Psychometric Engineering Redux: Making a PROMIS®

David Thissen
L.L. Thurstone Psychometric Laboratory
The University of North Carolina at Chapel Hill

To those wise men of China who,
thousands of years ago,
invented the psychological test.

Dedication of
A History of Psychological Testing
by Philip H. DuBois

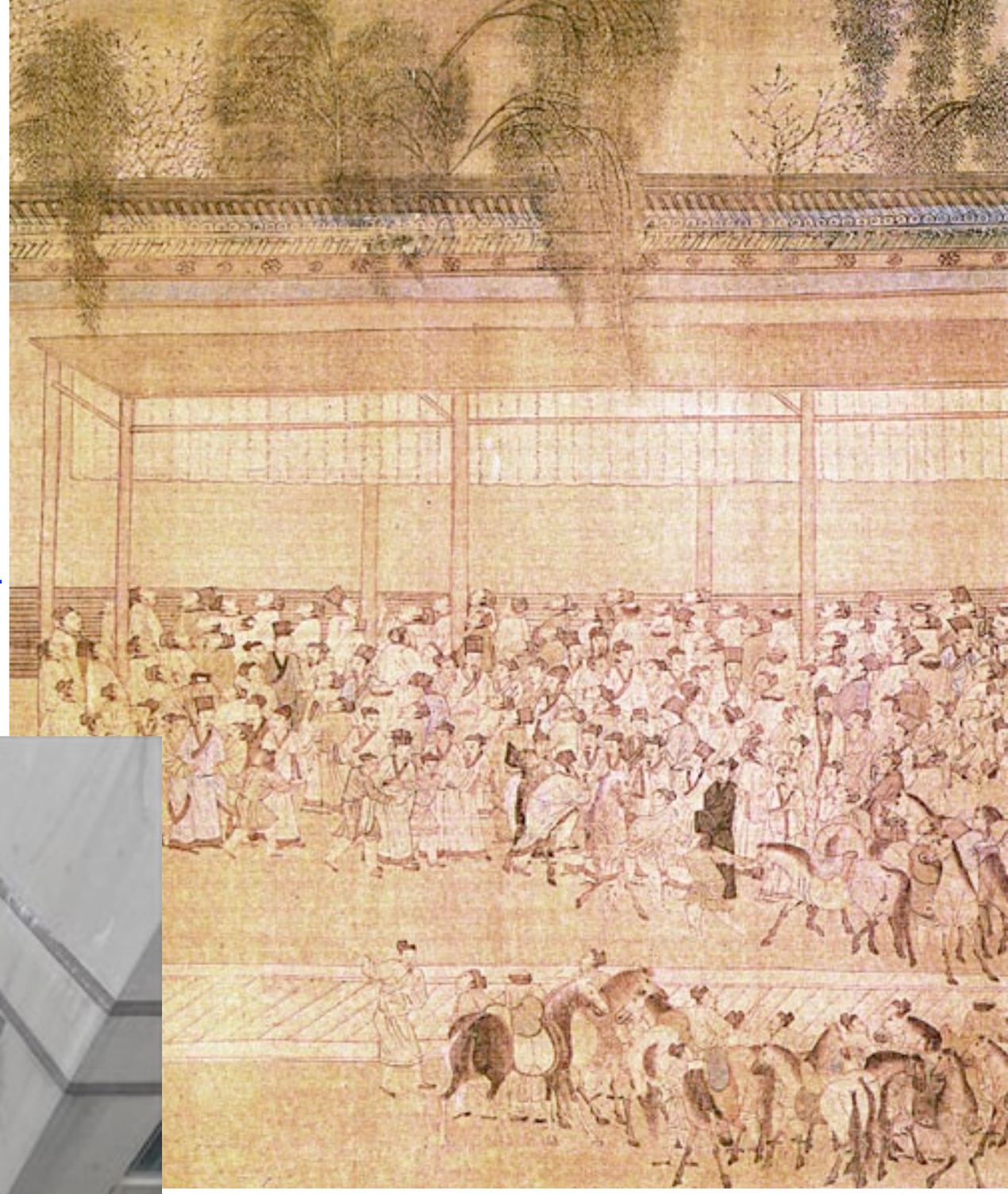
Captioned *Hundreds of Individual Civil Service Examination Rooms at Nanjing, China*, a cropped version of this photograph is the frontispiece for Philip H. DuBois' *A History of Psychological Testing*. This (original) is from "New China and the Printed Page" by Paul Hutchinson, *National Geographic*, June, 1927.

This photo is captioned *Civil Service Examination Halls at Beijing, China, about 1900*, in Philip H. DuBois' *A History of Psychological Testing*. This scanned file is from en.wikipedia.org, which identifies it as "Civil service examination halls. Examination hall with 7500 cells, Guangdong, 1873."

"Handmade oil painting reproduction of Civil Service Exam under Emperor Jen Tsung (fl. 1022) from a history of Chinese emperors, a painting by Anonymous Artist."

www.1st-art-gallery.com

Right: "Scholars waiting for the results of the civil service exam to be posted. Handscroll attributed to Qiu Ling (active 1530-1552) (Natinal Palace Museum, Taiwan)"
www.rightreading.com



Left: Candidates for the civil service examination, November 2009, Wuhan.
en.ce.cn

Army Alpha and Beta
1917-18

MEMOIRS

OF THE

NATIONAL ACADEMY OF SCIENCES

Volume XV

WASHINGTON
GOVERNMENT PRINTING OFFICE
1921

NATIONAL ACADEMY OF SCIENCES.

Volume XV.

PSYCHOLOGICAL EXAMINING IN THE UNITED STATES ARMY

EDITED BY

ROBERT M. YERKES.

GROUP EXAMINATION *a* IN A HOSPITAL WARD, CAMP LEE, OCTOBER, 1917.

SCORING GROUP EXAMINATION *a*, CAMP LEE, OCTOBER, 1917.

The transparent celluloid stencils used in scoring are shown at the near end of the table.

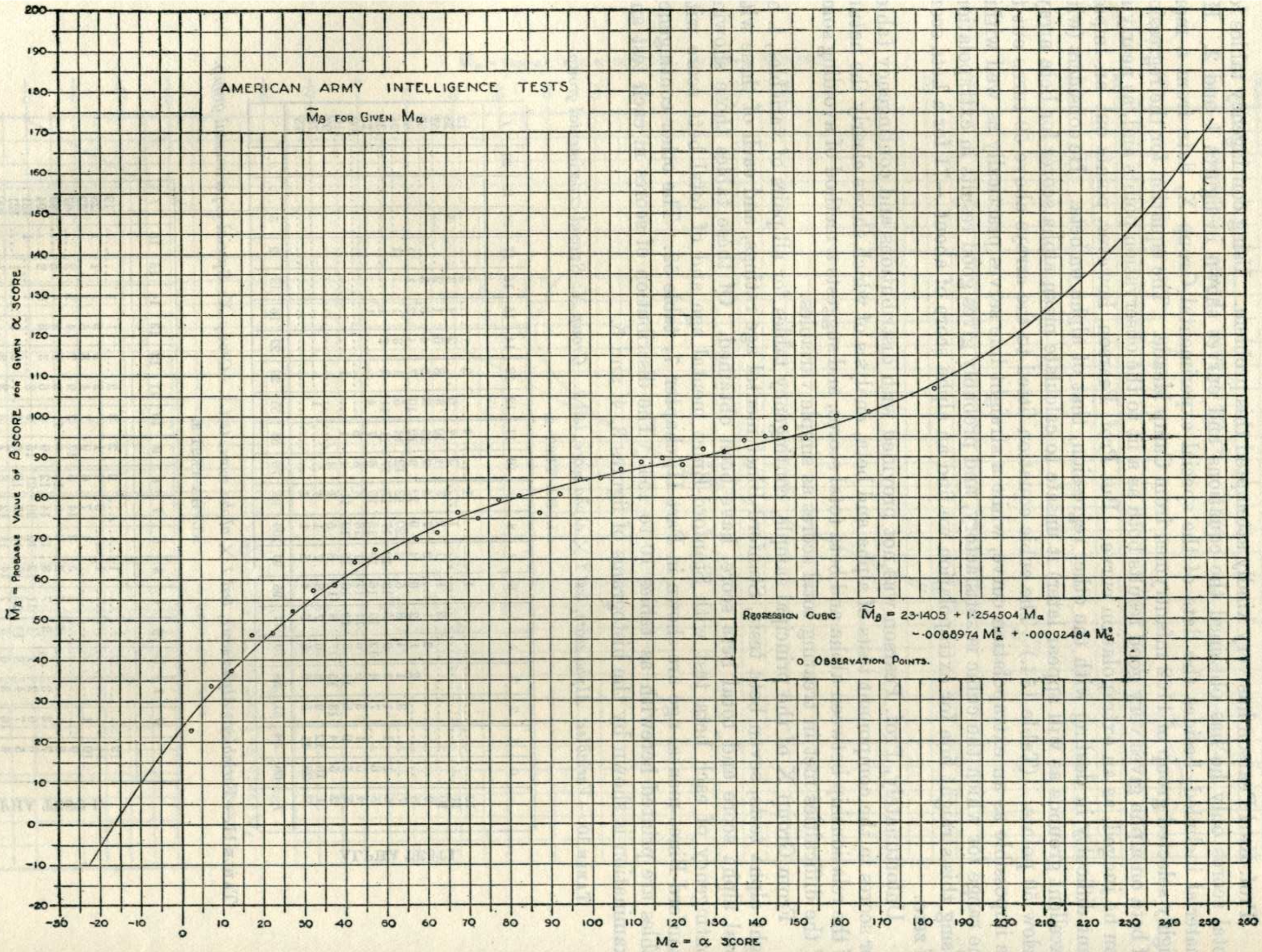
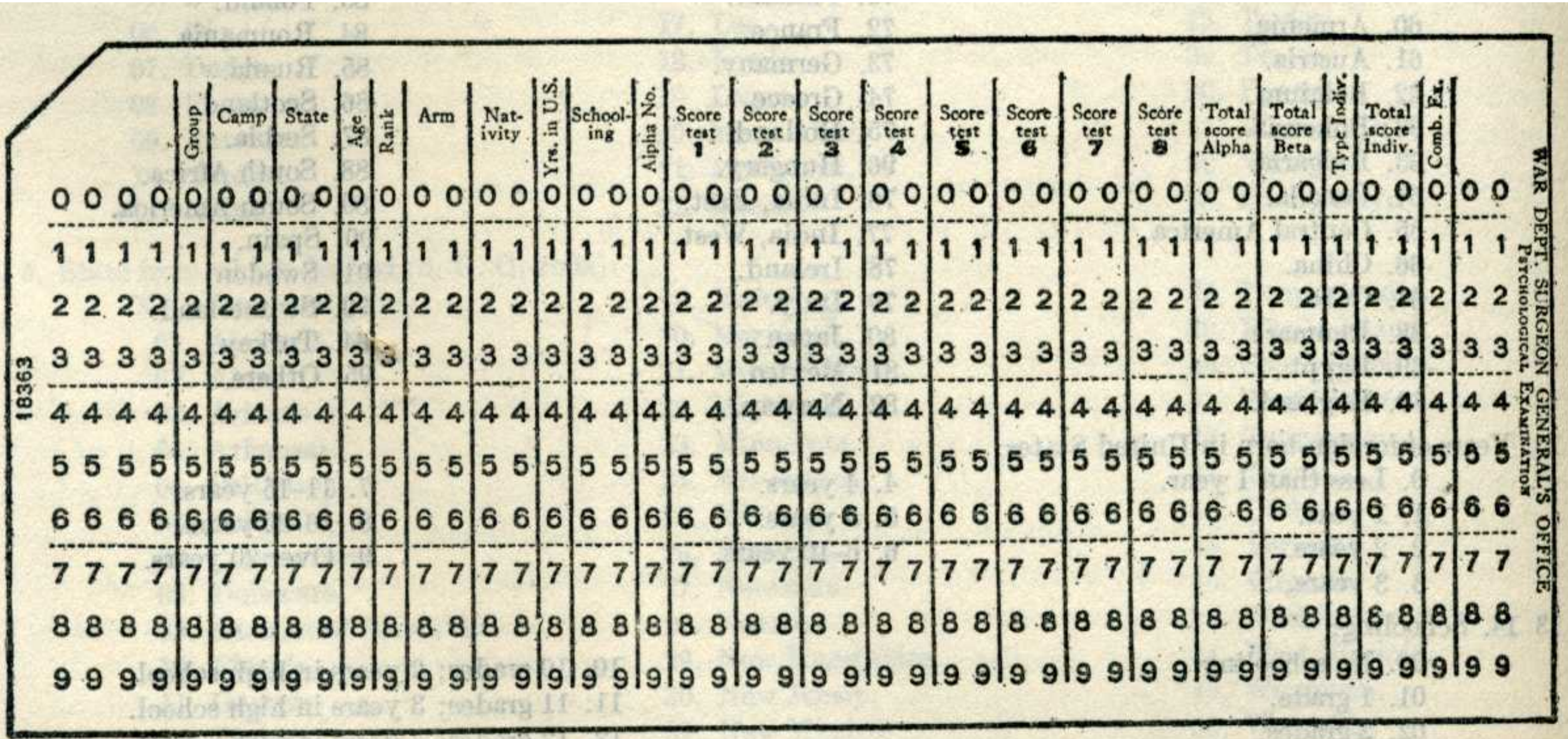


