MULTIVARIATE BEHAVIORAL RESEARCH, *41*(1), 55–64

A Cautionary Note on Using G^2(dif) to Assess Relative Model Fit in Categorical Data Analysis

Albert Maydeu-Olivares
University of Barcelona & Instituto de Empresa

Li Cai
University of North Carolina at Chapel Hill

The likelihood ratio test statistic G^2(dif) is widely used for comparing the fit of nested models in categorical data analysis. In large samples, this statistic is distributed as a chi-square with degrees of freedom equal to the difference in degrees of freedom between the tested models, but only if the least restrictive model is correctly specified. Yet, this statistic is often used in applications without assessing the adequacy of the least restrictive model. This may result in incorrect substantive conclusions as the above large sample reference distribution for G^2(dif) is no longer appropriate. Rather, its large sample distribution will depend on the degree of model misspecification of the least restrictive model. To illustrate this, a simulation study is performed where this statistic is used to compare nested item response theory models under various degrees of misspecification of the least restrictive model. G^2(dif) was found to be robust only under small model misspecification of the least restrictive model. Consequently, we argue that some indication of the absolute goodness of fit of the least restrictive model is needed before employing G^2(dif) to assess relative model fit.

The two most widely used statistics for assessing the goodness of fit of a model fitted to a contingency table are Pearson's χ^2 statistic and the likelihood ratio statistic G^2. Under the null hypotheses that the tested model holds in the popula-

This research was supported by the Department of Universities, Research and Information Society (DURSI) of the Catalan Government, and by grant BSO2003–08507 of the Spanish Ministry of Science and Technology.

Correspondence concerning this article should be sent to Albert Maydeu-Olivares, Faculty of Psychology, University of Barcelona, P. Valle de Hebrón, 171, 08035 Barcelona, Spain. E-mail: amaydeu@ub.edu

tion, the distribution of both statistics is well approximated by a chi-square distribution in large samples. However, at least since Cochran (1952) it is well known that when some cell probabilities are small the chi-square approximation to χ^2 and G^2 yields incorrect p-values. Yet, if the contingency table is large then the probabilities for some cells *must* necessarily be small as the cell probabilities must add up to one (Bartholomew & Tzamourani, 1999). In other words, if the number of cells is large the chi-square approximation to χ^2 and G^2 can not be used to test the overall fit of the model, regardless of sample size. Of course, having a small sample size only makes matters worse. Nevertheless, as Thissen and Steinberg (1997) pointed out, when the number of categories is five or more, the approximation becomes invalid for any model as soon as the number of variables is greater than six "with any conceivable sample size" (p. 61). Thus, researchers are faced with a serious problem as they generally wish to model much larger contingency tables.

For larger models, one can use resampling methods such as the parametric bootstrap to obtain more accurate p-values for χ^2 and G^2 (Bartholomew & Knott, 1999; Bartholomew & Tzamourani, 1999; Collins, Fiddler, Wugalter, & Long, 1993). However, a large number of resamples is needed to obtain a p-value with reasonable accuracy. So, the procedure is time consuming, particularly if one is interested in comparing the fit of several models. Maybe for this reason, to date the use of resampling to test models for categorical data is not widespread. Furthermore, a recent simulation study by Tollenaar and Mooijart (2003) revealed that the p-values for χ^2 and G^2 obtained using bootstrap need not be accurate.

Given these difficulties to assess *absolute* model fit (i.e., the fit of the model against the data), often times researchers only assess *relative* model fit (i.e., the fit of one or more constrained versions of a model relative to the most general model considered). In categorical data analysis perhaps the most popular procedure for comparing the fit of two models M_0 and M_1, with M_0 a special case of M_1, is to use the likelihood ratio statistic G^2(dif). In large samples, this statistic can be approximated by a chi-square distribution with degrees of freedom equal to the difference of degrees of freedom between the models. The popularity of G^2(dif) stems from a study by Haberman (1977) who showed that when M_1 holds, the chi-square approximation to G^2(dif) is valid even for large models and small sample. However, when M_1 does not hold, the above large sample reference distribution for G^2(dif) is no longer appropriate. Rather, the actual large sample distribution of G2(dif) will depend on the degree of model misspecification of M_1. To illustrate this, a simulation study is performed to show the behavior of G^2(dif) with different amounts of model misspecification in M_1.

