Polytomous IRT Analysis

By Abdullah Alfarwan

Introduction

We are moving from dichotomously (two category) scored items to polytomously scored items. Items that are scored in multiple-ordered categories are known as polytomously scored items. The probability of an examinee favoring a specific score category can be described by one of the polytomous IRT models, which can be the partial credit model (PCM), its' generalized partial credit model (GPCM), and/or the graded response model (GRM). The PCM requires equality constraint on slope, but in the GPCM different item discrimination is allowed. These polytomous models are generalized from dichotomous IRT models (Tang & Service, 1995). Four items from the Spirituality test are being examined. These are personal attitude items, so it intends to try to measure examinees latent traits(spirituality) and locate which category the examinee falls on, based on the ordered scale responses from 0 to 5. I chose four items (134 = I have a religious practice in my life, 135 = I have a spiritual practice in my life, 136 = I feel the presence of a higher power, 137 = My spirituality or religion brings me comfort) to use in this analysis. Three models (Partial Credit Model, Generalized Partial Credit Model, and Graded Response Model) were used to estimate item parameters in this analysis. The estimation method used with all the models is the Marginal Maximum Likelihood method. This means that it is sample dependent. Furthermore, each one of the five the items 134, 135, 136, 137, & 138 has 6 levels of response options consisting of (0 1 2 3 4 5), respectively.

Model Description

Partial Credit Model

This model is common in proficiency assessment. There are More than two response categories where each category represents a "step" towards a fully (corrected/endorsed) response. Partially

correct/endorsed responses can provide information for estimating a person's location. Progress to a fully correct/endorsed response must move sequentially through all prior steps (i.e., you cannot go from step 0 to step 2, progress is 0, 1, 2).

PCM Modeling

This model decomposes responses into a series of ordered pairs of adjacent categories (category scores) and successively applies a dichotomous model to each pair. This model assumes there is a point on the latent continuum below which an individual provides a response (e.g., x = 0 and above which they provide the next higher response, e.g., x = 1). This point then represents the transition point (h) from one score category to the next score category. A polytomous scored item has multiple ordered response categories with adjacent categories, each separated by a transition point.

Transition Location Parameters

Each transition point, h, on the latent continuum for item j is represented by delta. For a 2-step problem with x = (0,1,2) category scores, the two transition points are delta and delta.

Aggregating Separate Probabilities

Next, the separate probabilities are aggregated into their ordinal relationship. If an individual fails the transition from x=0 to x=1 and the probability of this event is 0. Dividing each of the three terms above by phi (the sum of three mutually exclusive outcome) gives the probability category scores (0, 1, 2)

PCM

The PCM specifies the conditional probability that an examinee with latent location θ obtains a category score x_j. δ_{jh} is a *transition location parameter* for item *j* and reflects the relative difficulty (step difficulty) of endorsing category *h* over category *h*-1. Notice that the category index *m* has a *j* subscript, e.g., m_j; thus different items can have different numbers of categories. δ_{jh} are conditionally defined and cannot be interpreted as independent pieces of an item. It is possible that δ_{j1} is more difficult that δ_{j2}, e.g., δ_{jh} do not need to be in increasing order. In the PC Model, the response categories must be ordered, this does NOT mean the transition locations must be ordered. The second transition is "easier" so the δ_{jh} are

not in increasing order. The PC Model is predicated on a unidimensional latent construct and all the items discriminate among respondents to the same degree.

Generalized Partial Credit Model

The GPCM model relaxes the assumption in the PCM that all items on the instrument have equal discriminations. Muraki (1992) did this by extending Master's (1982) Rasch-based PC model by moving to the 2PL model version, where α_j is the item discrimination, δ_{jh} is the transition location parameter between the h^{th} and h^{th}-1 category and $\delta_{j1}$is set to zero (0), delta's do not have to be in sequential order, and m_j categories (m_j-1 transitions) and k=(1, … , m_j). Modeling the probability of providing a response in item's k^{th} category, x_{jk} given θ(theta), α_j (item discrimination), δ_{jh} (item difficulty transition parameter between h^{th} and h-1 category). There are m category points, m-1 transition points. There can be category reversals, eg., δ_{jh} do not need to follow sequential order

Graded Response Model

Graded Response Model (GRM)

Samejima developed an alternative to the Rasch-based PCM & RSM and the 2PL GPC model based on Thurstone's method of successive intervals. The GR model focuses on establishing a boundary above which a person is expected to obtain certain category score(s) (as opposed to lower category score(s)). For example, given a multistep problem: (6/3) + 2 = X. We can determine the probability of a score of 1 or higher versus a score of 0, or the probability of a score of 2 versus a score of 1 or 0. This has the effect of turning the polytomous score into a series of cumulative responses below a particular category versus *at or above* this category.

2PL GRM

This version of the GRM model specifies the probability of a person responding with a category score, x_j *or higher* versus responding in a lower category score. The 2 PL GR model can be used for both rating-scale data (SA, A, D, SD) and partial credit data. Higher categories represent more of the latent trait. The probability of obtaining a category score (x_j) or higher, where α_j is item discrimination, δ_{xj} is the category boundary location for category score x_j. This model allows for different number of categories

across different items, where δ_{xj} are always in increasing order (no reversals). To calculate the probability of an individual obtaining a particular category score x_j or responding in a particular category k, (p_k) we need to calculate the difference between two cumulative probabilities for adjacent categories.

Results

It can be seen from the Table 1 that the items are unidimensional scale with the highest value of the first raw (item 1) of Eigenvalues = 4.03 and the rest of the raw are lower than 1, which means they are presenting just one construct, which is spirituality. So, this is one of the indications of the test and the item's reliability.

Table 1.

Eigenvalues of the Polychoric Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.03871739	3.70717052	0.8077	0.8077
2	0.33154687	0.02859199	0.0663	0.8741
3	0.30295488	0.10848435	0.0606	0.9346
4	0.19447054	0.06216022	0.0389	0.9735
5	0.13231032		0.0265	1.0000

Regarding the item's response options frequencies, it can be seen from Tables 2 through 5 that response option 0 was endorsed by the lowest percentage of the examinees, response options 1 through 4 had almost the same response percentage of examinee endorsement, and response option 5 has the highest endorsement from the examinees. This indicates that the majority of the examinees have the highest traits (theta value) above zero to endorse the items toward agree and strongly agree if we presume that the items options are (0 = not at all to 5 = yes at all).

