Goodness of fit and precision of the Graded Response Model estimates of a psychometric scale at varying number of categories.

Yuliana Mora Cedeño*
School of Statistics, University of Costa Rica, San Jose, Costa Rica –
yulianamoracedeno22@gmail.com

Abstract

This study aims to evaluate the absolute goodness of fit of Graded Response Model (GRM), proposed by Samejima in 1969, when the number of categories of a polytomous scale is changed, and to analyze the precision of GRM estimates of the latent trait. These polytomic scales are frequently used in psychometric measurements of skill levels, such as quantitative skills or reading skills. The analysis performs a simulation study design to assess these objectives. It was simulated a polytomous scale derived from normal variables based on a one-factor model. It was computed 100 replications of 1000 cases and 14 items for each scale of 3, 4, 5, and 6 categories. The GRM is estimated for each replication. The absolute goodness of fit is evaluated using the Likelihood Ratio Test (LRT) that contrasts the model with the saturated model: the G^2 test. I analyze precision computing the Mean Bias and the Root-Mean-Square-Error (RMSE) of the GRM theta parameter. The results show that, based on the G^2 test, the best fit of the GRM is obtained with a scale of 5 categories, and the fit worsens with a 6 categories scale. There were no good predictions of the skill level with any of the 4 scales because all of them had RMSE of 0.97 and Mean Biases of 0.87. The study results suggests that it is better to operationalize skill levels with a 5-categories scale.

Keywords: absolute goodness; polytomous scale; One factor model.

1. Introduction

The research project evaluates the goodness of fit of Graded Response Model (GRM), when a one-factor model is estimated from a scale that measures a latent trait θ. I evaluate the GRM goodness of fit when the scale´s number of categories varies. This kind of scales are often used in psychometric analyses of skills, such as reading or quantitative reasoning. Gomez, Artes & Deumal (1989) explain that the number of alternatives in a response scale has an effect on the estimation of the items' parameters and on the information function that describes the latent trait. Hernández, Muñiz and García (2000) express that, in Item Response Theory (IRT) models, there is an important loss of information due to reducing the number of levels in the response scale; these authors propose the hypothesis that an increasing the number of categories may produce a better fit model and more precise estimates of the latent trait.

IRT models for dichotomous data lead the way for the development of models for polytomic items, such as the Partial Credit Model, the Generalized Partial Credit Model and GRM. These models were needed due to the extended utilization of items with more than two categories; e.g, Likert scales, evaluation tests, etc. GRM was proposed by Samejima (1969), and was designed for ordinal polytomic items, thus generaralizing the two-parameter logistic model. Attorresi, Abal, Galibert, Lozzia & Aguerri (2011) define the GRM as the model that "describes the behavior of an item *i* with a discrimination parameter (a_i) and a set of threshold parameters (b_{ij} = *1, ..., m*) located between the continuous categories of a polytomic item ($j = 0, \dots, m$) ." (p. 234). They state that one of GRM's main goals is to determine the place of the threshold values b_{ij} in the latent variable continuum, and to represent the probability that an individual's response to a specific item i would be greater or equal to the threshold b_{ij} given a level θ of the trait. This probability is expressed as:

$$P^*(x \geq j \,/\, \theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ij})}}$$

My main goal is to evaluate GRM's absolute fit if the scale´s number of categories is varied, and to analyze the precision of the estimates of the latent trait level using Monte Carlo simulation of polytomic data that assumes a one-factor model.

In the different scenarios for the data simulation, the number of categories varies between 3 and 6 levels, assuming a Likert response scale. This kind of scales "has a response scaling, usually between 1 and 5, where 1 represents the lowest category, usually labeled 'Disagree Very Much', and 5 represents the highest category, usually labeled 'Agree Very Much' " (López-Pina, 2005). Given that the main goal of this project is to fit a model for polytomic data, I decided to start with a 3-category scale, and increase the number of categories by 1, up to a 6-category scale in order to keep the scale's symmetry, that is, 3 low categories and 3 high categories. The model requires to select a factorial weight implicit in the analysis. The ideal value of this factorial weight was derived from the literature. Hernández, Muñiz & García (2000) performed a similar simulation project; however, they also had empirical data to compare with the simulated results. They use a factorial weight of 0.75 to simulate the data; this value was derived from a factor analysis of the empirical data. Avendaño, Duarte & Campo (2006) computed a factorial weight of 0.556 in the evaluation of the APGAR (adaptability, partnertship, growth, affection, resolve) family functioning scale. Given that 0.75 seemed relatively high, I use a sampling weight of 0.60 in this simulation study.

2. Method

The simulation is performed using RStudio, version 0.98.1091. GRMs fit is analyzed with the R package called “mirt: Multidimensional Item Response Theory”, version 1.8, created by Phil Chalmers, Joshua Pritikin , Alexander Robitzsch and Mateusz Zoltak.

