DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2

Article 26

11-1-2004

A New Goodness-of-Fit Test for Item Response Theory

John H. Neel
Georgia State University, epsjhn@gsu.edu

Recommended Citation
Neel, John H. (2004) "A New Goodness-of-Fit Test for Item Response Theory," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 2 , Article 26.
DOI: 10.22237/jmasm/1099268760
Available at: http://digitalcommons.wayne.edu/jmasm/vol3/iss2/26

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 581-593

1538 – 9472/03/$95.00

A New Goodness-of-Fit Test for Item Response Theory

John H. Neel
Educational Policy Studies
Georgia State University

Chi-square techniques for testing goodness-of-fit in item response theory are shown to give incorrect results. A new measure, CB, based on cumulants is proposed which avoids the arbitrary nature of interval creation found in chi-square techniques. The distribution of CB is estimated using Monte Carlo techniques and critical values for testing goodness-of-fit are given.

Key Words: Goodness-of-fit, item response theory, item fit

Introduction

Item response theory (IRT) posits a functional relationship between the probability of success on a test item and an unobserved latent variable. Although one may wish for robustness, how well the many applications of IRT function is determined at least in part, and certainly in some cases completely, by how well the model fits observed data. Model fit to data on a particular test item has been judged by various chi-square techniques. Yen (1981) reviewed these techniques, found similarity between several, and recommended Q1. Modifications of Q1 have been implemented in various computer programs such as Bilog (Mislevy, R.J. & Bock, R.D., 1990) and BilogMG-3 (Zimowski et al, 2004).

In this article, I review the use of chi-square in examination of item fit and show that the chi-square statistic is misleading in that it shows items to not fit when one might in fact consider the items to fit well and that it shows items to fit when one might in fact consider the item to not fit well. Next, I explain why the various variants of chi-square have these difficulties. Then I propose a new measure of item fit based on cumulants, show why this new technique is not susceptible to the problems of the chi-square techniques, and find critical points for this technique via Monte Carlo investigation of their distribution. Finally, I list some remaining research needs on this technique.

Use of Chi-square in Item Fit

Stone (2000) summarized the typical procedures for testing fit of IRT models: "(a) Item and ability parameters are estimated; (b) A small number of ability subgroups are formed (e.g., 10) to approximate the continuous ability distribution; (c) An observed score response distribution is constructed by cross-classifying examinees using their ability estimates and score responses. Using the IRT model, the item parameter estimates and an ability level representing the discrete ability subgroups (e.g., midpoint of ability subgroup), an expected score response distribution across score categories for an item is obtained; (e) These predictions are then compared with the observed score response distribution. This comparison generally involves computing a goodness-of-fit or chi-square statistic for each individual item (e.g. Bock

John H. Neel is Department Chair in Educational Policy Studies. His research interests are in statistical power, item response theory, and C programming. Contact him at epsjhn@gsu.edu

1972;Yen 1981), and/or an examination of residuals (Hambleton & Swaminathan, 1985).”

Notation

Following common notation θ is defined as ability and $P_i(\theta_j)$ as the probability of passing item i for ability θ_j. The three-parameter logistic model and its variant two- and one-parameter models are assumed for $P_i(\theta_j)$ throughout this article:

$$P_i\left(\theta\right)=c_i+(1-c_i)\frac{1}{1+e^{-1.702a_i(\theta-b_i)}}\ .$$

Further, U_{ij} is defined as 1 if examinee j has a correct answer to item i and 0 if not. Some additional notation is:

N - number of examinees
n_j - number of examinees with common ability θ_j
K - the number of unique ability levels

See Hambleton, Swaminathan and Rogers (1991) for further model and notation explanation.

Chi-square techniques are misleading

Like many statistical techniques the goodness-of-fit technique is susceptible to increasing sample size. As sample size increases, the tests become ever more powerful and more and more items are rejected. Figure 1 is a histogram showing the upper tail p-values associated with chi-square tests of goodness-of-fit for 1000 items. These tests come from simulated data on 20 tests of 50 items each. A three-parameter model with a lognormal distribution for b, the logistic model location parameter; an exponential distribution for a, the logistic model slope parameter, a beta distribution for c, the lower asymptote, and ability normally distributed with mean 0 and standard deviation 1 was used to create item responses for 2000 examinees on each test. Discussion and justification for the use of these distributions may be found in Baker (1992).

A three-parameter model was then fit using BilogMg. The p-values are the values from the chi-square goodness-of-fit for the items. It is clear from the figure that the p-values have positive skew. There should have been 50 (1000 x .05 = 50) p-values less than .05, however, there were 123, almost 2½ times as many as expected. Applying a test for proportions to these data to test whether the observed proportion, .123, of p-values less than .05, differs from the expected value of .05, we find a z value of 10.59 ($p<.0000000000000001$). In the sense that the data were created from the given model, we can view all items as fitting the model. The technique clearly rejects many more items as not fitting than should have been rejected. Similarly, testing at the .01 level we would expect to reject only 10 items but 40 would have been rejected for data that has adequate fit. Other conditions, for example, number of parameters in the IRT model, distribution of ability, size of calibration sample, will affect how many the items chi-square technique incorrectly identifies. In some cases the proportion of errors can be quite large. An exploration of these conditions is not the purpose of this study. Here it is only shown that the technique can in fact err on the side of identifying too many items that do not fit. The chi-square test thus does show items not to fit when one might in fact consider the items to fit well; i.e., 123 rejections when only 50 were expected.