Table 2.

Item 134 Response Options Frequencies.

hwa134	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	42	14.00	42	14.00
1	36	12.00	78	26.00
2	47	15.67	125	41.67
3	34	11.33	159	53.00
4	48	16.00	207	69.00
5	93	31.00	300	100.00

Table 3.

Item 135 Response Options Frequencies.

hwa135	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	29	9.67	29	9.67
1	40	13.33	69	23.00
2	51	17.00	120	40.00
3	37	12.33	157	52.33
4	50	16.67	207	69.00
5	93	31.00	300	100.00

Table 4.

Item 136 Response Options Frequencies.

hwa136	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21	7.00	21	7.00
1	26	8.67	47	15.67
2	63	21.00	110	36.67
3	43	14.33	153	51.00
4	56	18.67	209	69.67
5	91	30.33	300	100.00

Table 5.

Item 137 Response Options Frequencies.

hwa137	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	27	9.00	27	9.00
1	20	6.67	47	15.67
2	44	14.67	91	30.33
3	46	15.33	137	45.67
4	69	23.00	206	68.67
5	94	31.33	300	100.00

Item Statistics

From Table 6, it can be seen that all the items have a good mean of about 3, which is higher than the mean value = 2.5 if each of the response options has the same percentage of response. That tells us that we are slightly positive skewed toward the right on the x-axis (high ability). Also, each one of the items has a strong correlation to the total test items and have values of 0.81, 0.85, 0.75, 0.77, 0.77, respectively. The total test items correlation coefficient is strong (0.92).

Table 6.

Item Statistics.

		Item-Total Correlations	
Item	Mean	Unadjusted	Adjusted
hwa134	2.96333	0.88958	0.81389
hwa135	3.06000	0.91173	0.85405
hwa136	3.20000	0.84290	0.75700
hwa137	3.30667	0.85834	0.77831
hwa138	3.27667	0.85231	0.77164
Total N=300, Cronbach's Alpha=0.9203			

Partial Credit Model

There are six response categories where each category represents a "step" towards a fully (corrected/endorsed) response. Partially correct/endorsed responses can provide information for estimating a person's location. From Table 7, the step point represents the transition point (h) from one score category to the next score category. Progress to a fully correct/endorsed response are moving sequentially through all prior steps (there is no movement from step 1 to step 3, progress is 1, 2, 3). Step 1 has the lowest value of parameter estimate (-0.93), which means all the examinees who endorsed 0 are located in the lowest or the first category which is below the first threshold of -0.93 and other examinees have traits that are higher than -0.93 and are located in the other category. They may be in the second one or in the fifth one depending on their level of trait. When the level of trait is higher, students with the same amount if highest located there. Also, we can see there are standard error for each one of the parameter step estimates, when the standard error gets lower and close to zero, then the estimate is more accurate. Also, it can be seen each polytomous scored item has multiple ordered response categories with adjacent categories, each separated by a transition point from – 0.93 to +23. Furthermore, the slope of each item is the same = 1.81. The PC Model is predicated on a unidimensional latent construct and all the items discriminate among respondents to the same degree(1.81).

Table 7.

Item Parameter Estimates for PCM.

Item	Parameter	Estimate	Standard Error
hwa134	Step 1	-0.93584	0.15689
	Step 2	-0.70975	0.14581
	Step 3	-0.05670	0.14354
	Step 4	-0.08651	0.14336
	Step 5	0.23823	0.12563
	Slope	1.81116	0.12410
hwa135	Step 1	-1.32717	0.17366

Item	Parameter	Estimate	Standard Error
	Step 2	-0.76975	0.14231
	Step 3	-0.09990	0.13920
	Step 4	-0.08553	0.13988
	Step 5	0.25107	0.12462
	Slope	1.81116	0.12410
hwa136	Step 1	-1.44441	0.20320
	Step 2	-1.25117	0.16218
	Step 3	-0.14400	0.13106
	Step 4	-0.10647	0.13326
	Step 5	0.31481	0.12247
	Slope	1.81116	0.12410
hwa137	Step 1	-1.11777	0.19683
	Step 2	-1.22598	0.17822
	Step 3	-0.44598	0.13793
	Step 4	-0.25941	0.12760
	Step 5	0.35510	0.11734
	Slope	1.81116	0.12410

