The Key to Objective Measurement

Geoff N Masters
Australian Council for Educational Research[1]

In his 1960 book *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish mathematician Georg Rasch introduced a statistical model for the analysis of test data. This model is now widely referred to as the 'Rasch Model'.

Rasch recognised that his model for tests had a remarkable property. Convinced of the significance of this property for the improvement of measurement in the social and behavioural sciences, Rasch spent the remaining twenty years of his life investigating this property and its implications.

The property that Rasch recognised can be understood by considering two persons *A* and *B* with imagined abilities β_A and β_B. If these two persons attempt a set of test items, and a tally is kept of the number of items N_{10} that person *A* answers correctly but *B* answers incorrectly, and of the number of items N_{01} that person *B* answers correctly but *A* answers incorrectly, then under the Rasch model, the difference β_A-β_B in the abilities of these two persons can be estimated as:

$$\ln (N_{10} / N_{01}).$$

What is remarkable about this fact is that this relationship between the parameterised difference β_A-β_B and the tallies N_{10} and N_{01} of observed successes and failures applies to *any* selection of items when test data conform to Rasch's model. In other words, provided that the responses of persons *A* and *B* to a set of items are consistent with the model, the difference β_A-β_B can be estimated by simply counting successes and failures *without having to know or estimate the difficulties of the items involved*. Any subset of items (eg, a selection of easy items, hard items, even-numbered items, odd-numbered items) can be used to obtain an estimate of the relative abilities of persons *A* and *B* from a simple tally table:

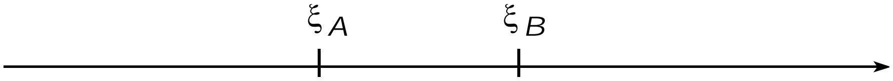
		Person B	
		Wrong	Right
Person A	Right	$\mathbf{N_{10}}$	N_{11}
	Wrong	N_{00}	$\mathbf{N_{01}}$

[1] 19 Prospect Hill Road, Camberwell, Victoria, Australia 3124 email: masters@acer.edu.au
October 2001. www.rasch.org/key.pdf

The possibility of obtaining an estimate of the relative abilities of persons *A* and *B* that is not dependent upon the details of the items used is referred to here as the possibility of objective comparison. The possibility of objective comparison is the key to objective measurement.

The Rasch Model

In its most general form, the Rasch model begins with the idea of a measurement variable upon which two objects *A* and *B* have imagined locations ξ_A and ξ_B:



The possibility of estimating the relative locations of objects *A* and *B* on this variable depends on the availability of two observable events:

- an event X indicating that ξ_B exceeds ξ_A
- an event Y indicating that ξ_A exceeds ξ_B

The Rasch model relates the difference between objects *A* and *B* to the events X and Y that they govern:

$$\xi_B - \xi_A \quad = \quad \ln (P_X/P_Y) \qquad [1]$$

where P_X is the probability of observing X and P_Y is the probability of observing Y.

Notice that, under the model, the odds P_X/P_Y of observing X rather than Y is dependent only on the direction and distance of ξ_B from ξ_A, and is uninfluenced by any other parameter.

Estimating $\xi_B - \xi_A$

An estimate of the difference between objects *A* and *B* on the measurement variable can be obtained if there are multiple independent opportunities to observe either event X or event Y. Under these circumstances, $\xi_B - \xi_A$ can be estimated as:

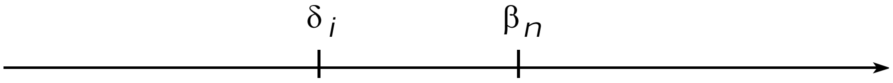
$$\ln (p_X/p_Y) = \ln (N_X/N_Y)$$

where p_X and p_Y are the proportions of occurrences of X and Y, and N_X and N_Y are the numbers of times X and Y occur in N_X+N_Y observation opportunities.