That the chi-square techniques can show items to fit when the items do not fit can occur when proportions passing the items are different within the same interval on the ability scale. When this happens in the same interval, proportions that are too high are combined with proportions that are too low and the items thus seem to fit. This is discussed somewhat further in the next section.

Why are the chi-square techniques misleading?

Moore (1986) lists reasons that the chi-square techniques have problems. Among these are the “arbitrariness introduced by the necessity to choose cells” and “the discarding of information within the cells”. The arbitrariness of the cells is one of the main problems in the use of chi-square. As used in such statistics as Q1, equal intervals are created along the ability scale and a value of $P_i(\theta_j)$ is selected to represent the probability of success throughout the interval.

Figure 1. Upper tail p-values associated with chi-square tests of goodness-of-fit for 1000 items.

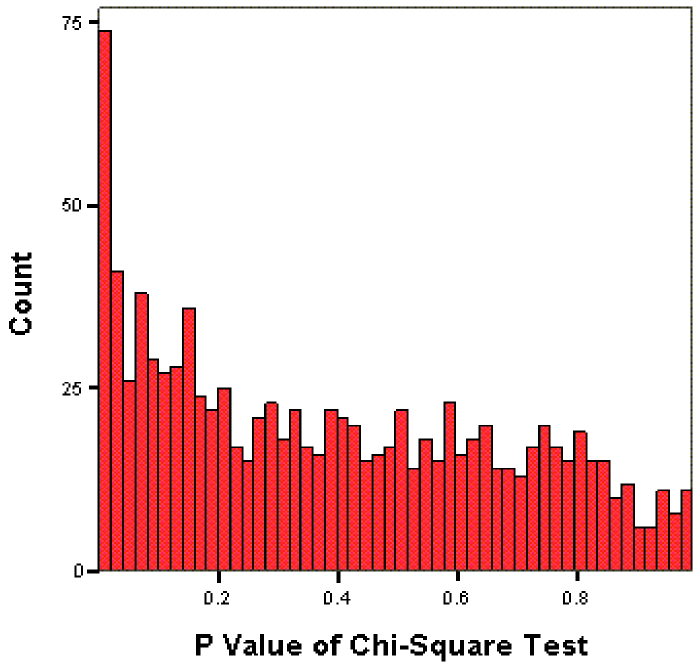
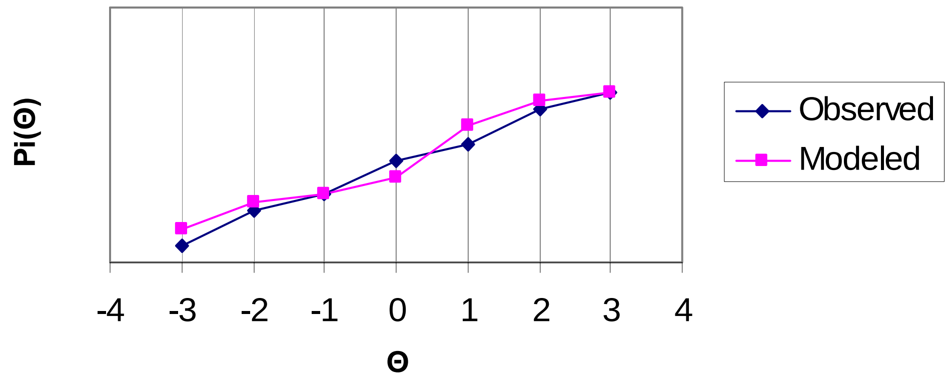


Figure 2
Forms the Areas Can Assume



How these intervals are created is arbitrary as is the length of the interval. In a particular case, the intervals that give a particular value of chi-square might give a different value if the intervals were either of a different length, began at a different point, or were both of a different length and began at different points.

A second problem is that Q1 uses the P_i value of the midpoint of the interval on the θ scale (other values, such as the maximum, minimum, or mean, might and have been used). In using this single value to represent all points in the interval, the possibly different probabilities throughout the interval are ignored. Treating all points in the interval as having the same $P_i(\theta_j)$ discards the information from the unequal $P_i(\theta_j)$ that exist across the interval due to the different values of θ_j. This is only worsened when intervals are combined, due to low sample size as is often done in chi-square goodness-of-fit tests, because a single value of $P_i(\theta_j)$ must then represent an even larger interval across the $P_i(\theta_j)$ scale.

Moreover, differences in observed proportions passing can be masked by the selection of intervals. This can happen if the first of two adjacent regions on the ability scale show a low proportion passing while the second shows a high proportion passing. If these two successive regions are included in the same interval, the total proportion passing could be very close to the appropriate and correct value.