The remaining parts of this article are organized as follows: First, the statistics used to test absolute and relative model fit in categorical data analysis are described in some detail. Next, a simulation study is reported where G^2(dif) is used to

assess the relative fit of competing item response theory (IRT) models. For an overview of these models see Embretson and Reise (2000); Hambleton and Swaminathan (1985); Hambleton, Swaminathan, and Rogers (1991); or van der Linden and Hambleton (1997). More specifically, G^2(dif) is used to compare the relative fit of a one-parameter logistic model versus a two parameter logistic model under correct and incorrect model specification of the least restrictive model. For correct model specification, the true model is a one-parameter logistic model. For incorrect model specification, the true model is a three parameter logistic model with equal slope parameters. Increasingly large lower asymptote parameters are used to specify models with increasing degree of model misspecification. The final section discusses the simulation results and offers some guidelines for the use of G^2(dif) in applied work.

χ^2, G^2, AND G^2(dif)

Consider a contingency table with C cells. Let the proportion of respondents in cell c be denoted as p_c, and the corresponding probability π_c. A model for the contingency table may be written as $\boldsymbol{\pi}(\boldsymbol{\theta})$, where $\boldsymbol{\pi}$ is a the vector of all C probabilities which are assumed to depend on a vector of q parameters, $\boldsymbol{\theta}$. The two most widely used statistics for assessing the absolute goodness of fit of the model are Pearson's χ^2 statistic and the likelihood ratio statistic G^2. Letting N denote sample size, Pearson's statistic can be written as

$$\chi^2 = N\sum_{c=1}^{C}\frac{\left(p_c - \pi_c\right)^2}{\pi_c};$$

and the likelihood ratio statistic can be written as

$$G^2 = 2N\sum_{c=1}^{C} p_c \log\left(\frac{p_c}{\pi_c}\right).$$

Under the null hypothesis that the true probabilities equal $\boldsymbol{\pi}(\boldsymbol{\theta})$, both statistics are distributed in large samples as a chi-square distribution with $C - q - 1$ degrees of freedom. However, as the size of the model (C) and the degree of sparseness of the data (N/C) increase, the empirical Type I errors of χ^2 and G^2 will not match its expected rates under the large sample reference distribution (Koehler & Larntz,

1980; Larntz, 1978). Also, of the two statistics, χ^2 is known to be less adversely affected by the size of the model and degree of sparseness of the data.

Consider now two alternative models M_0 and M_1, with M_0 a special case of M_1, and degrees of freedom $(C - q_0 - 1)$ and $(C - q_1 - 1)$, respectively. In this situation, model M_0 is said to be nested within M_1. Because M_0 is simpler than M_1, $(C - q_0 - 1) > (C - q_1 - 1)$ and when M_0 holds M_1 must necessarily hold. For testing the relative fit of M_0 with respect to M_1 the most commonly used test statistic is

$$G^2\left(\text{dif}\right) = G_0^2 - G_1^2, \tag{1}$$

where G_1^2 and G_0^2 are the G^2 statistics obtained when estimating M_1 and M_0, respectively. Under the null hypotheses that the model probabilities under M_1 and M_0 are equal, the distribution of G^2(dif) can be approximated by a chi-square distribution with $q_0 - q_1$ degrees of freedom in large samples.

G^2(dif) is the likelihood ratio statistic for testing the additional restrictions imposed by the more restrictive model (M_0) over those imposed by the less restrictive model (M_1), given that the latter holds. To emphasize this fact, Agresti (1990) uses the expression $G^2(M_0|M_1)$ to refer to this statistic. Also, since $G_0^2 \geq G_1^2$, $G^2(\text{dif}) \geq 0$ and G^2(dif) is large when M_0 fits poorly relative to M_1 (Agresti, 1990).