> The term 'objectivity' refers to the fact that the result of any comparison of two objects within some specified frame of reference is… *independent of everything else within the frame of reference other than the two objects which are to be compared and their observed reactions*. (Rasch, 1977; p 77, italics in original)

Application #1: Dichotomous Test Items

The most common application of Rasch's model is to tests in which responses to items are recorded as either wrong (0) or right (1). Each person *n* is imagined to have an ability β_n and each item *i* is imagined to have a difficulty δ_i, both of which can be represented as locations on the variable being measured:



In this case, observable event X is person *n*'s success on item *i*, and observable event Y is person *n*'s failure on item *i*:

$\xi_B - \xi_A$	**observation opportunity**	**observable event**	
		X	**Y**
$\beta_n - \delta_i$	person *n* attempts item *i*	1	0

The Rasch model applied to this situation is:

$$\beta_n - \delta_i \quad = \quad \ln (P_1/P_0) \qquad [2]$$

Estimating $\beta_n - \delta_i$

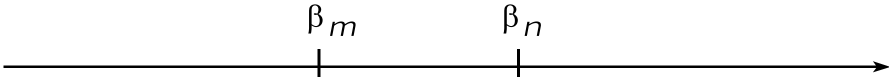
If person *n* could have multiple independent attempts at item *i*, then the difference $\beta_n - \delta_i$ between person *n*'s ability and item *i*'s difficulty could be estimated as:

$$\ln (p_1/p_0) \ = \ \ln (N_1/N_0).$$

Although this is true in theory, and this method could be useful in some situations, it is not a practical method for estimating $\beta_n - \delta_i$ from test data because test takers are not given multiple attempts at the same item (and if they were, they would not be independent attempts). To estimate the difference $\beta_n - \delta_i$ from test data, it is necessary to estimate β_n from person *n*'s attempts at a number of items, and to estimate δ_i from a number of persons' attempts at that item. In other words, the difficulties of a number of test items and the abilities of a number of test takers must be estimated simultaneously.

Comparing and measuring persons

In the application of Rasch's model to tests, every person has an imagined location on the variable being measured. Two persons *m* and *n* have imagined locations β_m and β_n:



It follows from [2] that, if persons *m* and *n* attempt the same item and their attempts at that item are independent of each other, then the modelled difference between persons *n* and *m* is:

$$\beta_n - \beta_m = \ln (P_{10}/P_{01}) \qquad [3]$$

where P_{10} is the model probability of person *n* succeeding but *m* failing the item, and P_{01} is the probability of person *m* succeeding but *n* failing that item.

It can be seen that [3] is Rasch's model [1] applied to the comparison of two persons on a measurement variable. The two observable events involve the success of one person but failure of the other in their attempts at the same item:

		observable event	
$\xi_B - \xi_A$	**observation opportunity**	**X**	**Y**
$\beta_n - \beta_m$	persons *n* and *m* independently attempt the same item	1,0	0,1

In this comparison of persons *m* and *n*, nothing was said about the difficulty of the item being attempted by these two persons. This is because [3] applies to every item. The odds of it being person *n* who succeeds, given that one of these two persons succeeds and the other fails, is *the same for every item* and depends only on the relative abilities of persons *m* and *n*.

Estimating $\beta_n - \beta_m$

Because the modelled odds P_{10}/P_{01} are the same for every item, the difference $\beta_n - \beta_m$ can be estimated as:

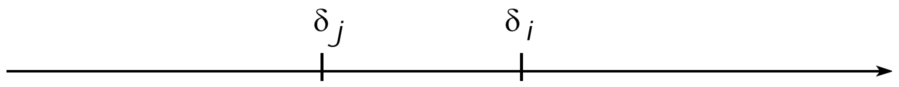
$$\ln (N_{10}/N_{01})$$

where N_{10} is the number of items that person *n* has right but *m* has wrong, and N_{01} is the number of items that person *m* has right but *n* has wrong.

When test data conform to the Rasch model, the relative abilities of two persons can be estimated in this way using any selection of items without regard to their difficulties (or any other characteristics). By making multiple pairwise comparisons of this kind, it is possible to estimate the relative locations of a number of persons on the same measurement variable.

Comparing and calibrating items

In the application of Rasch's model to tests, every item has an imagined location on the variable being measured. Two items *i* and *j* have imagined locations δ_i and δ_j:



It follows from [2] that, if items *i* and *j* are attempted by the same person and this person's attempts at items *i* and *j* are independent of each other, then the modelled difference between items *i* and *j* is:

$$\delta_i - \delta_j = \ln(P_{01}/P_{10}) \qquad [4]$$

where P_{10} is the model probability of the person succeeding on item *i* but failing item *j*, and P_{01} is the probability of the person succeeding on item *j* but failing item *i*.

