

**Rasch Measurement
Training Seminars**

WINSTEPS

and

Facets

Conducted by:

John Michael Linacre, Ph.D.

University of Sydney
Australia

info@winsteps.com
www.winsteps.com

P.O. Box 811322
Chicago IL 60681-1322

Winsteps

Winsteps is intended for solving practical measurement problems, quickly and conveniently. It constructs Rasch measures from simple rectangular data sets, usually of persons and items. After initial familiarization, it is straightforward to use in combination with other software. Item types that can be combined in one analysis include dichotomous, multiple-choice, and multiple rating-scale and partial credit items. Paired comparisons and rank-order data can also be analyzed. Missing data is no problem. Winsteps is designed as a tool that facilitates exploration and communication. The structure of the items and persons can be examined in depth. Unexpected data points are identified and reported in numerous ways. Powerful diagnosis of multidimensionality through principal components analysis of residuals detects and quantifies substructures in the data. The working of rating scales can be examined thoroughly, and rating scales can be recoded and items regrouped to share rating scales as desired. Measures can be fixed (anchored) at pre-set values. Winsteps is intended for practitioners who must make practical and quick decisions along the path to constructing effective tests, and who must then communicate their results usefully to end users. The developer of Winsteps uses the program daily in his own work, and is continually adding new features as a result of his own experience and feedback from users. Typical applications include educational tests, psychological assessments, attitude surveys, patient performance protocols, and calibrating adaptive-test item banks. Winsteps can process up to 1,000,000 persons, 30,000 items, and each item can have a rating scale of up to 255 categories. Conceptually, it originated around 1983 in the program Microscale, which was conceptualized by Ben Wright and written by Mike Linacre. It was the first Rasch program to flexibly accommodate missing data, also the first to run on a personal computer. That program was further developed as Mscale, then Bigscale, Bigsteps and now Winsteps.

Facets

Facets is designed to handle really tough applications of unidimensional Rasch measurement. It constructs measures from complex data involving heterogeneous combinations of examinees, items, tasks, judges along with further measurement and structural facets. It handles flexibly combinations of items of different formats in one analysis. These include dichotomies, rating scales with up to 255 categories, Poisson counts and Bernoulli trials. Multiple different measurement models can be included in the same analysis, including paired comparisons, rank-order, rating scales, partial credit and dichotomizations involving from 1 to 255 facets. Measures can also be fixed (anchored) individually or by group mean, facilitating equating and linking across test sessions. Quality-control fit evaluation of all measures is provided. Unexpected data points are identified. Bias, differential item functioning and interactions can be measured. Weighting schemes can be implemented. Up to 1 million examinees etc. can be included in one analysis. Typical applications have over 90% missing data. Since interpreting Facets output can be challenging, it is recommended that simpler Rasch approaches be tried first. Facets has been used successfully to construct measures for medical staff performance, patient performance, public speaking, sports performance by individuals and teams, and consumer preferences. It originated in 1986 in order to service the complex, high-stakes practical examinations of the American Society of Clinical Pathologists. In these examinations, the competence of laboratory technicians to make slides of sections of the human body is assessed by senior pathologists using a complex judging and equating design. Facets was successfully applied to these data, and continues to be. Since then Facets has been used in many rating and assessment endeavors in the educational and medical fields, such as the continuing evaluation of the Advance Placement English Literature and Composition Program for the College Board and Educational Testing Service. The Georgia (USA) High School Writing Test employs Facets as do the Minneapolis (USA) Public Schools for “prompt” equating, and a research project at the University of Northern British Columbia (Canada) for evaluating “Consistency of Writing Prompt Difficulties Across Elementary School Grades.”

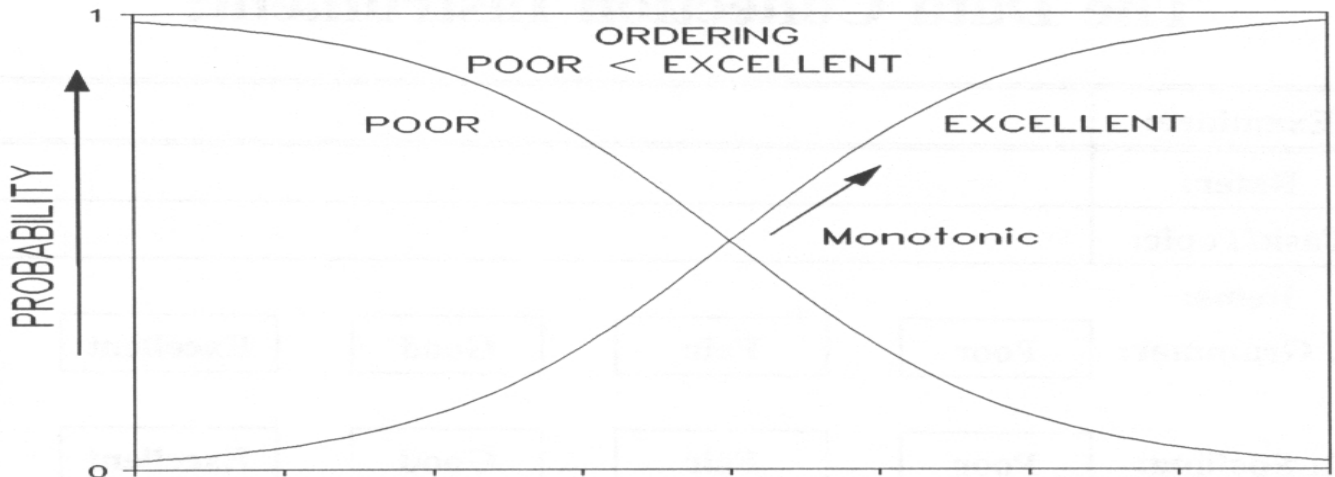
Rasch Measurement

Rasch analysis is the necessary and sufficient means of constructing linear measures from ordinal observations. Around 1953, Danish mathematician, Georg Rasch, was faced with the challenge of constructing measures from raw scores on intelligence and reading tests. He perceived that the Poisson distribution could be used to convert ordinal counts into linear measure. From this insight has blossomed a whole family of “Rasch models”. His initial work is described in his book “Probabilistic Models for some Intelligence and Attainment Tests.” For 20 years, from 1960 till his death in 1980, Georg Rasch and Ben Wright, of the University of Chicago, cooperated closely in developing both the theory and the practical use of Rasch models. Important books are Wright & Stone, “Best Test Design”, and Wright and Masters, “Rating Scale Analysis”. Students of Georg Rasch, colleagues in Europe, and particularly students of Ben Wright in the USA and Australia have propagated Rasch measurement around the world. A recent useful books are “Applying the Rasch Model” by Bond & Fox and “Introduction to Rasch Measurement” edited by E.V. Smith Jr. and R.M. Smith.