There are two reasons for the popularity of G^2(dif). First, this statistic is easy to compute. For maximum likelihood estimation, G_1^2 and G_0^2 are routinely printed in the output by computer programs when models are estimated using grouped data (i.e., sample proportions for observed response patterns). For large models, however, it is more convenient for computational reasons to perform maximum likelihood estimation using individual observations rather than sample proportions. In this case, G_1^2 and G_0^2 are generally not printed in the output, only the maxima of the loglikelihood functions are printed. Yet, G^2(dif) is obtained simply by taking the difference of those function maxima multiplied by –2. Second, Haberman (1977) showed that when M_1 holds, the chi-square approximation to G^2(dif) is valid even for large models and small samples. That is, G^2(dif) can be trusted for testing the relative fit of M_0 versus M_1, even when the model is so large and G_1^2 and G_0^2 cannot be trusted to assess the absolute goodness of fit of each of the models separately. Haberman's results are most useful in the context of log-linear models (the focus of his article). In these models when C is not too large it is possible to estimate a saturated model; that is, a model whose absolute fit is perfect. Because in log-linear models it is possible to assess the fit of nested models relative to the saturated model, no assessment of the absolute fit is needed. For other classes of models (such as the IRT models considered here) for which there is no simple counterpart of the log-linear saturated model it is necessary to assess the absolute goodness of

fit of the least restrictive model being compared. This is illustrated in the following simulation study.

SIMULATION STUDY

A simulation study was performed to illustrate the small sample behavior of G^2(dif) for testing the relative fit of M_0 versus M_1 with increasing degrees of model misspecification of the larger model (M_1). The one, two, and three parameter logistic IRT models (see Lord & Novick, 1968) with a normally distributed latent trait are used in these simulations. These models are widely used in educational research to model binary contingency tables. Given N respondents to p educational items these models assume that the conditional probability of endorsing an item, Y_i, is a function of an unobserved latent trait η. For the one-parameter model (also known as Rasch model), this conditional probability is

$$\Pr\left(Y_i = 1 \middle| \eta\right) = \frac{1}{1 + \exp\left[-a\left(\eta - b_i\right)\right]}. \tag{2}$$

For the two-parameter model is

$$\Pr\left(Y_i = 1 \middle| \eta\right) = \frac{1}{1 + \exp\left[-a\left(\eta - b_i\right)\right]}. \tag{3}$$

And finally for the three-parameter model is

$$\Pr\left(Y_i = 1 \middle| \eta\right) = c_i + \left(1 - c_i\right)\frac{1}{1 + \exp\left[-a_i\left(\eta - b_i\right)\right]}. \tag{4}$$

The three models are nested. The most general model is the three-parameter logistic model. This model reduces to the two-parameter model when all the low asymptotes c_i equal zero. In turn the two-parameter model reduces to the one-parameter model when the slope parameters a_i are all equal.

In these simulations two sample size conditions, three model size conditions, and four degrees of model misspecification were investigated. There were 1,000 replications used in each cell of the 2 × 3 × 4 factorial design. The sample sizes were 1,000 and 5,000 observations, and the three different model sizes were 5, 10, and 30 variables. The sample sizes were chosen to be medium to large for typical

applications in educational research, whereas the model sizes were chosen to be small to medium also for typical applications.

In all cases the models tested were a one-parameter versus a two-parameter logistic model. Also, in all cases data were simulated according to a three-parameter logistic model with equal slope parameters (a_i) and equal lower asymptote parameters (c_i) across items. The slope parameter a used to generate the data was 0.8 in all cases. For the 5 item tests, the intercept parameters were $\mathbf{b} = (-2, -1.25, 0, 1.25, 2)'$. The 10- and 30-item tests were obtained by repeating these **b** parameters twice and six times, respectively. Four sets of lower asymptote parameters were used: correct model specification of the least restrictive model ($c = 0$), minor model misspecification ($c = 0.01$), small model misspecification ($c = 0.05$), and moderate model misspecification ($c = .25$).