It can be seen that [4] is Rasch's model [1] applied to the comparison of two items on a measurement variable. The two observable events involve the person's success on one item but failure on the other:

		observable event	
$\xi_B - \xi_A$	**observation opportunity**	**X**	**Y**
$\delta_i - \delta_j$	items *i* and *j* independently attempted by the same person	0,1	1,0

In this comparison of items *i* and *j*, nothing was said about the ability of the person attempting them. This is because [4] applies to every person. The odds of success on item *i* given success on one item but failure on the other is *the same for every person* and depends only on the relative difficulties of items *i* and *j*.

Estimating $\delta_i - \delta_j$

Because the modelled odds P_{01}/P_{10} are the same for every person, the difference $\delta_i - \delta_j$ can be estimated as:

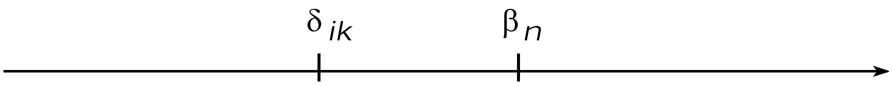
$$\ln(n_{01}/n_{10})$$

where n_{10} is the number of persons with item *i* right but *j* wrong, and n_{01} is the number of persons with *j* right but *i* wrong.

When test data conform to the Rasch model, the relative difficulties of two items can be estimated in this way using any group of persons without regard to their abilities (or any other characteristics). By making multiple pairwise comparisons of this kind, it is possible to estimate the relative locations of a number of items on the measurement variable.

Application #2: Polytomous Test Items

We consider now the application of Rasch's model to tests in which responses to items are recorded in several ordered categories labelled 0, 1, 2, … K_i. Each person *n* is imagined to have an ability β_n and each item *i* is imagined to have a set of K_i parameters δ_{i1}, δ_{i2}, … δ_{iK_i} each of which can be represented as a location on the variable being measured. For example,



where δ_{ik} governs the probability of scoring *k* rather than *k-1* on item i:

$\xi_B - \xi_A$	**observation opportunity**	**observable event**	
		X	**Y**
$\beta_n - \delta_{ik}$	person *n* attempts item *i*	k	k-1

The Rasch model applied to this situation (Masters, 1982) is:

$$\beta_n - \delta_{ik} \quad = \quad \ln (P_k / P_{k-1}) \qquad [5]$$

In polytomous test items, objective comparison (and thus objective measurement) continues to depend on the modelling of the relationship between two imagined locations on the variable and two observable events. This comparison is 'independent of everything else within the frame of reference'—including other possible outcomes of the interaction of person *n* with item *i*. The conditioning out of other possible outcomes to focus attention only on the two observable events that provide information about the relative locations of the two parameters of interest is a fundamental feature of Rasch's model.

The conditioning on a pair of adjacent response alternatives has parallels with McFadden's (1974) assumption that a person's probability of choosing to travel by car *rather than* by bus should be independent of the availability of other options (eg, train). McFadden refers to this as the assumption of 'independence from irrelevant alternatives'. In a similar way, it is assumed in this application of the Rasch model that a person's probability of choosing or scoring *k rather than k-1* is independent of all other possible outcomes.

Estimating $\beta_n - \delta_{ik}$

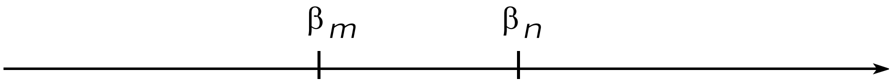
As for dichotomously-scored items, if person *n* could have multiple independent attempts at item *i*, then the difference $\beta_n - \delta_{ik}$ could be estimated from proportions or counts of occurrences of *k* and *k-1*:

$$\ln (p_k / p_{k-1}) \ = \ \ln (N_k / N_{k-1}).$$

However, because multiple independent attempts at test items usually are not possible, this method is not feasible in practice.