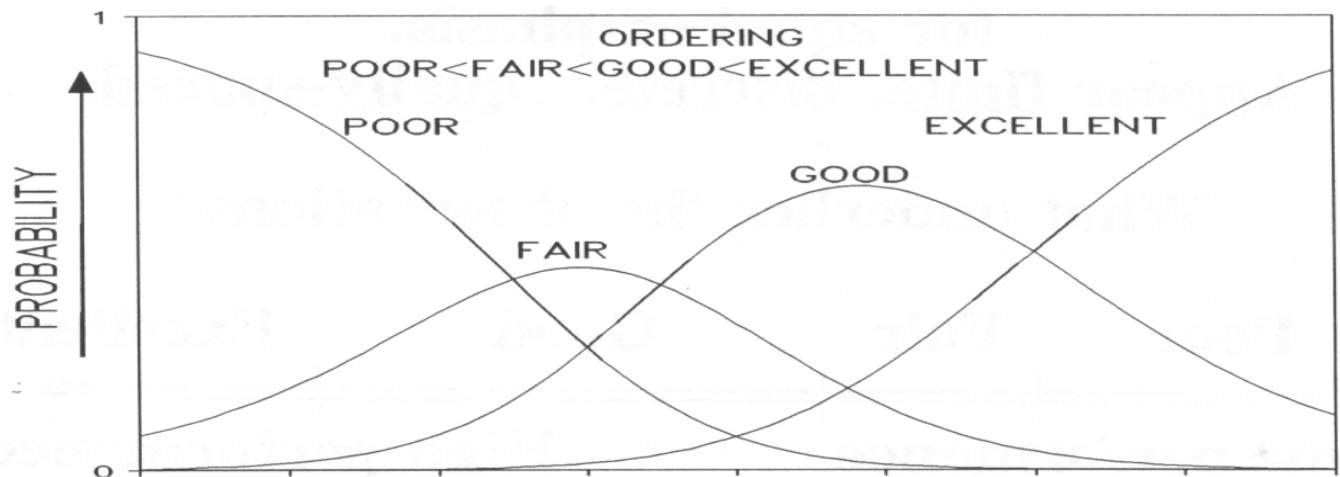
John Michael Linacre, Ph.D., M.A., C.D.P., C.C.P., D.M.

John "Mike" Linacre is the developer of software widely used in constructing objective measures from ordinal observations. Mike has an M.A. in Mathematics from Cambridge University, and a Ph.D. in Psychometrics from the University of Chicago. He worked closely with Benjamin D. Wright, the leading advocate of Rasch measurement, for over 15 years as a Research Associate at the University of Chicago. He has been editor of *Rasch Measurement Transactions* for more than 15 years. Mike recently relocated to Australia, but continues to develop measurement software and to consult internationally on educational and psychological measurement problems. He is an adjunct Professor at the University of the Sydney. He has authored well over 100 published articles and conference papers. His comments regarding improving the quality of Olympic Figure Skating Judging were recently quoted in the Toronto (Canada) Globe-Mail.

PROBABILITY ORDERING MONOTONICITY



⇒ INCREASING COMPETENCE ⇒



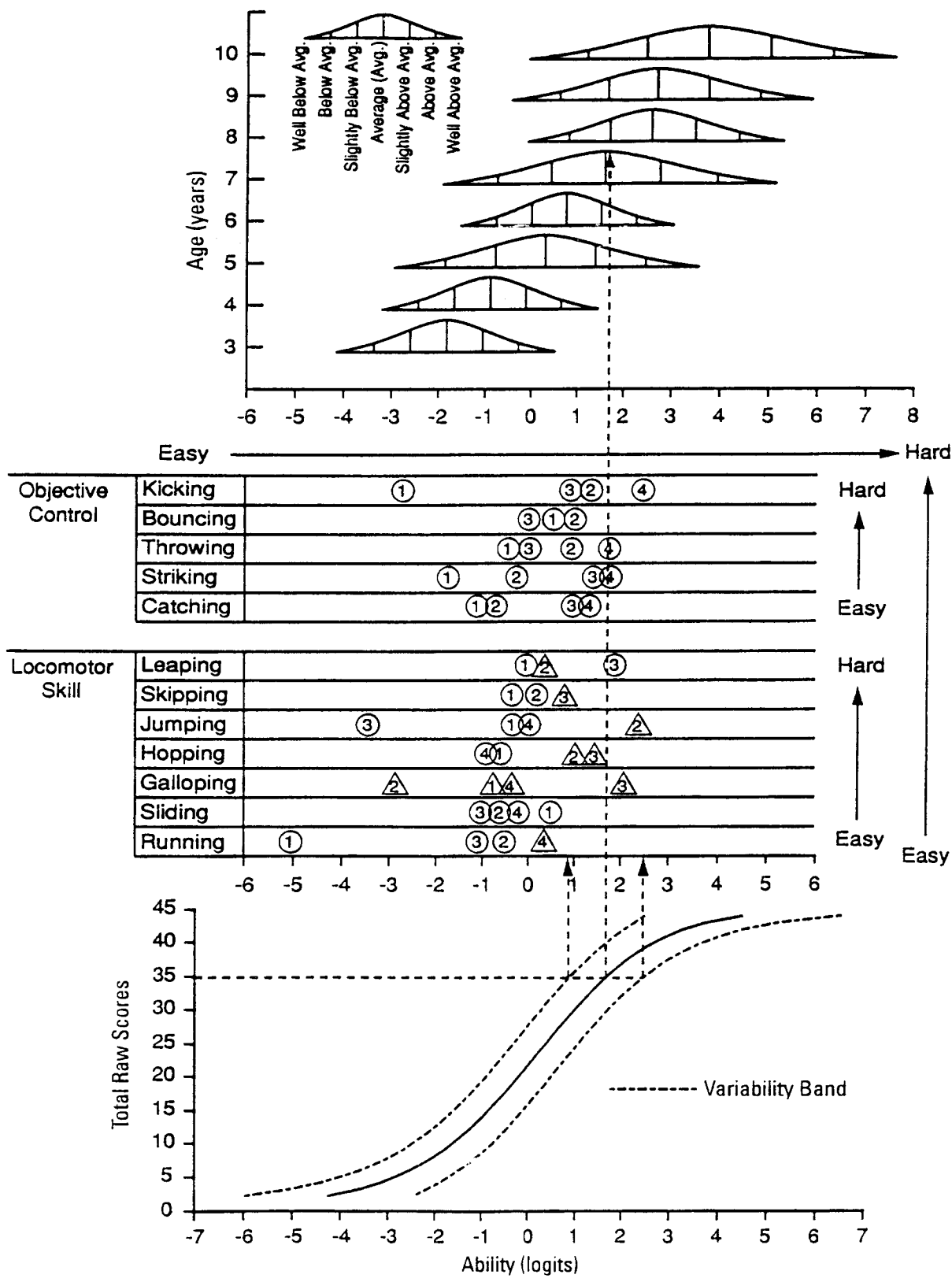


Figure 4. A proposed scoring-reporting sheet.

