DOCUMENT RESUME

ED 364 573 TM 020 793

AUTHOR Linacre, John M.
TITLE Generalizability Theory and Many-Facet Rasch Measurement.
PUB DATE Apr 93
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Equations (Mathematics); Estimation (Mathematics); Evaluators; *Generalizability Theory; Item Response Theory; *Mathematical Models; *Measurement Techniques; *Scores; *Test Reliability
IDENTIFIERS *Many Faceted Rasch Model; *Rasch Model

ABSTRACT

Generalizability theory (G-theory) and many-facet Rasch measurement (Rasch) manage the variability inherent when raters rate examinees on test items. The purpose of G-theory is to estimate test reliability in a raw score metric. Unadjusted examinee raw scores are reported as measures. A variance component is estimated for the examinee distribution. Other variance components, due to item and rater distributions, interaction effects and random noise, are accumulated as examinee score error. Rasch computes a measure for each examinee, adjusted for the particular items and raters met by that examinee, that is more fair than the raw score. Rasch test reliability is higher than G-theory reliability because Rasch error variance excludes item and judge variance. (Contains 4 references.) (Author)

Generalizability Theory and Many-facet Rasch Measurement

John M. Linacre
University of Chicago

Paper presented at the
1993 Annual Meeting of the
American Educational Research Association
Atlanta, Georgia, April 13, 1993

Running head:
GENERALIZABILITY THEORY AND RASCH MEASUREMENT

Abstract

Generalizability theory (G-Theory) and many-facet Rasch measurement (Rasch) manage the variability inherent when raters rate examinees on test items. The purpose of G-Theory is to estimate test reliability in a raw score metric. Unadjusted examinee raw scores are reported as measures. A variance component is estimated for the examinee distribution. Other variance components, due to item and rater distributions, interaction effects and random noise, are accumulated as examinee score error. Rasch computes a measure for each examinee, adjusted for the particular items and raters met by that examinee, that is more fair than the raw score. Rasch test reliability is higher than G-Theory reliability because Rasch error variance excludes item and judge variance.

Key-words: Generalizability Theory, Many-facet Rasch measurement.

Generalizability Theory and Many-facet Rasch Measurement

INTRODUCTION

Performance assessment requires a rater to evaluate the quality of an examinee's performance on some task or performance item. The rater's evaluation is expressed as a numerical rating on a rating scale. The numerical ratings obtained by each examinee are summed into a raw score. This raw score summarizes the examinee's performance level for decision-makers, but the raw score is skewed by rater severity, item difficulty, examinee-rater-item interactions and random noise. This skew must be managed in order for valid, fair decisions to be made.

"The score on which the decision is to be based is only one of many scores that might serve the same purpose... The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his *universe* score... Knowing that observed score and universe score are not identical, the decision maker will want to take the discrepancy into account. One way to do this is to accompany each [observed score] by an expression of uncertainty. Another possibility is to *correct* the observed score in some manner so that it better approximates the universe score; this corrected value will also have uncertainty" (Cronbach et al. 1972 p.15-16).

Cronbach identifies two approaches to raw score discrepancies. Generalizability theory (G-theory) follows Cronbach's first way. Many-facet Rasch measurement (Rasch) follows Cronbach's second possibility. But, in following these alternative approaches, G-theory and Rasch address two different, but related problems. The decisions-maker's choice of technique must be guided by which problem is more pressing.

G-theory concerns itself with discovering how similar the observed raw scores might be to any other raw scores the examinees might obtain under very similar circumstances. G-theory considers the examinees as a group. Its aim is to estimate the error variance associated with examinee raw scores, but not to adjust any examinee's raw score for the particular raters, tasks or items that the examinee encountered. The finalé of a G-theory analysis is a reliability coefficient, "the generalizability coefficient", that summarizes the correlation between the examinees' unobservable "universe" scores and the scores that happen to have been observed. This reliability is only generalizable to very similar situations because all the variance components (main effects, interactions and random error) must maintain their values. This is the "observation is a sample from a defined universe" requirement (Cronbach et al. p.366).