FIG. 2. Regression of examination beta on examination alpha prepared by Prof. Karl Pearson.



Statistical record card used in mechanical sorting by Hollerith system. All information is coded in numerical values and holes punched in the corresponding numbers on the card. A two-digit number is represented by two holes in adjacent columns. The figure is reduced from 7⅜ inches long.

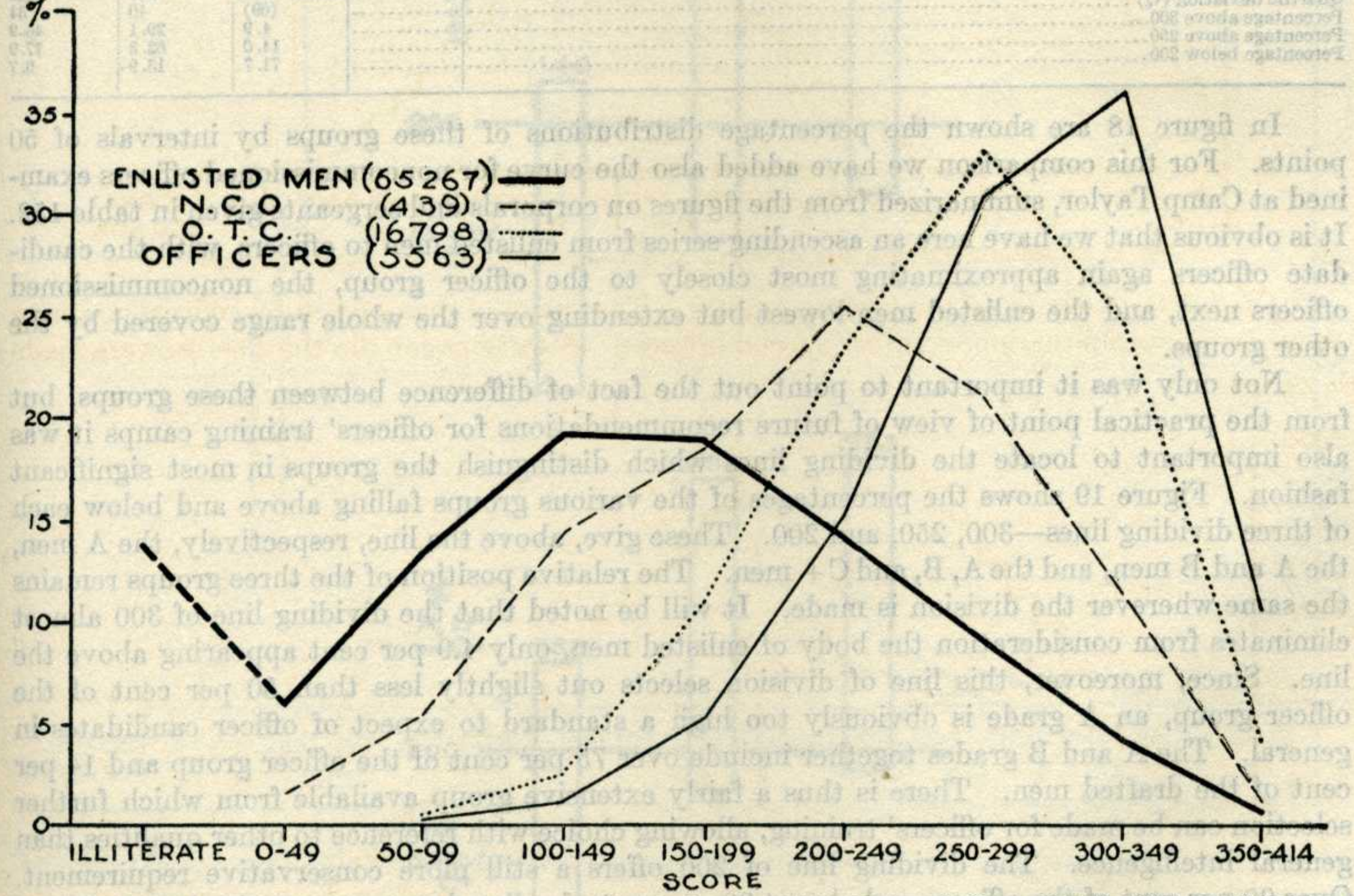


FIG. 18. Percentage distribution of enlisted men, noncommissioned officers, candidate officers, and officers. Note that "illiterates" for whom percentage is given in case of enlisted men do not properly fall on scale of abscissæ.

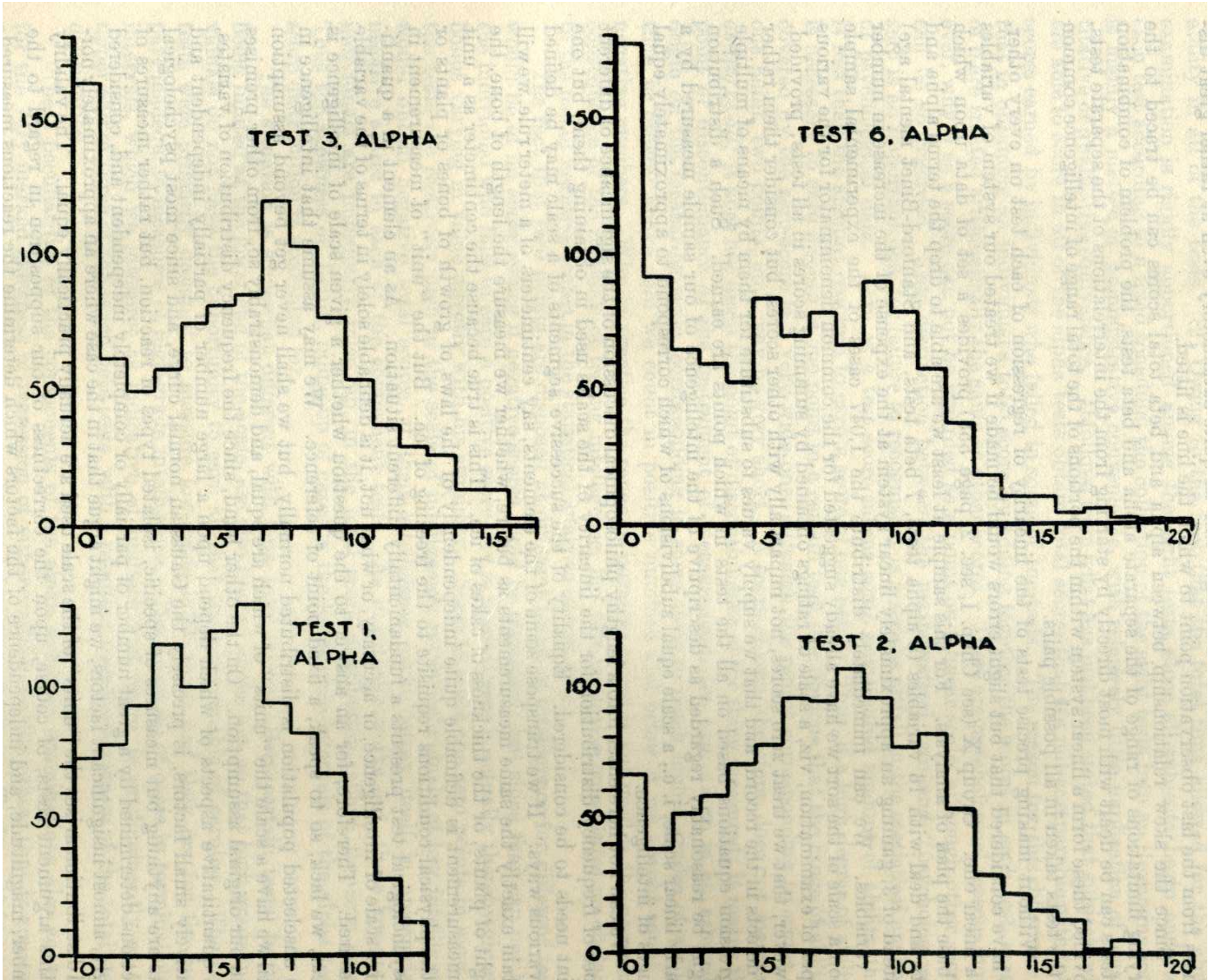


FIG. 3a. Histograms of scores in alpha tests. Special experimental Group X. 1,047 cases. Horizontal scale is score in test; vertical scale is number of cases.