Proposed Measure

In an attempt to bypass the difficulty of Q1 and similar grouped statistics, the modeled cumulative proportion passing an item is contrasted to the observed cumulative proportion passing. Consider that a given test was taken by N examinees resulting in ability estimates that are arranged in order from the smallest to the largest. Some of these ability estimates may be equal for different examinees and thus we might consider that we have J unique ability estimates and that we label these as $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_J; J \leq N$ with the general element being labeled as $\hat{\theta}_j$. We then let n_j be the number of equal ability estimates at $\hat{\theta}_j$; n_j will often be 1. Using the appropriate IRT model fit from the data, $P_i\left(\hat{\theta}_j\right)$ is the modeled probability of a correct response on item i at $\hat{\theta}_j$ and $n_j P_i\left(\hat{\theta}_j\right)$ is the modeled expected number of correct answers at $\hat{\theta}_j$. The cumulative modeled expected number of correct responses up to and including $\hat{\theta}_j$ is $\sum_{k=1}^{j} n_k P_i\left(\hat{\theta}_k\right)$.

In order to bring this cumulative modeled expected number of correct responses into a common range regardless of the difficulty of the item or the number of examinees taking the test, each of these values is divided by their maximum value,

$$MAX = \sum_{k=1}^{j} n_k P_i\left(\hat{\theta}_k\right),$$

thus setting the range of these values from 0 to 1 and these values represent the modeled cumulative proportion passing the item, $MCPP_j$:

$$MCPP_j = \frac{\sum_{k=1}^{j} n_k P_i\left(\hat{\theta}_k\right)}{MAX}.$$

$MCPP_j$ can be compared to the observed cumulative proportion passing, $OCPP_j$, by counting the number of examinees who got the item correct at each ability level, cumulate these counts at the ability levels, and divide by the MAX. Note that dividing by MAX only brings the maximum value of $OCPP_j$ to one if the total number of observed correct responses to the item is exactly equal to the cumulative modeled expected number of correct responses. This is unlikely in practice. Thus, the maximum value of $OCPP_j$ will be less than one when fewer than the total number of correct responses is obtained and it will be greater than one when more than the total number of correct responses is obtained.

The proposed measure is based upon comparisons of the differences between $MCPP_j$ and $OCPP_j$. The basic idea is to examine the area between two lines. One line is formed by plotting $MCPP_j$ at each level of ability and then connecting these points with straight lines. The second line is formed by plotting $OCPP_j$ at each level of ability and this second set of points is also connected using straight lines. Thus two lines are created each formed from a series of straight lines. The area between the lines is then taken as a measure of how much the lines diverge. If the area between the lines is zero, the two lines must coincide everywhere. In that case $MCPP_j$ equals $OCPP_j$ at every value of $\hat{\theta}_j$. As the lines diverge from each other, the area will grow larger. This is illustrated in Figure 2. The points in Figure 2 were selected for illustration purposes. In practice, the values of θ would not be evenly spaced and would likely not have integer values. For a typical test, there would be hundreds or thousands of unequally spaced θ values. In Figure 2, there are six areas bound by the vertical lines at -3, -2, -1, 0, 1, 2, and 3. These areas are of 3 types:

Trapezoid – bounded by (-3,-2) & (1,2)
Triangle – bounded by (-2,-1), (-1,0), &
(2,3)
Two triangles – as bounded by (0,1)

Formulas for the areas of these figures are well known. The only thing perhaps not well known is to find the point where the two triangles touch in the interval (0,1). This is a simple process of the simultaneous solution of the two intersecting lines, usually a topic in a beginning algebra course. A caution is to be sure that any area calculated is given a positive sign. Some areas could become negative if in finding a length of a side or an altitude, a larger value were subtracted from a smaller one. In any case, with this caution to pay attention to the signs of numbers, finding the area between the lines is a simple application of formulas for the areas of two common figures, trapezoids and triangles. The individual areas can be found and then added to obtain the total area between the two lines. I have labeled this area as CB, for Clifford Blair or the area Caught Between the lines.

I define this measure by two sources. First, CB is an area measure similar to the DIF measure defined by Raju (1988). Second, CB is an area measure that combines information from each ability level. There is no discarding of information and there is no arbitrariness of interval location or length because there are no intervals. The discarding of the intervals has been managed by the use of the cumulants.

An Example

Table 1 lists some created data to be used as an example to illustrate the proposed techniques. Table 1 contains 7 unique values of $\hat{\theta}_j$ with 20 examinees distributed across the $\hat{\theta}_j$ values. The number of examinees at each value of $\hat{\theta}_j$ is listed under n_j. The 20 examinees were distributed across the 7 ability levels to be suggestive of a normal distribution. $P_i\left(\hat{\theta}_j\right)$ is tabled for each value of $\hat{\theta}_j$ using a one-parameter model with b=0. The expected number of passes at each ability level is the number of examinees at that ability level times the probability of success at the ability level. These are listed under

$$n_j P_i\left(\hat{\theta}_j\right).$$

The cumulative expected number of passes at each ability level is the sum of the expected number of passes up to that ability level. These are listed under

$$\sum_{k=1}^{j} n_k P_i\left(\hat{\theta}_k\right).$$

As discussed earlier these values are divided by their maximum value, MAX, which is the last value of

$$\sum_{k=1}^{j} n_k p_i\left(\hat{\theta}_k\right).$$

The u_{ij} values listed in Table 1 were selected for the subjects so that the observed number of passes was always within one unit of the expected number of passes. In the sense that the observed number of passes could not be made any closer, we can say that these data fit the model. The observed cumulative proportion of passes, OCPPj, was found by cumulating the number of passes up to and including an ability level and then dividing by MAX.