The sample size for each of the simulations is 1000 cases and 14 items. There are 100 replicates for each scale of 3, 4, 5 and 6 categories. As a result, there are 400 matrices of polytomic responses. These are used to estimate the skill parameter θ assuming the GRM. The first step for generating the data is to simulate normally distributed continuous variables from a one-factor model with the following structure:

$$y = 0{,}60F_1 + \varepsilon$$

The unique factor F_1 and the constant error term ε are simulated as normally distributed random variables with mean equal to 0 and standard deviation equal to 1. For every one of the 1000 cases, there are 14 variables generated according to this structure. These variables represent the hypothetical subjects' responses. Thus, the model assumes that each polytomous scale represents a continuous non-observable trait. The next step is to establish thresholds to transform these continuous variables in polytomous ordinal variables. Table 1 shows the rules followed to perform these transformations.

I estimate the GRM for each of the 400 matrices. I compute the G2 statistic and its p-value for each item in order to evaluate the absolute fit; the mean p-value is the selected measure of global fit. Additionally, I record the mean number of items that have a bad fit, controlling for the number of response categories.

Table 1
Rules to recode simulated continuous variables into polytomous scales.

Number of categories	Category					
	1	2	3	4	5	6
3	$-6 \leq y \leq -1$	$-1 < y \leq 1$	$1 < y \leq 6$			
4	$-6 \leq y \leq -1{,}5$	$-1{,}5 < y \leq 0$	$0 < y \leq 1{,}5$	$1{,}5 < y \leq 6$		
5	$-6 \leq y \leq -2$	$-2 < y \leq -1$	$-1 < y \leq 1$	$1 < y \leq 2$	$2 < y \leq 6$	
6	$-6 \leq y \leq -4$	$-4 < y \leq -2$	$-2 < y \leq 0$	$0 < y \leq 2$	$2 < y \leq 4$	$4 < y \leq 6$

In order to study the precision of the estimates of the latent trait level, I record each of the $\hat{\theta}$ predicted by the model, and simulate a vector of 1000 values from a standard normal distribution based on the pre-defined values of θ. Based on these values, I compute the Mean Bias and the Root-Mean-Square-Error (RMSE) using the following formulas:

$$RMSE = \sqrt{\frac{\sum_{r=1}^{100}(\theta_j - \hat{\theta}_j)^2}{R}} \qquad Mean\ Bias = \frac{\sum_{r=1}^{100}|\theta_j - \hat{\theta}_j|}{R}$$

Where:

θ = pre-defined true value of the latent parameter
$\hat{\theta}$= predicted value of the latent parameter
R= Number of replicates
r=Replicate identifier
j=Individual identifier

Some of the replicates could not generate estimates due to lack of convergence in the iteration process. I delete these latter replicates, and compute the fit measures with the remaining replicates only.

3. Results

The evaluation of GRM's absolute fit shows a progressive increment in the mean p-value of the G^2 statistic, between the scales of 3 categories and the scales of 5 categories, with values that range between 0.36 and 0.46, the latter representing the best fit. However, the fit worsens with 6-category scales (Figure 1).

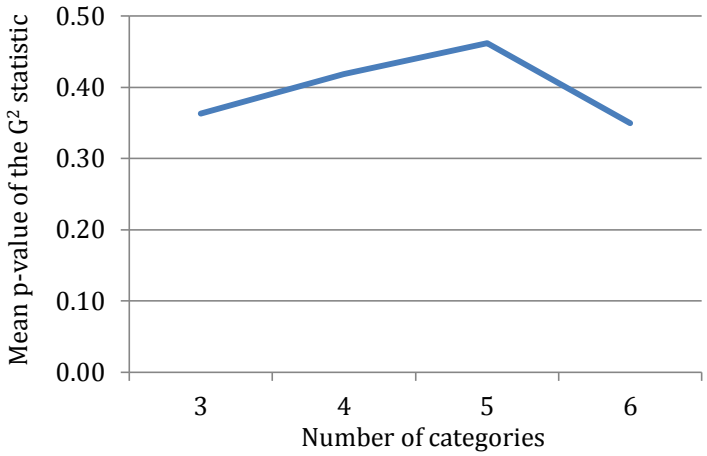

Figure 1. Mean p-value of the G^2 statistic of simulated GRMs, by number of response categories.

The analysis of the mean number of items with poor fit shows that the goodness of fit improves when the number of categories increase from 3 to 4, given that the mean number of items with bad fit decreases from 3 to 1. It is still 1 with 5-category scales, but it increases again (showing worse fit) with 6-category scales.