"We may assume that intelligence in the unselected population is distributed normally, but we shall never get beyond assumption until we have a scale the "units" of which are equal, and demonstrably so, from other premises than our original assumption.

"On the other hand, since the frequency distribution of variates the quantitative aspects of which depend upon a large number of partially independent and relatively small factors, is precisely the Gaussian normal curve, and since most psychological tests are anything but measures of specific, isolated types of reaction, but rather measures of reactions determined by a great number of practically or completely independent and, considered singly, almost insignificant factors, we might argue that in the case where an approximately normal distribution arises the "units" of the scale used are actually practically equal."

Psychological Examining in the United States Army, p. 623

TABLE 155.—*Coefficients of correlation for all possible pairs of variables dealt with. For method of calculation see text.*

	Alpha test.								Beta test.							Stanford-Binet mental age.	Alpha total score.	Beta total score.
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7			
Alpha tests:																		
1	1	.7304	.5897	.7093	.6862	.6802	.6697	.6584	.4213	.5764	.5920	.6286	.5600	.5428	.5536	.6810	.7729	.6900
2	.7304	1	.7455	.7915	.7634	.7730	.7357	.7425	.4437	.5979	.6591	.6823	.6425	.5673	.5228	.7788	.8445	.7236
3	.5897	.7455	1	.8055	.7543	.6129	.6714	.7750	.3641	.4477	.5269	.6795	.5982	.5451	.5827	.6932	.8609	.6957
4	.7093	.7915	.8055	1	.8343	.6814	.7304	.8608	.4274	.4948	.5590	.6422	.5455	.5595	.5412	.8240	.9373	.6831
5	.6862	.7634	.7543	.8343	1	.6736	.7784	.8227	.4091	.5366	.5544	.6388	.5859	.5589	.4972	.7219	.8974	.6919
6	.6802	.7730	.6129	.6814	.6736	1	.7036	.6930	.4294	.5978	.6046	.6003	.6055	.5273	.6039	.7109	.8405	.7369
7	.6697	.7357	.6714	.7304	.7784	.7036	1	.6718	.4231	.4024	.5589	.6276	.5569	.4890	.5491	.7261	.8769	.6814
8	.6584	.7425	.7750	.8608	.8227	.6930	.6718	1	.3670	.4585	.5761	.6663	.5804	.6076	.5379	.7158	.9289	.7001
Beta tests:																		
1	.4213	.4437	.3641	.4274	.4091	.4294	.3231	.3670	1	.4844	.4957	.4795	.4082	.5011	.5050	.4650	.4460	.6190
2	.5764	.5979	.4477	.4948	.5366	.5978	.4024	.4585	.4844	1	.6272	.5493	.5187	.5702	.5849	.6108	.4373	.7889
3	.5920	6591	.5269	.5590	.5544	.6046	.5589	.5761	.4957	.6272	1	.6796	.6465	.5878	.5764	.6140	.6176	.8215
4	.6286	.6823	.6795	.6422	.6388	.6003	.6276	.6663	.4795	.5493	.6796	1	.7520	.6438	.5634	.6571	.6418	.8876
5	.5600	.6425	.5982	.5455	.5859	.6055	.5568	.5804	.4082	.5187	.6465	.7520	1	.5629	.5523	.6080	.7027	.8377
6	.5428	.5673	.5451	.5595	.5589	.5273	.4890	.6076	.5011	.5702	.5878	.6438	.5629	1	.5694	.5860	.5135	.7808
7	.5536	.5228	.5827	.5412	.4972	.6039	.5491	.5379	.5050	.5849	.5764	.5634	.5523	.5694	1	.6026	.5872	.7313
Stanford-Binet mental age	.6810	.7788	.6932	.8240	.7219	.7109	.7261	.7158	.4650	.6108	.6140	.6571	.6080	.5860	.6026	1	.8111	.7270
Alpha total score	.7729	.8445	.8609	.9373	.8974	.8405	.8769	.9289	.4460	.4373	.6176	.6418	.7027	.5135	.5872	.8111	1	X
Beta total score	.6900	.7236	.6957	.6831	.6919	.7369	.6814	.7001	.6190	.7889	.8215	.8876	.8377	.7808	.7313	.7270	X	1

Correlations computed with
corrections for restriction of range

WAIS Digit-Symbol

7. DIGIT SYMBOL

1	2	3	4	5	6	7	8	9
–	⊥	⊐	L	∪	o	^	X	=

SCORE

SAMPLES

2	1	3	7	2	4	8	1	5	4	2	1	3	2	1	4	2	3	5	2	3	1	4	6	3

1	5	4	2	7	6	3	5	7	2	8	5	4	6	3	7	2	8	1	9	5	8	4	7	3

6	2	5	1	9	2	8	3	7	4	6	5	9	4	8	3	7	2	6	1	5	4	6	3	7

9	2	8	1	7	9	4	6	8	5	9	7	1	8	5	2	9	4	8	6	3	7	9	8	6

Army Beta Digit-Symbol

1	2	3	4	5	6	7	8	9
–	И	⊐	L	∪	o	^	X	=

1

3	1	2	1	3	2	1	4	2	3	5	2	9	1	4

2

6	3	1	5	4	2	7	6	3	8	7	2	9	5	4

3

6	3	7	2	8	1	9	5	8	4	7	3	6	9	5

4

1	9	2	8	3	7	4	6	5	9	4	8	5	7	6

5

9	3	8	6	4	1	5	7	2	6	2	4	8	1	3

6

4	9	5	1	7	5	2	6	9	3	7	8	4	1	8

Plate XII. Beta Test 4: Digit-Symbol.

Woodworth's Personal Data Sheet

1919

1. Do you usually feel well and strong? YES NO
2. Do you usually sleep well? YES NO
3. Are you often frightened in the middle of the night? YES NO
4. Are you troubled with dreams about your work? YES NO
5. Do you have nightmares? YES NO
6. Do you have too many sexual dreams? YES NO
7. Do you ever walk in your sleep? YES NO
8. Do you have the sensation of falling when going to sleep? YES NO
9. Does your heart ever thump in your ears so you cannot sleep? YES NO
10. Do ideas run through your head so you cannot sleep? YES NO

… (106 more questions)

The entire questionnaire is reproduced in
A History of Psychological Testing, Philip H. DuBois (1970), Allyn and Bacon.

Examination of Emotional Fitness for Warfare.
R.S. Woodworth

In the hope of providing a means of quickly sifting out from the draft and holding for individual examination at the hands of the neuropsychiatrist, those of neurotic tendencies, a questionnaire was made up from symptoms believe to indicate such tendencies. When a given symptom was reported by twenty-five percent or more of an unselected group, it was eliminated as not being sufficiently diagnostic. … where the average college student reports about ten out of the hundred symptoms inquired about, the average neurasthenic or hysteric recognized at Camp Upton scored over forty. …

Pscyhological Bulletin, February, 1919

Woodworth's research on the Personal Data Sheet was described at the Annual Meeting of the American Psychological Association in Baltimore in December of 1918.

Presentations at that meeting were almost all about "war problems."

A new Ph.D. named L.L. Thurstone presented a paper at that meeting on "The selection and training of telegraphers."

