Psychology 405: Psychometric Theory Validity

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

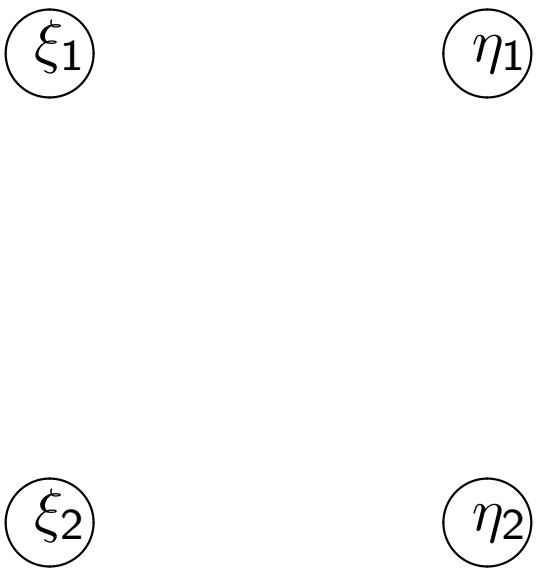
May, 2018

Outline

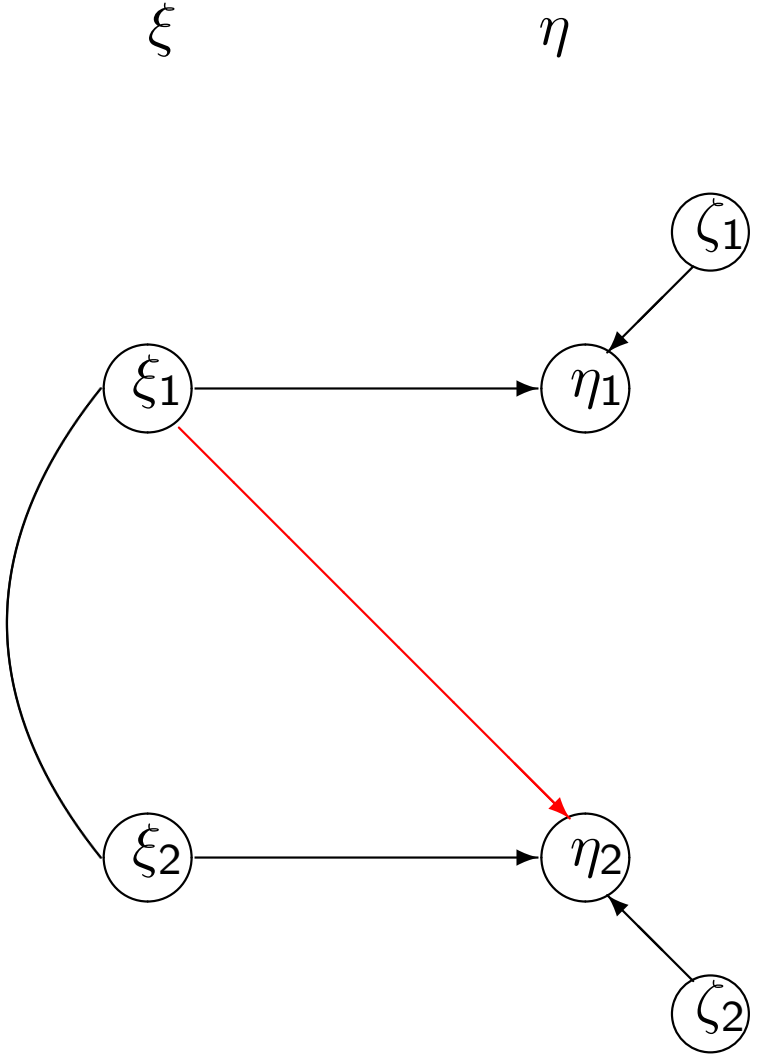
Observed Variables

Latent Variables

ξ η

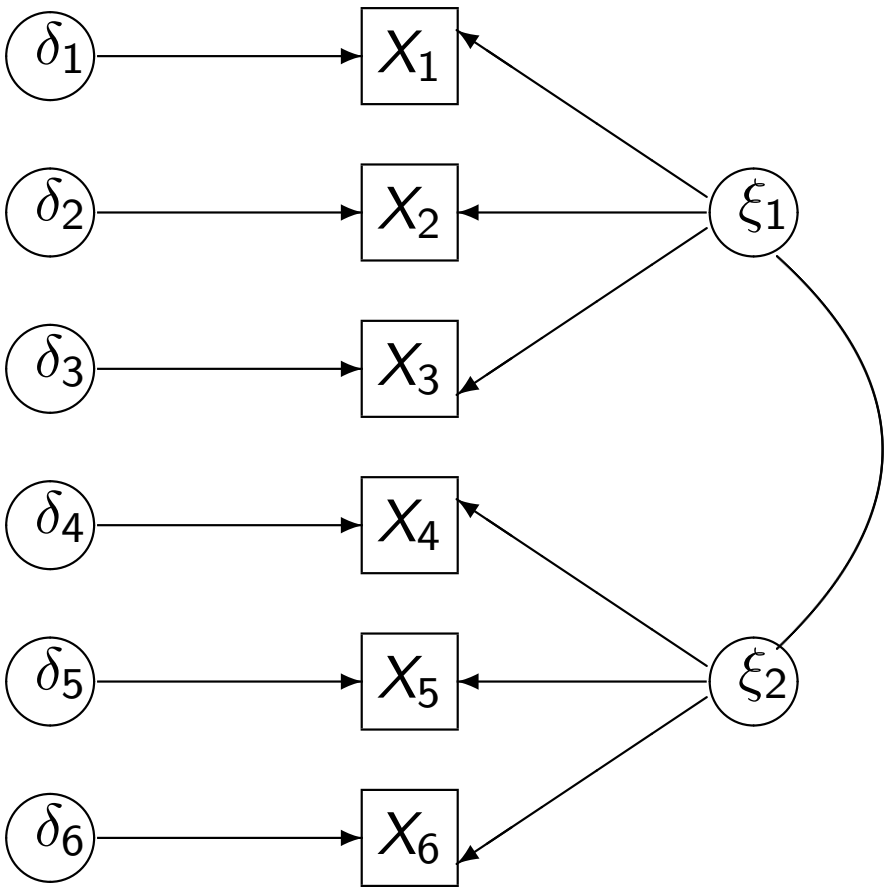


Theory: A regression model of latent variables

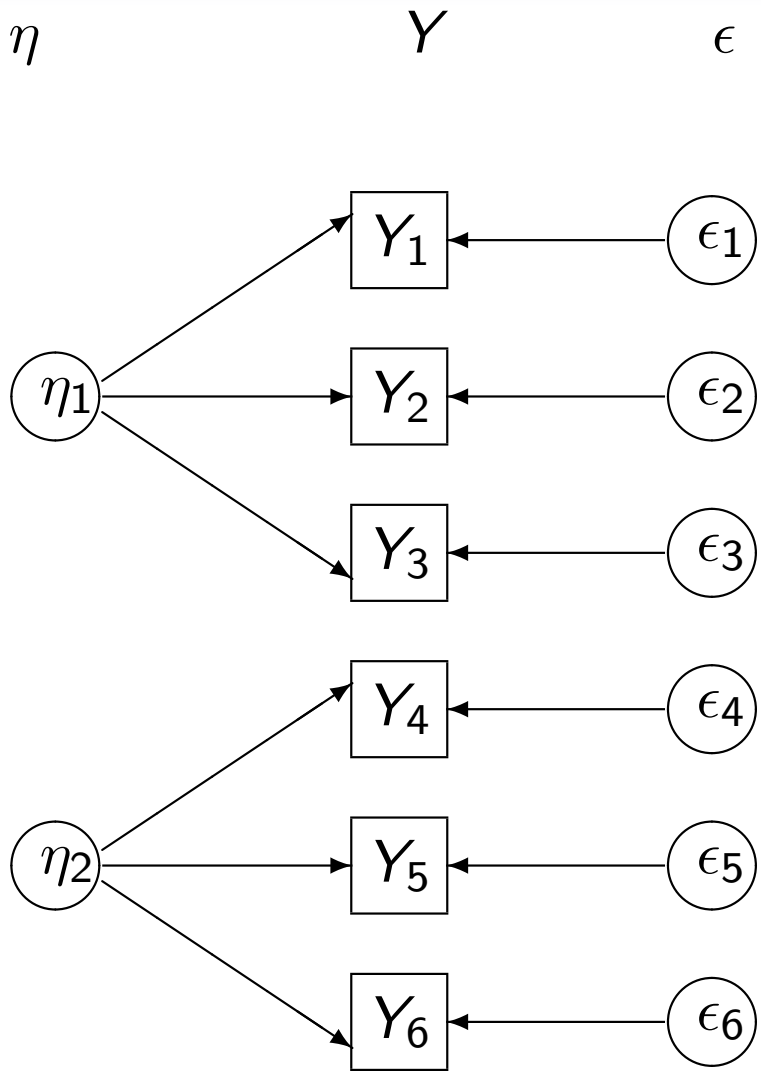


A measurement model for X – Correlated factors

