The Spanish Journal of Psychology
2005, Vol. 8, No. 1, 100-110

1138-7416

Fitting Rasch Model using Appropriateness Measure Statistics

José Antonio López Pina and M. Dolores Hidalgo Montesinos
University of Murcia

In this paper, the distributional properties and power rates of the Lz, Eci2z, and Eci4z statistics when they are used as item fit statistics were explored. The results were compared to t-transformation of Outfit and Infit mean square. Four sample sizes were selected: 100, 250, 500, and 1000 examinees. The abilities were uniform and normal with mean 0 and standard deviation 1, and uniform and normal with mean –1 and standard deviation 1. The pseudo-guessing parameter was fixed at .25. Two ranges of difficulty parameters were selected: ±1 logits and ±2 logits. Two test lengths were selected: 15 and 30 items. The results showed important differences between the T-infit, T-outfit, Lz, Eci2z, and Eci4z statistics. The T-oufit, T-infit, and Lz statistics showed poor standardization with estimated parameters because their distributional properties were not close to the expected values. However, the Eci2z and Eci4z statistics showed satisfactory standardization on all conditions. Further, the power rates of Eci2z and Eci4z were 5% to 10% higher than the power rates of Lz, T-outfit, and T-infit to detect items that do not fit Rasch model.
Keywords: Rasch model, item response theory, appropriateness measure, item fit statistics

El objetivo de este trabajo fue estudiar la potencia y propiedades distribucionales de tres estadísticos de medida de la adecuación cuando se utilizan como estadísticos de ajuste de los ítems. Los estadísticos sometidos a comparación fueron: Lz, Eci2z y Eci4z. Los resultados obtenidos se compararon con los estadísticos T-outfit y T-infit. Se seleccionaron cuatro tamaños muestrales: 100, 250, 500 y 1000 sujetos. Se sometieron a estudio distintas distribuciones de habilidad: uniforme y normal, con media 0 y desviación típica 1, y uniforme y normal con media –1 y desviación típica 1. El parámetro de pseudo-azar fue fijado en .25. Para los parámetros de dificultad se utilizaron dos distribuciones uniformes de ±1 logits y ±2 logits. Por ultimo, se consideraron dos longitudes de tests: 15 y 30 ítems. Los resultados mostraron que los estadísticos Lz, T-outfit y T-infit no tienden a los valores esperados cuando se calculan con parámetros estimados, mientras que los estadísticos Eci2z y Eci4z mantuvieron mejor las propiedades de sus distribuciones teóricas. Además, la potencia de estos dos últimos estadísticos para detectar ítems no ajustados al modelo de Rasch estuvo entre un 5% y un 10% más que la potencia de los estadísticos Lz, T-outfit y T-infit.
Palabras clave: modelo de Rasch, teoría de la respuesta al ítem, medida de la adecuación, estadísticos de ajuste de ítems

Correspondence should be addressed to: José A. López Pina, Depto. de Psicología Básica y Metodología, Facultad de Psicología, Campus de Espinardo, 30100-Murcia (Spain). Phone: 968-363478. Fax: 968-364115. E-mail: jlpina@um.es
Translation: Virginia Navascués Howard

Psychological measurement has changed from the massive administration of the classic test model (Gulliksen, 1950; Lord & Novick, 1968) to the use of mathematical models that demand severe restrictions of the data to justify that the test designed measures the attribute it is meant to measure. One of the models that has generated the most research ever since its publication in 1960 has been the model of Rasch (Fisher & Molenaar, 1995; Rasch, 1960; Van der Linden & Hambleton, 1997; Wright & Stone, 1979), which shares with the other item response theory (IRT) models the assumptions of unidimensionality and local independence. Unidimensionality is evidence that the items essentially measure one and only one attribute (Stout, 1987), whereas local independence is evidence that the responses to an item are not influenced by the responses to prior or subsequent items, or that the responses of a group of subjects of the same ability are not related to each other (Hambleton & Swaminathan, 1995). If these assumptions are met, then Rasch's probabilistic model guarantees a unidimensional scale of the attribute measured, where the separability of item parameters and examinee abilities is a reality instead of a mere hypothesis assumed in the model (Bond & Fox, 2001). But, as occurs with the rest of the models in the IRT framework, Rasch's model (Rasch, 1960; Wright &Stone, 1979) does not assume that unidimensionality and local independence are mere hypotheses that are deduced after the model has been fitted, but instead they should be empirically proved. That is, before stating that a set of items met the Rasch model's expectations, studies of fit must be performed, both on the items and on examinee response patterns, in order to determine the extent to which the responses obtained follow the pattern expected in the model.

Several statistics have been proposed to prove that item response patterns and/or examinee patterns met the model characteristics. Some fit statistics have been specifically developed for Rasch's model, such as the residual statistics of Wright and Stone (1979), whereas statistics based on the likelihood function (Drasgow & Levine, 1986; Levine & Rubin, 1979) and on the comparison of item characteristic curves—ICCs— (Harnish & Tatsuoka, 1983; Tatsuoka, 1984) can be used in any of the dichotomic item response models: Rasch's model (Rasch, 1960; Wright & Stone, 1979), the 2-p logistic model (Birnbaum, 1968; Lord, 1980), and the 3-p model (Birnbaum, 1968; Lord, 1980). Generally, these statistics have standardized versions under the normal curve that allow making decisions about fit with specific significance levels.