Pscyhological Bulletin, February, 1919

Aviation Psychology Program

1940s

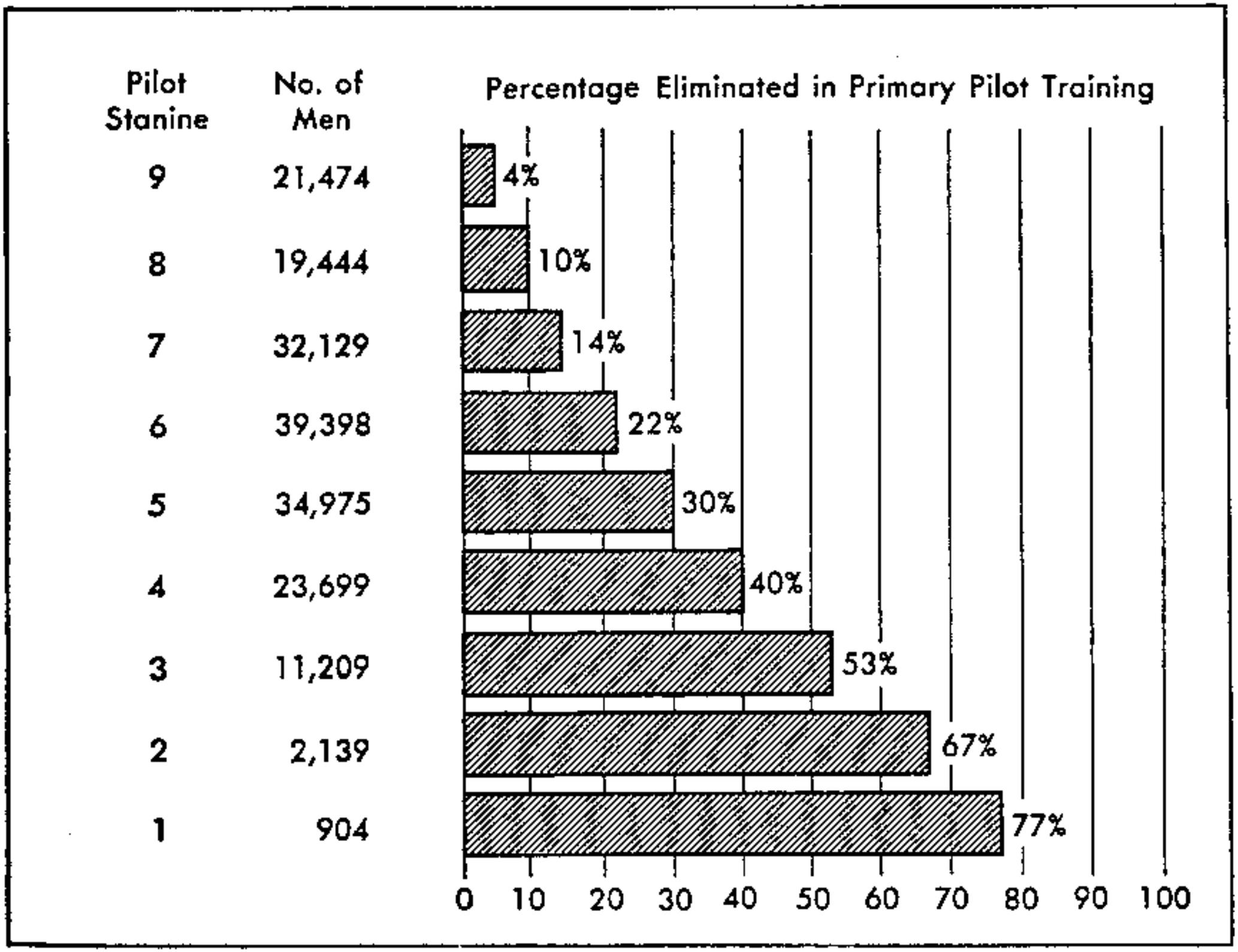


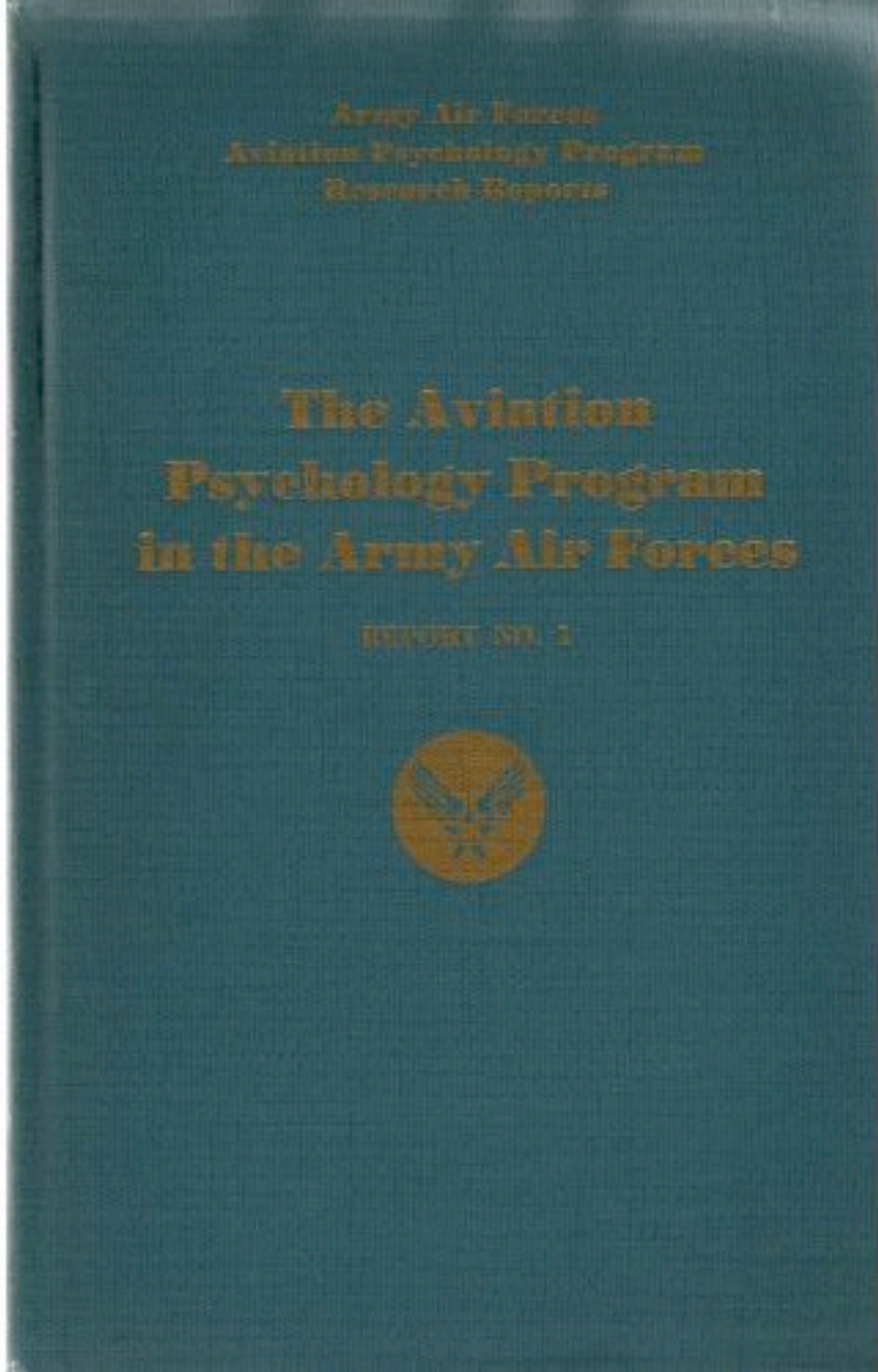
Figure 4-7. Expectancy Chart Showing Relation between Performance on Pilot Selection Battery and Elimination from Primary Flight Training.

For example,

Thorndike, R. L. (Ed.). (1947). *Research Problems and Techniques* (Report No. 3). Washington, DC: US Printing Office.

became

Thorndike, R. L. (1949). *Personnel Selection: Test and Measurement Techniques*. New York: Wiley.



The Assessment of Men

1940s

The story of around 5000 military and civilian personnel tested during the second half of World War II to become agents of the Office of Strategic Services—the OSS—predecessor of the (postwar) CIA.

Spies.

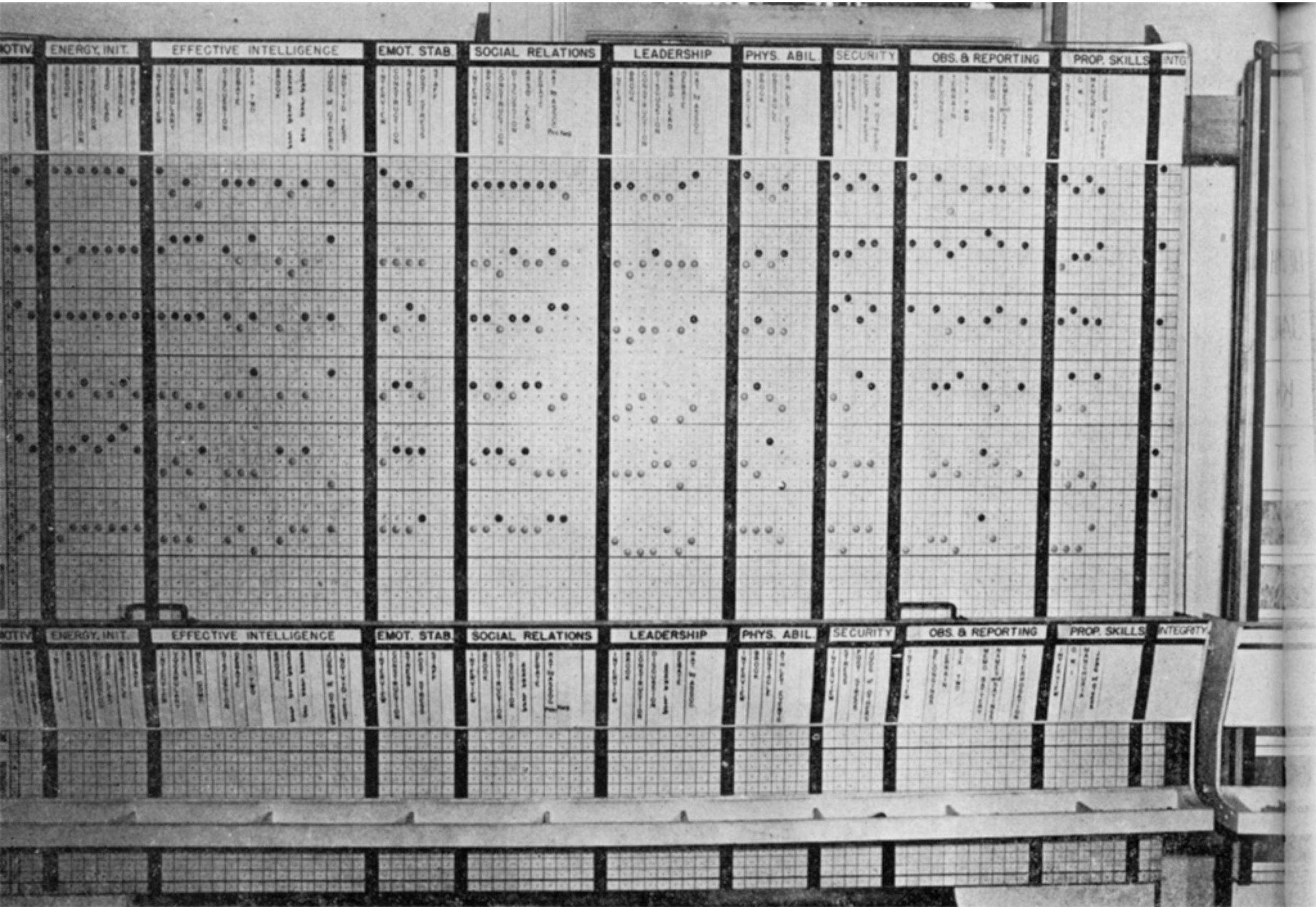
Designed in a meeting at "The Farm" in N. Virginia (see *The Recruit*), carried out largely at Station S (and Station W and WS and a couple others).