In all cases the estimation was performed using maximum marginal likelihood estimation via an EM algorithm (see Bock & Aitkin, 1981). The parameter estimation subroutines were written in GAUSS (Aptech Systems, 2003) and produced identical results as MULTILOG (Thissen, 2003) in trial runs. To ensure numerical accuracy of the estimates, 81 quadrature points, equally spaced between –5 and 5 were used in the numerical integration of response pattern probabilities. In all conditions but the combination of $c = 0.25$, $p = 5$, and $N = 1{,}000$, all 1,000 replications converged. Even in the only condition where convergence was a problem, it was a relatively rare event (only 8 of 1,000 times). A case was deemed nonconvergent if the maximum intercycle change in the parameter estimates did not drop below the threshold of 0.0001 after 200 E-step iterations, which is a much more stringent criterion than MULTILOG's default settings.

RESULTS

Correct Model Specification ($c = 0$)

In this case data were generated using a one-parameter logistic model and G^2(dif) was used to test the fit of the one-parameter model against a two-parameter model. Because in this case the larger model M_1 holds, from results in Haberman (1977), G^2(dif) was expected to wrongly reject the null hypothesis that the model probabilities under M_1 and M_0 are equal according to a chi-square distribution with $q_0 - q_1$ degrees of freedom.

As can be seen in Table 1 this is the case across all conditions. For instance, for $p = 5$ and $\alpha = 1\%$, G^2(dif) rejects the null hypotheses that the two models fit equally well 0.8% of the times when $N = 1{,}000$ and 1.1% of the times when $N = 5{,}000$. Thus, when the larger model is correctly specified, G^2(dif) can be used to test the relative fit of these nested IRT models even in large models and very sparse contingency tables.

TABLE 1
Empirical Rejection Rates for G^2(dif)

c	*0*						*0.01*						*0.05*						*0.25*					
p	*5*		*10*		*30*		*5*		*10*		*30*		*5*		*10*		*30*		*5*		*10*		*30*	
N	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*	*1000*	*5000*
$\alpha = 0.01$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.03	0.02	0.06	0.02	0.19	0.11	0.92	0.05	0.45	0.26	0.99	0.96	1.00
$\alpha = 0.05$	0.05	0.04	0.06	0.05	0.07	0.05	0.07	0.06	0.05	0.07	0.07	0.10	0.08	0.17	0.10	0.41	0.27	0.98	0.11	0.70	0.47	0.99	0.99	1.00
$\alpha = 0.10$	0.10	0.08	0.11	0.16	0.13	0.09	0.11	0.11	0.10	0.13	0.13	0.19	0.13	0.26	0.19	0.53	0.40	0.99	0.27	0.79	0.59	1.00	1.00	1.00
$\alpha = 0.20$	0.21	0.19	0.21	0.22	0.21	0.20	0.21	0.21	0.21	0.24	0.23	0.32	0.24	0.42	0.33	0.69	0.57	1.00	0.42	0.87	0.73	1.00	1.00	1.00
$\alpha = 0.25$	0.26	0.25	0.27	0.27	0.27	0.26	0.27	0.27	0.26	0.29	0.28	0.39	0.31	0.48	0.39	0.74	0.62	1.00	0.48	0.90	0.79	1.00	1.00	1.00

Note. The number of replications was 1,000 for each combination of c, p, and N except for $c = 0.25$, $p = 5$, and $N = 1{,}000$ where only 992 replications converged. The number of degrees of freedom available for testing is {4, 9, and 29} for p = {5, 10, and 30}.

On the other hand, G^2 can only be safely used to test the absolute goodness of fit of the one-parameter model when the model is small. For instance, for $p = 5$ the empirical rejection rates at $\alpha = \{0.01, 0.05, 0.10, 0.20, \text{and } 0.25\}$ are $\{0.005, 0.05, 0.09, 0.20, \text{and } 0.26\}$ when $N = 1{,}000$ and $\{0.01, 0.04, 0.08, 0.17, \text{and } 0.21\}$ and $N = 5{,}000$. When $p = 10$ or 30 the empirical rejection rates are zero at these alpha levels. In other words, G^2 will always retain the null hypothesis and the test becomes essentially useless.