Item Characteristics Curve

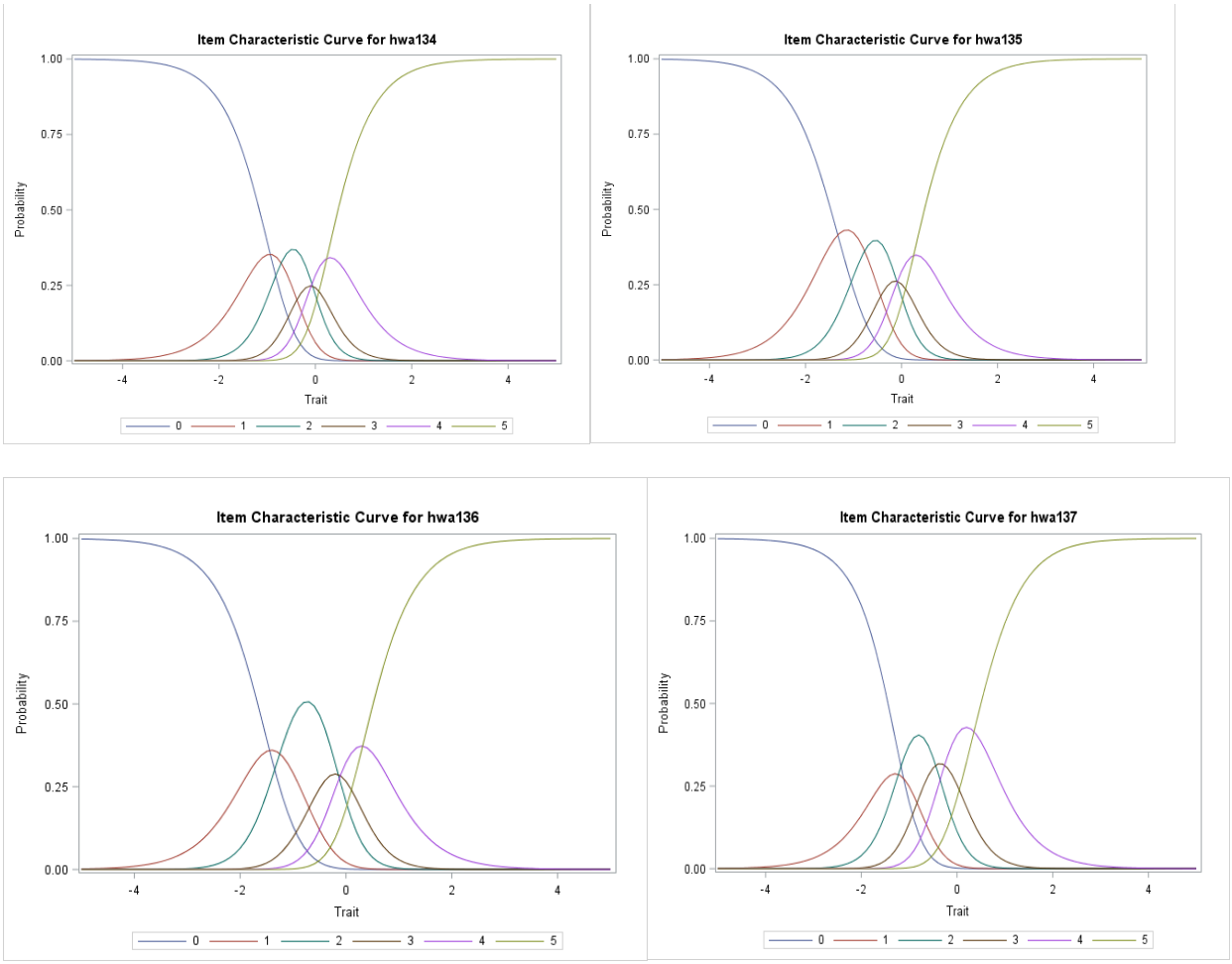
The ICCs illustration presenting visualization forms of what has been discussed in these tables above. It can be seen where the location of each transaction point is where the lines intercept with each other. What do polytomous models look like? Polytomous models have a line that dictates each possible response. The line for the highest point value is typically S-shaped like a dichotomous curve such as green line number (5) in figure 1 which present the examinees locations with highest level of spirituality, all the area above its intercept with line number 4 is where the examinee with highest traits are located. The line for the lowest point value is typically sloped down like the 1-P dichotomous curve. Point values in the middle typically have a bell-shaped curve. In item 134 that scored 0 to 5 points. Only students with theta >0.23 are likely to get the full points (green(5)), while students -0.93>theta are likely to receive 0 points (blue).

The Item Response Category Characteristic Curve shows the likelihood of respondents selecting a certain score on the scale (1-6) at various levels of the latent trait. An item is better at discriminating between individuals when the curves are peaked and dispersed across all levels of the latent trait. For example, an item with high discrimination would have 6 peaks dispersed from low levels of the latent trait to high levels of the latent trait. Normally, you would generate a graph for all 25 of the items, but for demonstration purposes, we will just look at the curves for the five items on the agreeableness scale.

To graphically describe the association between the probability to respond in a particular category k for any item and any given levels of the latent trait a plot of the item category characteristic curves (ICCs) can be constructed. The higher the discrimination parameter the steeper the curve, thus denoting a stronger association between the item and the latent trait. Flat ICCs provide evidence that the probability to score in a higher category does not change in a relevant way as the level of the latent trait increases.

Figure 1.

Item Characteristic Curve for Items 134, 135, 136, and 137.

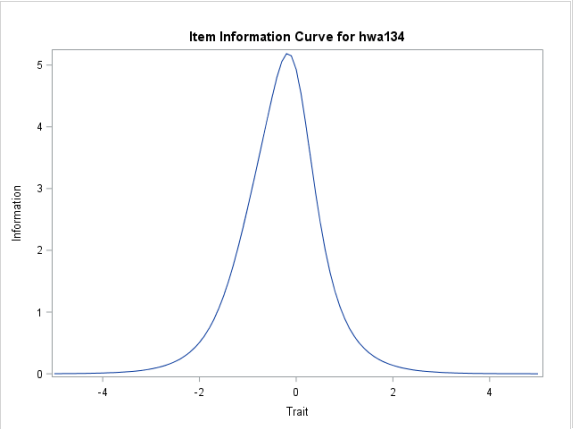


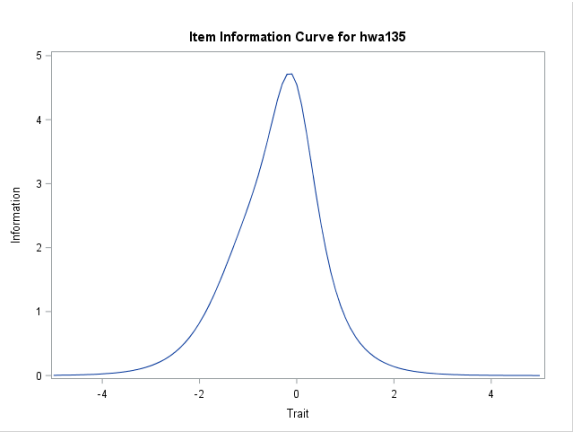
Using the Item Information Curve

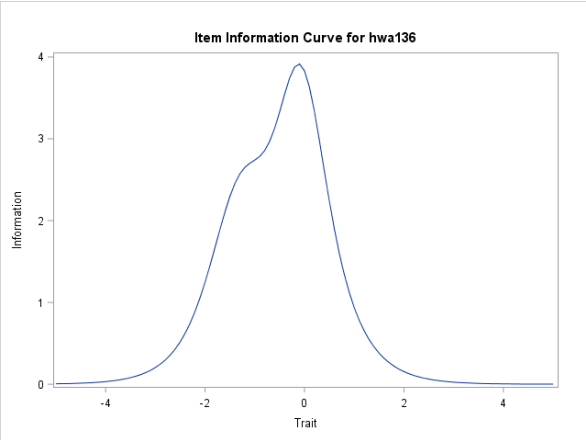
The Item Information Curve shows how well and precisely each item measures the latent trait at various levels of the attribute. Certain items may provide more information at low levels of the attribute, while others may provide more information at higher levels of the attribute. Again, we need to evaluate our data separately for each of the scales, owing to the fact that we have six latent variables within our dataset. And from this