Station
S

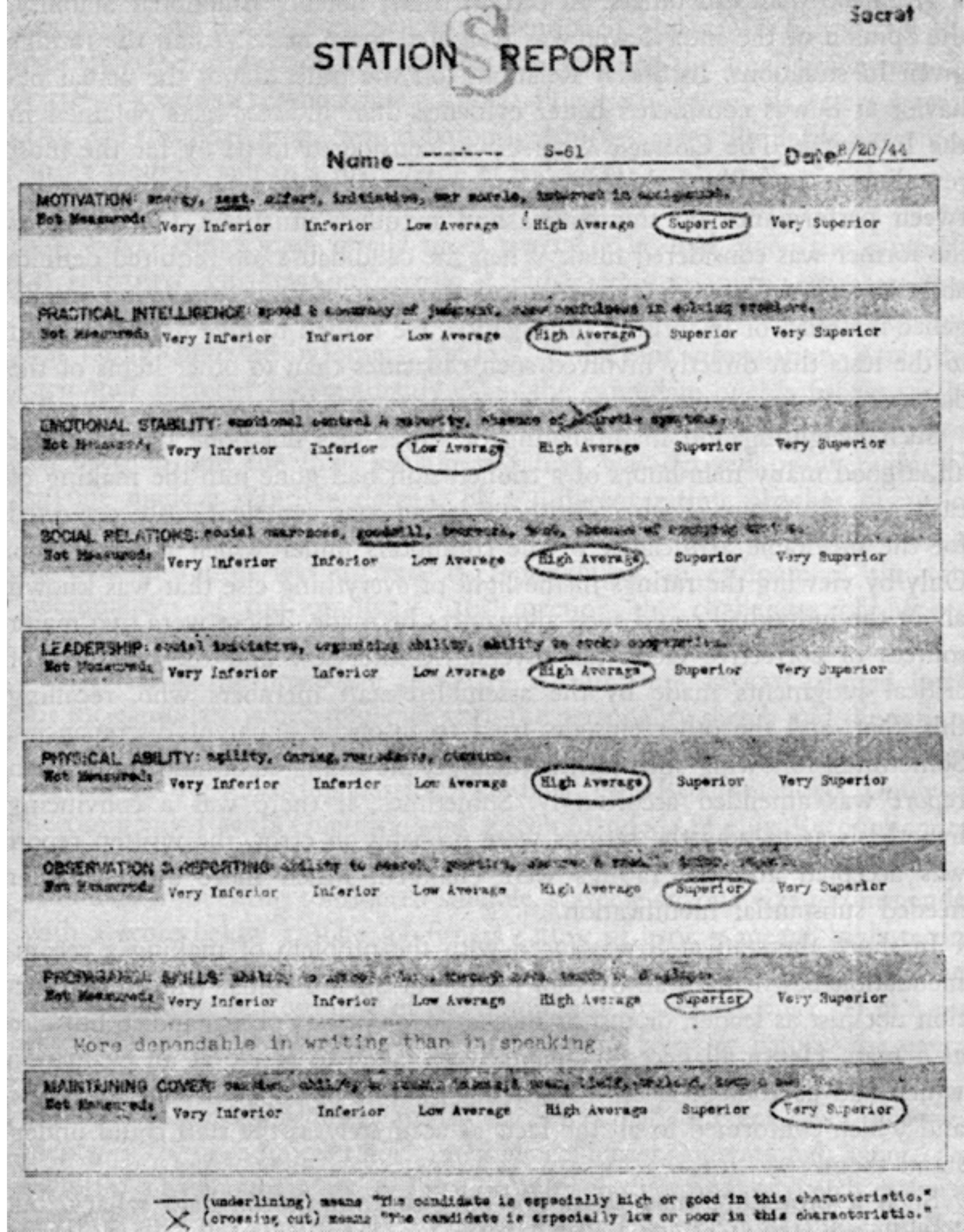
The Construction Situation

Staff Conference at S

The Rating Board



An example of a report from Station S



Secret

STATION S REPORT

Name ------- S-61 Date 8/20/44

MOTIVATION: energy, zest, effort, initiative, war morale, interest in assignment.
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

PRACTICAL INTELLIGENCE: speed & accuracy of judgment, resourcefulness in solving problems.
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

EMOTIONAL STABILITY: emotional control & maturity, absence of neurotic symptoms.
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

SOCIAL RELATIONS: social awareness, goodwill, teamwork, tact, absence of annoying traits.
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

LEADERSHIP: social initiative, organizing ability, ability to evoke cooperation.
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

PHYSICAL ABILITY: agility, daring, ruggedness, stamina.
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

OBSERVATION & REPORTING: ability to search, [illegible]
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

PROPAGANDA SKILLS: [illegible]
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

More dependable in writing than in speaking

MAINTAINING COVER: [illegible]
Not Measured: Very Inferior | Inferior | Low Average | High Average | Superior | Very Superior

——— (underlining) means "The candidate is especially high or good in this characteristic."
X (crossing out) means "The candidate is especially low or poor in this characteristic."

The American Soldier studies 1940s

Volume IV of *Studies in Social Psychology in World War II*,

Measurement and Prediction

by

Samual A. Stouffer

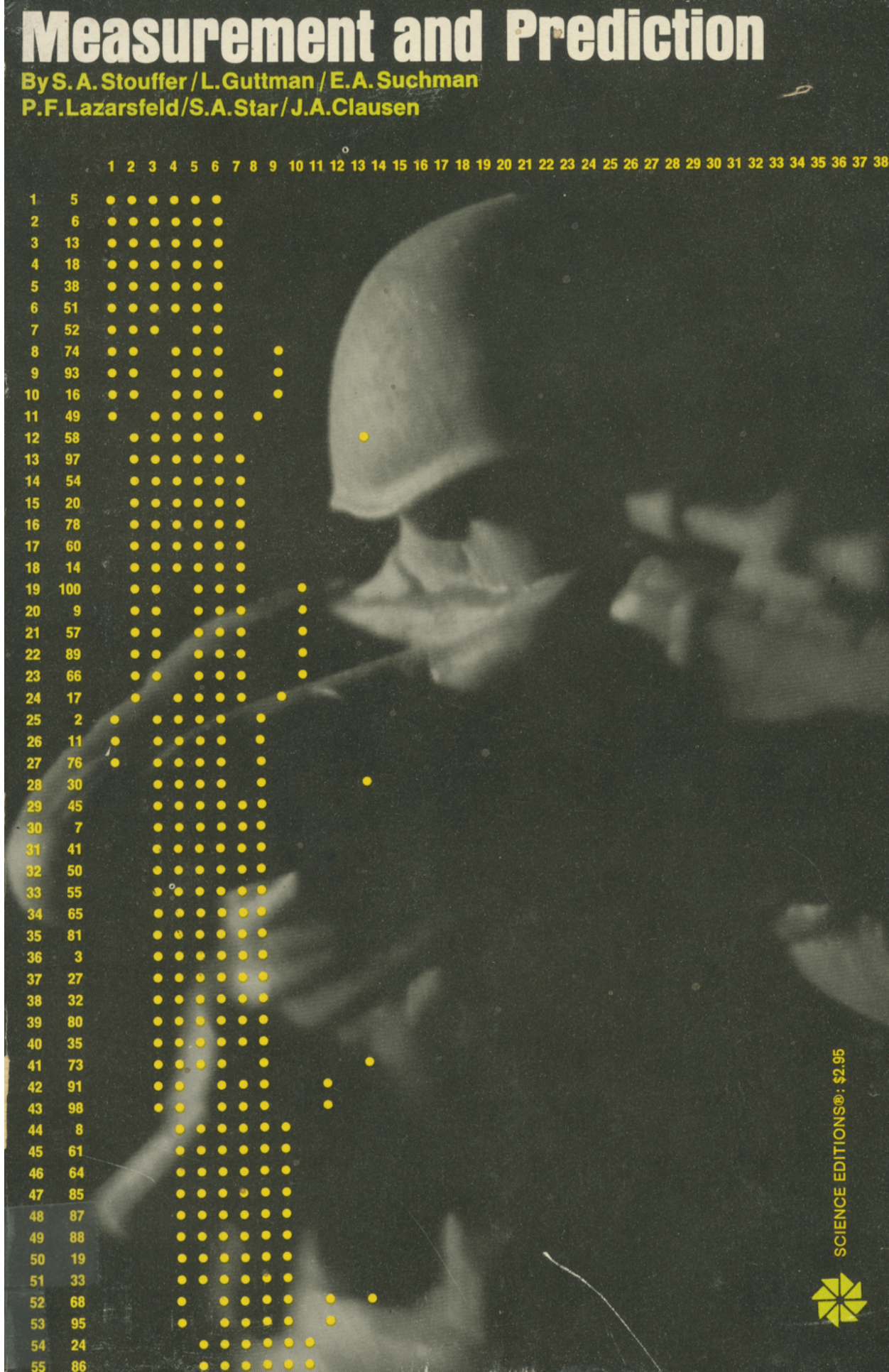
Louis Guttman

Edward A. Suchman

Paul F. Lazarsfeld

Shirley A. Star

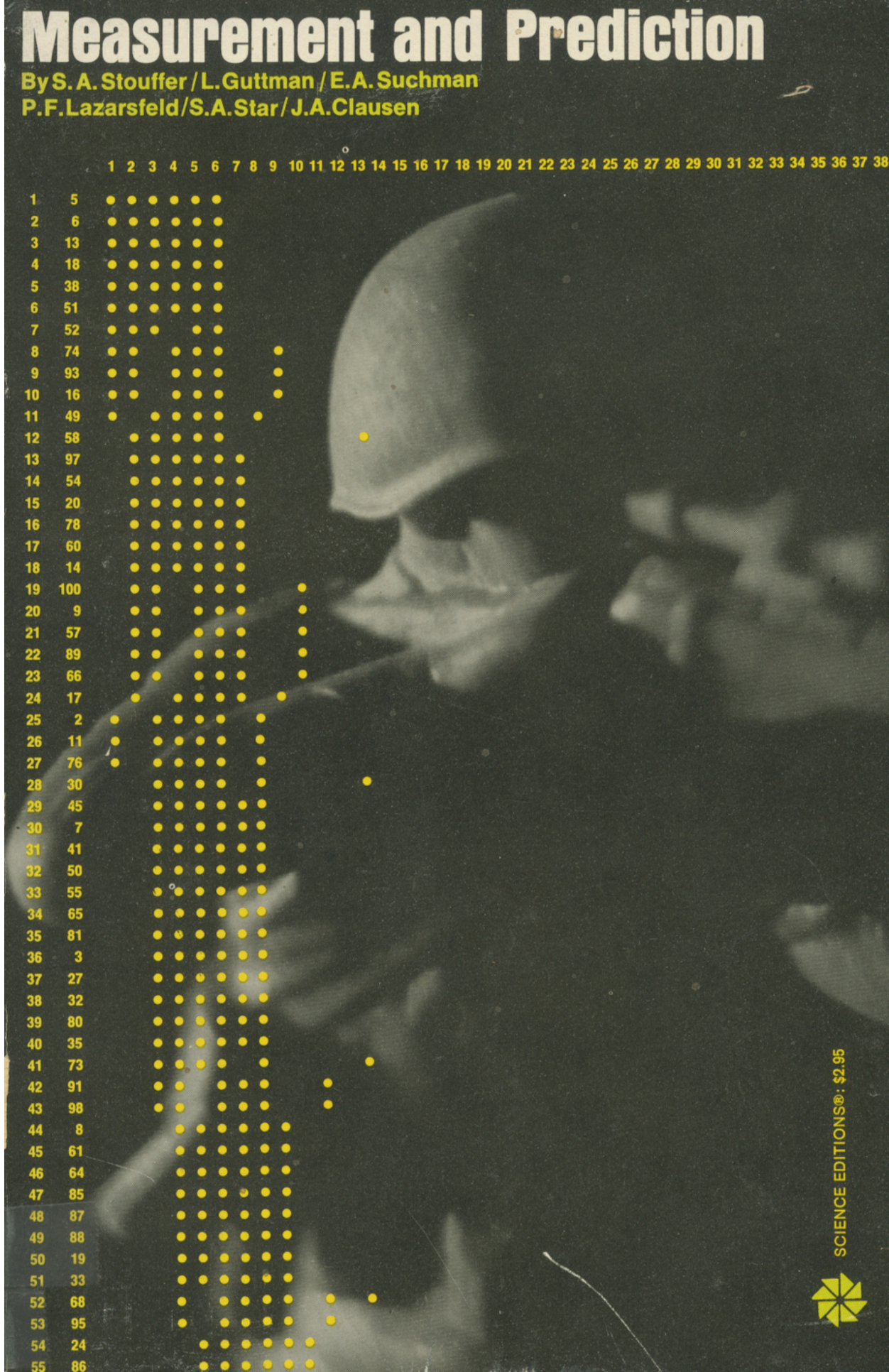
John A. Clausen



“Guttman Scales”

