55(1):137 – 150.

Lubinski, D. and Benbow, C. P. (2006). Study of Mathematically Precocious Youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, 1(4):316–345.

Lubinski, D., Webb, R., Morelock, M., and Benbow, C. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology*, 86(4):718–729.

McPherson, W. B. (1901). Gideon's water-lappers. *Journal of the American Oriental Society*, 22:70–75.

Möttus, R., Sinick, J., A.Terracciano, Hřebíckova, M., Kandler, C., and Jang, J. A. . . . K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, 117(4).

Möttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costani, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszesksi, A., Murray, A., Revelle, W., Wright, A. G., Yarkoni, T., Ziegler, M., and Zimmerman, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality*, 34(6).

Oden, M. (1968). *The fulfillment of promise: 40-year follow-up of the Terman gifted group*, volume 77. Stanford University Press.

OSS Assessment Staff (1948). *Assessment of Men: Selection of personnel for the office of strategic services*. Rinehart, New York.

Revelle, W. (2023). The seductive beauty of latent variables: ISSID award for distinguished contribution to the study of individual differences. Belfast. International Society for the Study of Individual Differences.

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences*, 221:112552.

Revelle, W. (2025). *psych:Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, https://CRAN.r-project.org/package=psych, 2.5.3 edition. R package version 2.5.3.

Revelle, W. and Condon, D. M. (2018). Reliability. In Irwing, P., Booth, T., and Hughes, D. J., editors, *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. John Wily & Sons, London.

Revelle, W. and Condon, D. M. (2019). Reliability: from alpha to omega. *Psychological Assessment*, 31(12):1395–1411.

Revelle, W., Dworak, E. M., and Condon, D. M. (2021). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, 169.

Revelle, W. and Garner, K. M. (2023). Measurement: Reliability, construct validation, and scale construction. In Harry T. Reis, T. W. and Judd, C. M., editors, *Handbook of Research Methods in Social and Personality Psychology (in press)*.

Taylor, H. C. and Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, 23(5):565 – 578.

Terman, L. M. and Oden, M. (1947). *Genetic studies of genius*. Stanford University Press; Oxford University Press., Palo Alto, CA.

Terman, L. M. and Oden, M. (1959). *The gifted group at mid-life: Thirty-five years' follow-up of the superior child*, volume 5. Stanford Univ Pr.

Underwood, E. (2014). Starting young. *Science*, 346(6209):568–572.

Wiggins, J. S. (1973). *Personality and prediction: principles of personality assessment*. Addison-Wesley Pub. Co, Reading, Mass.

Wolfe, T. (1970). *The Right Stuff*. Straus & Giroux„ New York: Farrar.

Zola, A., Condon, D. M., and Revelle, W. (2021). The Convergence of Self and Informant Reports in a Large Online Sample. *Collabra: Psychology*, 7(1).