Misspecified Models ($c > 0$)

In this case data were generated using a three parameter logistic model with common slopes and non-zero lower asymptote parameters and G^2(dif) was used to test the fit of a one parameter model against a two parameter model. As can be seen in Table 1, when the larger model is misspecified, the empirical rejection rates of G^2(dif) are higher than the expected rejection rates under a chi-square distribution with $q_0 - q_1$ degrees of freedom. The discrepancy between the empirical rejection and expected rejection rates increases when the degree of model misspecification increases, sample size increases, and model size increases. In fact, the empirical rejection rates reasonably match the expected rejection rates only under minor model misspecification provided the model and sample size is not too large. Even under small model misspecification the empirical rejection rates reasonably match the expected rejection rates if the model and sample size are sufficiently small. In the most extreme case considered (largest model size, largest sample size, and largest model misspecification) G^2(dif) always favors the largest model (two parameter logistic model) over the more parsimonious model (one parameter logistic model).

DISCUSSION

The reported simulation study illustrates that G^2(dif) can be safely used to assess model fit assessment for large models and under conditions of extreme data sparseness provided the least restrictive model holds. Indeed, when the least restrictive model holds relative fit assessment can be safely performed for much larger models than those whose absolute fit can be tested with the χ^2 and G^2 statistics. However, when the least restrictive model being compared is misspecified, statistical inferences based on G^2(dif) can be misleading. This is because in this case a chi-square distribution is no longer the appropriate large sample reference distribution for this statistic. Therefore, some indication of the goodness of fit of the largest model to the data is needed to justify the use of G^2(dif).

In the field of IRT modeling several methods have been proposed that enable researchers to obtain an indication of the absolute goodness of fit of the model being

fitted. For instance, Reiser (1996) proposed inspecting residuals for univariate and bivariate margins (see also Bartholomew & Tzamourani, 1999). Also, Drasgow, Levine, Tsien, Williams, and Mead (1995) proposed using graphical methods, as well as χ^2 statistics for single items, pairs of items, and triplets of items in IRT models. It may well be that an indication of absolute goodness of fit of the least restricted model by one of these (or similar methods) suffices to justify the use of G^2(dif) to assess relative model fit. This is because the G^2(dif) statistic is somewhat robust to small model misspecification of the least restrictive model. In this regard, Maydeu-Olivares, Morera, and D'Zurilla (1999) provided an example of the use of the graphical methods proposed by Drasgow et al. to assess absolute model fit followed by the use of G2(dif) to assess relative model fit in an IRT context.

In closing, some form of absolute goodness of fit testing should precede testing relative model fit. Comparing the fit of alternative models without some indication of overall model fit of the least restrictive model may result in misleading substantive conclusions in applications. It is clear that further research is needed to improve the assessment of overall goodness of fit in categorical data problems when model size is large and/or the contingency table is sparse.

Also, although the present study is confined to categorical variables, readers should bear in mind that a similar problem occurs when modeling continuous variables using structural equation modeling (SEM). In a recently published study, Yuan and Bentler (2004) clearly showed that when a likelihood ratio statistic is used to compare two nested models but the least restrictive model is misspecified inflated Type I errors are obtained. Their results thus concur with those presented here.

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Aptech Systems (2003). *GAUSS (version 6.0.8)* [Computer program]. Maple Valley, WA: Author.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociolological Methods and Research*, *27*, 525–546.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihodd estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315–345.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. (1993). Goodness of fit testing for latent class models. *Multivariate Behavioral Research*, *28*, 375–389.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polychotomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143–165.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, *5*, 1148–1169.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Koehler, K., & Larntz, K. (1980). An empirical investigation of goodness of fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*, 336–344.

Larntz, K. (1978). Small sample comparison of exact levels for chi-squared goodness of fit statistics. *Journal of the American Statistical Association*, *73*, 253–263.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maydeu-Olivares, A., Morera, O., & D'Zurilla, T. J. (1999). Using graphical methods in assessing measurement invariance in inventory data. *Multivariate Behavioral Research*, *34*, 397–420.

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, *61*, 509–528.

Thissen, D. (2003). *MULTILOG user's guide*. Chicago: SSI International.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–66). New York: Springer Verlag.

Tollenaar, N., & Mooijart, A. (2003). Type I errors and power of the parametric goodness-of-fit test. Full and limited information. *British Journal of Mathematical and Statistical Psychology*, *56*, 271–288.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer Verlag.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737–757.

Accepted January 2005