Comparing and measuring persons

In the application of Rasch's model to tests in which responses to items are recorded in several ordered categories, every person has an imagined location on the variable being measured:



It follows from [5] that, if persons *m* and *n* attempt the same item and their attempts at that item are independent of each other, then the modelled difference between persons *n* and *m* is:

$$\beta_n - \beta_m = \ln\left(P_{k,k-1}/P_{k-1,k}\right) \qquad [6]$$

where $P_{k,k-1}$ is the model probability of person *n* scoring *k* but *m* scoring *k-1*, and $P_{k-1,k}$ is the probability of person *m* scoring *k* but *n* scoring *k-1* on that item.

It can be seen that [6], which applies for all values of *k* (k=1, 2, …K_i), is Rasch's model [1]:

$\xi_B - \xi_A$	**observation opportunity**	**observable event**	
		X	**Y**
$\beta_n - \beta_m$	persons *n* and *m* independently attempt the same item	k,k-1	k-1,k

If one of persons *m* and *n* scores *k* on an item, and the other scores *k-1*, then the probability of it being person *n* who scores *k* is *the same for every item* and depends only on the relative abilities of persons *m* and *n*.

Estimating $\beta_n - \beta_m$

Because the modelled odds $P_{k,k-1}/P_{k-1,k}$ are the same for every item, the difference $\beta_n - \beta_m$ can be estimated as:

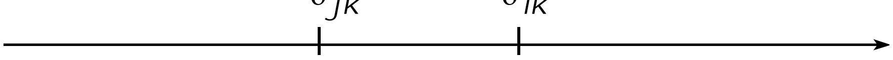
$$\ln\left(N_{k,k-1}/N_{k-1,k}\right)$$

where $N_{k,k-1}$ is the number of items on which person *n* scores *k* and *m* scores *k-1*, and $N_{k-1,k}$ is the number of items on which person *m* scores *k* and *n* scores *k-1*.

Once again, when test data conform to the Rasch model, the relative abilities of two persons can be estimated in this way using *any* selection of items. And by making multiple pairwise comparisons of this kind, it is possible to estimate the relative locations of a number of persons on the measurement variable.

Comparing and calibrating items

In polytomous items, each item parameter δ_{ik} (k=1,2,…K_i) is a location on the variable being measured. The parameters δ_{ik} and δ_{jk} from two different items *i* and *j* can be compared on this variable:



It follows from [5] that, if items *i* and *j* are attempted by the same person and this person's attempts at items *i* and *j* are independent of each other, then the modelled difference between parameters δ_{ik} and δ_{jk} is:

$$\delta_{ik} - \delta_{jk} = \ln (P_{k-1,k}/P_{k,k-1}) \qquad [7]$$

where $P_{k,k-1}$ is the probability of the person scoring *k* on item *i* but *k-1* on item *j*, and $P_{k-1,k}$ is the probability of the person scoring *k* on item *j* but *k-1* on item *i*.

It can be seen that [7], which applies for all values of *k* (*k*=1, 2, …K_i), is Rasch's model [1]:

$\xi_B - \xi_A$	**observation opportunity**	**observable event**	
		X	**Y**
$\delta_{ik} - \delta_{jk}$	items *i* and *j* independently attempted by the same person	k-1,k	k,k-1

In this comparison of items *i* and *j*, nothing was said about the ability of the person attempting them. This is because [7] applies to every person. When a person attempts items *i* and *j*, the probability of the person scoring *k* on item *i* given that they score *k* on one item and *k-1* on the other is *the same for every person*.

Estimating $\delta_{ik} - \delta_{jk}$

Because the modelled odds $P_{k-1,k}/P_{k,k-1}$ are the same for every person, the difference $\delta_{ik} - \delta_{jk}$ can be estimated as:

$$\ln (n_{k-1,k}/n_{k,k-1})$$

where $n_{k,k-1}$ is the number of persons scoring *k* on item *i* but *k-1* on item *j*, and $n_{k-1,k}$ is the number of persons scoring *k* on item *j* but *k-1* on item *i*.

When test data conform to Rasch's model, the difference $\delta_{ik} - \delta_{jk}$ can be estimated in this way using *any* group of persons without regard to their abilities (or any other characteristics).