Rasch (Linacre 1989) concerns itself with obtaining for each examinee a measure from which the details of the examinee's particular raters, items and tasks have been removed. Rasch considers each examinee as an individual, and attempts to liberate, statistically, each examinee's measure from the distributional details of the other examinees who happen to be included in the same analysis. The analytic aim is to transform each raw score from its

inevitably non-linear form into a linear measure, correcting it for the particular raters, tasks or items that the examinee encountered. A local error variance is also estimated for each examinee measure. The finalé of a Rasch analysis is thus a linear measure of each examinee's performance level, qualified by its standard error and quality-control fit statistics. Rasch also provides reliability coefficients. Rasch measures generalize to qualitatively similar, but quantitatively different, situations, such as adding slightly harder items to a math test, or including some more lenient raters on the judging panel. This is the "unidimensionality" requirement (Andrich 1988 p.9).

The decision-maker's criteria for selection of analytic approach are clear:

A. If it is important to estimate the similarity between the observed raw scores of this group of examinees and the raw scores that similar groups of examinees might obtain under other, similar circumstances, then G-theory is helpful.

B. If it is important to estimate for each examinee a measure as free as possible from the particularities of the components that generated the raw score, then Rasch is essential.

AN INSTRUCTIVE EXAMPLE

Kenderski (1983), reported in Shavelson and Webb (1991 p.8), presents a data set shown in Table 1, with fictitious identifying names added. Nine-year-old children were tape-recorded while solving mathematics problems in class, and again three weeks later. Two raters counted, independently, the number of times each child asked for help from other children. While solving the problems, each child asked between 0 and 4 times.

G-theory analysis

The data is modelled as a linear composition of components in the raw-score metric.

$$X_{nor} = \mu_{gm} + \mu_n + \mu_o + \mu_r + \mu_{no} + \mu_{nr} + \mu_{nor,e} \quad (1)$$

where

X_{nor} = the observed count for child c on occasion o by rater r
μ_{gm} = the grand mean of the counts
μ_n = the effect for child n
μ_r = the effect for rater r
μ_o = the effect for occasion o
μ_{nr} = the interaction effect between child n and rater r
μ_{no} = the interaction effect between child n and occasion o
μ_{ro} = the interaction effect between rater r and occasion o
$\mu_{nro,e}$ = the interaction effect between child n, rater r and occasion o, combined with, and indistinguishable from, random error, e.

Applying equation (1) to the data in Table 1 yields the raw score components shown in Figure 1. Though these components do, in fact, provide an estimate of each child's propensity to ask for help, it is not the individual component values, but their distribution that is of essence to G-theory.

G-theory analyses these data in two stages. The first stage is the generalizability study (G-study). This estimates variance components from a pilot data set. In the G-study equation (1) from the basis for an analysis of variance (ANOVA). The results are shown in Table 2. The ANOVA mean-squares agree with the Figure 1. For instance, the variance of the occasion effect is much greater than of the child-occasion interaction. But this is deceptive. The mean-squares are themselves composites in which the variance of each effect has been inflated by variance identified with interactions and error. Thus Table 2 identifies how the expected mean-squares, estimated by the observed mean-squares, decompose into underlying variance components. After decomposition, the variance of the occasion effect is less than that of the child-occasion interaction.

It is the variance components, now separated from other sources of variance, that are hypothesized to be generalizable. Computing variance components requires detailed knowledge of the experimental design. Design misspecification or data not in accord with the design invalidate the variance component estimates.