The residual statistics were developed to study the fit of the items and of examinee response patterns to the model (Wright & Masters, 1982; Wright & Stone, 1979), whereas the statistics based on the likelihood function (Drasgow & Levine, 1986) and those that use the comparison of ICCs (Harnish & Tatsuoka, 1983) were developed exclusively to study the fit of the examinee response pattern to the proposed model, generating a research field known as appropriateness measure (Hulin, Drasgow, & Parsons, 1983). However, as with the residual statistics, appropriateness measure statistics can be applied as item fit statistics and vice versa, item fit statistics as statistics to study the degree of aberration of examinee response patterns (Reise, 1990).

Outfit Statistic

The unweighted total fit statistic (Outfit) is based on the residual obtained from subtracting the probability predicted by the model as a function of the estimated parameters from the observed response. It is calculated as:

$$MS(UT) = \frac{1}{N}\sum_{i=1}^{N}\frac{(U_i - P_{ij})^2}{w_{ij}} \tag{1}$$

where U_{ij} is the observed response for the subject i in the item j, P_{ij} is the probability of a correct response according to the Rasch model, $w_{ij} = P_{ij}\,(1 - P_{ij})$, and N is the sample size. Note that w_{ij} is the information function of the item defined in the model. The standard deviation of this statistic can be estimated by:

$$\sigma(UT) = \frac{\left[\sum_{i=1}^{N}\frac{1}{w_{ij}} - 4N\right]^{1/2}}{N} \tag{2}$$

The unweighted total fit statistic (MS(UT)) follows a χ^2 distribution with one degree of freedom. Its mathematical expectation is 1 and its standard deviation is obtained by Equation 2. Smith (1991) and Smith, Schumacker and Bush (1998) showed that there is no unique critical value to study item fit, but instead it depends on sample size and on information function. Moreover, this statistic is highly affected both by high-ability subjects' unexpected incorrect responses to easy items and by low-ability subjects' unexpected correct responses to difficult items.

Infit Statistic

The weighted total fit statistic (Wright & Masters, 1982) has the following form:

$$MS(UT) = \frac{\sum_{i=1}^{N}(U_{ij} - P_{ij})^2}{\sum_{i=1}^{N}w_{ij}} \tag{3}$$

where U_{ij}, P_{ij}, and w_{ij} are interpreted as in Equation 1. In this statistic, the residual is weighted by the information function, which reduces the influence of extreme values. Its standard deviation is:

$$\sigma(UT) = \frac{\left[\sum_{i=1}^{N} w_{ij} - 4 \sum_{i=1}^{N} w_{ij}^{2}\right]^{1/2}}{\sum_{i=1}^{N} w_{ij}} \qquad (4)$$

Both statistics, Outfit and Infit, have been standardized under the normal distribution by the following transformation:

$$t = \left[(MS^{1/3} - 1)\left(\frac{3}{\sigma}\right)\right] + \left[\frac{\sigma}{3}\right] \qquad (5)$$

where *MS* is the mean square of Equations 1 or 3, σ is the standard deviation of Equations 2 or 4. As Lz and the indexes ECI2z and ECI4z of Tatsuoka (1984) are also standardized under the normal distribution, in this study, we will use the *t* transformation of the Outfit and Infit mean square, which we shall call T-outfit and T-infit. Values of T-outfit and T-infit of less than –2 indicate less variation than expected by the model, which means that the response pattern is fairly close to the expected Guttman pattern, whereas values of T-out and T-infit higher than +2 indicate that the response pattern obtained has more randomness than expected by the model.

Lz Statistic

The Lz statistic is calculated by:

$$l_z = \frac{l(\theta) - \epsilon[l(\theta)]}{\{Var\,[l(\theta)]\}^{1/2}} \qquad (6)$$

where

$$l(\theta) = \sum_{i=1}^{N}[U_{ij}\,(\mathrm{In}P_{ij}) + (1\text{–}U_{ij})(\mathrm{In}Q_{ij})]$$

$$\epsilon[l(\theta)] = \sum_{i=1}^{N}[P_{ij}\,(\mathrm{In}P_{ij}) + Q_{ij}(\mathrm{In}Q_{ij})]$$

$$Var[l(\theta)] = \sum_{i=1}^{N} P_{ij}Q_{ij}\left[\mathrm{In}\left(\frac{P_{ij}}{Q_{ij}}\right)\right]^{2}$$

where U_{ij} and P_{ij} are defined as in the Outfit and Infit statistics, and $Q_{ij} = 1 - P_{ij}$. This statistic follows a standardized normal distribution when calculated with true item and subject parameters (Reise, 1990). Negative Lz values are associated with unlikely response patterns, whereas positive values are associated with more consistent response patterns than expected by the model.