Table 1. Data Illustrating Good Fit.

j	$\hat{\theta}_j$	n_j	$P_i(\hat{\theta}_j)$	$n_j P_i(\hat{\theta}_j)$	$\sum_{k=1}^{j} n_k P_i(\hat{\theta}_k)$	$MCPP_j = \frac{\sum_{k=1}^{j} n_k P_i(\hat{\theta}_k)}{MAX}$	u_{ij}	$\sum_{k=1}^{j} u_{ki}/MAX$	AREA
1	-3	1	.0000	.0060	.0060	0.0006	0	.0000	
2	-2	3	.0322	.0965	.1025	0.0106	000	.0000	0.006
3	-1	4	.1542	.6168	.7194	0.0745	0010	.1036	0.020
4	0	5	.5000	2.5000	3.2194	0.3335	10101	.4143	0.055
5	1	3	.8458	2.5374	5.7567	0.5963	101	.6215	0.053
6	2	3	.9678	2.9035	8.6602	0.8970	111	.9322	0.030
7	3	1	.9940	0.9940	9.6542	1.0000	1	1.0358	0.036
		J=7			MAX = $\sum_{k=1}^{j} n_k P_i(\hat{\theta}_k) =$ 9.6542				CB= .200

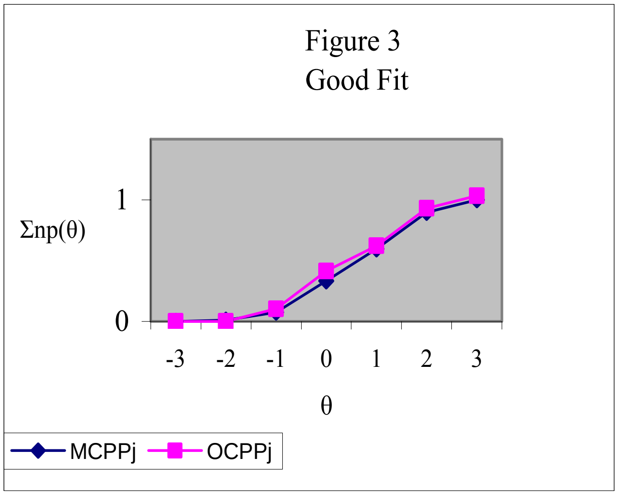


Figure 3

Good Fit

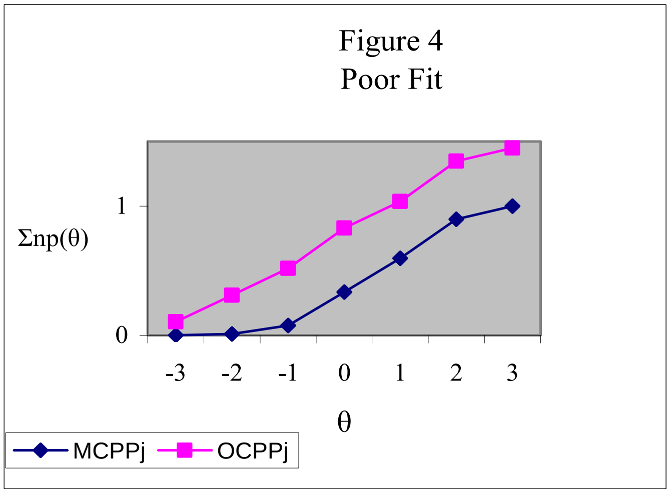


Figure 4

Poor Fit

Figure 3 represents a plot the MCPPj and OCPPj on the vertical axis with $\hat{\theta}_j$ on the horizontal axis. The lines are formed by connecting the MCPPj and OCPPj Points.

Figure 3 represents rather good fit of the data in that the observed number of passes was selected to be within one unit of the expected number of passes for each ability level. This is in contrast to Figure 4. Figure 4 was created from the data of Table 2 just as Figure 3 was created from Table 1.

Table 2 presents data created to show poor model fit by changing the u_{ij} values while keeping the same abilities and one-parameter model as Table 1. The u_{ij} values at the first three levels were selected to represent more passes than the model indicates. Accordingly the areas for the two situations differ. The CB area is found in both Table 1 and Table 2 by finding the area for the various trapezoids and triangles and then adding these areas for the CB area. The CB area for the good fit of Table 1 and Figure 3 is .200 while the CB area for the poor fit of Table 2 and Figure 4 is 2.40. This is in the direction expected. CB should be less when the fit is good and greater when the fit is poor. Comparing these two areas brings up the question of when is the fit good and when is it poor? One answer to this question is to test the hypothesis that the fit is good. In order to test that hypothesis, the probability distribution of CB needs to be known. To determine the probability distribution of CB, the distribution of the area was simulated under known conditions.