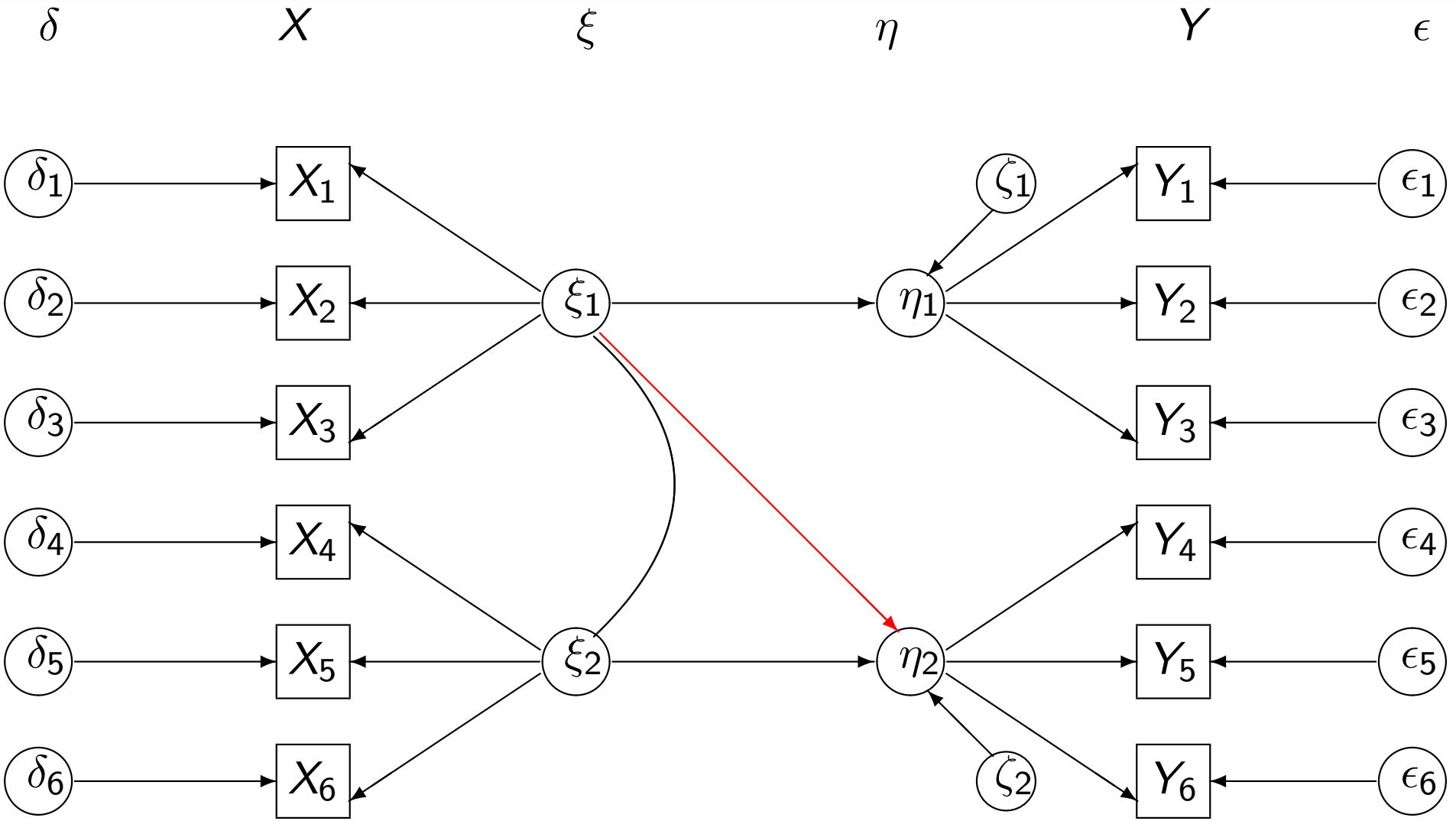
δ X ξ



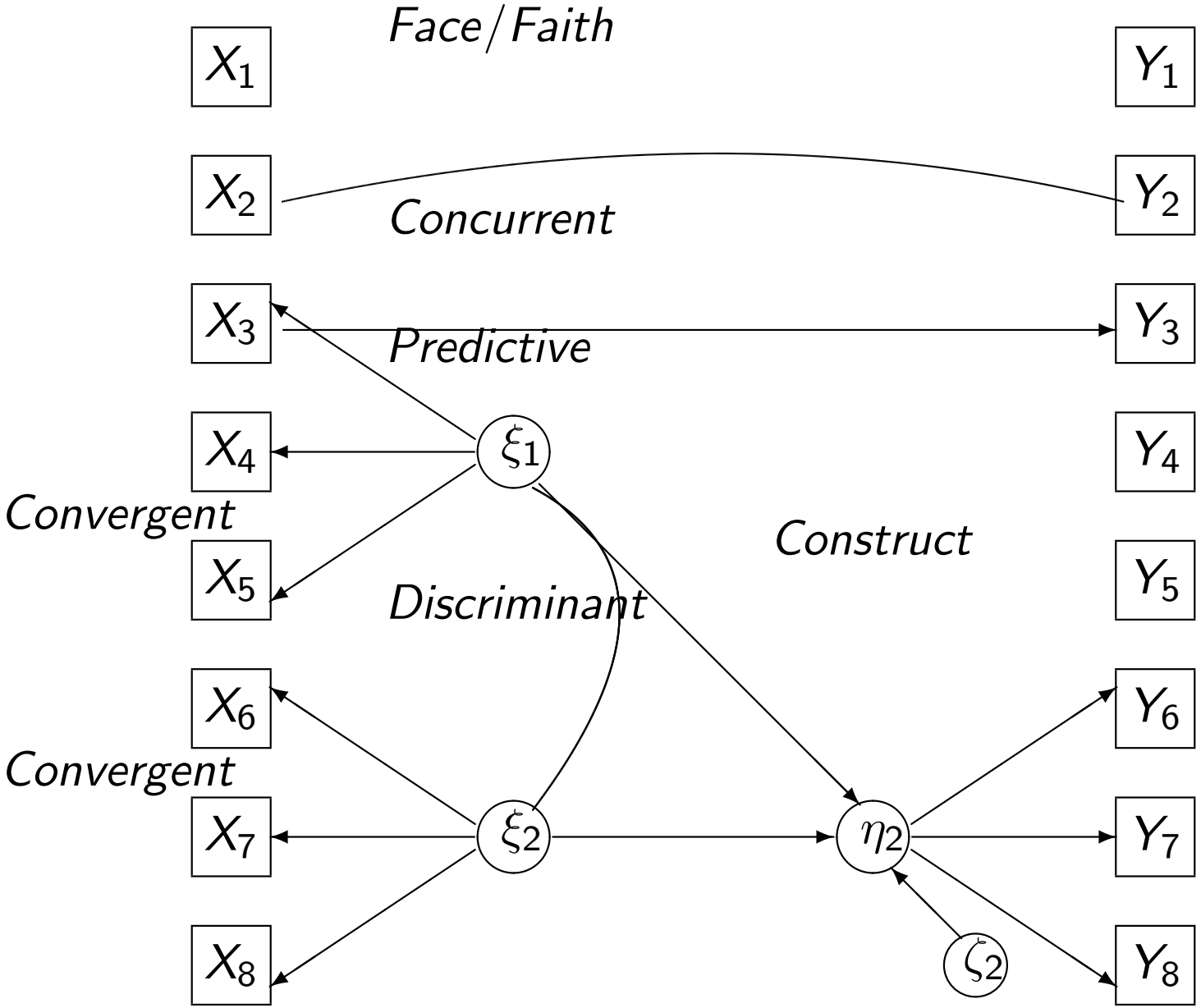
A measurement model for Y - uncorrelated factors



A complete structural model



Types of Validity



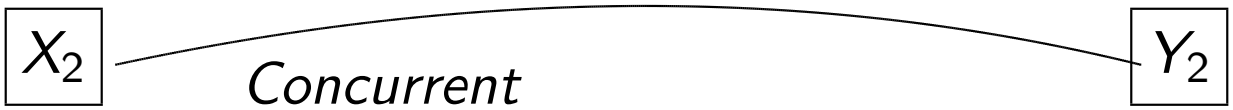
Face Validity

$\boxed{X_1}$ *Face*/*Faith*

Representative Content

Seeming relevance

Concurrent Validity

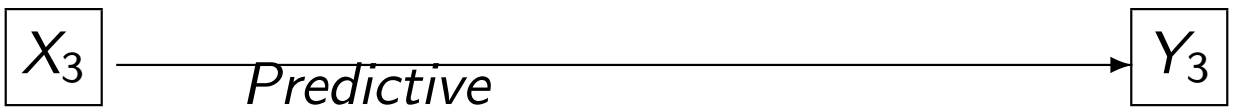


Does a measure correlate with the criterion?