FIMSM Cognitive Items

MEASURE PATIENT HERE

Circle Sum & Draw Lines

N Comprehension
 O Expression
 P Social Interaction
 Q Problem Solving
 R Memory

FIM at +1 S.E.
 FIM at -1 S.E.
 FIM Raw Score
 Linear FIMITs
 SE FIMITs

KeyFIM Patient Record

Level	N	O	P	Q	R				
	7	7	7	7	7				
	7	7	7	7	7				
	6	6	6	6	6				
	5	5	5	5	5				
	4	4	4	4	4				
	3	3	3	3	3				
	2	2	2	2	2				
	1	1	1	1	1				
	0	0	0	0	0				
	34	35	34	33	32				
	33	34	35	34	33				
	32	33	34	35	34				
	31	32	33	34	35				
	30	31	32	33	34				
	29	30	31	32	33				
	28	29	30	31	32				
	27	28	29	30	31				
	26	27	28	29	30				
	25	26	27	28	29				
	24	25	26	27	28				
	23	24	25	26	27				
	22	23	24	25	26				
	21	22	23	24	25				
	20	21	22	23	24				
	19	20	21	22	23				
	18	19	20	21	22				
	17	18	19	20	21				
	16	17	18	19	20				
	15	16	17	18	19				
	14	15	16	17	18				
	13	14	15	16	17				
	12	13	14	15	16				
	11	12	13	14	15				
	10	11	12	13	14				
	9	10	11	12	13				
	8	9	10	11	12				
	7	8	9	10	11				
	6	7	8	9	10				
	5	6	7	8	9				
	100	118	100	95	90				
	18	13	10	8	7				
	85	80	75	70	65				
	80	75	70	65	60				
	10	8	7	6	5				
	5	4	3	2	1				
	45	40	35	30	25				
	5	4	3	2	1				
	50	45	40	35	30				
	5	4	3	2	1				
	55	50	45	40	35				
	60	55	50	45	40				
	6	5	4	3	2				
	65	60	55	50	45				
	70	65	60	55	50				
	7	6	5	4	3				
	75	70	65	60	55				
	8	7	6	5	4				
	80	75	70	65	60				
	10	8	7	6	5				
	85	80	75	70	65				
	90	85	80	75	70				
	13	10	8	7	6				
	95	90	85	80	75				
	18	13	10	8	7				

For Rating Unexpectedness: 1 S.E. ≈ 15 FIMITs

RATE
PATIENT
HERE

Circle
Rating

Sum=

Measurement challenges:

**INFINITY
CONTINUITY
IRREGULARITY**

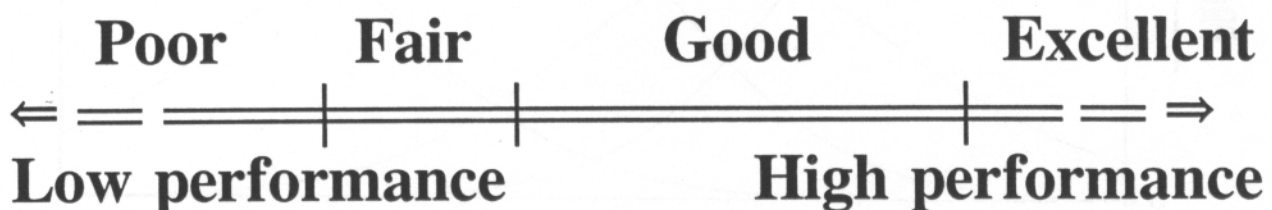
The Data Collection Instrument:

Examinee:				
Rater:				
Task/Topic:				
Items:				
1. Grammar:	Poor	Fair	Good	Excellent
2. Spelling:	Poor	Fair	Good	Excellent

**Rating categories printed equally spaced
for equal emphasis.**

Appear finite, discrete, equally-spaced.

What underlies the observations?



An infinite, continuous, irregular variable

Measurement challenges:

**ADDITIVITY
INTERVAL SCALING
ESTIMABILITY
INCOMPLETENESS**

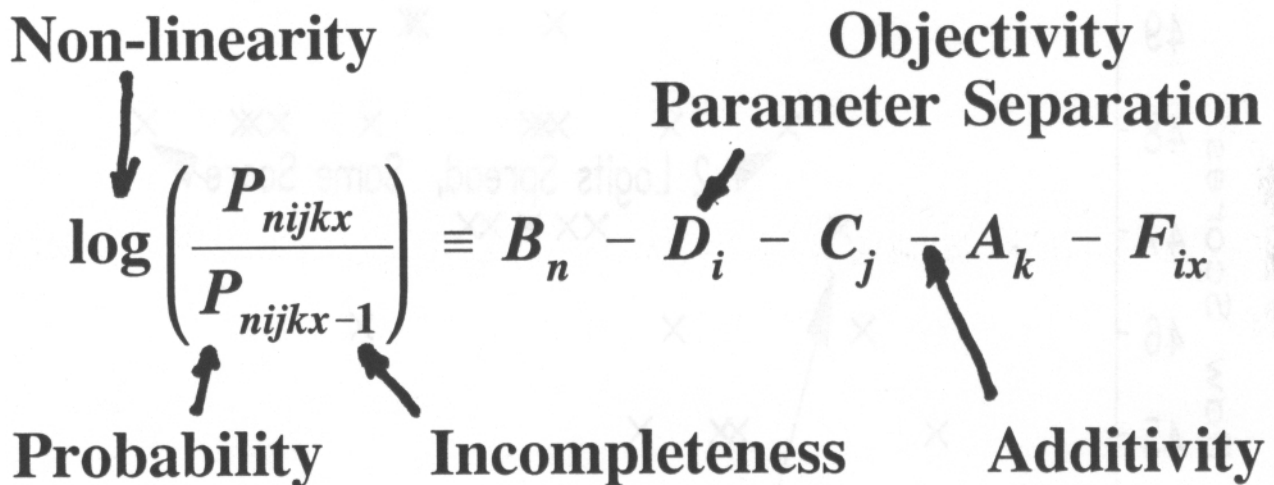
We want each datum to be explained, stochastically, by an additive combination of person n 's ability (B_n) to perform relative to item i 's difficulty (D_i) in view of rater j 's severity (C_j) when encountering task k 's difficulty (A_k) rated on the calibrated rating scale ($\{F_{ix}\}$):