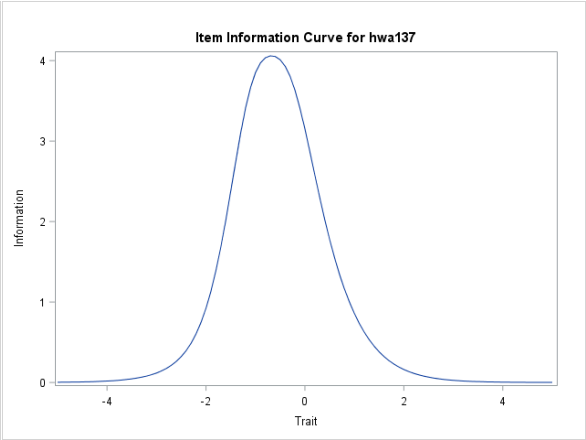
Figure 2 .

Items Information Curve 134, 135, 136, and 137.







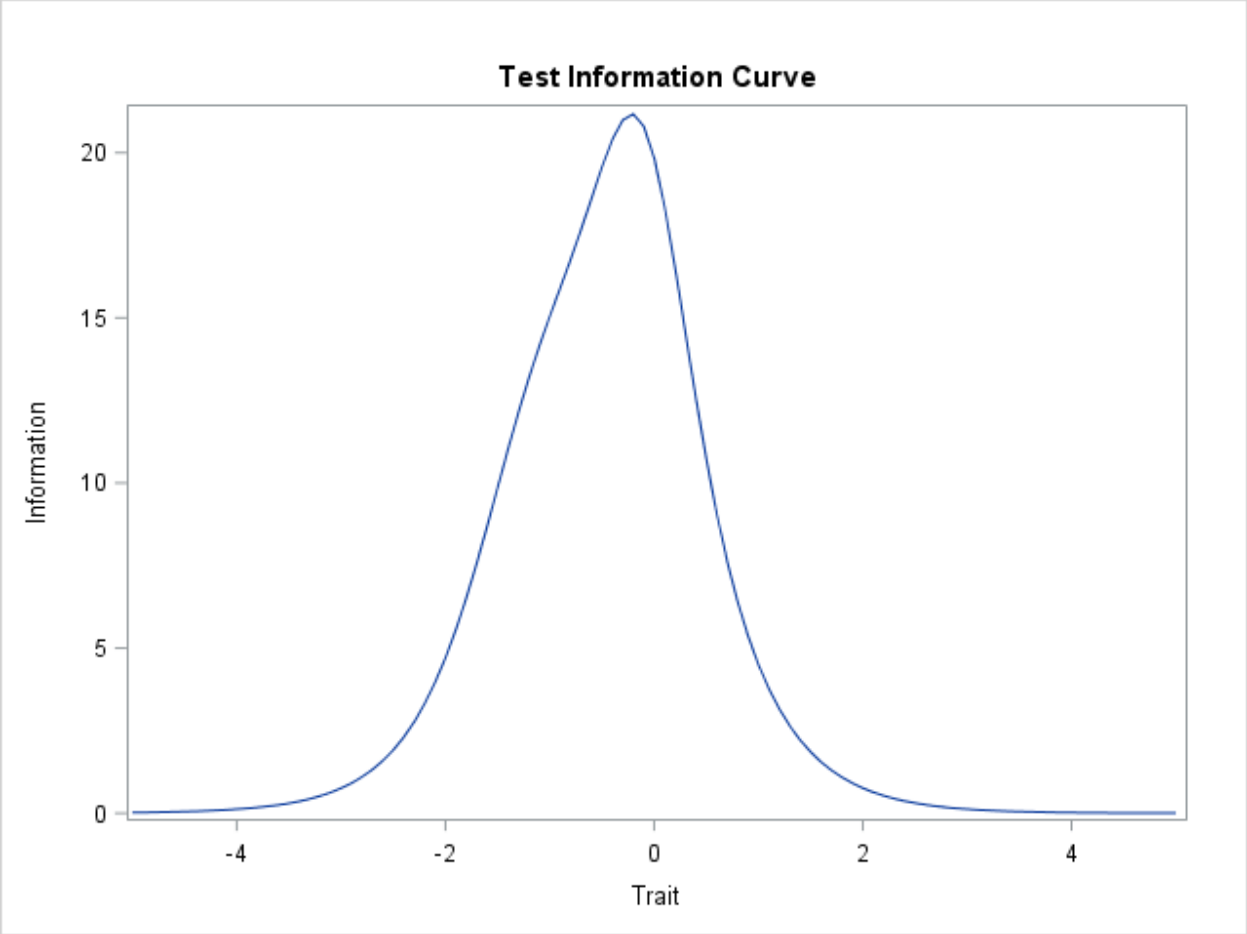


Using the Test Information Curve

The Test Information Function Curve aggregates the Item Information Curves across all the items. It tells us how well the test measures the latent trait at various levels of the attribute. Ideally, we would want this line to peak at about the mean of the sample because that is where the highest number of individuals would be. And this curve do.

Figure 3.

Test Information Curve



.