Simulations of Null Distributions

Because each of the measures proposed here is based on cumulative passing rates, there is a dependence between the $OCPP_j$ values and the $MCPP_j$ values. This means that finding probability distributions of these statistics through an analytic solution is difficult because of the dependencies introduced by the cumulants. Consequently a Monte Carlo simulation of the probability distributions is often used to estimate percentage points of such distributions. See Stephens (1986) for such a study. A Monte Carlo study was conducted to estimate percentage points of the distributions of the statistic proposed here for its null distributions; i.e. using data that were generated from known models under the null hypothesis that the data fit. Since the data were created from known models these data thus always fit the model so that the null hypothesis that the data fit was always true. I simulated data for one-, two- and three-parameter logistic IRT models over all combinations of the following numbers of items and number of subjects:

Numbers of items: 10, 20, 30, 50, 75, 100, 150, 300
Number of examinees: 100, 200, 300, 500, 800, 1000, 2000, 3000, 4000

There are 240 combinations of model, number of items, and number of subjects, 3 x 8 x 10. The programming was done such that each of these 240 combinations could be run without intervention. Each combination was termed a "run". For each run data was simulated until 50,000 items were available. For each test, I created the 1- 0, pass-fail, item data for the given model, estimated item parameters using BILOGMG (Mislevy, R.J. & Bock, R.D, 1990), calculated CB, and saved these statistics along with appropriate identifying information to a file. I wrote a program to find the percentage points 1, 2, . . . , 99, 99.5, 99.9, and 99.99 from these files and tabled the resulting points.

In creating the 1- 0, pass-fail, data I used a standard normal distribution for abilities; a lognormal distribution for b, the logistic model location parameter; an exponential distribution for a, the logistic model slope parameter; and a beta distribution for c, the logistic model lower slope asymptote. I checked the accuracy of the implementation of these distributions by comparing sample values from each with values from the SPSS functions for these distributions. Agreement to 4 decimal places or beyond was found in each case.

I adapted a program by Wu (1997) to use as a random number generator. I added a 1000 number shuffling routine (Press et al, 1988) to the random number generator. Without shuffling, Wu's random number generator has a period of approximately 2.3 x 10^18, more than sufficient to not repeat for the numbers used here. Addition of the shuffler increases the

Table 2. Data Illustrating Poor Fit.

j	$\hat{\theta}_j$	n_j	$P_i(\hat{\theta}_j)$	$n_jP_i(\hat{\theta}_j)$	$\sum_{k=1}^{j} n_k P_i(\hat{\theta}_k)$	MCPP$_j$ = $\dfrac{\sum_{k=1}^{j} n_k P_i(\hat{\theta}_k)}{\text{MAX}}$	u_{ij}	$\sum_{k=1}^{j} u_{ki}$/MAX	AREA
1	-3	1	.0000	.0060	.0060	0.0006	1	0.1036	
2	-2	3	.0322	.0965	.1025	0.0106	011	0.3107	0.202
3	-1	4	.1542	.6168	.7194	0.0745	0110	0.5179	0.372
4	0	5	.5000	2.5000	3.2194	0.3335	1010	0.8287	0.469
5	1	3	.8458	2.5374	5.7567	0.5963	101	1.0358	0.467
6	2	3	.9678	2.9035	8.6602	0.8970	111	1.3466	0.444
7	3	1	.9940	0.9940	9.6542	1.0000	1	1.4501	0.449
	J=7				MAX = $\sum_{k=1}^{j} n_k P_i(\hat{\theta}_k)$ = 9.6542				CB= 2.40

period of the random number generator and, more importantly, removes lag correlation from the generated data.

Use of the Tables

Tables 3, 4, and 5 give the .05, .01, and .001 upper area points of CB for one-, two-, and three-parameter models. These values can be used for a hypothesis test for the goodness-of-fit at significance levels of .05, .01, and .001. To conduct the test, calculate CB for a given item and then compare the item to tabled value. If CB exceeds the tabled value, then fit is rejected at the significance level for that value. If CB does not exceed the value, then fit is not rejected. As an example, if a 50 item test is calibrated on a sample of 1000 examinees and CB for an item is found to be .015, then fit for that item would be rejected at the .05 level ($CB_{.05}$ = .0142), but would not be rejected at the .01 or the .001 levels ($CB_{.01}$ = 0.0186, $CB_{.001}$ = .0255). Complete tables for numbers of items equal to 10, 20, 30, 50, 75, 100, 150, and 300; calibration sample sizes of 100, 200, 300, 500, 800, 1000, 2000, 3000, and 4000; and for one, two, and three parameter models may be obtained from the author. These tables list the percentage points 1-99 (in increments of .01), 99.5, 99.9, and 99.99. Four point interpolation within the table should work well so that the tables should

Table 3. Monte Carlo Estimated Upper Area Points of CB for One-Parameter Models.