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

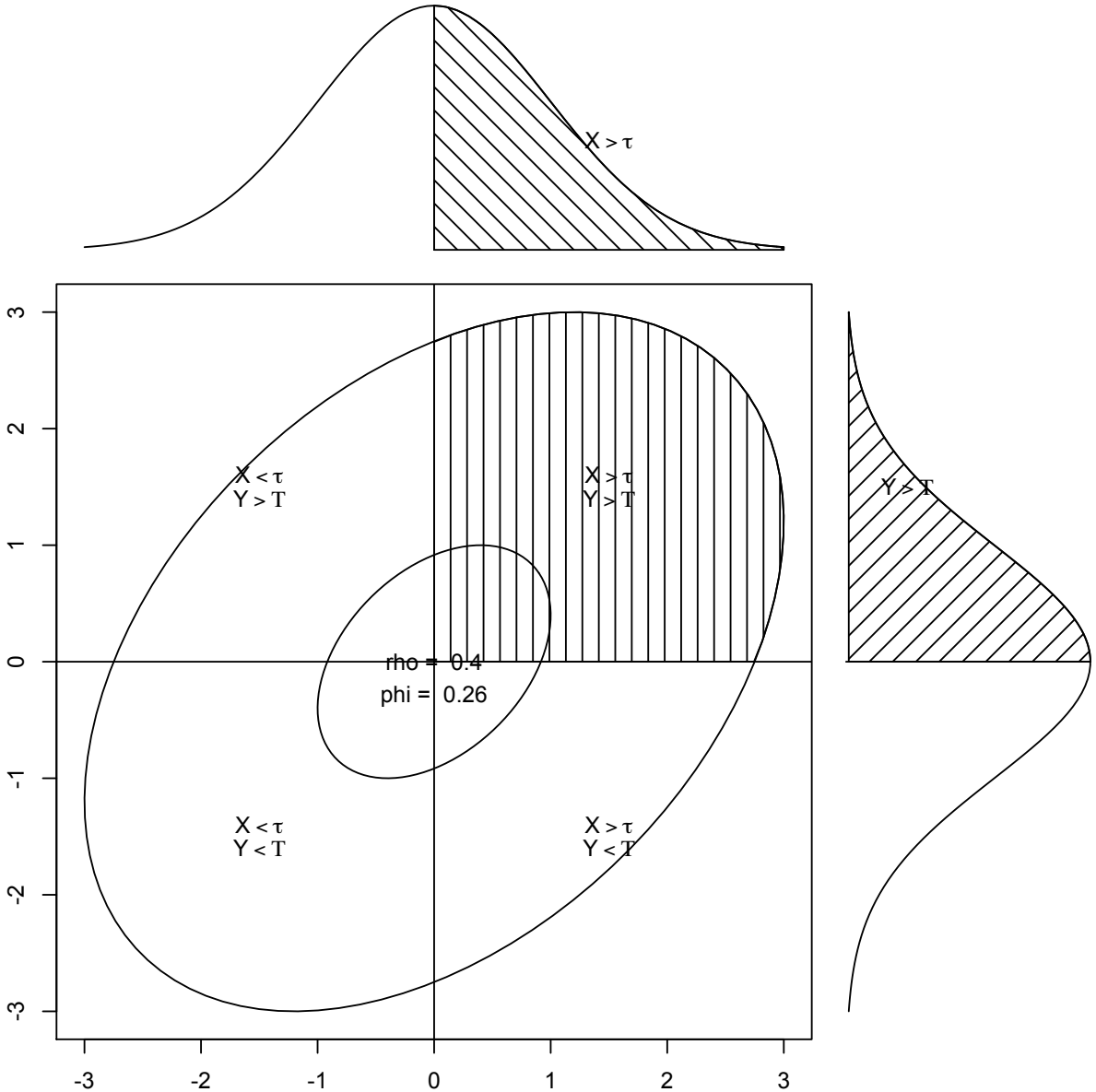
1. Continuous predictor, continuous criterion
   - Regression, multiple regression, correlation
   - Slope of regression implies how much change for unit change in predictor
2. Continuous predictor, dichotomous criterion
   - point bi-serial correlation
3. Dichotomous predictor, dichotomous outcome
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

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

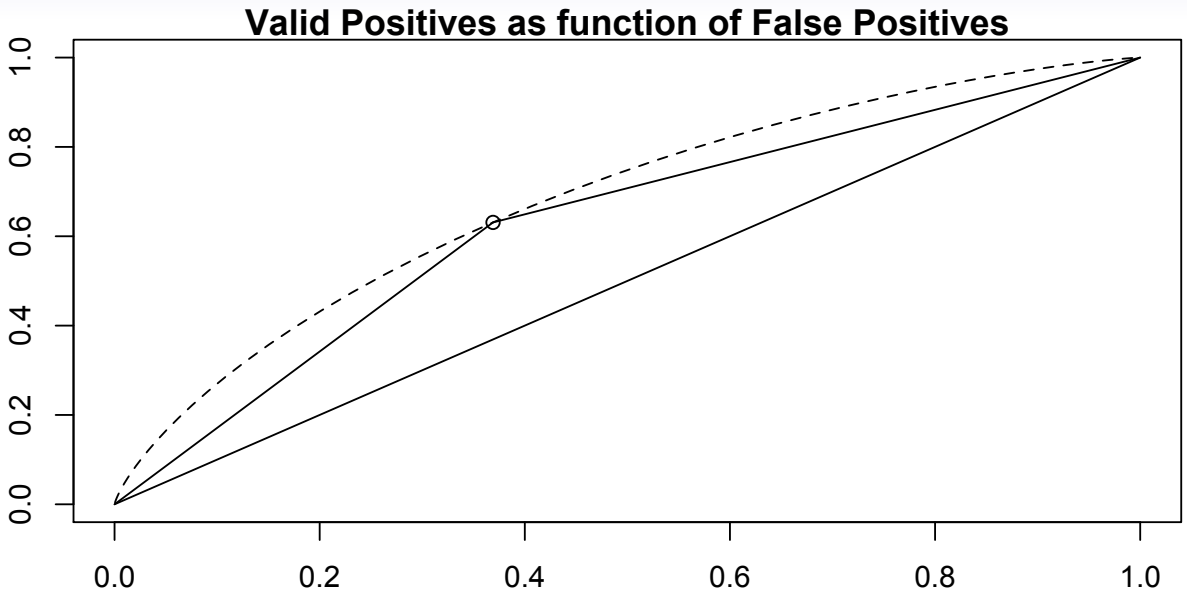
   - Therefore, the number of valid positives is

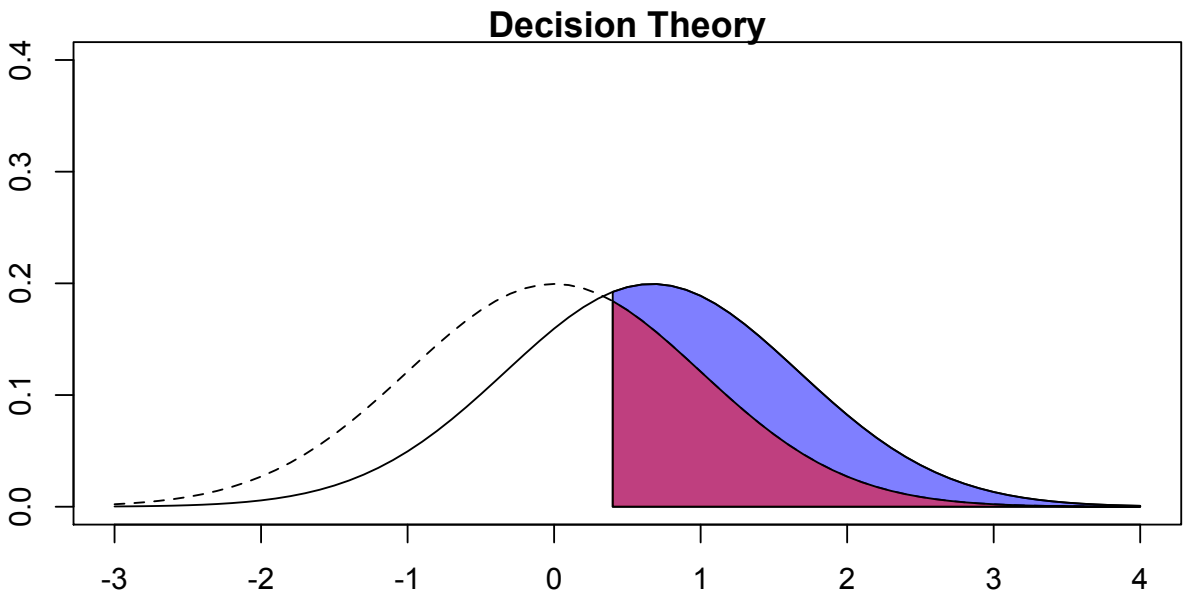
$$VP = BR * SR + \phi\sqrt{(BR)(1 - BR)(SR)(1 - SR)} \tag{2}$$