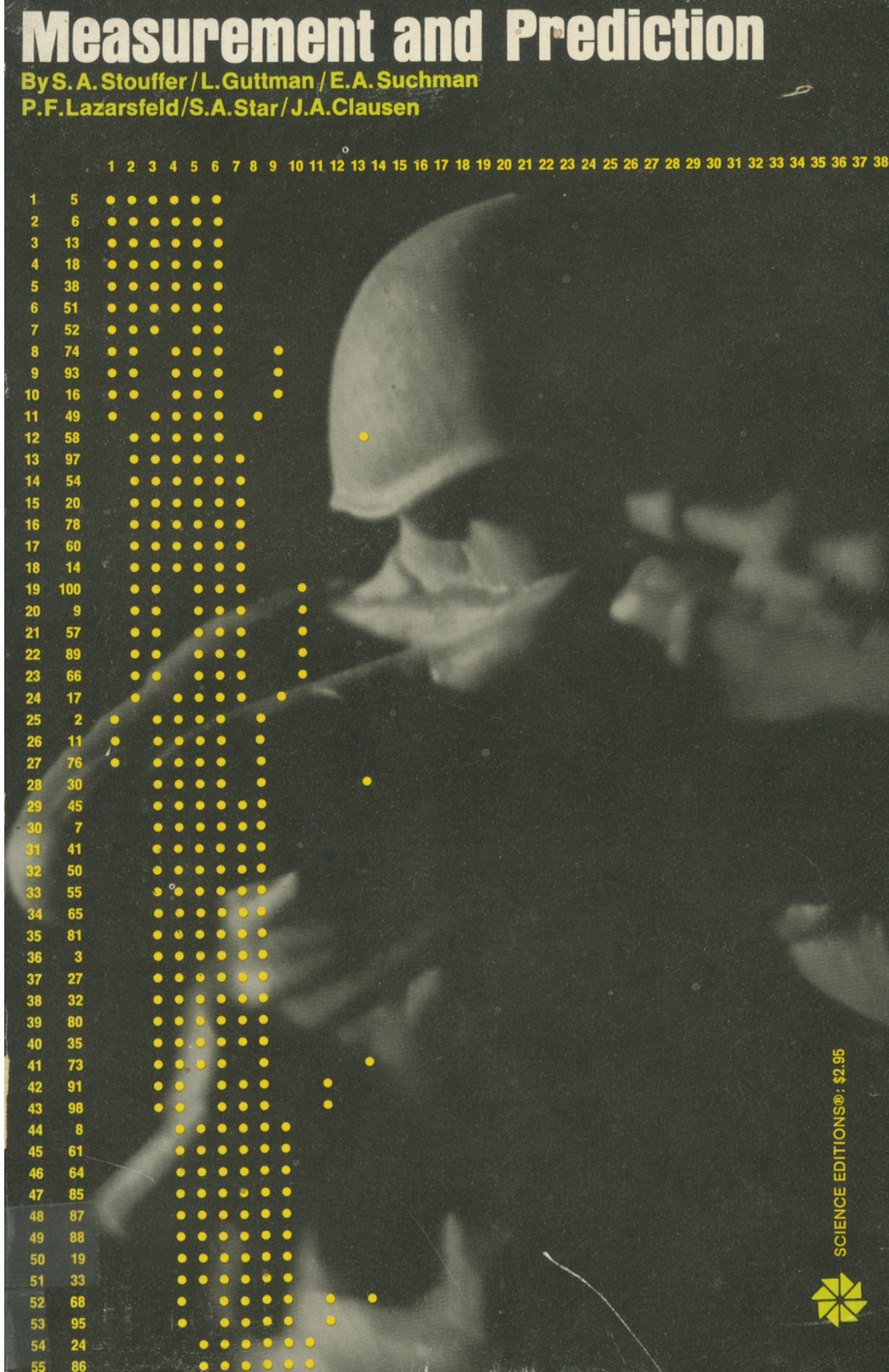
Louis Guttman’s

Ch. 2 *The Problem of Attitude and Opinion Measurement,*
Ch. 3 *The Basis for Scalogram Analysis,*
Ch. 6 *The Relation of Scalogram Analysis to other Techniques,*
Ch. 8 *Problems of Reliability,*
Ch. 9 *The Principal Components of Scale Analysis*



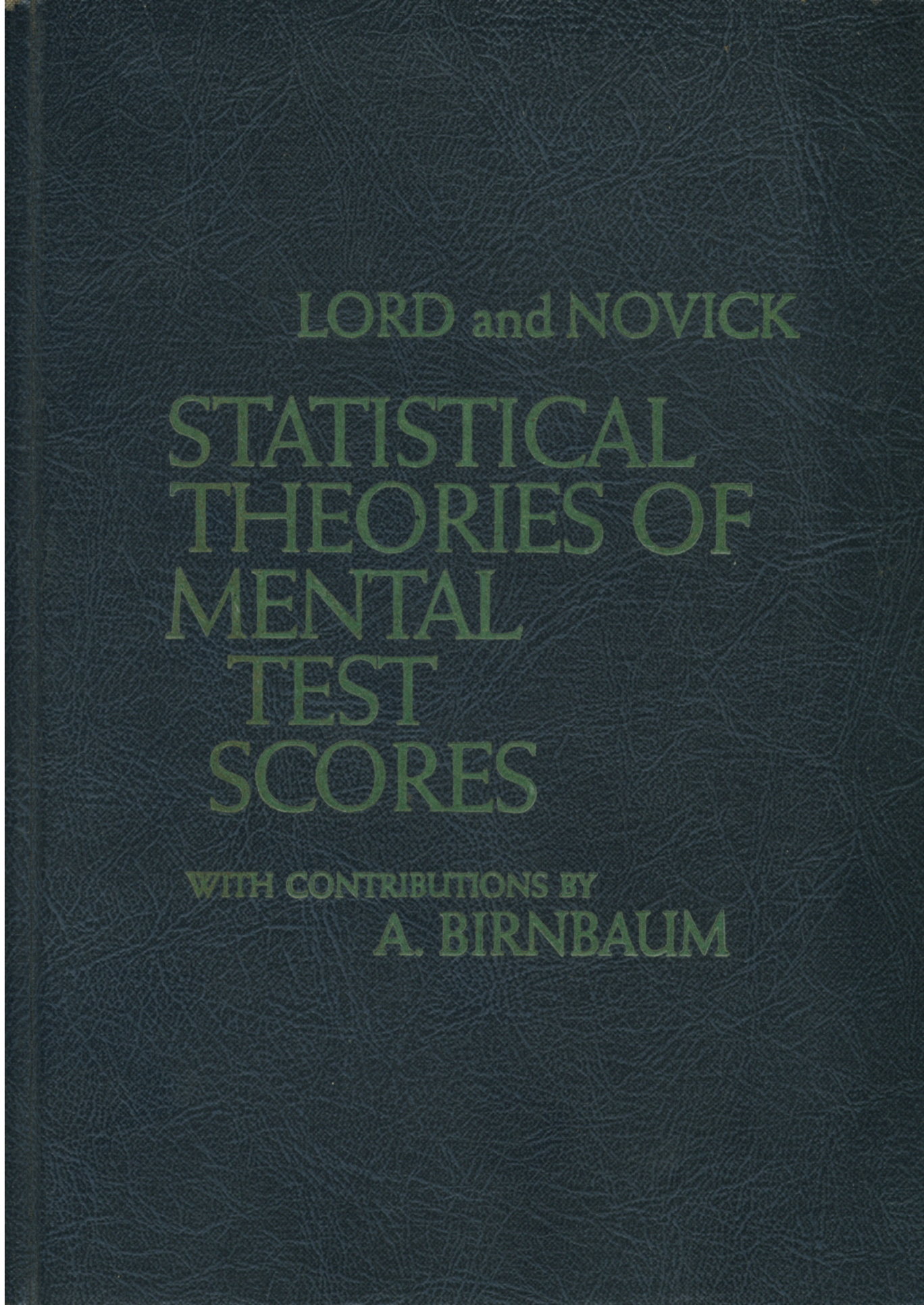
The Statistics of Latent Variable Models

Paul F. Lazarsfeld's
Ch. 10 *The Logical and Mathematical Foundation of Latent Structure Analysis,*
Ch. 11 *The Interpretation and Computation of some Latent Structures*



U.S. Office of Naval Research
&
Air Force Office of Scientific Research
1960s-80s

Lord, F.M. & Novick, M.R. (1968)
Statistical Theories of Mental Test Scores.
With contributions by
A. Birnbaum.
Reading, MA: Addison-Wesley.



Some examples of basic research from the ONR-AFOSR programs:

Frederic Lord—*Tailored Testing and aspects of IRT.*

David Weiss (& students and colleagues)—
Computerized adaptive testing (the name, as well as countless systems).

Mark Reckase (& colleagues at ACT)—*CAT research and many of the first forays into MIRT.*

Some examples of basic research from the ONR-AFOSR programs:

Fumiko Samejima—*Research on many item response models.*

R. Darrell Bock, Robert Gibbons, & Eiji Muraki—*Full-information item factor analysis.*

Robert Gibbons & Donald Hedeker—*Full-information item bi-factor analysis.*
[more on this later]

The "New Design" of the National Assessment of Educational Progress (NAEP) 1980s-90s

Some examples of basic research associated with NAEP:

Robert Mislevy—*Estimating the parameters of latent population distributions with IRT.*

Robert Mislevy—*Randomization-based inference with IRT in complex samples (plausible values)* [also ONR]

Robert Mislevy, Kathleen Sheehan, Neal Thomas—*M-GROUP–the beginning of multilevel IRT.*

Some examples of basic research associated with NAEP:

John Donoghue, Matthew Shultz, Edward Ip—
"Domain" or "Expected" or "Marginal" trace lines for MIRT models.