					K				
		10	20	30	50	75	100	150	300
N	α								
100	.05	.0502	.0368	.0339	.0348	.0371	.0394	.0447	.4275
	.01	.0643	.0456	.0416	.0430	.0471	.0508	.0638	.6450
	.001	.0804	.0555	.0510	.0540	.0640	.0702	.0883	.8136
200	.05	.0495	.0333	.0274	.0252	.0260	.0269	.0289	.2141
	.01	.0634	.0421	.0339	.0307	.0322	.0341	.0384	.5251
	.001	.0774	.0519	.0414	.0374	.0413	.0467	.0584	.7110
300	.05	.0494	.0325	.0250	.0215	.0215	.0221	.0233	.1839
	.01	.0632	.0410	.0308	.0262	.0266	.0280	.0314	.4283
	.001	.0782	.0525	.0384	.0316	.0334	.0372	.0461	.5183
500	.05	.0494	.0314	.0231	.0182	.0175	.0177	.0184	.0522
	.01	.0636	.0401	.0285	.0219	.0216	.0223	.0245	.1644
	.001	.0779	.0501	.0351	.0260	.0273	.0307	.0369	.2668
800	.05	.0490	.0312	.0219	.0160	.0149	.0147	.0149	.0356
	.01	.0629	.0396	.0271	.0194	.0184	.0188	.0205	.3474
	.001	.0765	.0502	.0327	.0233	.0235	.0272	.0319	.3727
1000	.05	.0493	.0309	.0217	.0152	.0139	.0136	.0137	.0173
	.01	.0635	.0393	.0262	.0183	.0172	.0178	.0191	.2408
	.001	.0783	.0490	.0318	.0219	.0228	.0247	.0312	.3012
1500	.05	.0490	.0307	.0211	.0140	.0124	.0119	.0118	.0130
	.01	.0630	.0388	.0258	.0166	.0155	.0159	.0173	.1137
	.001	.0768	.0482	.0307	.0194	.0212	.0228	.0276	.1954
2000	.05	.0494	.0307	.0207	.0133	.0115	.0111	.0108	.0116
	.01	.0633	.0384	.0250	.0159	.0147	.0149	.0162	.0234
	.001	.0771	.0470	.0297	.0187	.0191	.0221	.0266	.2245
3000	.05	.0491	.0306	.0204	.0126	.0106	.0101	.0098	.0097
	.01	.0636	.0385	.0244	.0149	.0135	.0139	.0150	.0164
	.001	.0762	.0466	.0287	.0171	.0183	.0212	.0244	.0301
4000	.05	.0490	.0306	.0203	.0122	.0101	.0096	.0092	.0091
	.01	.0626	.0386	.0240	.0142	.0129	.0137	.0148	.0163
	.001	.0760	.0466	.0282	.0165	.0181	.0211	.0243	.0285

N - Number of examinees in the calibration sample
α - Upper tail area
K - number of items on the test

The tabled value is the Monte Carlo estimated point that cuts off an area of α in the upper tail of the distribution of CB when the item and ability parameters were estimated for a one-parameter logistic IRT model with a calibration sample of size N on a K item test.

Table 4. Monte Carlo Estimated Upper Area Points of CB for Two-Parameter Models.

		10	20	30	50	75	100	150	300
N	α								
100	.05	.0464	.0344	.0362	.0416	.0458	.0493	.0536	.0594
	.01	.0600	.0446	.0487	.0585	.0666	.0720	.0798	.0923
	.001	.0768	.0639	.0689	.0872	.1009	.1094	.1293	.1480
200	.05	.0478	.0286	.0253	.0272	.0303	.0318	.0351	.0390
	.01	.0635	.0363	.0324	.0365	.0423	.0445	.0507	.0585
	.001	.0802	.0480	.0436	.0520	.0589	.0651	.0714	.0909
300	.05	.0489	.0277	.0221	.0220	.0242	.0258	.0277	.0310
	.01	.0652	.0356	.0278	.0289	.0328	.0364	.0394	.0458
	.001	.0846	.0486	.0358	.0380	.0448	.0504	.0561	.0686
500	.05	.0497	.0281	.0201	.0176	.0188	.0200	.0217	.0244
	.01	.0656	.0371	.0251	.0230	.0254	.0276	.0305	.0355
	001	.0831	.0490	.0332	.0305	.0335	.0381	.0435	.0516
800	.05	.0503	.0284	.0191	.0151	.0158	.0165	.0176	.0198
	.01	.0662	.0377	.0245	.0195	.0213	.0228	.0253	.0297
	.001	.0848	.0492	.0323	.0265	.0294	.0319	.0365	.0443
1000	.05	.0505	.0288	.0191	.0142	.0145	.0152	.0163	.0182
	.01	.0659	.0378	.0247	.0186	.0199	.0214	.0237	.0272
	.001	.0837	.0485	.0332	.0255	.0280	.0322	.0361	.0408
1500	.05	.0504	.0296	.0190	.0131	.0129	.0136	.0141	.0160
	.01	.0657	.0384	.0245	.0173	.0178	.0196	.0211	.0244
	.001	.0812	.0509	.0322	.0253	.0258	.0298	.0336	.0392
2000	.05	.0508	.0292	.0190	.0126	.0121	.0124	.0133	.0149
	.01	.0665	.0387	.0245	.0170	.0167	.0180	.0205	.0232
	.001	.0824	.0490	.0339	.0251	.0253	.0278	.0328	.0386
3000	.05	.0510	.0295	.0187	.0120	.0115	.0116	.0121	.0134
	.01	.0667	.0384	.0243	.0167	.0168	.0174	.0190	.0215
	.001	.0838	.0483	.0328	.0258	.0258	.0287	.0316	.0368
4000	.05	.0512	.0296	.0190	.0118	.0109	.0112	.0121	.0131
	.01	.0661	.0385	.0244	.0167	.0161	.0176	.0194	.0212
	.001	.0832	.0498	.0318	.0252	.0255	.0277	.0326	.0369