Generalized Partial Credit Model

The GPCM model relaxes the assumption in the PCM that all items on the instrument have equal discriminations and it can be seen from table 8 that the slope is differ from item to others. The slope of Item 134 = 2.29 and item 135 = -3.08. As mentioned in the model description in GPCM, Rasch-based PC model moving to the 2PL model version in GPCM, where α_j is the item discrimination, δ_{jh} is the transition location parameter (item 134 step1 = -0.69) between the h^{th} and h^{th}-1 category and $\delta_{j1}$is set to zero (0). In this model delta's are not in sequential order. It can be seen also that the probability estimates of providing a response in item's category (one of the six categories which are before -0.96 to above 0.30 in items 134) , x_{jk} given θ(theta), α_j (item discrimination), δ_{jh} (item difficulty transition parameter between h^{th} and h-1 category). Such as, some of the examinee are located in the category above 0.30 on the scores scale their location base on the item 134' difficulty, slope = 2.19 and examinee spirituality level, which means the group of the examinees' have the highest level of spirituality among other examinees. level is the There are m category points, m-1 transition points. There can be category reversals, eg., δ_{jh} do not need to follow sequential order.

Table 8.

Item Parameter Estimates.

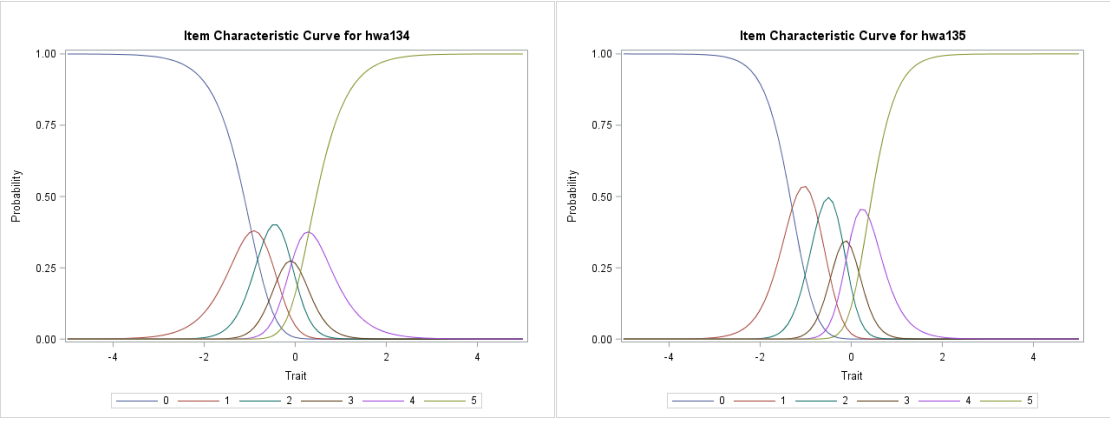
Item	Parameter	Estimate	Standard Error
hwa134	Step 1	-0.96172	0.13986
	Step 2	-0.69051	0.12861
	Step 3	-0.09693	0.12674
	Step 4	-0.06584	0.12536
	Step 5	0.30097	0.11722
	Slope	2.19427	0.31982
hwa135	Step 1	-1.29826	0.13162
	Step 2	-0.72942	0.10424
	Step 3	-0.18673	0.10025

Item Parameter Estimates.

Item	Parameter	Estimate	Standard Error
	Step 4	-0.04229	0.10008
	Step 5	0.37947	0.09753
	Slope	3.08423	0.49400
hwa136	Step 1	-1.41815	0.24406
	Step 2	-1.37699	0.20498
	Step 3	-0.07096	0.16429
	Step 4	-0.13890	0.16358
	Step 5	0.22652	0.15260
	Slope	1.39468	0.17722
hwa137	Step 1	-1.08552	0.21214
	Step 2	-1.26133	0.19593
	Step 3	-0.44600	0.14713
	Step 4	-0.27569	0.13703
	Step 5	0.33900	0.12590
	Slope	1.66193	0.21832

Figure 4.

Item Information Curve for Items 134, 135, 136, and 137.



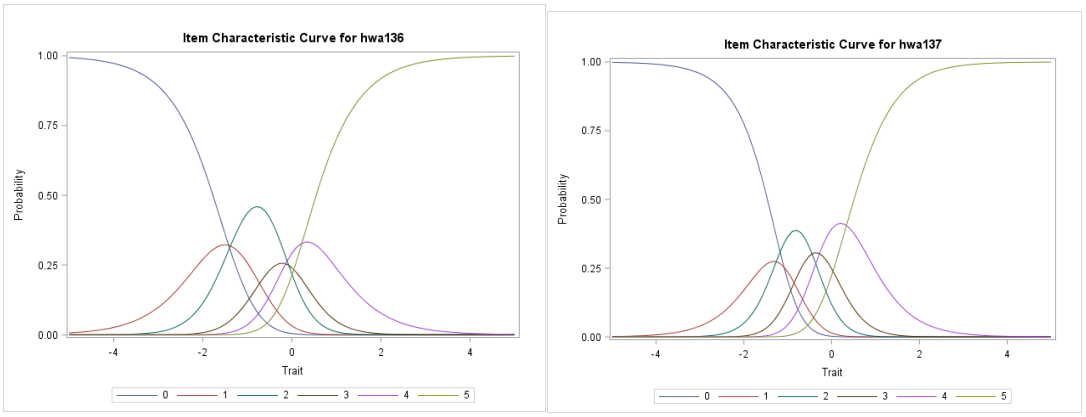
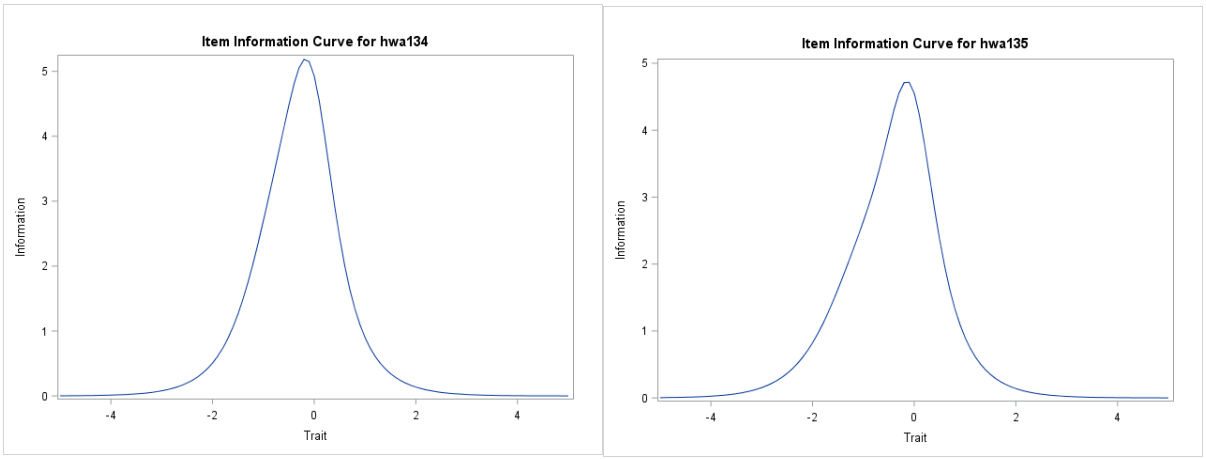