The second stage of a G-theory analysis is the decision study (D-study). This uses the variance components from the G-Study applied to new (or the same) data to make decisions. The experimental design of the D-study can be different from the G-study, but its variance components must be estimable from G-study variance components. Thus the G-study may have two raters for each examinee's performance, but the D-study may have only one rater. The D-study includes consideration of alternative data collection strategies, as depicted in Table 3. This captures the relationship between data collection effort and raw score reliability. The "generalizability coefficient" is analogous to the Cronbach alpha, and is the reliability of relative decisions among child raw scores. The "index of dependability" is the reliability of absolute decisions relative to fixed raw scores.

G-Theory, itself, is not concerned with the substantive results of the study. The G-study enables an optimal data collection design to be hypothesized during the D-study. Data for decision-making will then be collected according to this design. These data will be reported and analyzed as uncorrected raw scores, assumed to have the same reliability as that estimated for this design from the G-study.

Under G-theory, pass-fail decisions for individual examinees are based solely on their raw scores. G-theory only provides information as to the overall level of correct classification for all examinees at all possible raw-score pass-fail points.

Rasch analysis compared to G-theory

The data is modelled to be the stochastic outcome of the logit-linear probability model:

$$\log\left(\frac{P_{norx}}{P_{norx-1}}\right) = B_n - C_r - D_o - F_x \qquad (2)$$

where

P_{norx} = the probability of child n being given by rater r on occasion o a rating of x
P_{norx-1} = the probability of child n being given by rater r on occasion o a rating of x-1
B_n = the propensity to ask for child n
C_r = the resistance to noticing asking of rater r
D_o = the difficulty of asking on occasion o
F_x = the incremental psychological obstacle overcome in asking x times compared with x-1 times.

Since the intention is to obtain child measures generalizable to any similar raters and occasions, interaction terms are not parameterized. Any interactions inflate the measurement error and are also detected by quality-control fit statistics.

Figure 2 summarizes the results of the Rasch analysis. A measure for each element (child, rater or occasion) of each facet (children, raters and occasions) is estimated and located in a common frame of reference. Tables 4, 5, and 6 list the measures, standard errors and fit statistics for the three facets.

In contrast to the G-Theory analysis, the Rasch analysis permits substantive investigation into the behavior of individual children. Table 1 suggests several such questions. Figure 2 provides the answers. Fred clearly asks significantly more often than the other children. Children asked significantly less often on the second occasion, three weeks after the first occasion. This important result, buried in a variance component in Table 2, begs for further investigation. Do the children no long need help? Have the children discovered that other children's help doesn't help? Has the teacher requested children not to ask for help?

Rasch does not have an explicit counter-part of the D-study because every G-study is also a D-study. Every data collection is for decision making. Nevertheless the results of one study may be used to guide the experimental design of another study. If the expected spread of person measures is expected to be the same in both studies, then, to estimate the reliability of the new study,

$$R_T = \frac{C_T * R_O}{C_T * R_O + C_O * (1-R_O)} \quad (4)$$

where

R_T = Target (wanted) reliability
R_O = Observed (previous) reliability
C_T = Target (wanted) average number of observations per examinee
R_O = Observed (previous) average number of observations per examinee.

or, to estimate average number of observations per examinee,

$$C_T = \frac{C_O * R_T * (1-R_O)}{(1-R_T) * R_O} \quad (5)$$

where the parameterization is the same as before.

Rasch reliability is independent of the source of the extra observations, e.g., more occasions or more raters, provided the extra observations are independently made and relevant. Consequently, the extra observations can be made in any way that is convenient. Further, there is no requirement that each examinee be observed the same number of times, or that any statistically-motivated experimental design be enacted precisely.

Unlike G-theory, Rasch does not require exactly formulated experimental designs. G-theory requires these designs in order for the complete set of variance components to be estimable. The constraints on Rasch data collection designs are that the measures be estimable unambiguously in one frame of reference or that the relationship between disjoint observations be specified. Unambiguous estimation implies that the measure of every element of every facet can be located unambiguously relative to every other element. In complete data collection designs, this requirement is always met.