Obstacles to Objectivity #1: 'Discrimination'

In some models proposed for the analysis of test data, in addition to a location β_n for each person *n* and a location δ_i for each item *i*, a 'discrimination' parameter α_i is proposed for each item *i*. An example of such a model is the 2-parameter Item Response Theory (IRT) model:

$$\alpha_i (\beta_n - \delta_i) = \ln (P_1/P_0) \qquad [8]$$

If we follow the steps outlined earlier and consider independent attempts of two persons *m* and *n* at item *i*, then for the 2-parameter IRT model we obtain:

$$\alpha_i (\beta_n - \beta_m) \quad = \quad \ln (P_{10}/P_{01}) \qquad [9]$$

where P_{10} is the probability of person *n* succeeding but *m* failing item *i*, and P_{01} is the probability of person *m* succeeding but *n* failing.

It can be seen from [9] that the odds of person *n* succeeding but *m* failing given that one of these two persons succeeds and the other fails is *not* the same for all items. Rather, the odds depend on the discrimination of the item in question.

To compare the locations of persons *m* and *n* on the measurement variable, it is not possible to ignore the particulars of the items involved and simply tally occurrences of (1,0) and (0,1). The comparison of β_n and β_m on the measurement variable is dependent not only on the two observable events (1,0) and (0,1) that they govern, but also on the details (viz, the discriminations) of the items these two persons take. For this reason, the 2-parameter IRT model does not permit 'objective' comparison in the sense described by Rasch (1977).

Because it is not possible to make an objective comparison of any two persons under the 2-parameter IRT model, it is not possible to condition the item details out of the procedure for estimating the locations of a number of persons on the measurement variable. In other words, the inability to make objective pairwise comparisons of individuals rules out the possibility of making objective measures on that variable.

Obstacles to Objectivity #2: Cumulative 'Thresholds'

In models for polytomously-scored items it is common to introduce the notion of a 'threshold' parameter, intended to divide all ordered response alternatives to an item up to and including alternative *k-1* from response alternatives *k* and above. Thurstone also referred to thresholds as 'category boundaries' (eg, Edwards and Thurstone, 1952).

The cumulative threshold notion is used as the basis for Samejima's (1969) model[1]:

$$\beta_n - \gamma_{ik} \quad = \quad \ln [(P_k+P_{k+1}+ \ldots P_{K_i}) / (P_0+P_1+\ldots P_{k-1})]$$

In this model, the item threshold γ_{ik} governs the probability of scoring *k or better* on item *i*. Figure 1 shows Samejima's Graded Response Model for an item with four ordered response alternatives labelled 0, 1, 2 and 3.

From Figure 1 it can be seen that the observable events in this model are *compound* events.
eg, event X: response in category 1 <u>or</u> 2 <u>or</u> 3
event Y: response in category 0.

The consequence is that the elementary equations in this model are not independent because $(P_1+P_2+P_3)/P_0 > (P_2+P_3)/(P_0+P_1) > P_3/(P_0+P_1+P_2)$. As a result, cumulative 'thresholds' are not independent, but are always ordered $\gamma_{i1} < \gamma_{i2} < \gamma_{i3}$.

[1] Samejima's model also includes an item discrimination parameter α_i which has been ignored here for simplicity.

The elementary equations in Samejima's model lead to the following expressions for the probabilities of person *n* scoring 0, 1, 2 and 3 on item *i*:

$$
\begin{aligned}
P_{ni0} &= 1 - \exp(\beta_n-\gamma_{i1})/[1+\exp(\beta_{\mathrm{n}}-\gamma_{i1})] \\
P_{ni1} &= \exp(\beta_n-\gamma_{i1})/[1+\exp(\beta_n-\gamma_{i1})] - \exp(\beta_n-\gamma_{i2})/[1+\exp(\beta_n-\gamma_{i2})] \\
P_{ni2} &= \exp(\beta_n-\gamma_{i2})/[1+\exp(\beta_n-\gamma_{i2})] - \exp(\beta_n-\gamma_{i3})/[1+\exp(\beta_n-\gamma_{i3})] \\
P_{ni3} &= \exp(\beta_n-\gamma_{i3})/[1+\exp(\beta_n-\gamma_{i3})]
\end{aligned}
$$

It is not possible to condition one set of parameters (either the person parameters or the item thresholds) out of the estimation procedures for the other in this model.