Tetrachoric and phi as function of cut points

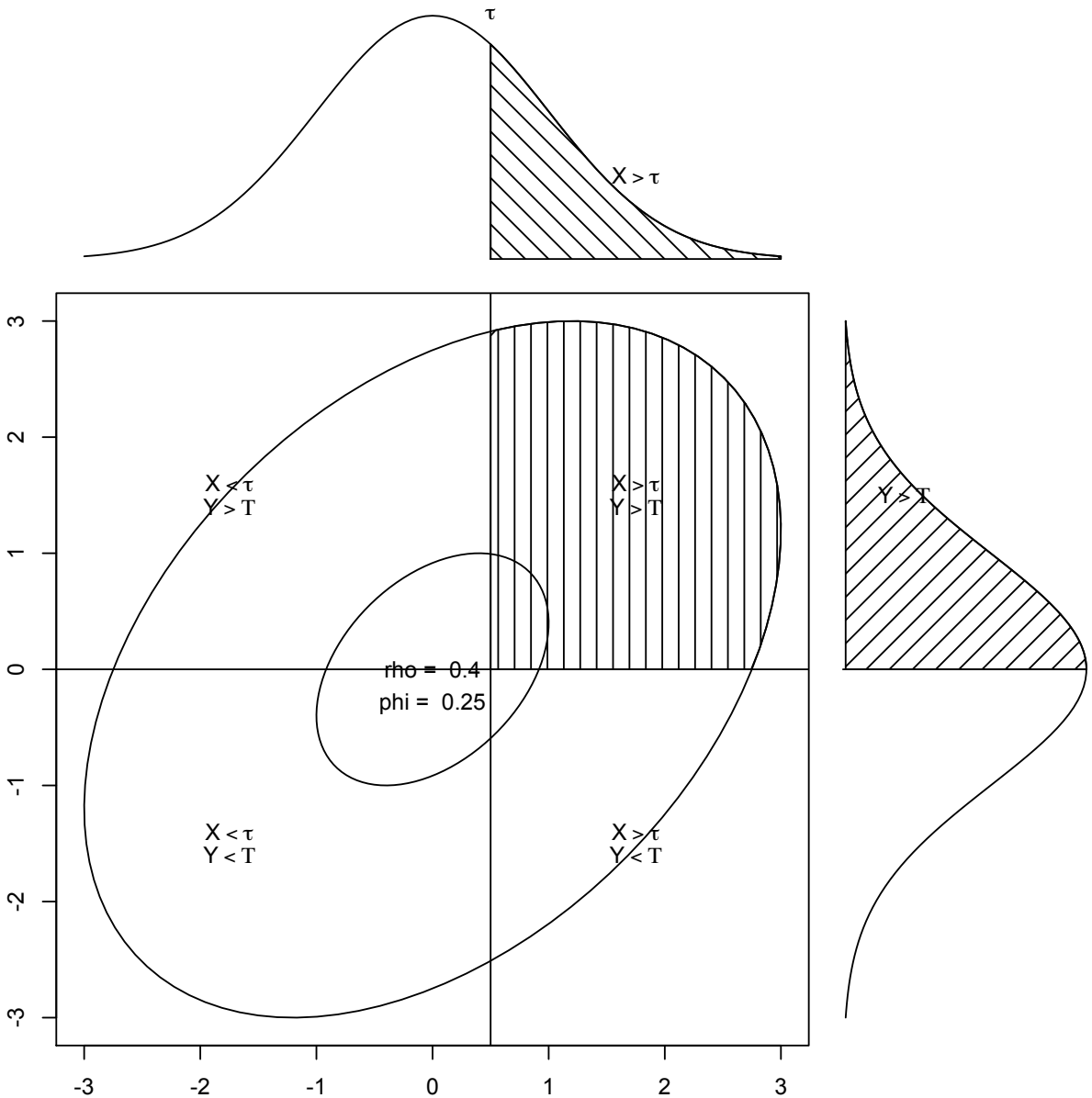


A decision theoretic approach

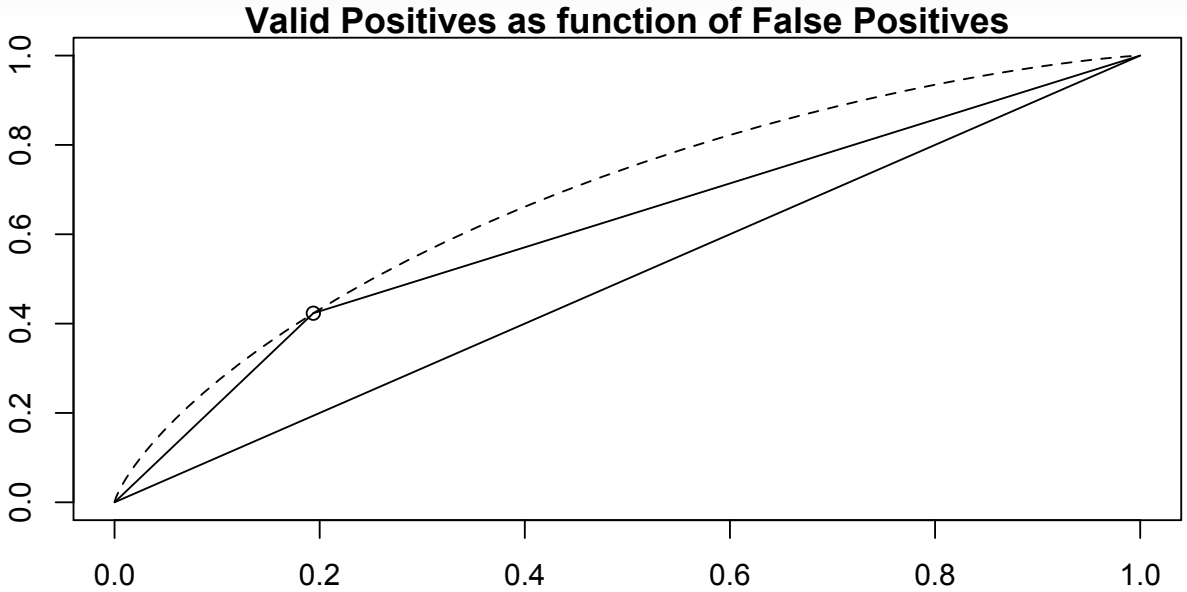


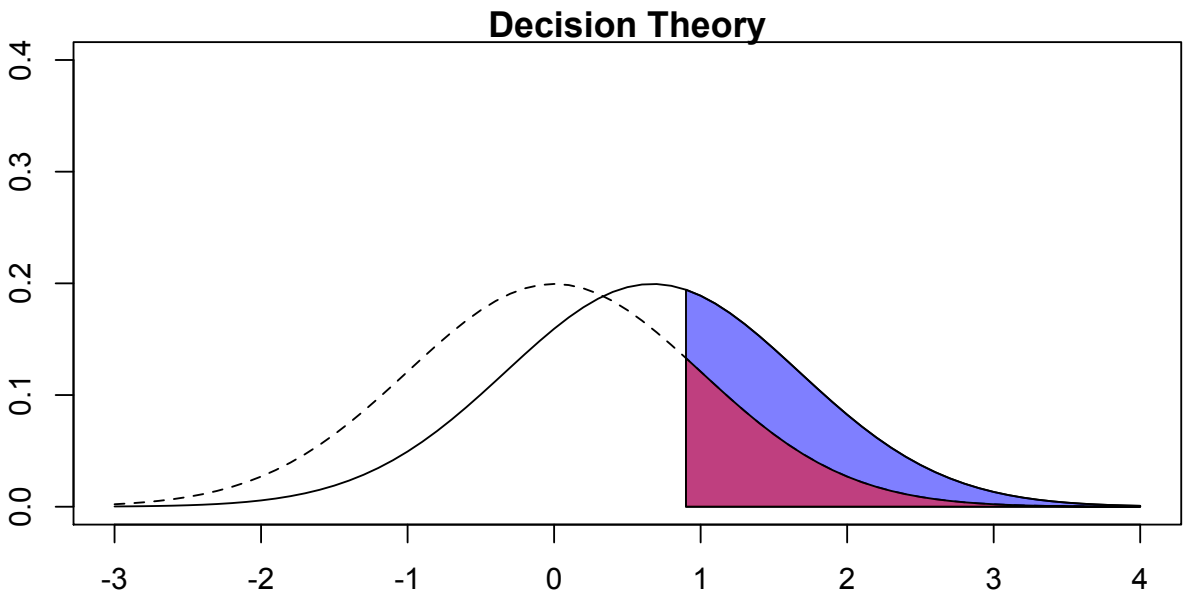


Tetrachoric and phi as function of cut points .5,0

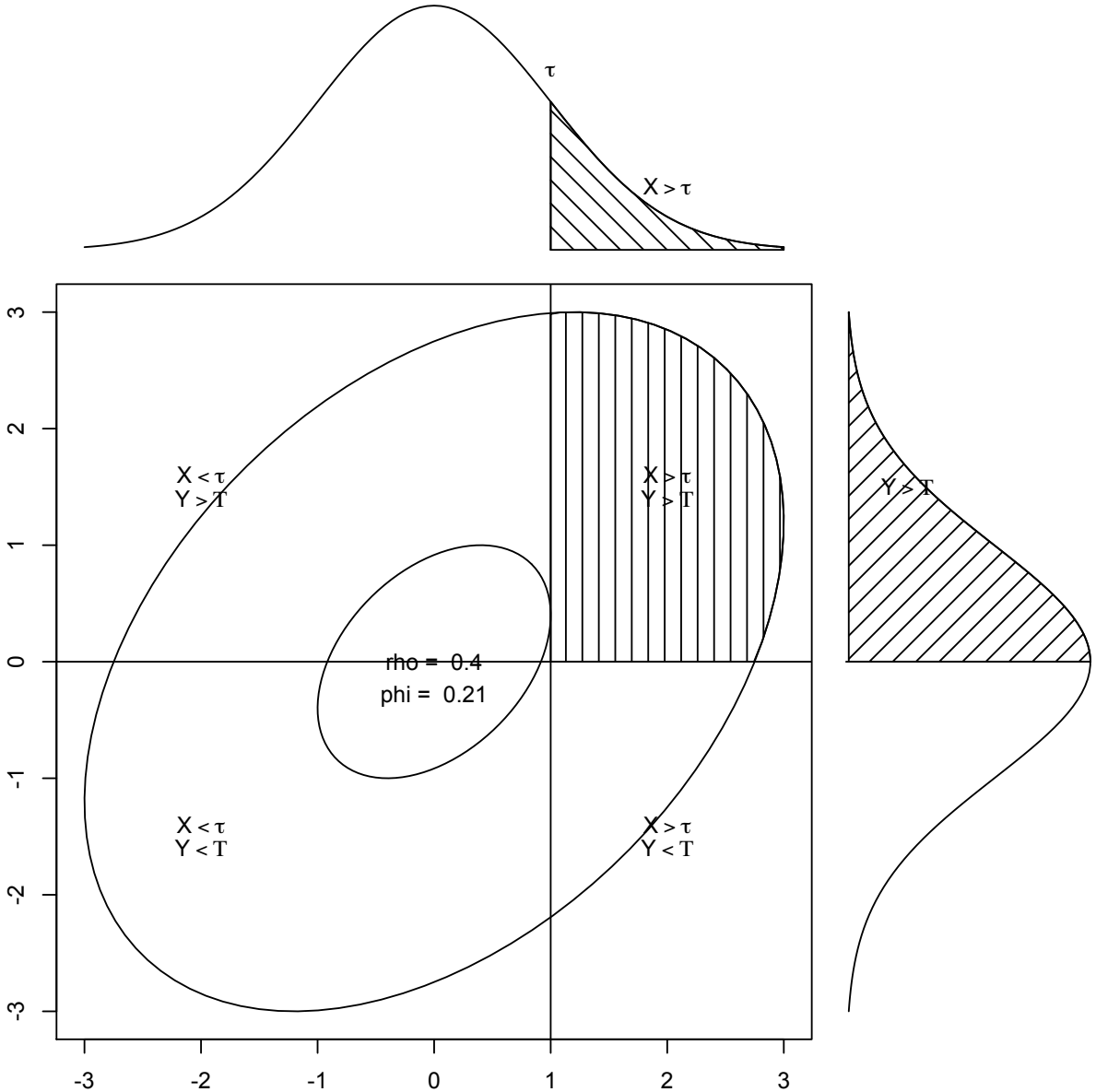