Ambiguous estimates occur when there are disjoint subsets. Suppose the California raters rate the California students, but the New York raters rate the New York students. If there is no overlap in students or raters, then we do not know if the California students were more successful because they were more able or their raters more lenient. An obvious solutions to this disjunction is to share raters or students across both test sites. Alternatively, assert that, as groups, California and New York students are equally able, or California and New York raters are equally lenient. This problem is often overlooked in the raw score metric that, in this case, automatically asserts that California and New York raters are equally lenient.

Rasch analysis challenges the normality assumption that underlies G-Theory analysis. The distribution of child measures is highly skewed. This is also visible in Figure 1, which also

identifies the lack of normality in other components, but this lack of normality is not identified in the ANOVA in Table 2. Dropping the two outliers, Fred and Betty, from the analysis, improves the normality of the data, but reduces the observed child variance below the error variance in both Rasch and G-Theory analyses. In the Rasch analysis, this merely indicates how homogeneous most children are in asking. There are simply not enough observations of each child to clearly distinguish their different proclivities to ask. In the G-theory analysis, this produces components with negative variance, indicative of either model misspecification or sampling error. This incongruity reinforces the "steady state" constraint of G-theory. D-study data collection must include outliers similar to Fred and Betty or else its results will be paradoxical - a theoretical impossibility (Shavelson & Webb 1991 p. 39).

Rasch goes beyond G-Theory not only by estimating a measure and a standard error for each element, but also by reporting quality control fit statistics. An assumption of G-theory is that the observed behavior is a random sample of all behavior. "To the extent that behavior is inconsistent across occasions, generalization form the sample of behavior collected on one occasion to the universe of behavior across all occasions is hazardous" (Shavelson and Webb 1991 p.7). G-theory collects such inconsistency into variance components that are modelled to apply equally everywhere. Inconsistent behavior, however, is either local to particular observations or particular elements, or is systematic in which case it has size and direction. In this data, most children were systematically inconsistent across occasions - they asked *less* the second time. But Archie was locally inconsistent in that he asked conspicuously *more* the second time. The systematic effect is quantified by the relative Rasch measures of the two occasions. Archie's local inconsistency is detected by his fit statistics. Archie's behavior inflates the child-occasion interaction variance. What is unusual about Archie that prompts this locally inconsistent behavior? Can this behavior be expected to be observed in the D-study in order to maintain the size of the child-occasion variance?

The aim of G-theory is to estimate and optimize the reliability of decisions emanating from the data collection. For this data, the G-theory reliability is the generalizability coefficient for relative child scores reported in the D-study, Table 3, for 2 raters on two occasions. The G-theory reliability value is 0.62. The equivalent Rasch reliability coefficient is report on Table 4. The Rasch reliability value is 0.87. In both cases the variance due to rater and occasion main effects is discounted.

G-theory reliability is more conservative because G-theory parameterizes interaction variance as structural and not part of random variance. The effect of random variance reduces directly with the number of replications of all types. Interaction variance reduces only with replications of those interactions. But interaction variance cannot be expected to maintain a steady state across D-studies or even within a study. Consequently, regarding interaction as random error, except when identified otherwise by fit statistics, is the more consistent approach. Interaction identified by misfit is a local inconsistency that falls outside the framework of both G-theory and Rasch.

CONCLUSION

The view that G-theory and Rasch are alternative solutions to the same problem is seen to be an illusion. G-theory provides a general summary of a process that is hypothesized to continue into the future in exactly the same way, apart from well-controlled alterations in experimental design. G-theory has no implications for the individual examinee apart from the number of observations that will be made. For each examinee, the raw score is the measure estimate.

Rasch concentrates on the individual examinee. For each examinee, a measure is estimated that is as independent as is statistically possible of the particularities of the raters, items, tasks etc, that the examinee encountered. Rasch thus provides a measure as fair and accurate as it is possible to derive from the data that was obtained. It is on such a measure that decisions relating to each examinee must be based.

REFERENCES

Andrich, D., (1988). Rasch models for measurement. Newbury Park, Ca.:Sage.