$$B_n - D_i - C_j - A_k - \{F_{ix}\} \Rightarrow \{Datum, X_{nijk}\}$$

This requires that, from the observed, inferentially incomplete data, measures be sufficiently estimated:

$$\{Data, X_{...}\} \Rightarrow \hat{B}_n - \hat{D}_i - \hat{C}_j - \hat{A}_k - \{\hat{F}_{ix}\}$$

Measurement solution:
MANY-FACET RASCH MODEL
 provides **LINEAR MEASURES**
 with their **PRECISION (S.E.)**
 and **COHERENCE (Data FIT)**



PRECISION of $B_n =$
SE = $1 / \sqrt{\Sigma}$ (Model Variance of n's Data)

COHERENCE of $B_n =$
FIT = $\Sigma (X_{nijk} - \text{Expectation})^2 / \text{Variance}$

This necessary and sufficient solution:
Invented by Georg Rasch in 1953,
Used successfully throughout the world,
to construct measures with
UTILITY and MEANING

Quality-Control Mean-Square Fit Statistics



Outfit: Outlier-sensitive. Conventional χ^2 divided by its degrees of freedom:

$$\text{Outfit}_n = \frac{1}{L} \sum_{i=1}^L \frac{(X_{ni} - E_{ni})^2}{V_{ni}}$$

where X_{ni} is datum, E_{ni} is expectation, V_{ni} is variance.

Infit: Inlying-pattern-sensitive. Information-weighted χ^2 divided by its degrees of freedom:

$$\text{Infit}_n = \frac{\sum_{i=1}^L (X_{ni} - E_{ni})^2}{\sum_{i=1}^L V_{ni}}$$

Responses: Easy--Items--Hard	Pattern Diagnosis	OUTFIT mean-square	INFIT mean-square	S.E. inflator
111 0110110100 000	Modelled/Ideal	1.0	1.1	1.0
111 1111100000 000	Guttman/Deterministic	≈ 0.3	≈ 0.5	1.8
000 0000011111 111	Miscode	≈ 12.6	4.3	3.5
011 111110000 000	Carelessness/Sleeping	≈ 3.8	1.0	1.9
111 1111000000 001	Lucky Guessing	≈ 3.8	1.0	1.9
101 0101010101 010	Response set/Miskey	≈ 4.0	≈ 2.3	2.0
111 1000011110 000	Special knowledge	0.9	≈ 1.3	1.1
111 1010110010 000	Imputed outliers †	≈ 0.6	1.0	0.8
Right Transition Wrong				
	OUTFIT sensitive to outlying observations	» 1.0 unexpected outliers	» 1.0 disturbed pattern	
high--low--high		« 1.0 overly predictable outliers	« 1.0 Guttman pattern	
	INFIT sensitive to pattern of inlying observations			
low--high--low				

† as when a tailored test is filled out by inputting all “right” response to easier items and all “wrong” to harder items.

Complete Judging Plan

Judge Essay Person	1 ABC	2 ABC	3 ABC	4 ABC	5 ABC	6 ABC	7 ABC	8 ABC	9 ABC	10 ABC	11 ABC	12 ABC
1	553	686	877	687	777	685	565	667	586	567	776	696
2	454	542	445	534	334	344	433	526	444	445	533	534
3	434	544	343	555	433	544	563	443	554	454	443	343
4	345	426	232	545	445	225	464	456	642	446	445	335
5	443	548	656	545	657	448	558	466	464	448	547	348
6	544	846	843	565	633	367	788	673	666	566	564	454
7	545	665	454	667	755	646	773	785	874	565	745	447
8	553	763	655	675	775	653	773	656	784	576	573	574
9	343	643	643	645	534	523	665	674	753	546	545	765
10	564	766	884	776	655	667	875	778	778	667	649	888
11	535	524	537	544	545	435	546	557	326	446	456	334
12	436	644	444	546	666	555	574	445	745	356	763	676
13	445	486	657	566	246	366	368	448	467	348	569	349
14	446	533	333	344	545	343	463	353	354	346	462	363
15	548	855	743	746	766	656	665	765	854	666	862	844
16	644	653	547	545	643	454	556	467	666	447	558	667
17	414	817	625	628	536	518	425	618	717	627	639	436
18	334	655	443	445	243	473	445	747	654	445	435	334
19	747	745	837	756	755	847	664	688	737	656	847	938
20	443	666	735	556	557	557	588	667	666	557	476	488
21	242	443	336	465	245	243	263	245	441	253	342	254
22	564	765	747	666	864	577	667	576	667	557	667	785
23	446	566	753	646	444	565	475	388	576	557	557	776
24	332	422	334	433	322	214	423	223	323	313	233	223
25	543	664	544	657	646	544	454	448	547	545	456	464
26	644	764	955	756	545	658	655	867	776	646	756	885
27	342	346	334	344	346	234	256	256	345	345	256	253
28	343	463	335	334	465	573	341	475	442	243	462	272
29	433	444	323	446	334	333	235	336	423	336	323	343
30	542	564	244	655	445	224	546	575	645	446	432	555
31	325	514	313	425	315	314	334	225	525	314	324	314
32	644	744	445	545	533	553	567	584	664	447	556	364

**Thought to be "ideal".
Gives precise measures.**

"Rotating Test-Book" Plan

Judge Essay Person	1 ABC	2 ABC	3 ABC	4 ABC	5 ABC	6 ABC	7 ABC	8 ABC	9 ABC	10 ABC	11 ABC	12 ABC
1	553	686										
2		542	445									
3			343	555								
4				545	445							
5					657	448						
6						367	788					
7							773	785				
8								656	784			
9									753	546		
10										667	649	
11											456	334
12	436											676
13	445						368					
14		533						353				
15			743						854			
16				545						447		
17					536						639	
18						473						334
19	747			756								
20		666			557							
21			336			243						
22				666			667					
23					444			388				
24						214			323			
25							454			545		
26								867			756	
27									345			253
28	343									243		
29		444									323	
30			244									555
31	grading performed by any available judges											
32	grading performed by any available judges											