A decision theoretic approach with low beta

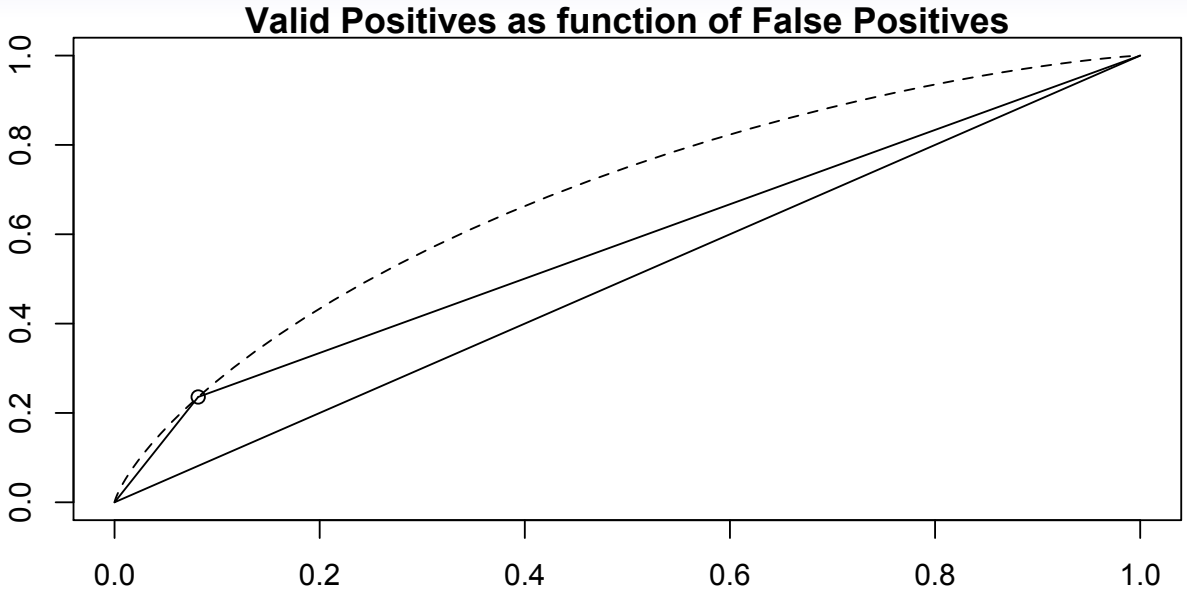


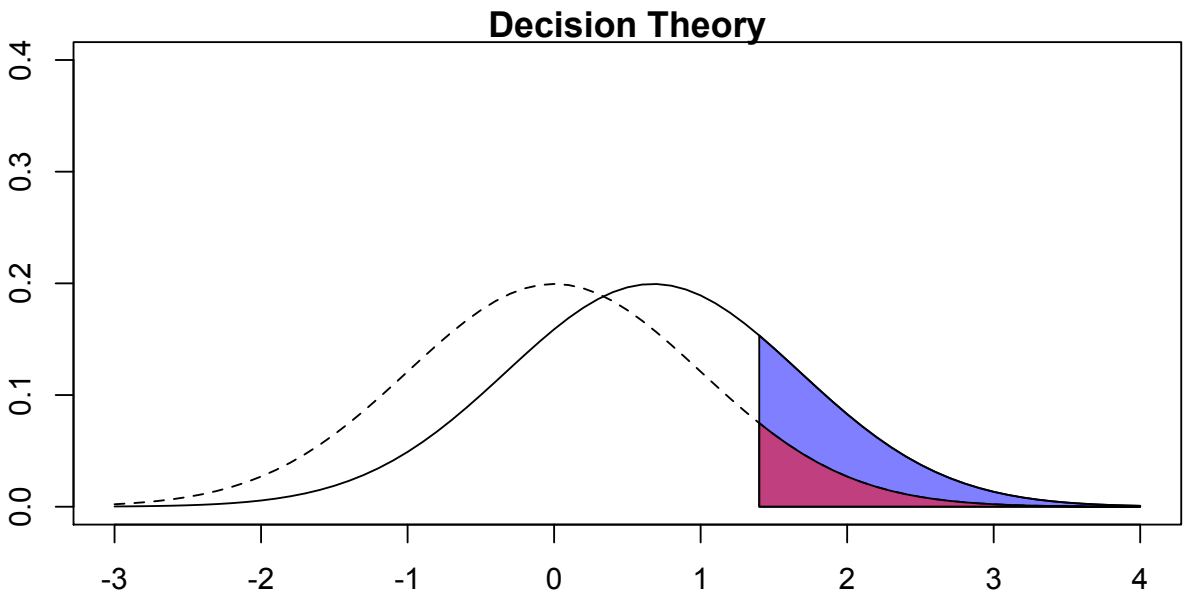


Tetrachoric and phi as function of cut points 1,0

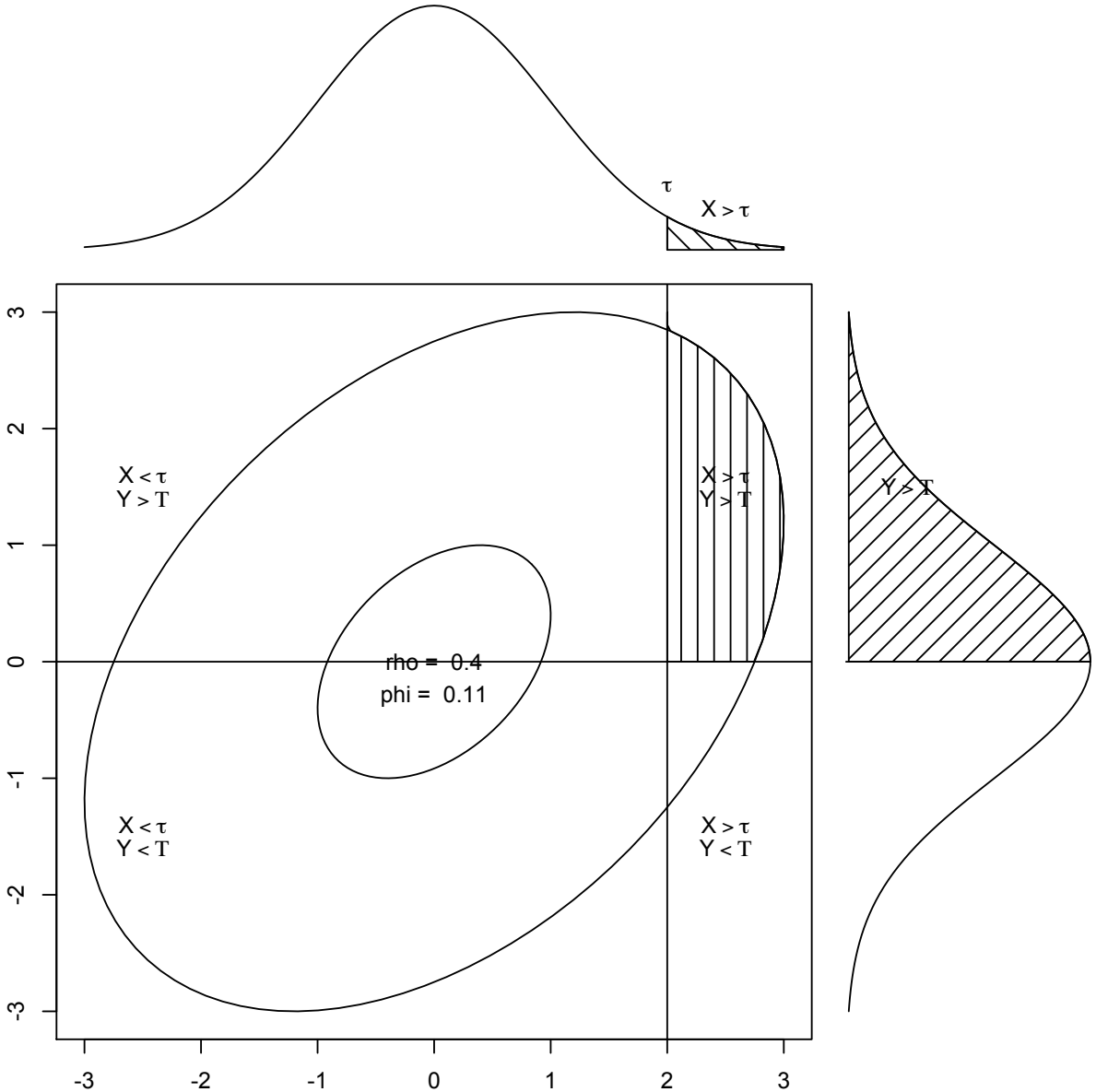


A decision theoretic approach with high beta

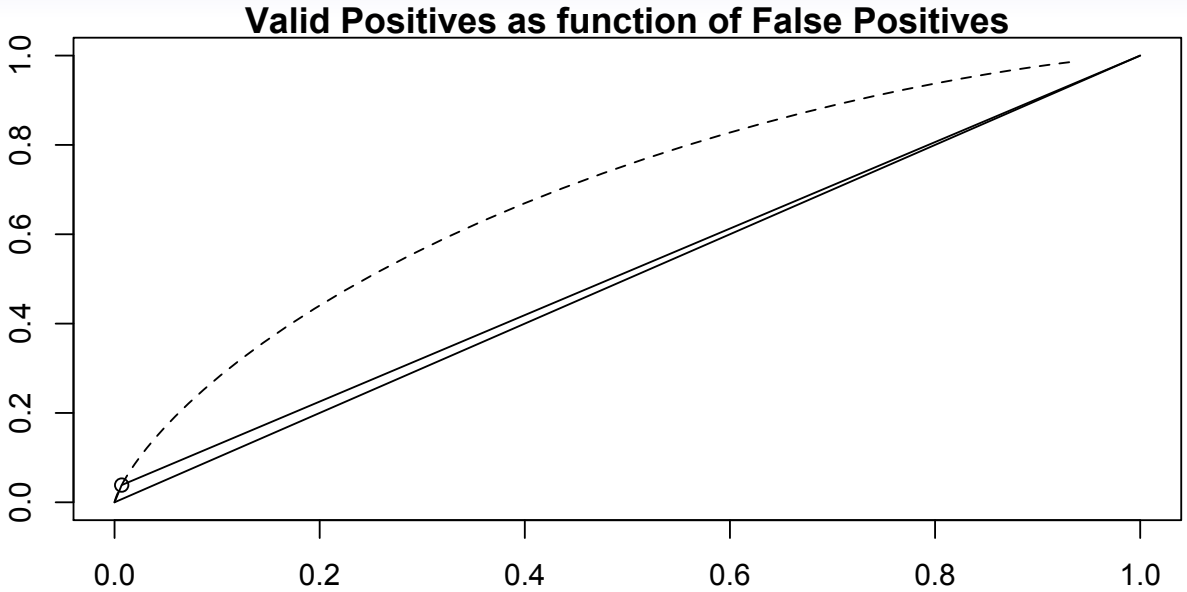