Cronbach, L.J, Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley.

Linacre, J.M. (1989) Many-facet Rasch measurement. Chicago: MESA Press.

Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: a primer. Newbury Park, Ca.: Sage.

TABLE 1

How many times does each 9-year old child ask for help
from other children to solve math problems in class (Kenderski 1983)?

	Occasion:	First		3 weeks later		← does time matter?
Child	Rater:	Sue	Lucy	Sue	Lucy	← are raters exchangeable?
1	Archie	0	1	1	2	
2	**Betty**	3	4	1	2	← does Betty ask signif. more?
3	Charles	2	2	1	0	
4	David	1	2	0	1	
5	Ethel	1	2	2	1	
6	**Fred**	4	4	3	4	← does Fred ask signif. most?
7	George	1	1	2	1	
8	Harriet	2	2	0	0	
9	Ida	1	1	1	2	
10	**Jenny**	1	1	1	0	← does Jenny ask signif. least?
11	Kate	1	2	1	1	
12	Luke	1	2	1	1	⇑
13	Mary	2	1	1	1	**Rasch questions**

TABLE 2

Generalizability Theory Generalizability Study
as described in R. Shavelson & N. Webb,
Generalizability Theory, 1991, Newbury Park CA, Sage

			ANOVA		G-Study
Source of Variation	Variance Component	Expected Mean Square	df	Mean Square	Estimated Variance Component
13 Children (*n*)	σ^2_p	$\sigma^2_{pro,e} + n_o\sigma^2_{pr} + n_r\sigma^2_{po} + n_r n_o\sigma^2_p$	12	2.5769	0.3974
2 Raters (*r*)	σ^2_r	$\sigma^2_{pro,e} + n_p\sigma^2_{ro} + n_o\sigma^2_{pr} + n_p n_o\sigma^2_r$	1	0.6923	0.0096
2 Occasions (*o*)	σ^2_o	$\sigma^2_{pro,e} + n_p\sigma^2_{ro} + n_r\sigma^2_{po} + n_p n_r\sigma^2_o$	1	3.7692	0.1090
nr	σ^2_{pr}	$\sigma^2_{pro,e} + n_o\sigma^2_{pr}$	12	0.3590	0.0673
no	σ^2_{po}	$\sigma^2_{pro,e} + n_r\sigma^2_{po}$	12	0.8526	0.3141
ro	σ^2_{ro}	$\sigma^2_{pro,e} + n_p\sigma^2_{ro}$	1	0.3077	0.0064
nro,e	$\sigma^2_{pro,e}$	$\sigma^2_{pro,e}$	12	0.2244	0.2244

TABLE 3

Generalizability Theory Decision Study
as described in R. Shavelson & N. Webb,
Generalizability Theory, 1991, Newbury Park CA, Sage

		G Study	Alternative D Studies		
Raters =		1	2	2	4
Occasions =		1	2	4	2
Source of Variation	$\hat{\sigma}^2$	Estimated Variance Component			
Children (*n*)	$\hat{\sigma}^2_p$	0.3974	0.3974	0.3974	0.3974
Raters (*r*)	$\hat{\sigma}^2_r$	0.0096	0.0048	0.0048	0.0024
Occasions (*o*)	$\hat{\sigma}^2_o$	0.1090	0.0545	0.0273	0.0545
nr	$\hat{\sigma}^2_{pr}$	0.0673	0.0337	0.0337	0.0168
no	$\hat{\sigma}^2_{po}$	0.3141	0.1571	0.0785	0.1571
ro	$\hat{\sigma}^2_{ro}$	0.0064	0.0016	0.0008	0.0008
nro,e	$\hat{\sigma}^2_{pro,e}$	0.2244	0.0561	0.0281	0.0281
Relative (child score) error variance (nr+no+nro,e)	$\hat{\sigma}^2_{Rel}$	0.6058	0.2469	0.1403	0.2020
Generalizability coefficient	$\hat{\rho}^2$	0.40	0.62	0.74	0.66
Absolute (child score) error variance (r+o+nr+no+ro+nro,e)	$\hat{\sigma}^2_{Abs}$	0.7308	0.3078	0.1732	0.2597
Index of dependability (reliability)	$\hat{\Phi}$	0.35	0.56	0.70	0.60