Less judge effort
Robust against mistakes

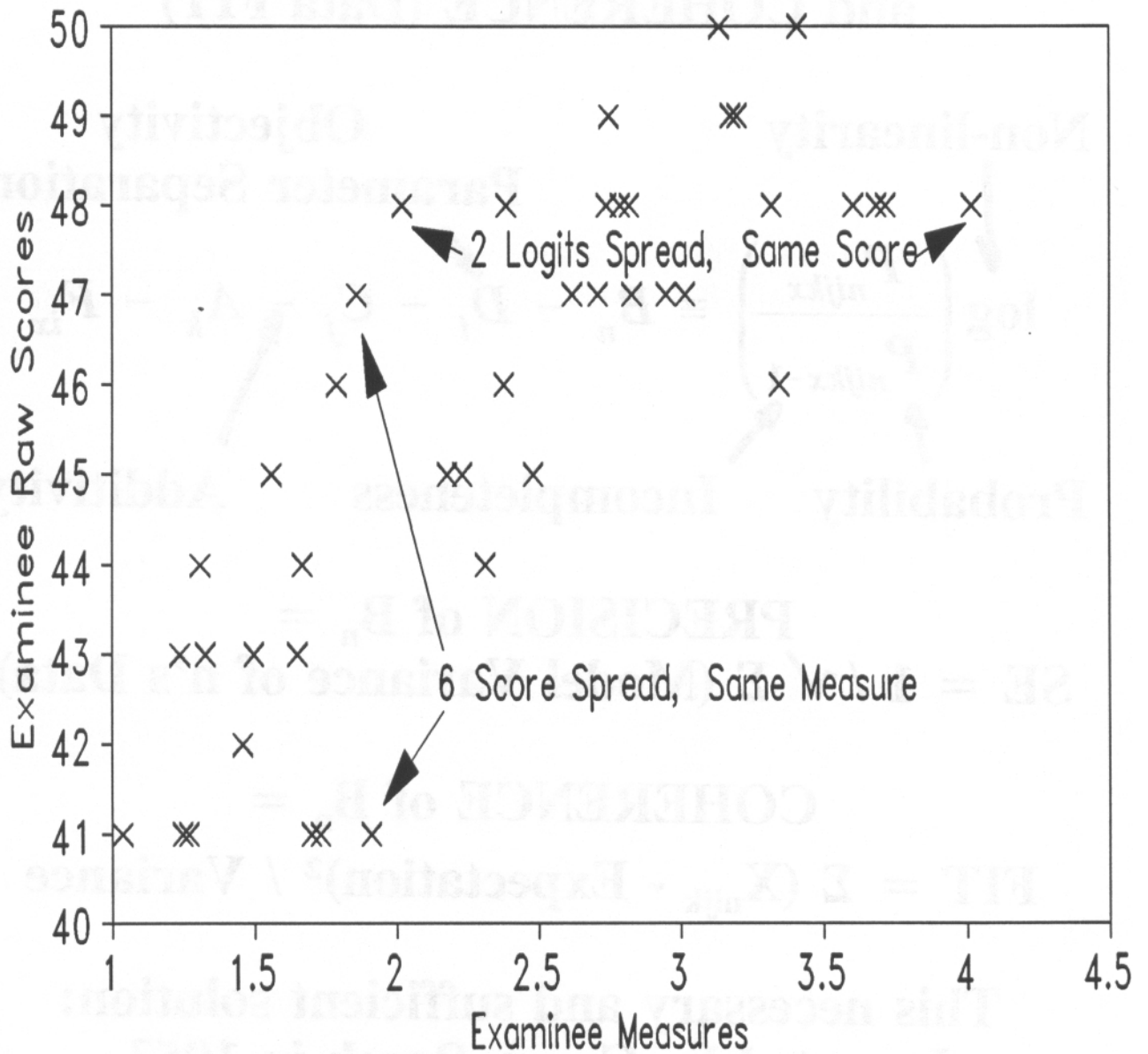
"Minimal Effort" Judging Pla

Judge Essay Person	1 ABC	2 ABC	3 ABC	4 ABC	5 ABC	6 ABC	7 ABC	8 ABC	9 ABC	10 ABC	11 ABC	12 ABC
1							6	7	5			
2				5	3					5		
3	4									5	3	
4		2				5				4		
5					7					4	5	
6	5	6									6	
7					5		3					4
8		6						6		5		
9					3		6	4				
10				7	5				7			
11										6		3
12	4			6							4	7
13	4			6					4			
14		6	3				4					
15	4								4	6		
16		6	4								8	
17			2	6	6							
18						4	4					4
19		7		6		4						
20			7	5					6			
21						2	6			3		
22			7		8				6			
23		6	7					8				
24						2	3					2
25						4	4					
26									7	6		8
27			3					2			6	
28	4	3										2
29			3					3				3
30	2							5			3	
31				2		4		2				
32			5		5	5						