Figure 5.

Items Information Curve 134, 135, 136, and 137.



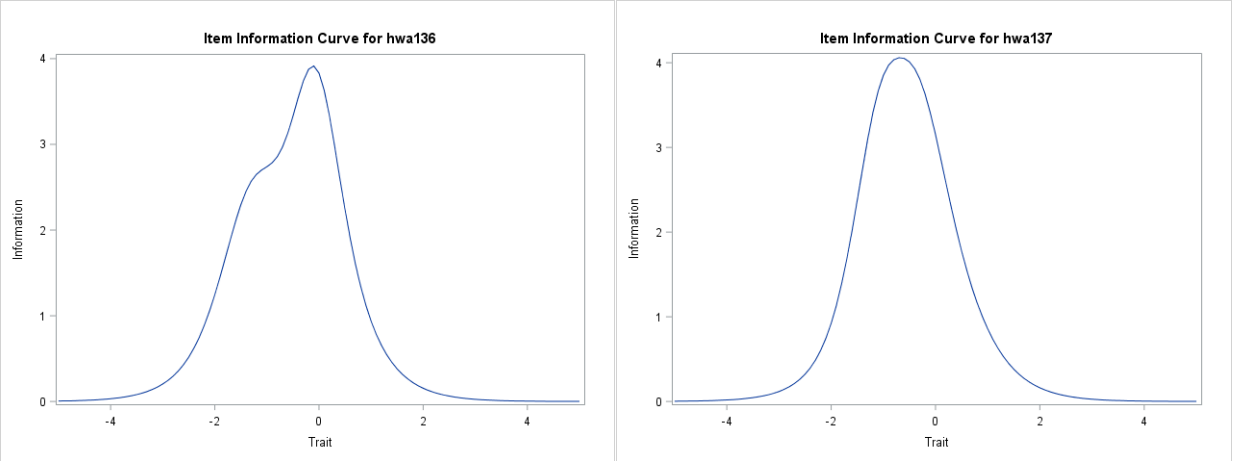
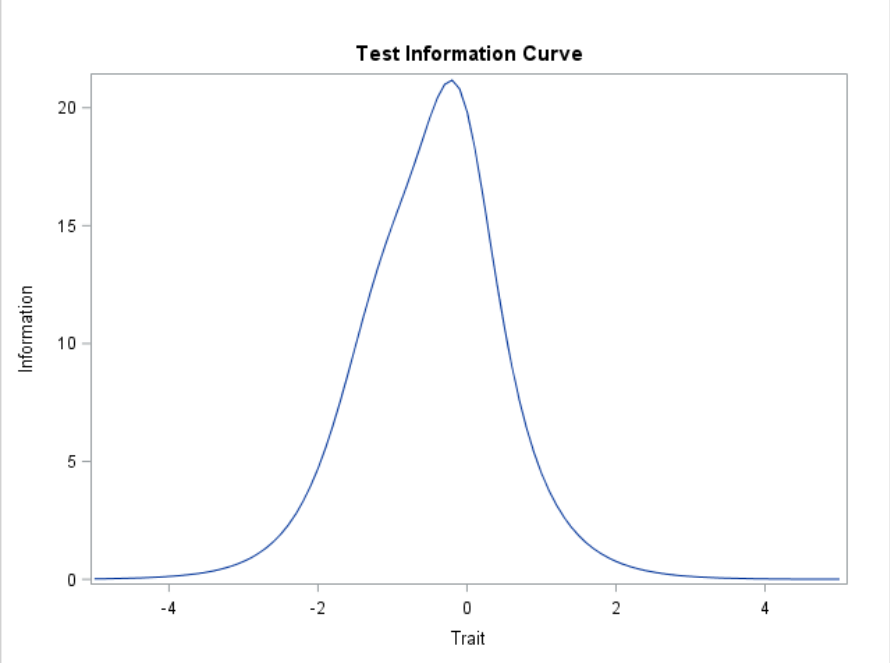


Figure 6.

Test information Curve.



Graded Response Model

The GR model focuses on establishing a boundary above which a person is expected to obtain certain category score(s) (as opposed to lower category score(s)). For example, from table 9, item 134 parameter boundary estimate in this model are broader than its boundary which were estimated by PCM

and GPCM. We can determine the probability of a score of 1 or higher versus a score of 0, or the probability of a score of 2 versus a score of 1 or 0. This has the effect of turning the polytomous score into a series of cumulative responses below a particular category versus *at or above* this category.

This model can be used for both rating-scale data and this items responses are (Not at all to Yes at all). From table 9, item 134 the higher categories which is above the threshold 5 = 0.52 represented more of the latent trait(spirituality). This model allows for different number of categories across different items, where δ_{xj} are always in increasing order (no reversals). Also, we should know that the the probability of an individual obtaining a particular category score x_j or responding in a particular category k, (p_k) result of calculating the difference between two cumulative probabilities for adjacent categories. the slope in this model is not constant, it varies based on the item items(134 =4.07 and 135 = 4.90).