TABLE 4

Child Measurement Report (Rasch)

Obsvd Score	Obsvd Count	Obsvd Average	Calib Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Child	
15	4	3.8	**3.97**	**0.98**	0.3	-1	0.2	0	**Fred**	**←Two outliers**
10	4	2.5	**1.42**	**0.65**	0.6	0	0.7	0	**Betty**	**←up S.D. from 0.59 to 1.75**
6	4	1.5	-0.68	0.83	0.9	0	1.0	0	Ethel	
5	4	1.3	-1.40	0.87	1.3	0	1.3	0	Charles	
5	4	1.3	-1.40	0.87	1.4	0	1.4	0	George	
5	4	1.3	-1.40	0.87	1.0	0	1.1	0	Ida	
5	4	1.3	-1.40	0.87	0.2	-1	0.2	-1	Kate	
5	4	1.3	-1.40	0.87	0.2	-1	0.2	-1	Luke	
5	4	1.3	-1.40	0.87	0.6	0	0.6	0	Mary	**Second time**
4	4	1.0	-2.17	0.89	**2.2**	1	**2.3**	1	**Archie**	**←asks more**
4	4	1.0	-2.17	0.89	0.7	0	0.7	0	David	**not less**
4	4	1.0	-2.17	0.89	1.8	0	1.8	0	Harriet	
3	4	0.8	-2.98	0.92	0.6	0	0.7	0	Jenny	
5.8	4.0	1.5	-1.01	0.86	0.9	-0.3	0.9	-0.2	Mean	
3.1	0.0	0.8	1.75	0.07	0.6	0.9	0.6	0.9	S.D.	

RMSE 0.87 Adj S.D. 1.52 Separation 1.75 Reliability 0.75
Fixed (all same) chi-square: 50.78 d.f.: 12 significance: .00 **←Children differ!**
Random (normal) chi-square: 11.71 d.f.: 11 significance: .39 **but only Fred & Betty, <u>not</u> others**

TABLE 5

Occasion Measurement Report (Rasch)

Obsvd Score	Obsvd Count	Obsvd Average	Calib Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Occasion
45	26	1.7	0.80	0.33	0.8	0	0.8	0	First
31	26	1.2	-0.80	0.34	1.1	0	1.1	0	3 weeks later
38.0	26.0	1.5	0.00	0.33	0.9	-0.4	0.9	-0.2	Mean
7.0	0.0	0.3	0.80	0.01	0.1	0.5	0.2	0.5	S.D.

RMSE 0.34 Adj S.D. 0.72 Separation 2.16 Reliability 0.82
Fixed (all same) chi-square: 11.29 d.f.: 1 significance: .00 ←**Occasions differ!**
Not randomly equivalent

TABLE 6

Rater Measurement Report (Rasch)

Obsvd Score	Obsvd Count	Obsvd Average	Calib Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Rater
41	26	1.6	0.35	0.33	0.9	0	0.9	0	Lucy
35	26	1.3	-0.35	0.34	0.9	0	1.0	0	Sue
38.0	26.0	1.5	0.00	0.33	0.9	-0.3	0.9	-0.2	Mean
3.0	0.0	0.1	0.35	0.00	0.0	0.1	0.0	0.1	S.D.