Fast, fair, but less precise

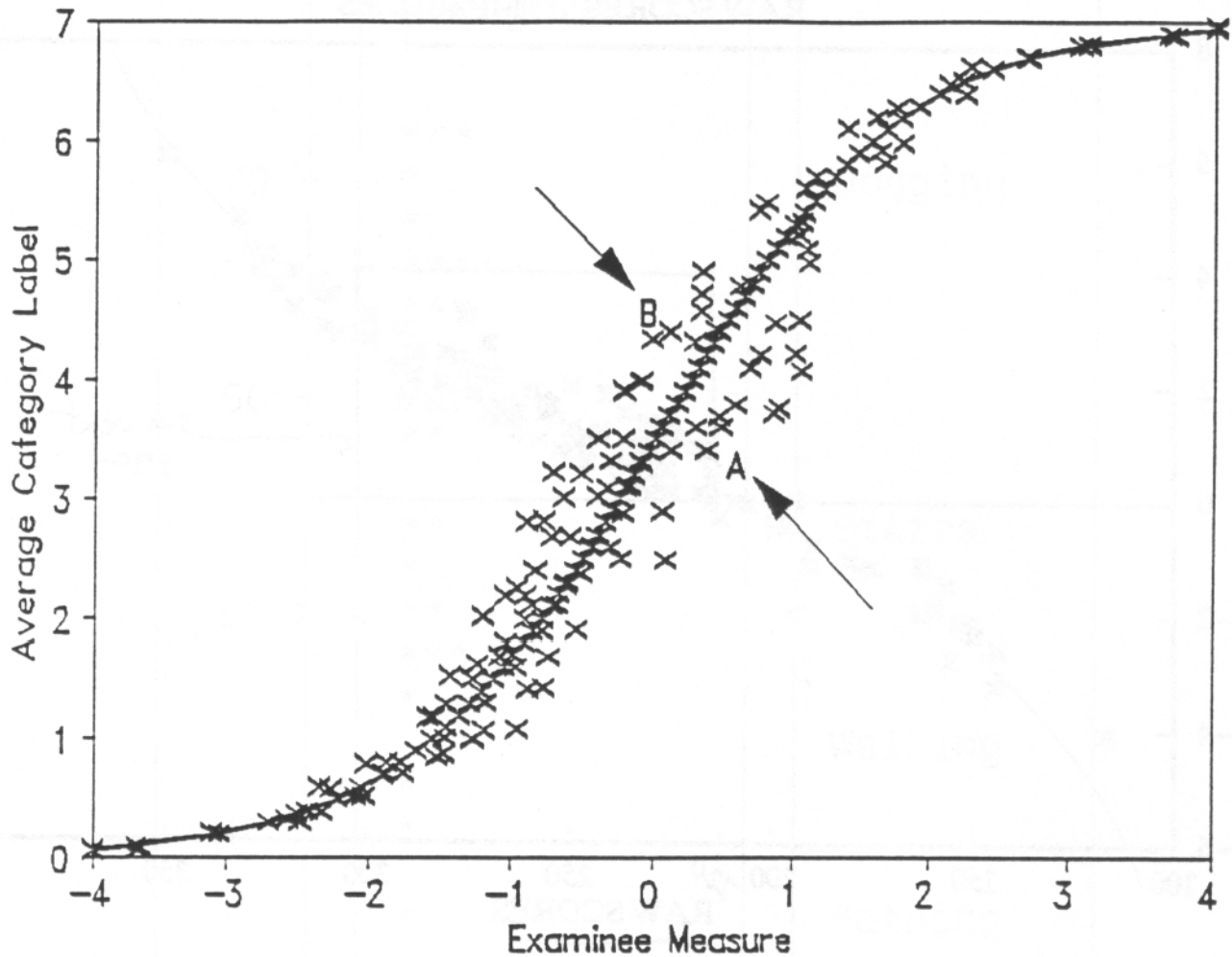
Georgia Spring 1990 Writing Data

Detail of Measure/Rating Plot



Form 74
121 Students
5 Items
72 Raters

Ratings to Measures conversion



Ogive = complete data

X = not rated by all judges

A = rated by severe judges

B = rated by lenient judges

		+Task (Easy)			
Measr	+Patient (Independent)		+Clinician (Lenient)		Scale
80	George *	Eating			4
70	*** Anne * **** *** *	Bathing			---
60	Mary **	Walking	Smith		3
	** * ** *** *	Stairs	Sato Wright Lopez		---
50	> < ** *** *	Jogging	Jones		< 2
40	* *				---
30	Igor				1
20		(Hard)			---
		+Task			0
Measr	+Patient (Dependent)			(Severe)	Scale
				+Clinician	

Measurement "Rulers" for patient independence.

FUNCTIONAL ASSESSMENT

CALI- BRATION	GENERAL PAIN	BACK PAIN	NECK PAIN
70	STANDING		
65		STANDING	<i>REACHING</i>
60	STAIRS	<i>SITTING</i>	
55	LOW SEAT		STANDING
50		LOW SEAT	SITTING
45			
40		STAIRS	LOW SEAT STAIRS
35	REACHING	REACHING	
30	SITTING		

Diagnosing Misfit

Classification	INFIT	OUTFIT	Explanation	Investigation
Hard Item	Noisy	Noisy	Bad item	Ambiguous or negative wording? Debatable or misleading options?
	Muted	Muted	Only answered by top people	At end of test?
Item	Noisy	Noisy	Qualitatively different item Incompatible anchor value	Different process or content? Anchor value incorrectly applied?
		?	Biased (DIF) item	Stratify residuals by person group?
		Muted	Curriculum interaction	Are there alternative curricula?
	Muted	?	Redundant item	Similar items? One item answers another? Item correlated with other variable?
Rating scale	Noisy	Noisy	Extreme category overuse	Poor category wording? Combine or omit categories?
	Muted	Muted	Middle category overuse	Wrong model for scale?
Person	Noisy	?	Processing error Clerical error Idiosyncratic person	Scanner failure? Form markings misaligned? Qualitatively different person?
High Person	?	Noisy	Careless Sleeping Rushing	Unexpected wrong answers? Unexpected errors at start? Unexpected errors at end?
Low Person	?	Noisy	Guessing Response set "Special" knowledge	Unexpected right answers? Systematic response pattern? Content of unexpected answers?
	Muted	?	Plodding Caution	Did not reach end of test? Only answered easy items?
Person/Judge Rating	Noisy	Noisy	Extreme category overuse	Extremism? Defiance?
	Muted	Muted	Middle category overuse	Conservatism? Resistance?
Apparent unanimity			Collusion?	
<p>INFIT: information-weighted mean-square, sensitive to irregular inlying patterns OUTFIT: usual unweighted mean-square, sensitive to unexpected rare extremes Muted: unmodelled dependence, redundance, error trends (MnSq<0.5) Noisy: unexpected unrelated irregularities (MnSq>1.5)</p>				

Polytomous Mean-Square Fit Statistics

Smith R.M. (1996) Polytomous Mean-Square Fit Statistics. Rasch Measurement Transactions 10:3 p. 516-517.