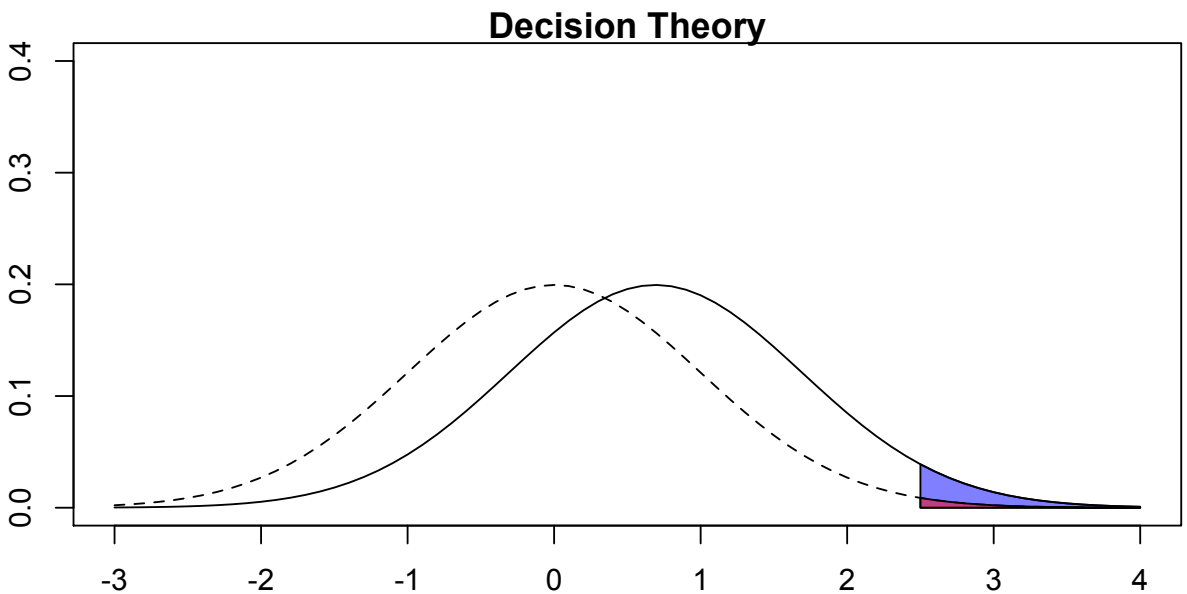


Tetrachoric and phi as function of cut points 2,0

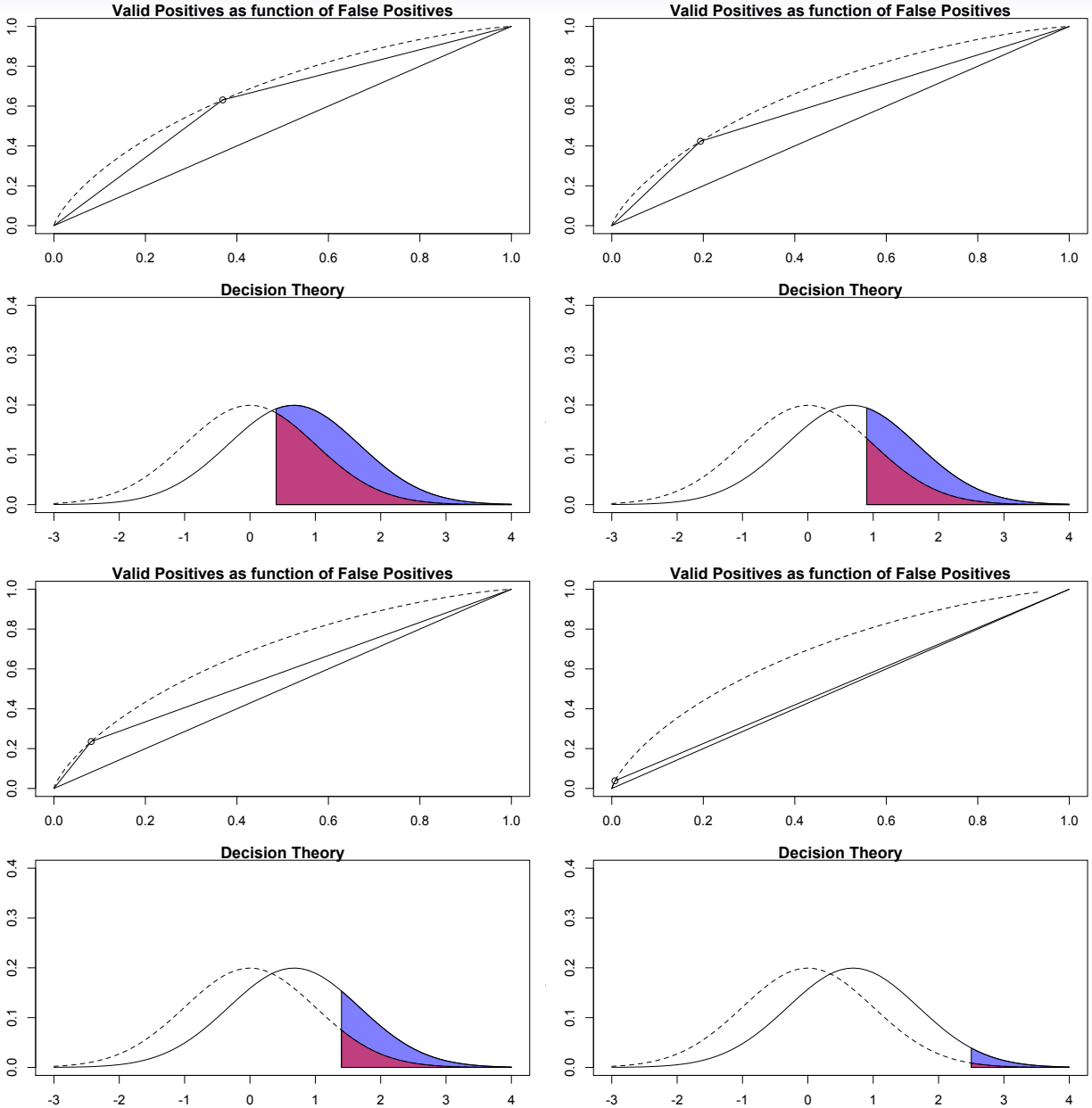


A decision theoretic approach with high beta





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 & 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