N - Number of examinees in the calibration sample
α - Upper tail area
K - number of items on the test

The tabled value is the Monte Carlo estimated point that cuts off an area of α in the upper tail of the distribution of CB when the item and ability parameters were estimated for a two-parameter logistic IRT model with a calibration sample of size N on a K item test.

Table 5. Monte Carlo Estimated Upper Area Points of CB for Three-Parameter Models

		K							
		10	20	30	50	75	100	150	300
N	α								
100	.05	.0954	.0797	.0761	.0766	.0801	.0832	.0893	.1009
	.01	.1112	.0911	.0886	.0919	.0990	.1044	.1168	.1374
	.001	.1274	.1039	.1034	.1161	.1325	.1414	.1610	.1894
200	.05	.0903	.0687	.0589	.0519	.0510	.0525	.0559	.0620
	.01	.1068	.0821	.0703	.0617	.0624	.0660	.0743	.0836
	.001	.1240	.0950	.0832	.0746	.0834	.0938	.1007	.1150
300	.05	.0884	.0668	.0553	.0442	.0413	.0412	.0434	.0457
	.01	.1054	.0815	.0682	.0537	.0507	.0534	.0570	.0606
	.001	.1253	.0953	.0820	.0655	.0654	.3142	.0764	.0806
500	.05	.0873	.0655	.0537	.0397	.0337	.0325	.0321	.0311
	.01	.1048	.0805	.0680	.0499	.0419	.0419	.0411	.0392
	001	.1253	.0955	.0826	.0639	.0566	.0550	.0573	.0573
800	.05	.0863	.0655	.0533	.0378	.0307	.0288	.0266	.0239
	.01	.1037	.0815	.0687	.0487	.0391	.0371	.0337	.0299
	.001	.1217	.0969	.0837	.0648	.0512	.0482	.0444	.0422
1000	.05	.0858	.0655	.0531	.0370	.0299	.0283	.0243	.0214
	.01	.1030	.0821	.0691	.0478	.0387	.0368	.0307	.0268
	.001	.1245	.0972	.0843	.0628	.0530	.2714	.0413	.0395
1500	.05	.0858	.0654	.0532	.0369	.0289	.0260	.0224	.0173
	.01	.1034	.0820	.0694	.0476	.0374	.0342	.0281	.0216
	.001	.1242	.0963	.0851	.0621	.0497	.2714	.0426	.0366
2000	.05	.0859	.0652	.0530	.0365	.0293	.0259	.0207	.0158
	.01	.1031	.0818	.0686	.0477	.0385	.0331	.0263	.0200
	.001	.1241	.0959	.0857	.0613	.0535	.0456	.0386	.0309
3000	.05	.0848	.0653	.0532	.0360	.0290	.0258	.0189	.0145
	.01	.1020	.0823	.0701	.0472	.0376	.0324	.0241	.0184
	.001	.1242	.0964	.0910	.0630	.0481	.0448	.0388	.0352
4000	.05	.0851	.0653	.0531	.0362	.0294	.0253	.0189	.0138
	.01	.1038	.0819	.0694	.0479	.0382	.0324	.0239	.0179
	.001	.1269	.0961	.0868	.0629	.0498	.3112	.0371	.0360

N - Number of examinees in the calibration sample
α - Upper tail area
K - number of items on the test

The tabled value is the Monte Carlo estimated point that cuts off an area of α in the upper tail of the distribution of CB when the item and ability parameters were estimated for a three-parameter logistic IRT model with a calibration sample of size N on a K item test.

provide adequate support for testing items on common sized tests and with common calibration sample sizes.

Future Research

How well these procedures work will depend on many factors. One such factor is how well the assumed distributions for ability and for the parameters of the one-, two-, and three-parameter logistic item response theory model match sample data. Accordingly, some studies of fit that examine CB for real data together with the distributional assumptions made here will be important. Although each set of three points, for α = .05, .01, and .001, is based on 50,000 items, a simulation with more items might be necessary to obtain better estimated upper area points.

This could be time consuming for it took about 180 days of 400 megahertz computer time to complete the Monte Carlo portion of this study. Another factor will be how well CB compares in terms of power to other procedures such as the Q1 procedure. Studies comparing the power of such procedures will help.

Yet another factor is how well the interpolation will work. That would require comparison of interpolated points from this study with values that are found by simulation just as these values were found. Finally, given the ever increasing speed of modern computing, it is probably possible to simulate any given observed situation and estimate the required percentage points required for each test of goodness-of-fit.