Response String Easy.....Hard	INFIT MnSq	OUTFIT MnSq	RPM Corr.	Diagnosis
I. modelled:				
33333132210000001011	.98	.99	.78	<i>Stochastically monotonic in form, strictly monotonic in meaning</i>
31332332321220000000	.98	1.04	.81	
33333331122300000000	1.06	.97	.87	
33333331110010200001	1.03	1.00	.81	
II. overfitting (muted):				
33222222221111111100	.18	.22	.92	Guttman pattern high discrimination low discrimination tight progression
33333222221111100000	.31	.35	.97	
32222222221111111110	.21	.26	.89	
32323232121212101010	.52	.54	.82	
III. limited categories:				
33333333322222222222	.24	.24	.87	high (low) categories central categories only 3 categories
22222222221111111111	.24	.34	.87	
33333222222222111111	.16	.20	.93	
IV. informative-noisy:				
32222222011111111130	.94	1.22	.55	noisy outliers erratic transitions noisy progression extreme categories
33233332212333000000	1.25	1.09	.77	
33133330232300101000	1.49	1.40	.72	
33333333330000000000	1.37	1.20	.87	
V. non-informative:				
22222222222222222222	.85	1.21	.00	one category central flip-flop rotate categories extreme flip-flop random responses
12121212121212121212	1.50	1.96	-.09	
01230123012301230123	3.62	4.61	-.19	
03030303030303030303	5.14	6.07	-.09	
03202002101113311002	2.99	3.59	-.01	
VI. contradictory:				
11111222332221111111	1.75	2.02	.00	folded pattern central reversal high reversal Guttman reversal extreme reversal
11111111122222222222	2.56	3.20	-.87	
22222222233333333333	2.11	4.13	-.87	
00111111122222222333	4.00	5.58	-.92	
00000000033333333333	8.30	9.79	-.87	

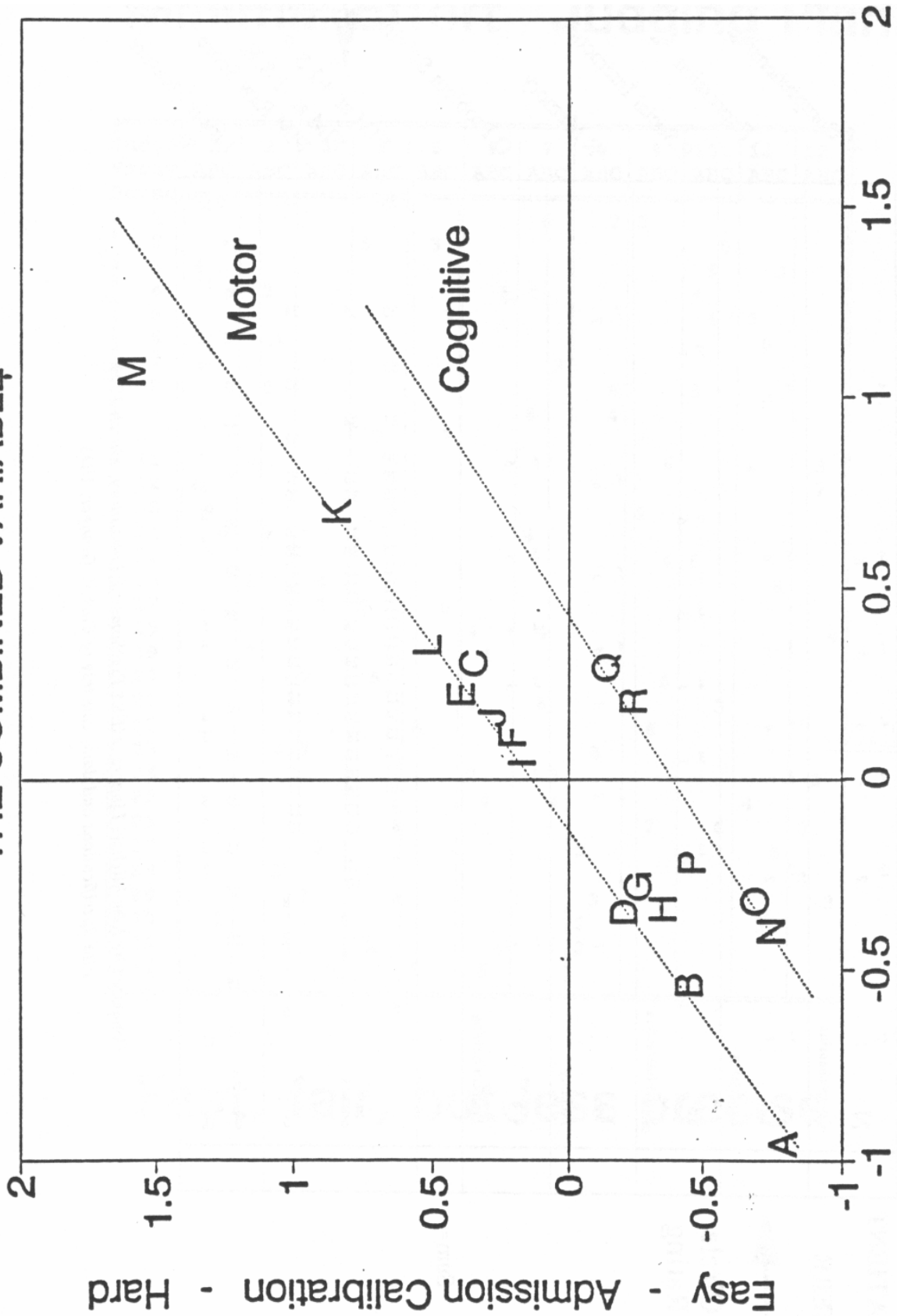
KeyMoth DIAGNOSTIC PROFILE

Basic Level: highest three consecutive correct responses before an error
 Ceiling Level: three consecutive errors