RMSE 0.33 Adj S.D. 0.09 Separation 0.26 Reliability 0.06 **But**
Fixed (all same) chi-square: 2.13 d.f.: 1 significance: .14 ←**Raters exchangeable!**

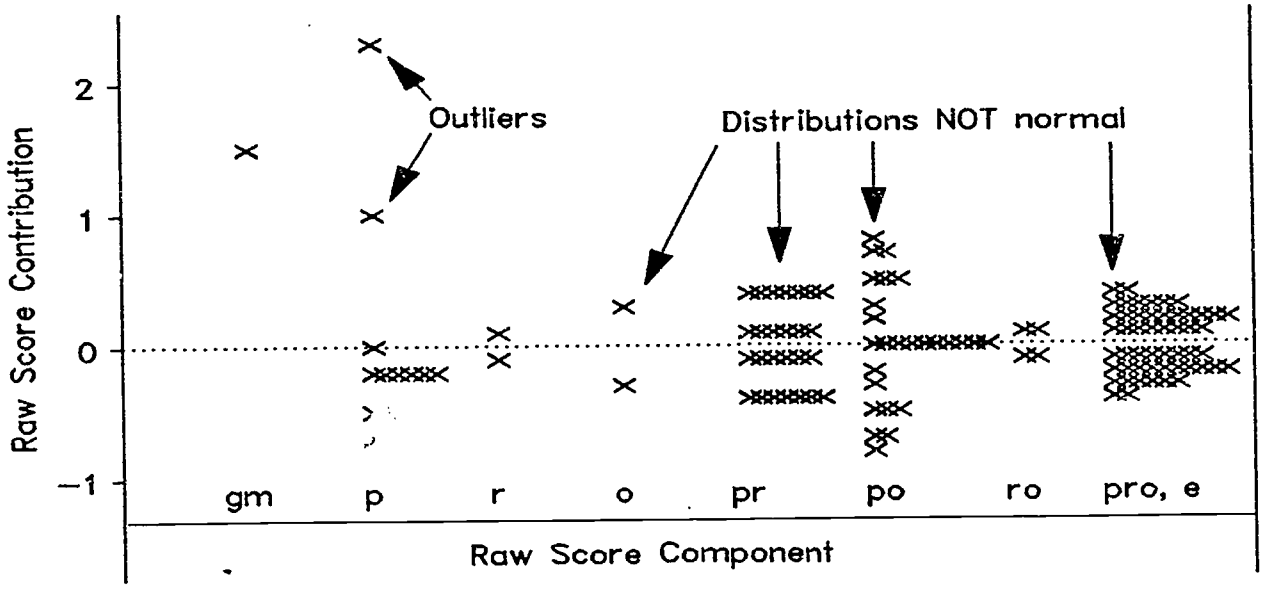

Figure 1. Decomposition of raw scores according to the G-Theory model.

All Facet Summary (Rasch) Non-linear scale ↓

Measure	+Child	Total Asks	+Rater	+Occasion	Times
4	+ Fred	*Asks most:* 15	+*Notices more*	*More asking*	+(4)
	| *1 S.E.*		|	|	|
3	+ ↓		+	+	+ 3.5
	|		|	|	|
2	+ ↑ *1 S.E.*		+	+	+ 3
	| Betty	10	|	|	| 2.5
1	+ ↓ *1 S.E.*		+	+ First	+
	|		| Lucy	↓ *1 S.E.*	| 2
0	* - ↑ *1 S.E.*- - - - - - - - - - - - - - -		* - - - - - - *	- - - - - -*	
	| Ethel	6	| Sue	↑ *1 S.E.*	| 1.5
-1	+ ↑ *1 S.E.*		+	+ Later	+
	| Charles George Ida Kate Luke Mary	5	|	|	|
-2	+ Archie David Harriet	4	+	+	+ 1
	|		|	*Less*	|
-3	+ Jenny	*Asks least:* 3	+*Notices less*+	*asking*	+(0)
	↑		↑	↑	Mean
	Fred - SE=.98 significantly most		Raters the same	Later less	
	Betty - SE=.65 significantly more		statistically	significantly	
	Others - SE=.9 much the same		SE=.35	SE=.35	

Figure 2. Estimated measures according to the many-facet Rasch model.