[more on this later]

Patient Reported Outcomes Measurement Information System (PROMIS®) 2004-2013

The PROMIS Network

The PROMIS initiative established a collaborative relationship between NIH and individual research teams. The broad objectives of the PROMIS Network were to:

- Develop and test a large bank of items measuring patient-reported outcomes

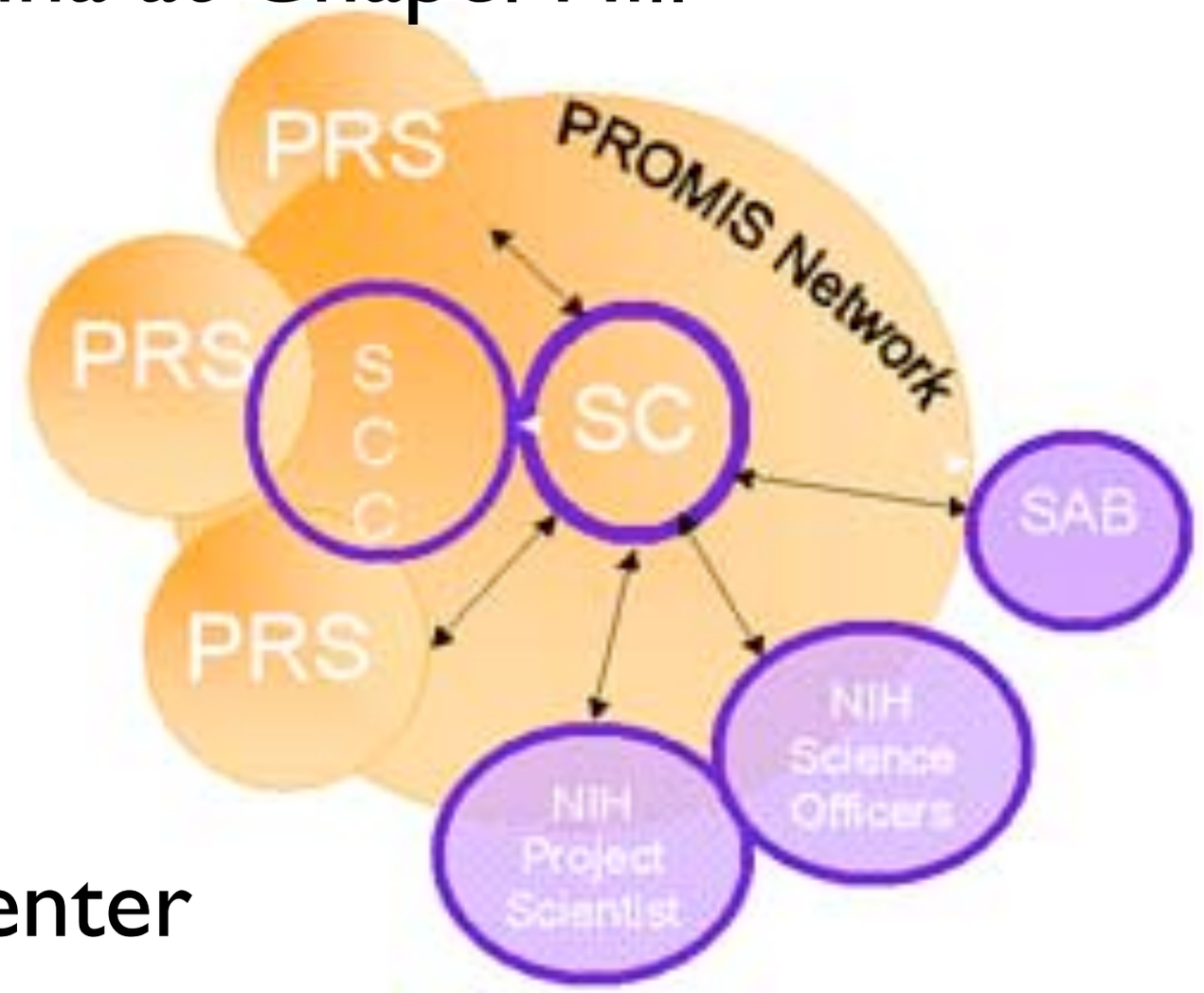
- Create a computerized adaptive testing system that permits efficient, psychometrically robust assessment of patient-reported outcomes in clinical trial research

- Create a publicly available system that allows clinical researchers to access a common repository of items and computerized adaptive tests

The PROMIS Network

Primary Research Sites:

- University of North Carolina at Chapel Hill
- University of Washington
- University of Pittsburgh
- Stony Brook University
- Stanford University
- Duke University



Other Units:

- Statistical Coordinating Center
- Advisory Board
- NIH Scientific Staff

Seven years ago we thought the only way was to make unidimensional scales …

Reeve, B.B., Hays, R.D, Bjorner, J.B., *et al*. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*, *45*, S22-31.

Even in that context, Seung Choi developed a new CAT stopping rule ...

Choi, S.W., Grady, M.W., & Dodd, B.G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71*, 37–53.

"The predicted standard error reduction (PSER) stopping rule uses the predictive posterior variance to determine the reduction in standard error that would result from the administration of additional items."

A cousin of PROMIS has been NIH SBIR funding for the development of new IRT software—the first entirely new since the 1980s?

IRTPRO

SSI—Scientific Software International, Inc.

EQSIRT

Multivariate Software, Inc.

These efforts have made possible implementation of new ideas …

Investigating dimensionality in the service of the development of unidimensional scales at first made use of bifactor models (Gibbons & Hedeker, 1992), but has been part of the inspiration of further development …

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

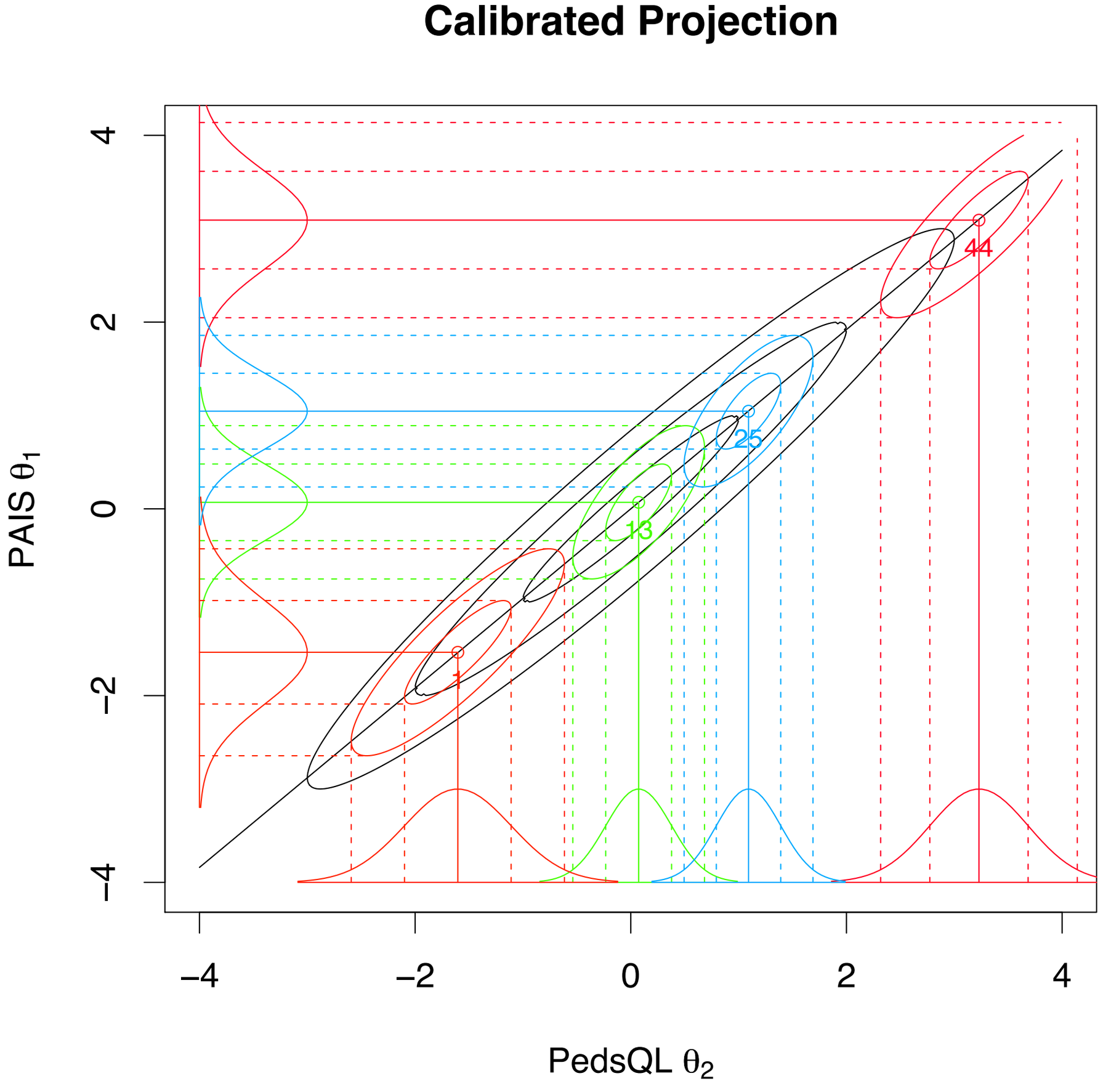
Cai, L, Yang, Ji Seung, & Hansen, M. (in press). Generalized full-information item bifactor analysis. *Psychological Methods*.

Among (many) other things, the two-tier development led to the possibility of *calibrated projection*, IRT linking between scales that are not jointly unidimensional …

Thissen, D., Varni, J.W., Stucky, B.D., Liu, Y., Irwin, D.E., & DeWalt, D.A. (in press). Using the PedsQL™ 3.0 Asthma Module to Obtain Scores Comparable with those of the PROMIS Pediatric Asthma Impact Scale (PAIS). *Quality of Life Research*.

Some interesting features of calibrated projection are that the EAPs are linearly related; indeed, $EAP[\theta_1] = r\ EAP[\theta_2]$ (in standard units).

The relation of the conditional variance $SD^2[\theta_1]$ with $SD^2[\theta_2]$ is more complex, because the posteriors change correlation / shape across the score scale.



IMPS 2011

At this conference, this becomes (a small) part of the …

Invited Symposium: *Special Models with Special Solutions: Statistical Issues in Hierarchical Item Factor Models*
Tuesday, 19 July, 1:30 p.m. -- 2:50 p.m., D1-LP-03

Calibration, Scaling, DIF, and Projection: A Common Framework Using Multidimensional IRT.
Moonsoo Lee, Mark Hansen, Li Cai

That symposium will also include...

The Lord-Wingersky Algorithm After 25+ Years: Version 2.0 for Hierarchical Item Factor Models.
Li Cai

IMPS 2011 at The Hong Kong Institute of Education 香港教育學院

Another example of basic research somewhat associated with PROMIS:

Brian Stucky's Dissertation (2011)—*Logistic Approximations of Marginal Trace Lines for Bifactor Item Response Theory Models*.