Table 9.

Item Parameter Estimates.

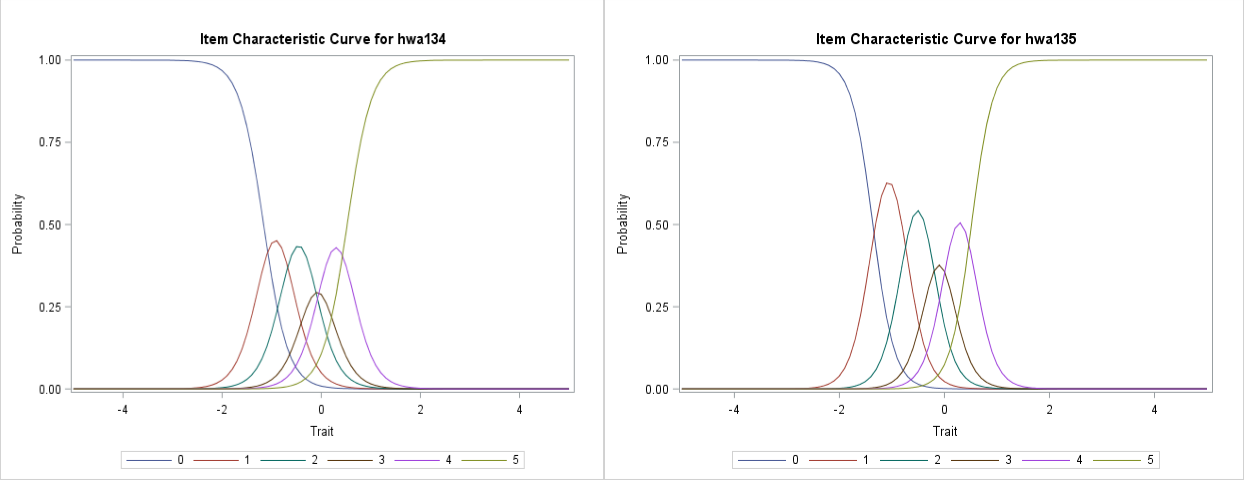
Item	Parameter	Estimate	Standard Error
hwa134	Threshold 1	-1.16434	0.10424
	Threshold 2	-0.68592	0.08473
	Threshold 3	-0.22752	0.07556
	Threshold 4	0.07006	0.07527
	Threshold 5	0.52158	0.08314
	Slope	4.07288	0.42764
hwa135	Threshold 1	-1.36298	0.11078
	Threshold 2	-0.75846	0.08437
	Threshold 3	-0.26264	0.07386
	Threshold 4	0.06070	0.07332
	Threshold 5	0.51505	0.08020
	Slope	4.90150	0.57494
hwa136	Threshold 1	-1.79193	0.15609
	Threshold 2	-1.19188	0.11629
	Threshold 3	-0.37746	0.08622

Item Parameter Estimates.

Item	Parameter	Estimate	Standard Error
	Threshold 4	0.03591	0.08282
	Threshold 5	0.60253	0.09420
	Slope	2.67575	0.26139
hwa137	Threshold 1	-1.53187	0.13418
	Threshold 2	-1.09748	0.10848
	Threshold 3	-0.53433	0.08635
	Threshold 4	-0.09642	0.07995
	Threshold 5	0.55546	0.08963
	Slope	3.02830	0.29940

Figure 7.

Item Characteristics Curve.



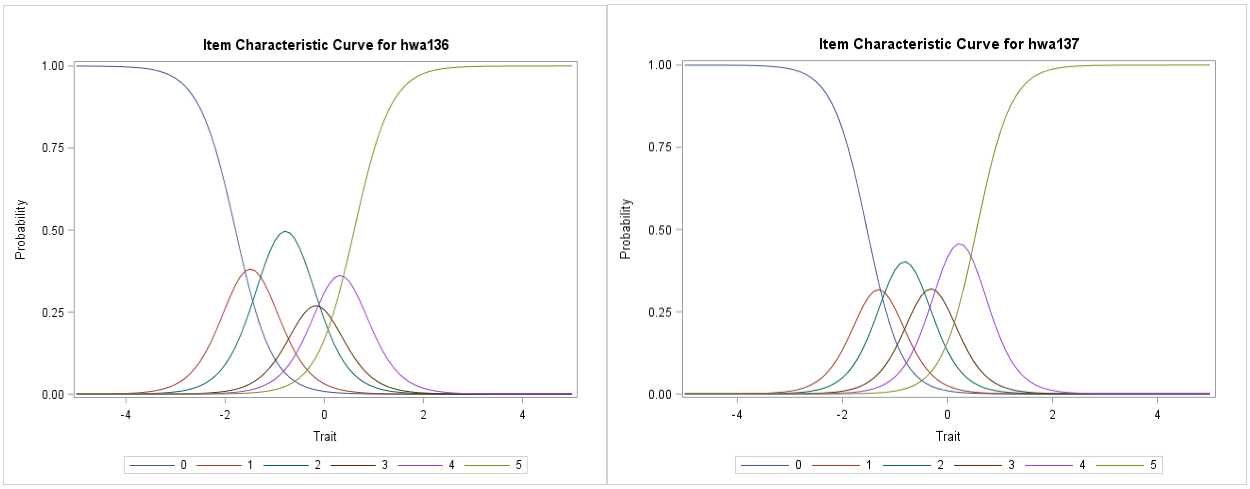
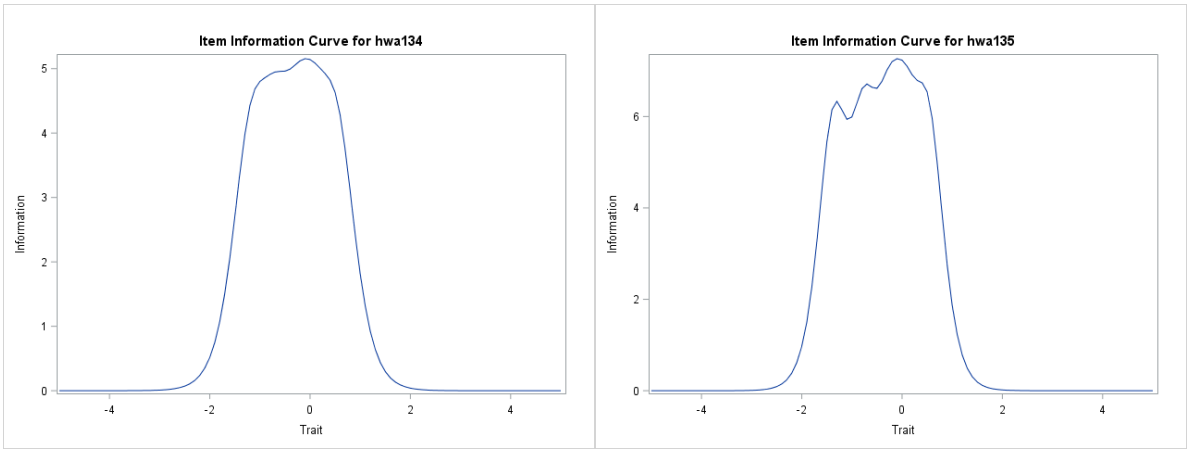