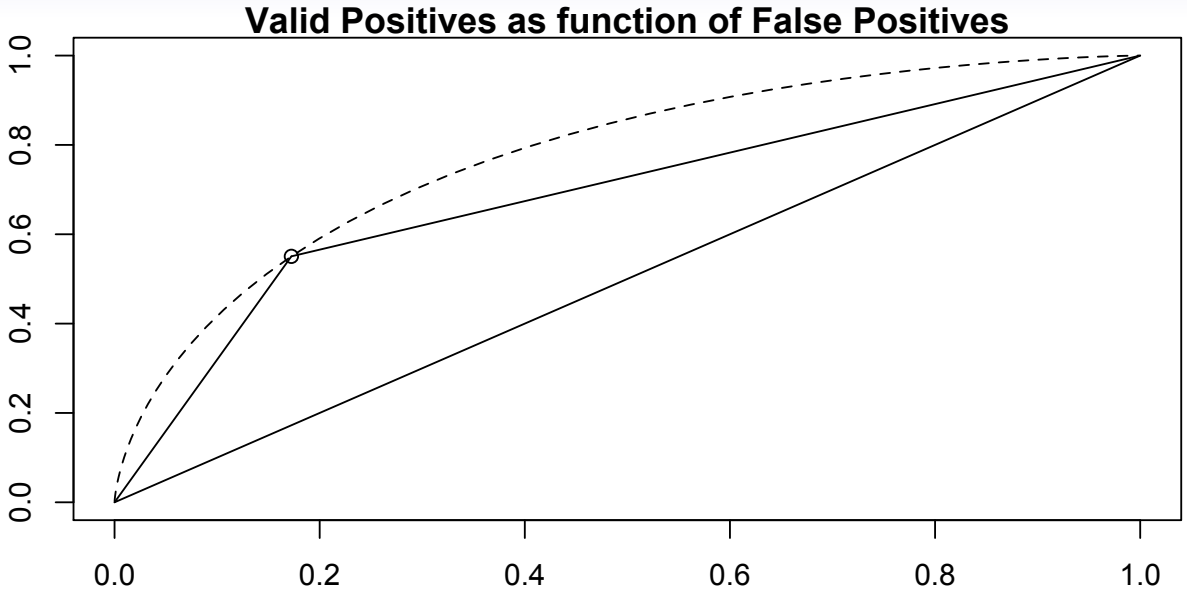
Accuracy = $.097 + .667 = .76$

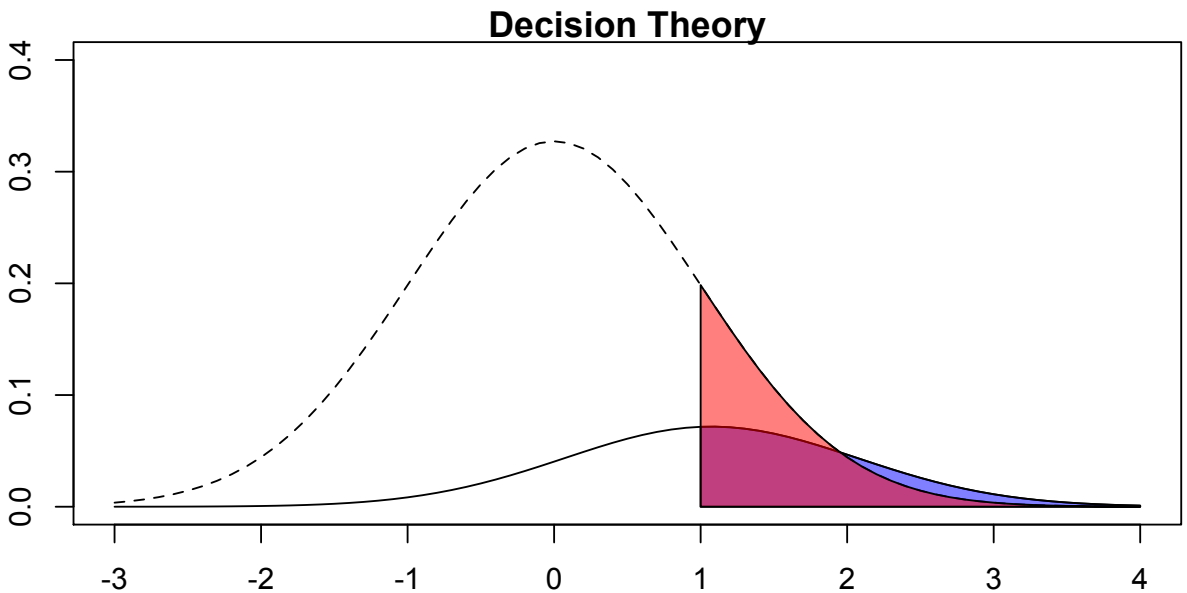
Sensitivity = $.097/(.097 + .079) = .55$

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

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

The Danielson and Clark data set as a decision problem



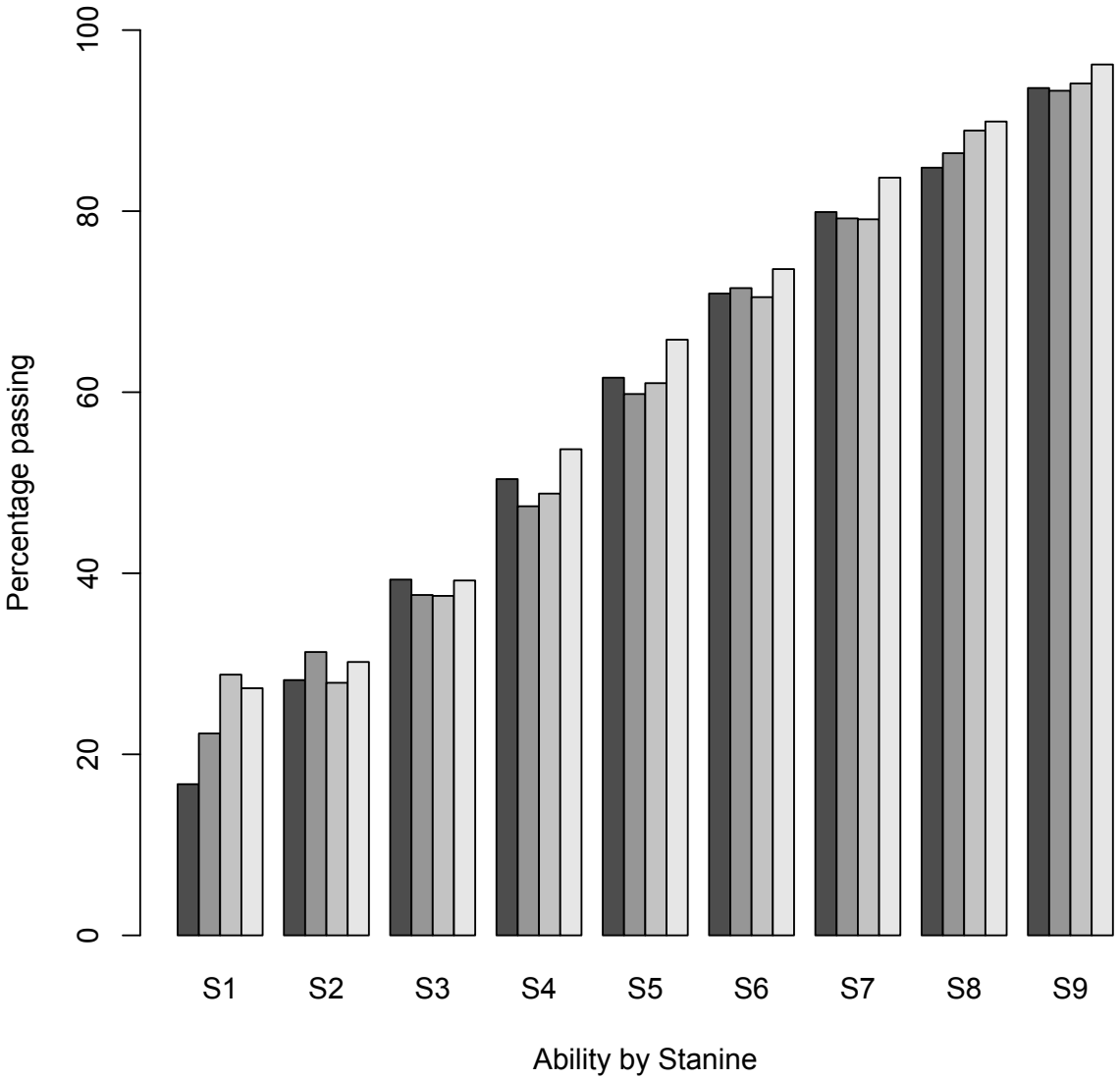


Classics in Prediction and selection

1. Gideon's selection of soldiers
2. OSS and Army Air Corps selection studies
3. **?** (1950) selection of psychology students (**?**)
4. Astronaut selection
5. Peace Corps selection

Gideon's assessment

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							
											Clinical Diagnosis		Individual Therapy		Research		Preference for Hiring	
	Rorschach	TAT	Sentence Compl.	Bender-Gestalt	Credentials	Objective Tests	Autobiographical Materials	Interview	Situations	Other	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 Interviewer					X						13	10	22	25*	14	21	09	25*
					X			X			21	04	22	16	13	20	09	14
Intensive Interviewer					X						16	07	16	14	24*	15	20	20
					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. Terman & Oden (1947, 1959); Oden (1968)
2. Kuncel, Campbell & Ones (1998); Kuncel, Hezlett & Ones (2001); Kuncel & Hezlett (2007) and graduate school prediction
3. Benbow, Lubinski & Stanley (1996); Lubinski & Benbow (2000); Lubinski, Webb, Morelock & Benbow (2001); Lubinski & Benbow (2006); Lubinski (2016)
4. Deary, Whiteman, Starr, Whalley & Fox (2004); Deary & Batty (2007); Deary, Strand, Smith & Fernandes (2007); Deary, Pattie & Starr (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
Time to complete[b,e]								

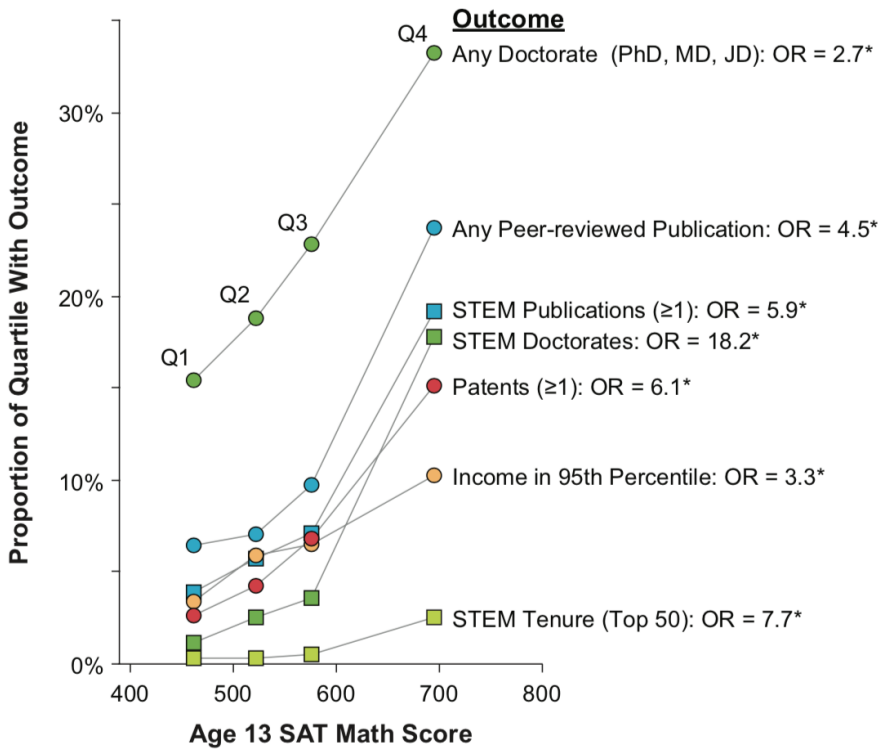
Kuncel et al. meta analysis predicting graduate school performance

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.