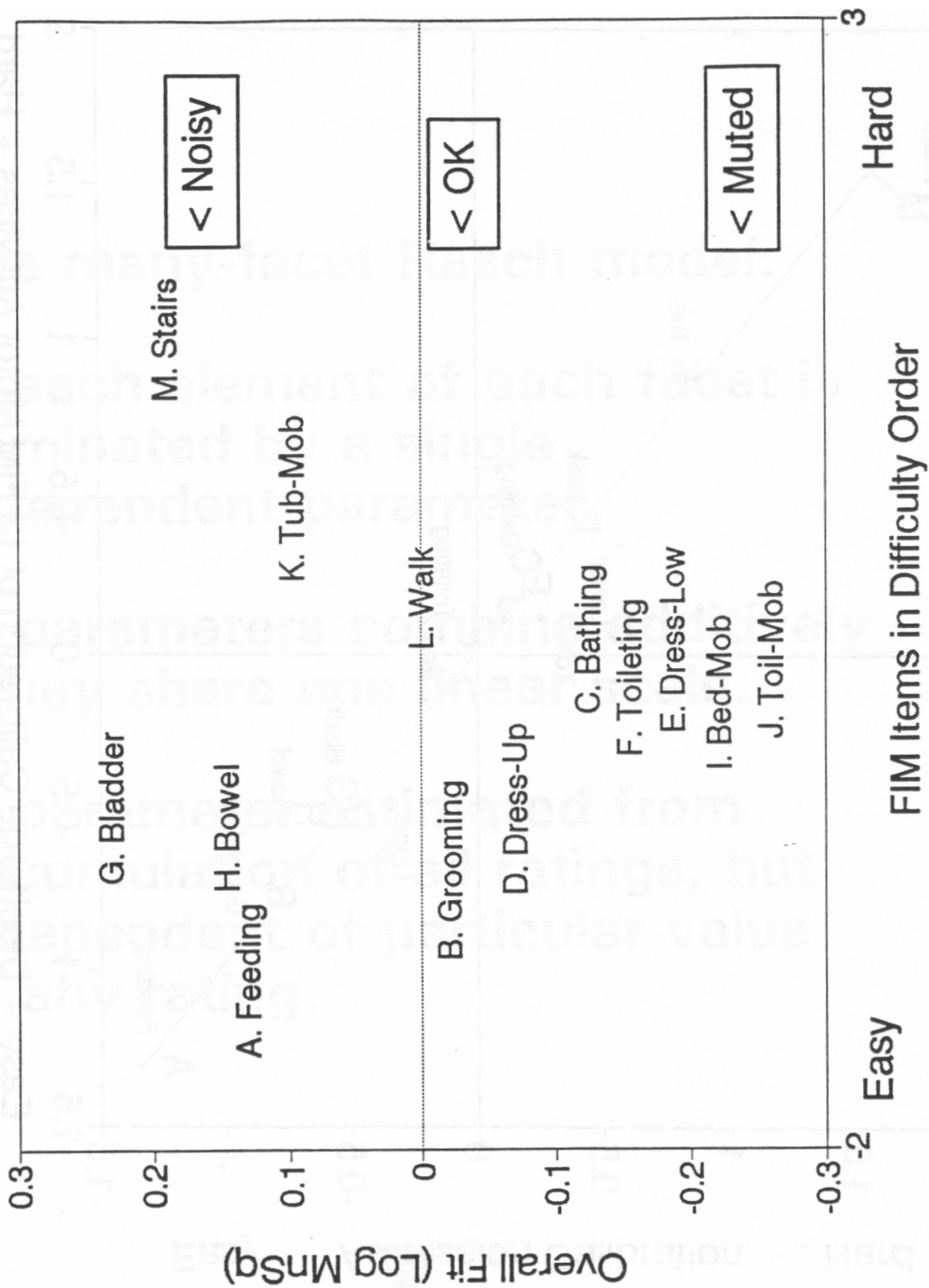
CONTENT



THE COMBINED VARIABLE



MOTOR VARIABLE (13 Items)



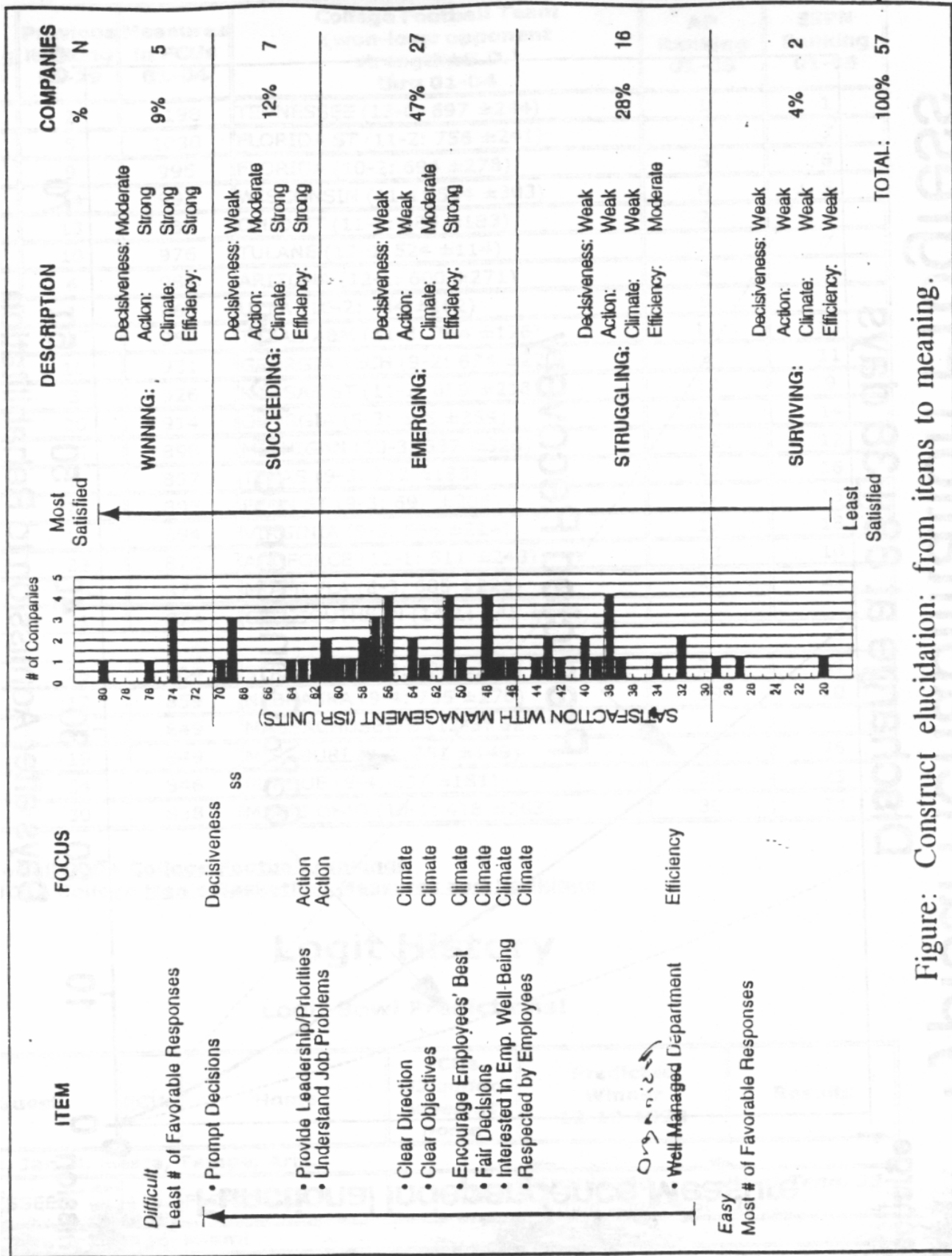


Figure: Construct elucidation: from items to meaning.