Figure 8.

Items Information Curve 134, 135, 136, and 137.



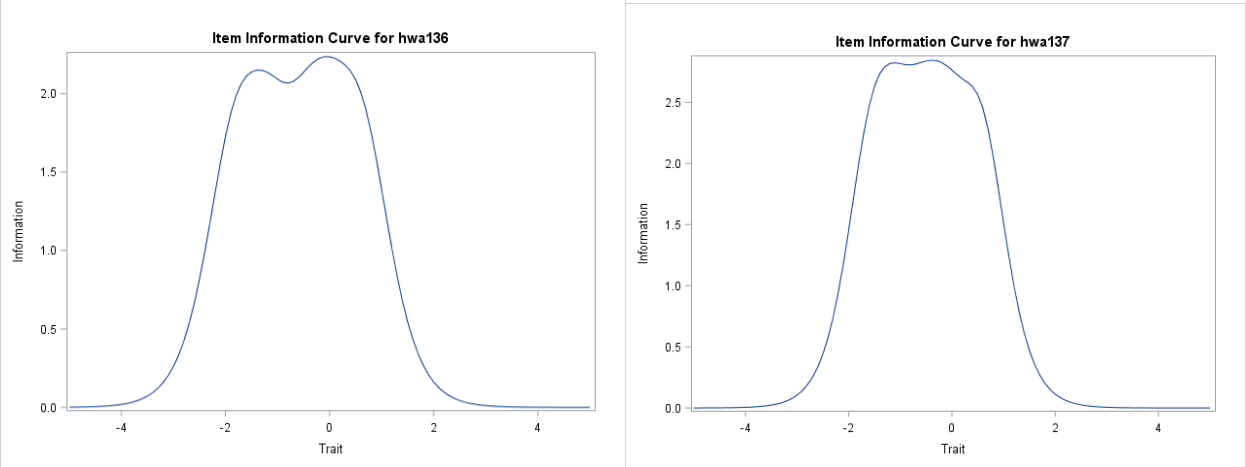
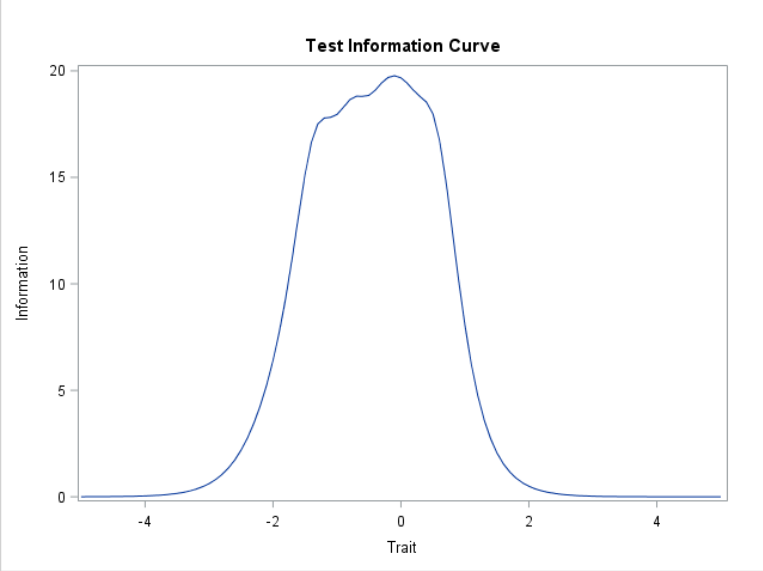


Figure 8.

Test information Curve.



Model fit

There are different methods to evaluate the models fit to the data. I used the Akaike information criterion (AIC) and the model with the lowest value of AIC is the model that fit the give data. From table 10, GRM is the best model that is fitting the data being used with AIC = 4079(Zwick, 1990).

Table 10.

Model Fit Statistics

Model Name	AIC
PCM	4134.72
GPCM	4122.57
GRM	4079.13

Conclusion

Each item and total item reliability correlation coefficient is strong that means the items from this test had high consistency. From the descriptive statistics, and the item parameter estimate I can tell these items are measuring the construct (spirituality) that it designed to measure, which is one of the validity indications. The categories toward the right on the score scale has the highest values of endorsement. Finally, GRM model fits the data better than PCM and GPCM.

References

Tang, K. L., & Service, E. T. S.-E. T. (1995). *Polytomous Item Response Theory Models and Their Applications in Large-Scale Testing Programs: Review of Literature* (p. 20). https://www.ets.org/research/policy_research_reports/publications/report/1996/ibtw

Zwick, R. (1990). When Do Item Response Function and Mantel-Haenszel Definitions of Differential Item Functioning Coincide? *Journal of Educational Statistics*, *15*(3), 185–197. https://doi.org/10.3102/10769986015003185