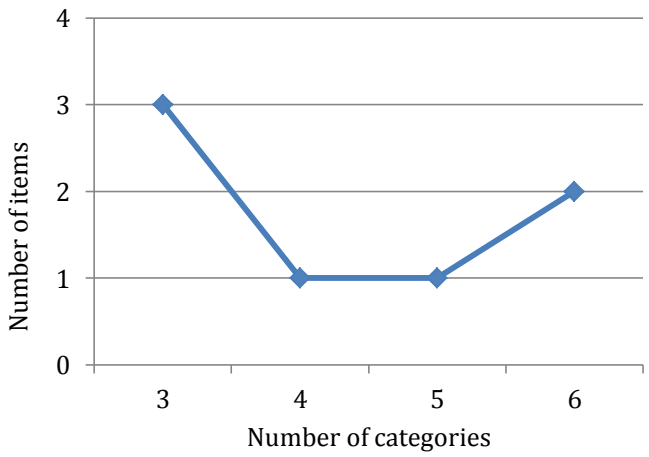

Figure 2. Mean number of items with bad fit, by number of response categories.

The estimates of the latent traits were almost identical across the different scales, with average RMSE very close to 1. From a relative perspective, the 6-category scale produced the most precise estimates, with a mean value of 0.96. Nevertheless, from an absolute perspective, none of the 4 scales has precise estimates because the RMSE values are not close to 0. The Mean Bias has the same pattern for the 4 scales, with mean values around 0.87; the 6-category scale is again the scale with the most precise estimates; however, neither measure shows that the model has good predictions of the latent trait.

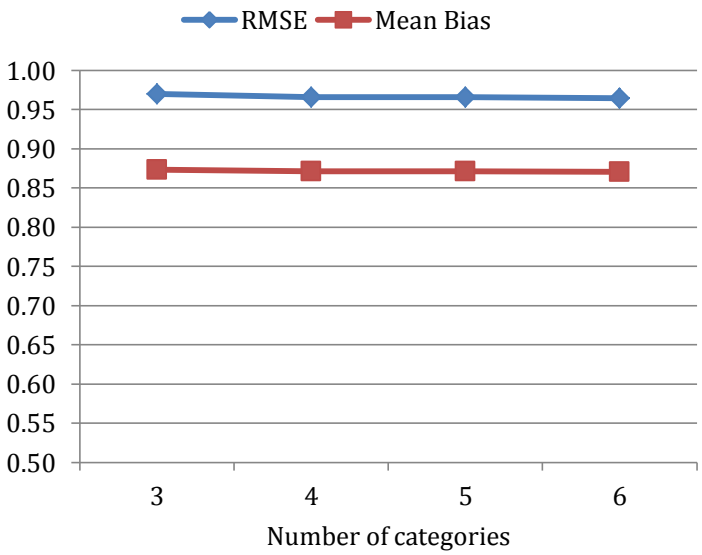

*Figure 3.*Mean Bias and average RMSE for the latent trait estimate, by number of categories.

4. Conclusions

The main goal of this article was to evaluate the absolute model fit of the GRM when varying the number of categories of a scale and to analyze the precision of the estimates of the latent trait, using Monte Carlo simulation of polytomous data derived from a one-factor model.

According to the results, the main conclusion is that the best fit is obtained with a 5-category response scale, based on the mean p-value of the G^2 statistic of a goodness of fit test and on the number of items with bad fit. However, the 4-category scale has also a good fit when compared to the 3-category scale. The 6-category scale has the worst fit. There are no good predictions for the latent trait level with any of the 4 scales because all of them have average RMSEs of 0.97 and Mean Biases of 0.87.

In the simulation procedure, some of the estimated models suffered from lack of convergence during the estimation process. A sizable number of replicates did not converge, and the iterative process was stopped at 500 iterations. I discarded 27, 36, 38 and 39 replicates for the scales with 3, 4, 5, and 6 categories, respectively.

References

Attorresi, H., Abal, F., Galibert, M., Lozzia, G. &Aguerri, M. (2011). Aplicación del modelo de respuesta graduada a una escala de voluntad de trabajo. *Interdisciplinaria.* (Vol. 28, n° 2, pp.231-244). Universidad de Buenos Aires.

Forero, L., Avendaño, C., Duarte, Z. & Campo, A. (2006). Consistencia interna y análisis de factores de la escala APGAR para evaluar el funcionamiento familiar en estudiantes de básica secundaria. *Revista Colombiana de Psiquiatría.* (Vol. XXXV, n° 1, pp.23-29). Universidad Autónoma de Bucaramanga.

Hernández, A.,Muñiz, J. &García, E. (2000).Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema.* (Vol. 12, n° 2, pp. 288–291). Universidad de Valencia y Universidad de Oviedo.

Lopéz, J. (2005). Ítems politómicos vs. dicotómicos: Un estudio metodológico. *Anales de psicología.* (Vol. 21, n° 2, pp.339–344). Universidad de Murcia.

Pérez, J. (2004). Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica. (Tesis de Licenciatura en Psicología). Universidad de Sevilla. Dpto. de Psicología Experimental.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.17 (4, Pt 2)

Tomás, J & Oliver, A. (1998). Efectos de formato de respuesta y método de estimación en análisis factorial confirmatorio. *Psicothema.* (Vol. 10, n° 1, pp.197–208). Universidad de Oviedo.