In contrast, the elementary equations for the Rasch model (see Figure 1) lead to the following expressions for the probabilities of person *n* scoring 0, 1, 2 and 3 on item *i*:

$$
\begin{aligned}
P_{ni0} &= 1 \;/\; \Psi \\
P_{ni1} &= \exp\,(\beta_n-\delta_{i1}) \;/\; \Psi \\
P_{ni2} &= \exp\,(2\beta_n-\delta_{i1}-\delta_{i2}) \;/\; \Psi \\
P_{ni3} &= \exp\,(3\beta_n-\delta_{i1}-\delta_{i2}-\delta_{i3}) \;/\; \Psi \qquad \text{where } \Psi \text{ is the sum of the numerators.}
\end{aligned}
$$

It is possible to condition the person parameters out of the estimation procedures for the item parameters, and vice versa, in this model.

Summary

When test data conform to Rasch's model it is possible to make 'objective comparisons' in the sense that the comparison of two person parameters does not depend on the details of the items used to make that comparison, and the comparison of two item parameters does not depend on the details of the persons taking those items. The possibility of objective comparison is the key to objective measurement. Estimates of the locations of a number of individuals on a measurement variable can be obtained by bringing together and rationalising the results of multiple pairwise comparisons (see Choppin, 1968).

In this paper it has been seen that:

- each parameter in the Rasch model is a *location* on an intended measurement variable;
- the model always is expressed in terms of the *difference* between two location parameters ($\xi_B - \xi_A$);
- the model requires two *observable events* X and Y, one indicating that ξ_B exceeds ξ_A, the other indicating that ξ_A exceeds ξ_B;
- the model specifies how the difference ($\xi_B - \xi_A$) governs the probability of observing X *rather than* Y; and
- to do this, the model *conditions out* all other parameters and all other observable events by considering the odds of X rather than Y (or in an alternative formulation, the conditional

probability of X given either X or Y). In the case of polytomous items, all other parameters and response categories *within the same item* are conditioned out of consideration.

In this paper, the application of Rasch's model to dichotomous test items [2] was presented separately because this is the best known application of the model. However, the application of Rasch's model to polytomous items [5] includes the model's application to dichotomous items (when K_i =1).

Finally, two sets of parameters sometimes included in models for test data—item discriminations and cumulative item thresholds—were shown to be obstacles to objective comparison and thus objective measurement as defined by Rasch (1977).

References

Choppin, BH (1968). An item bank using sample-free calibration. *Nature, 219*, 870-72.

Edwards, AL and Thurstone, LL (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17*, 169-80.

Masters, GN (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149 - 74.

McFadden, D (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarempka (Ed.) *Frontiers in Econometrics*. New York: Academic Press.

Rasch, G (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedogogiske Insitut.

Rasch, G (1977) On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Samejima, F (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement No. 17.*

	Thurstone/Samejima	**Rasch**
elementary equations (person n, item i, K_i=3)	$\beta_n-\gamma_{i1} = \ln [(P_1+P_2+P_3) / P_0]$ $\beta_n-\gamma_{i2} = \ln [(P_2+P_3) / (P_0+P_1)]$ $\beta_n-\gamma_{i3} = \ln [P_3 / (P_0+P_1+P_2)]$	$\beta_n-\delta_{i1} = \ln [P_1 / P_0]$ $\beta_n-\delta_{i2} = \ln [P_2 / P_1]$ $\beta_n-\delta_{i3} = \ln [P_3 / P_2]$
events being compared	compound (eg, response in category 1 or 2 or 3 rather than 0)	simple (comparison of adjacent response categories)
item parameters	global / unconditional each γ relates to all available response categories	local / conditional each δ relates to adjacent response categories only
relationship of elementary equations	dependent $(P_1+P_2+P_3) / P_0 >$ $(P_2+P_3) / (P_0+P_1) >$ $P_3 / (P_0+P_1+P_2)$	independent (eg, odds of response in category 1 rather than 0 is independent of odds of response in category 2 rather than 1)
implications for item parameters	$\gamma_{i1} < \gamma_{i2} < \gamma_{i3}$	δs are unfettered and free to take any value
model for ordered categories	When brought together, the elementary equations provide a model for ordered response categories in which the person parameters cannot be conditioned out of the estimation procedure for the items.	The elementary equations provide a model for ordered response categories in which the person parameters can be conditioned out of the estimation procedure for the items and vice versa.
objective measurement	not possible	possible

Figure 1. Comparison of Thurstone/Samejima and Rasch Models for Polytomous Items