Benbow and Lubinski: Beyond the threshold



Deary: the Scottish sample and test retest relability

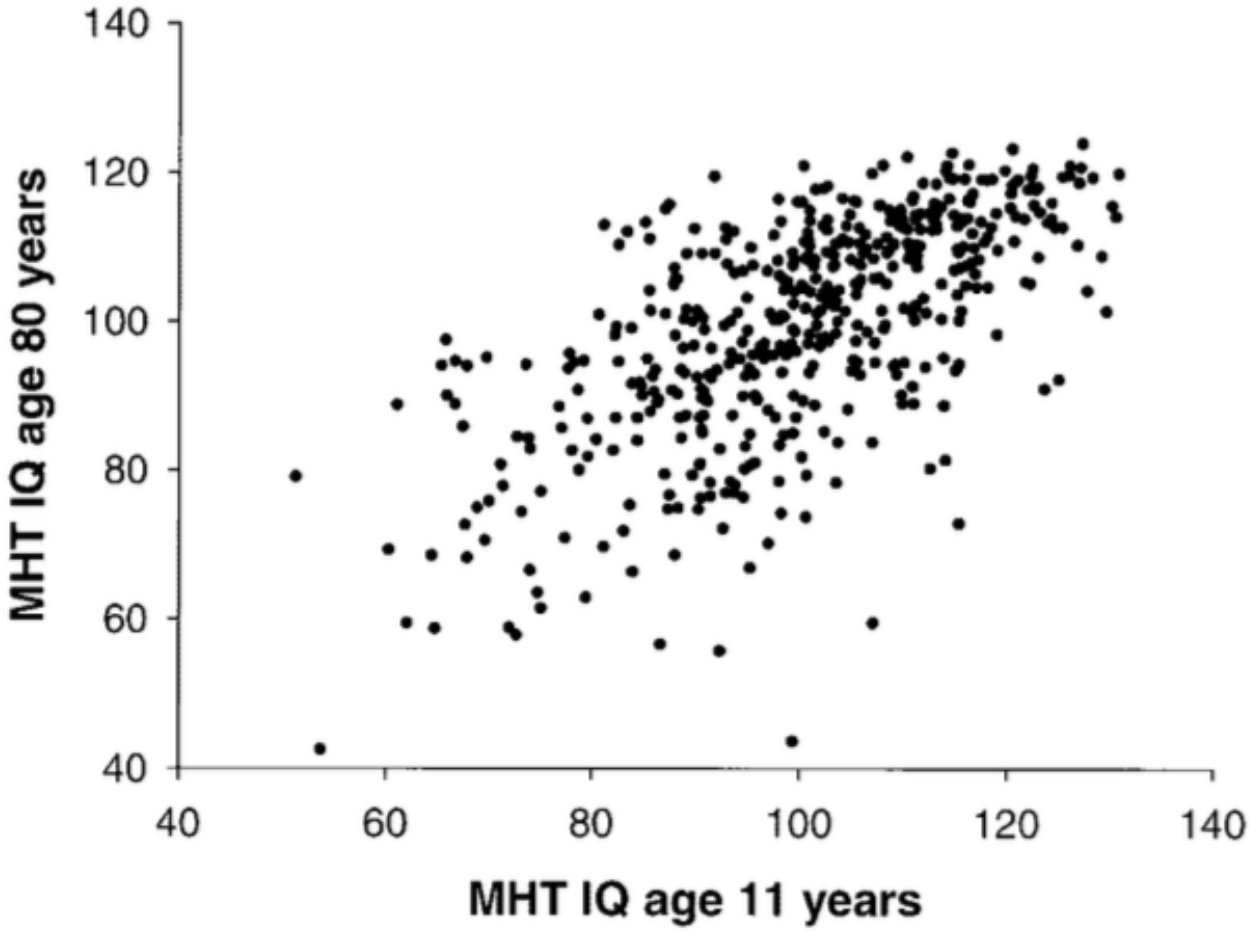
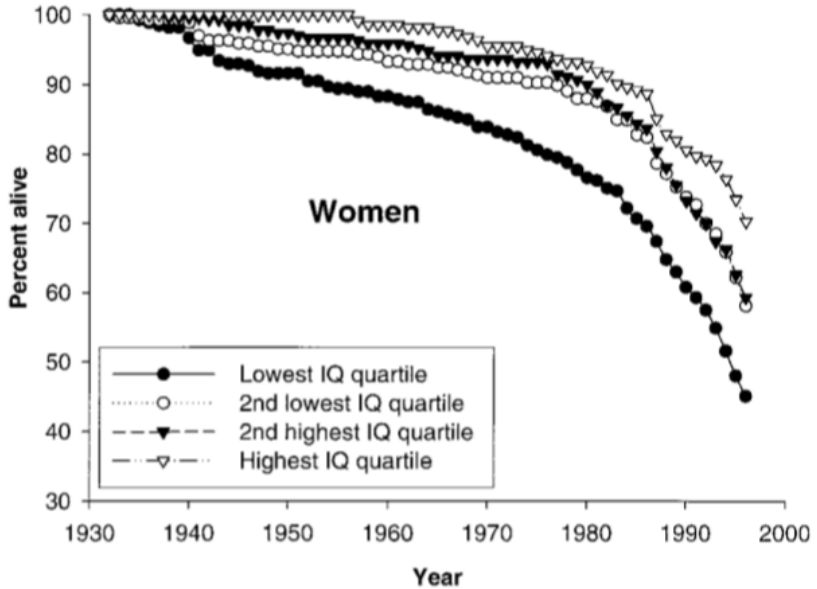
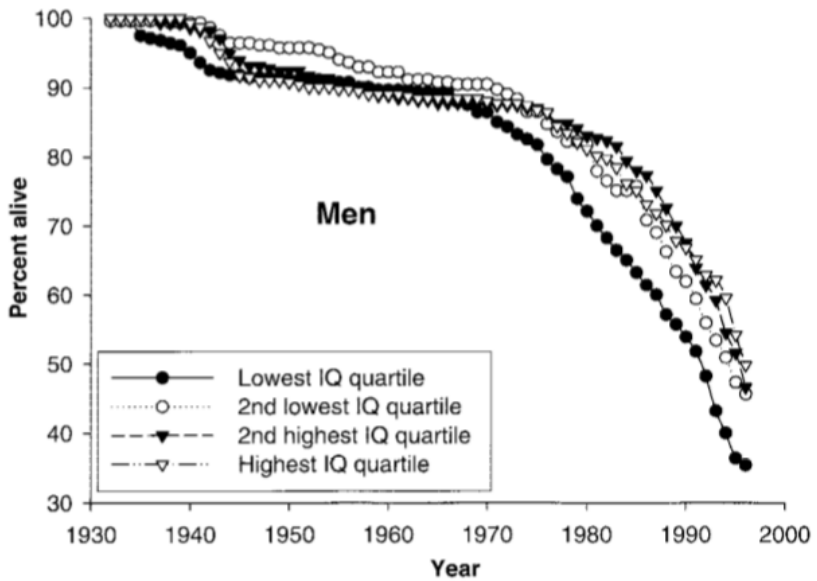


Figure 3. Scattergram of age-corrected Moray House Test (MHT) scores at age 11 and age 80 for participants in the Lothian Birth Cohort 1921 of the Scottish Mental Survey 1932.

Deary: the Scottish sample and mortality





Deary: the Scottish sample and mortality: a model

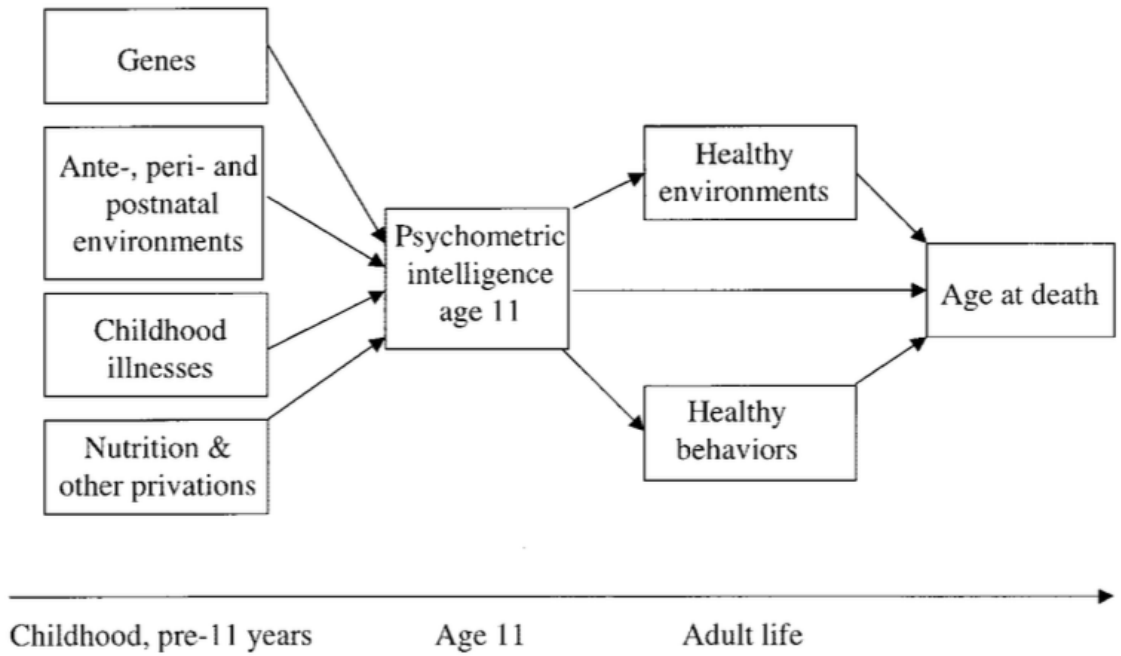


Figure 6. Some possible influences and pathways linking mental ability in childhood and survival. From *Brain and Longevity: Perspectives in Longevity* (p. 162, Figure 3), by C. Finch, J.-M. Robine, & Y. Christen (Eds.), 2003, Berlin: Springer. Copyright 2003 by Springer. Adapted with permission.

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

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

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

Deary, I. J., Pattie, A., & 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., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21.

Deary, I. J., Whiteman, M., Starr, J., Whalley, L., & 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.

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

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

Kuncel, N. R., Hezlett, S. A., & 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.

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

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

Lubinski, D. & 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., & 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.

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

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

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

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