For example, one might assume that the estimated ability levels in a given calibration sample were correct and then find the analogous points to those in this study for use in testing goodness-of-fit. The advantage of using the estimated abilities is that they should represent the distribution of ability and thus instead of assuming a distribution of abilities, such as was done in this study, the distribution of abilities is estimated from the observed data.

This should give a procedure that is stronger in the sense that it is not necessary to make one of the assumptions that was made here. It is also possible to make a similar use of the estimated logistic model parameters and obtain a similar benefit.

Conclusion

Weaknesses of traditional chi-square tests (e.g. Q1) of goodness-of-fit in item response theory are well known and have been shown here. An attempt to avoid these weaknesses was made by basing a statistic, CB, based on cumulants. Using cumulants avoided the arbitrary creation of intervals that causes difficulties in Q1 and thus might avoid the weaknesses of such chi-square statistics. Examples of CB were given under conditions of good and poor fit. Percentage points in the probability distribution of CB were estimated from a Monte Carlo study and an example given to show the use of these points. Suggestions were made regarding additional work with CB.

References

Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. New York: Marcel Dekker.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-Fit techniques*. New York: Marcel Dekker.

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

Mislevy, R. J. & Bock, R. D. (1990). BILOG [Computer software]. Chicago: Scientific Software, Inc.

Moore, D. S. (1986). Tests of chi-squared type. In D'Agostino, R. B. & Stephens, M. A. (Eds.), *Goodness- of-fit techniques* (pp. 63-96). New York: Marcel Dekker.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Betterling, W. T. (1988). *Numerical recipes in C*. New York: Cambridge University Press.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.

Snedecor, G. W. (1956). Statistical Methods. Ames, Iowa: Iowa State University Press.

Stephens, M. A. (1986). Tests based on regression and correlation. In D'Agostino, R. B. & Stephens, M. A. (Eds.), *Goodness-of-fit techniques* (pp. 63-96). New York: Marcel Dekker.

Stone, C.A. (2000). Monte-Carlo based null distribution for an alternative fit statistic. *Journal of educational measurement,* 37, 58-75.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *applied psychological measurement,* 5, 245-262.

Wu, P. (1997). Multiplicative, congruential random-number generators with multiplier +- 2^k1 +-2^k2 and modulus 2^p-1. *ACM transactions on mathematical software*, 23(2), 255-265.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2004). BILOG-MG 3 [Computer software]. Chicago: Scientific Software, Inc.

Appendix

I met R. Clifford Blair during my first years as a professor in the College of Education at the University of South Florida. There are two incidents I would like to relate about Cliff that may give the reader some insight into his character. The first time he was a student in my class at the University of South Florida, he explained that he was legally blind and asked if he could record the classes. I, of course, consented and he routinely recorded every class. I was concerned as to how a student with limited vision would handle some of the basic statistical formulas and other mathematics in the class. I thought of it with my own limitations and how difficult it would be for me not to be able to see things. I was not very sophisticated as to how other people used alternative methods to learn.

One of the courses Cliff took from me was a course in test construction for teachers. Students in such a course soon consider themselves great experts at test construction and are often very critical of the tests they have in that course. When I returned the first test and went over it with the class, one student became very upset at a particular multiple choice item he had missed. He said that I had said a particular thing in class and that made the item choice he had selected correct. I replied that I would never have said that because it was clearly wrong and he must have misunderstood me. Another student jumped in and said that no, I had stated it just as the first student said and he had it in his class notes. The conversation went on a bit and I was beginning to think that I really had made an error. At the time, I was too new to want to admit such a thing. I did not want to admit to myself that I had told the class anything wrong and certainly did not want to admit it to the class. Things were going worse for me as two other students began to support the first two when Cliff raised his hand and said, “Just a minute, I have it on tape here.” Now I was really in difficulty. He had the evidence and I would have to hear it in front of everyone. He pressed the play button and there it was in my own voice: exactly what I told the students I had said. They had both written it down incorrectly. I have respected and appreciated Cliff Blair ever since.

I left the University of South Florida and came to Georgia State University. After a few years I took a trip back and went to see some old friends. There was a faculty lounge that was about the size of a large classroom. The door was near one corner of the room and Cliff was seated at a table in the far corner when I walked in. He had not known that I was coming but after two or three steps into the room, he stood up, greeted me, and invited me to sit down with him. After a bit of discussion, I reminded him that he had not seen me for several years and that he did not know I was coming. “How could you recognize me”, I asked. He explained first that I was far enough away that his small area of useful vision could take in most of my body and that to him I have a characteristic walk and profile. From that, he recognized me.

Cliff is a surprising man who doesn’t seem to have limits. He was always an excellent student and just as good a friend. In the test question incident, he identified the class (it was three classes back as I remember) that contained the discussion, found the tape, rewound it to the right point, and had it ready to play in a very short time. He was extremely well organized in both his recall of the situation and in his collection of tapes. In the lounge incident, he showed me how well he could use the abilities he had. He has used them well and has had a productive and profitable career. He is a respected and sought instructor. I am proud to have been around as he started that career. So I am naming this technique for him as others have done (Snedecor, 1956, p. 244) to thank him for the privilege of knowing him all these years.