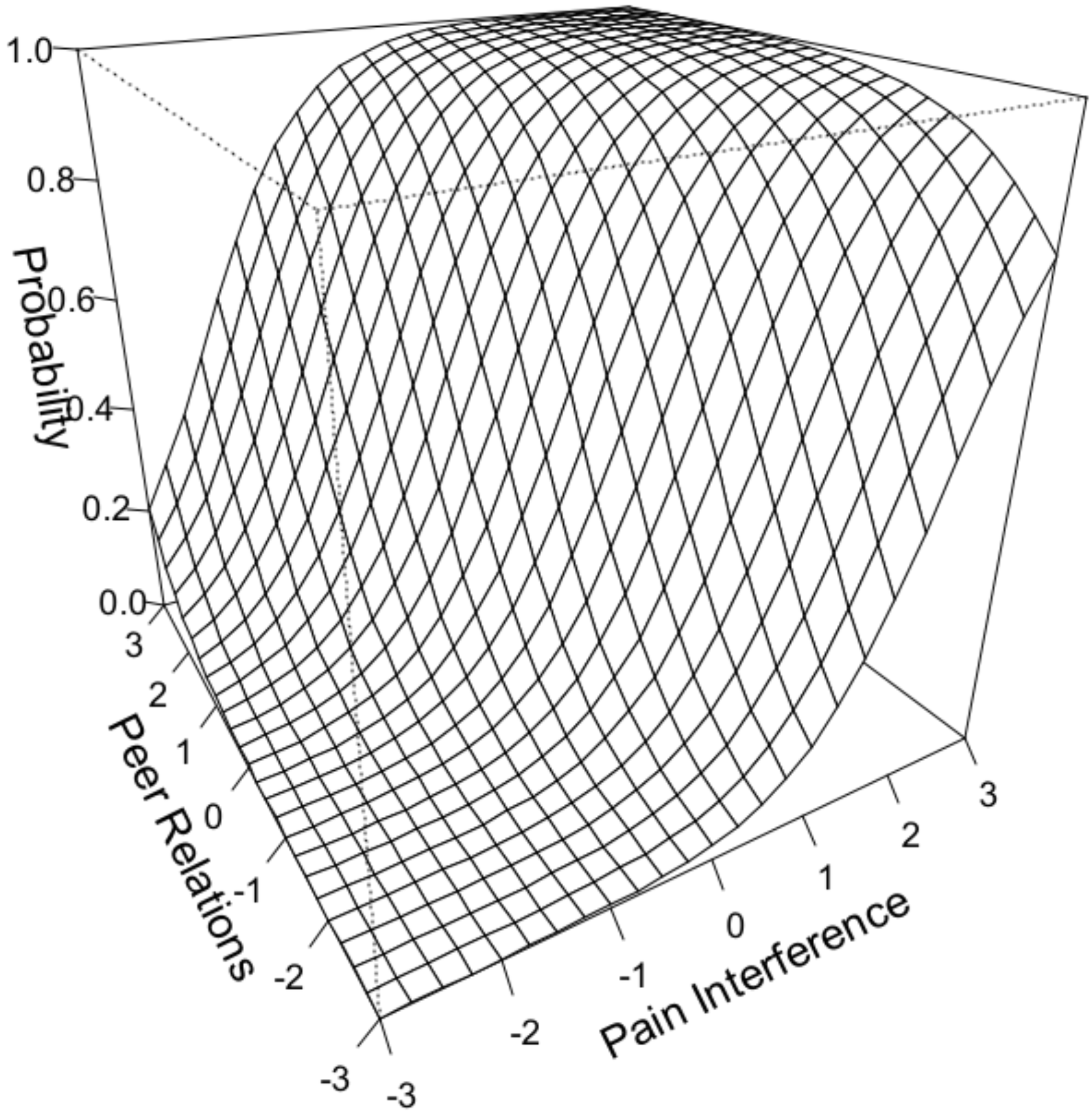
[The seeds of this were in the Donoghue-Shultz-Ip work with NAEP.]

Graded Model Slope Parameters for Seven Items with a Locally Dependent Pair

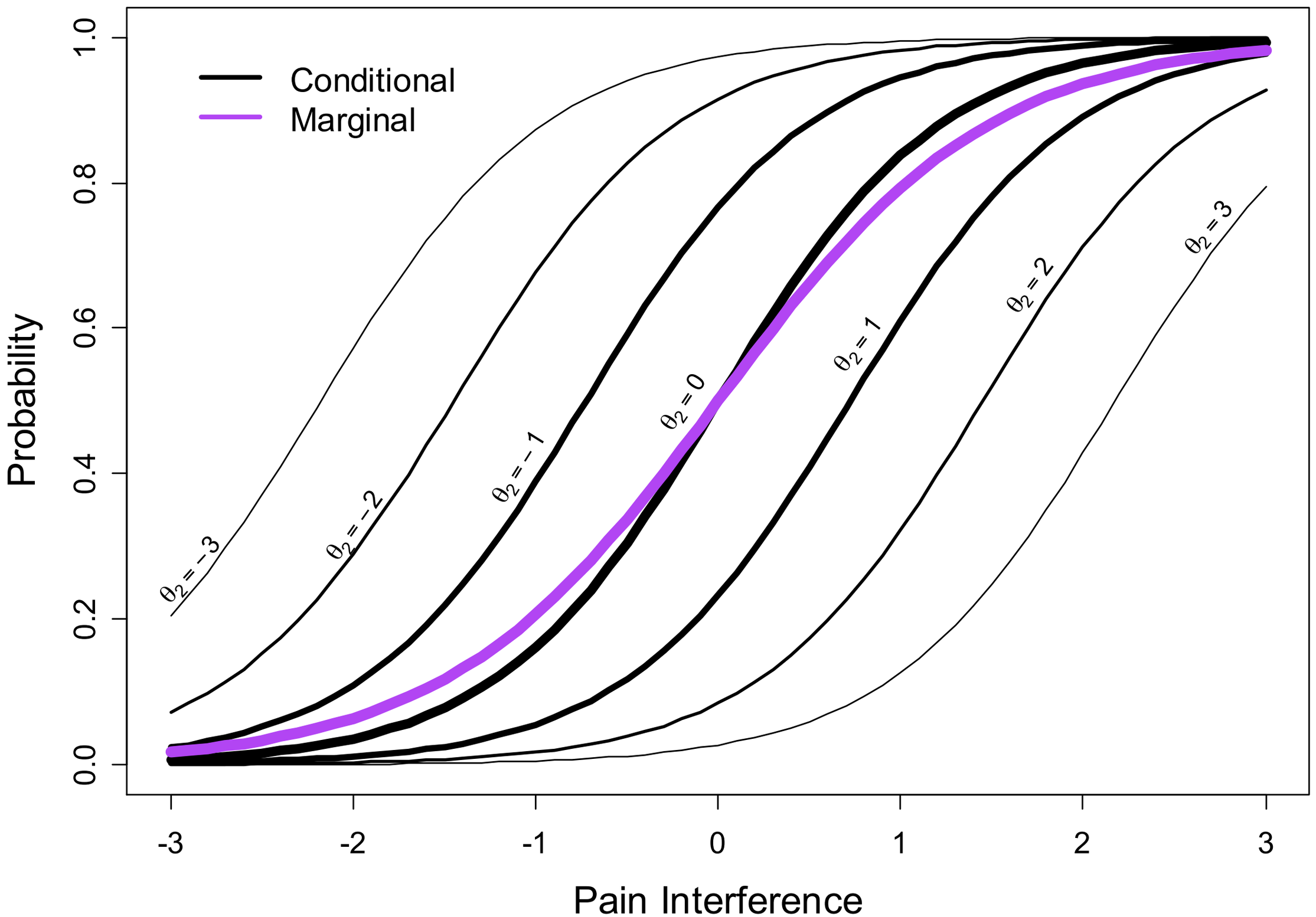
	1D	Bifactor	
Form 4 Items	IRT	Primary	Secondary
It was hard to get along with other people when I had pain.	1.51	1.64	1.19
I wanted to be alone when I had pain.	1.30	1.40	1.19
I hurt a lot.	1.37	1.41	---
It was hard for me to remember things when I had pain.	1.53	1.54	---
It was hard to do sports or exercise when I had pain.	1.60	1.66	---
I missed school when I had pain.	1.26	1.29	---
It was hard to stay standing when I had pain.	2.34	2.42	---

Thanks to Brian Stucky for these slides!

It was hard to get along with others when I had pain



Thanks to Brian Stucky for these slides!



Thanks to Brian Stucky for these slides!

Graded Model Slope Parameters for Seven Items with a Locally Dependent Pair

	ID	Bifactor			ID
Form 4 Items	IRT	Primary	Secondary	Marginal	*Separately*
It was hard to get along with other people when I had pain.	1.51	1.64	1.19	1.34	1.34
I wanted to be alone when I had pain.	1.30	1.40	1.19	1.15	1.13
I hurt a lot.	1.37	1.41	---		1.41
It was hard for me to remember things when I had pain.	1.53	1.54	---		1.50
It was hard to do sports or exercise when I had pain.	1.60	1.66	---		1.65
I missed school when I had pain.	1.26	1.29	---		1.26
It was hard to stay standing when I had pain.	2.34	2.42	---		2.35

Thanks to Brian Stucky for these slides!

IMPS 2011

At this conference, I expect Eddie Ip will say more about this …

Parallel Session: Item Response Theory - Methodology I
Tuesday, 19 July, 11:10 a.m. -- 12:30 p.m., D1-LP-07

Projective IRT for Purified Constructs

Edward Ip

IMPS 2011 at The Hong Kong Institute of Education 香港教育學院

Another example of basic research somewhat associated with PROMIS:

Reise, S.P. , Moore, T.M., & Haviland, Mark G.(2010). Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment*, *92*, 544-559.

"ECV" may be another abbreviation we'll all soon know.

IMPS 2011

Another example of basic research somewhat associated with PROMIS:

Parallel Session: *Differential Item Functioning & Local Independence*
Thursday, 21 July, 11:10 a.m. -- 12:30 p.m., D2-LP-10

Identifying Local Dependence with a Score Test Statistic Based on the Bifactor 2-Parameter Logistic Model.
Yang Liu, David Thissen

related in the same session:

A New Procedure for Detecting Departures from Local Independence in Item Response Models.
Michael Edwards, Carrie Houts, Li Cai

IMPS 2011 at The Hong Kong Institute of Education 香港教育學院

If we were starting PROMIS now, we might consider MIRT (calibration *and* scoring!) but for one challenge:

How will we make MIRT "scores" comprehensible, understandable, or usable, for consumers?

They are inherently multivariate, and only fully comprehensible in multivariate analyses.

How will we explain to users the difference between orthogonal, "residual," bifactor (hierarchical) second-tier scores and likely-correlated "simple structure" scores?

Good luck!

eHarmony®

Free to Review Your Matches

First Name:

I'm a: Woma seeking Men

Zip Code:

Country: United States

Email:

Note: Your email is used to log back in

Confirm Email:

Password:

Must be at least 5 characters

How did you hear about us? Please select...

Find My Matches

TRUSTe CERTIFIED PRIVACY · BBBOnLine RELIABILITY PROGRAM · VeriSign Trusted VERIFY

Love is out there.
We can help you find it.

Why Join eHarmony
Learn more

You've come to the right place. A place where millions of people are brought together based on the things that really matter. Like who you are on the inside. And for the last 10 years, we've focused on just one thing. Helping people find deep, meaningful love. So let's get started. Because love begins here.

David+Kimbra
Matched by eHarmony

See why eHarmony is responsible for nearly 5% of marriages in the U.S.*

*2009 U.S. survey conducted for eHarmony by Harris Interactive®

Thanks to all whose research I've "borrowed" (stolen) for this presentation!*

Preparation of this presentation was funded in part by the National Institutes of Health through the NIH Roadmap for Medical Research PROMIS initiative, Grants 1U01AR052181-01 and 2U01AR052181-06 from the National Institutes of Health.

*See Thissen, D. (2001). Psychometric engineering as art. *Psychometrika*, *66*, 473-486.