Eci2z and Eci4z Statistics

Tatsuoka and Linn (1983) developed six caution statistics (ECI1 to ECI6) to detect aberrant response patterns. These statistics were standardized and adapted under the IRT by Tatsuoka (1984). In this study, we will use two statistics (ECI2z and ECI4z) out of the six original ones, because Tatsuoka (1984) suggested that ECI4z and ECI6z have identical standardized forms, and the correlation between ECI1z and ECI2z was very close to 1.

Caution statistics are based on the ratio between two covariances. The numerator is the covariance between the observed item patterns and the test response patterns, whereas the denominator is the covariance between the pattern expected by the model and the Guttman pattern. The mathematical expression of the statistic ECI2z is:

$$ECI2z = \frac{\sum_{i=1}^{N}(P_{ij} - U_{ij})(G_i - \mu_G)}{\left[\sum_{i=1}^{N} P_{ij}Q_{ij}\,(G_i - \mu_G)^2\right]^{1/2}} \qquad (7)$$

and of the ECI4z statistic:

$$ECI4z = \frac{\sum_{i=1}^{N}(P_{ij} - U_{ij})(P_{ij} - \mu_P)}{\left[\sum_{i=1}^{N} P_{ij}Q_{ij}\,(G_i - \mu_P)^2\right]^{1/2}} \qquad (8)$$

where U_{ij}, P_{ij} y $Q_{ij} = 1 - P_{ij}$ were already commented on in the previous statistics $G_i = \frac{1}{n}\sum_{i=1}^{N} P_{ij}$, $\mu_G = \frac{1}{N}\sum_{i=1}^{N} G_i$, $\mu_P = \frac{1}{N}\sum_{i=1}^{N} P_i$, *n* is the number of items and *N* is the number of persons.

Thus, ECI2z compares the pattern of item scores with the mean probability through the test items, whereas ECI4z compares the pattern of item scores with the expected probability according to the Rasch model. Low values, around 0, of these statistics represent a good fit of the data to the proposed model (Birenbaum, 1986).

Previous Studies

When a test is developing, it is important to decide what kind of items to use as a latent construct indicator. In psychometric tests, closed-answer items are generally used; items with several options from which examinees must select one. If the test is an achievement, skill, or ability test, there will only be one correct choice, which means that some examinees, usually low-ability ones, will try to guess the correct answer randomly, thus artificially altering the parameters of these items (Meijer, 1996). In the Rasch model, the probability of a minimum-ability examinee guessing an item correctly is 0 (Rasch, 1960; Wright & Stone, 1979), so the item fit statistics employed in this model must detect the items that have received a higher than expected percentage of random responses. Moreover, the fit statistics should maintain their distributional characteristics even when the conditions under which the test is administered are not optimal.

Research carried out till now, however, has revealed that the fit statistics show problems in their distributions when the conditions of response-pattern evaluation are not optimal, that is, when these response patterns do not clearly fulfill the assumptions of the model. Thus, Meijer and Sijtsma (2001) stated that it is doubtful whether the *t* transformation of the Outfit and Infit mean squares follows a normal distribution, although Rogers and Hattie (1987) found that they were sensitive to guessing. Smith (1991) found that the *t* transformations only followed a normal standardized distribution when they were calculated from true parameters, but when calculated from item or examinee estimated parameters, or from both, this produced severe restrictions in the means and standard deviations, which affected Type I error rate. However, Smith (1991) found that these transformations were sensitive to random guessing of items.

Regarding the statistics used in the field of measurement of aberrant patterns, Molenaar and Hoijtink (1990, 1996) found that the statistic Lz only followed a normal standardized distribution when calculated from true parameters, and its variance, calculated from estimated parameters, was smaller than the one expected under normal distribution (Molenaar & Hoijtink, 1990; Nering, 1995, 1997; Reise, 1995). Noonan, Boss, and Gessaroli (1992) also found that the distribution of Lz was negatively skewed.

Drasgow, Levine, and McLaughlin (1987) stated that ECI4z was better standardized (Li & Olejnik, 1997) and had a higher detection rate than ECI2z. On the other hand, Noonan et al. (1992) said that ECI4z had means and standard deviations close to the normal distribution, although the distributions were positively skewed, and this skewness was less than one half of that of the other statistics (ECI2z and T-Infit) and, moreover, was less affected by test length.

The object of this investigation is to study the power of three statistics normally used in the field of appropriateness measure: Lz of Drasgow and Levine (1986), and Eci2z and Eci4z (Tatsuoka, 1984) as item fit statistics under the Rasch model, and to compare them with the statistics (Outfit and Infit), and their corresponding *t* transformations, generally used to study item fit in the context of this model. The distributional properties of these five statistics in the context of random item guessing will also be studied.

Method

Experimental Conditions

An item with *k* response options has a *1/k* probability of being randomly guessed correctly. In IRT, this probability is expressed in the pseudo-guessing parameter, defined as the probability of a low-ability level examinee correctly guessing an item randomly (Hambleton & Swaminathan, 1985; Lord, 1980). To evaluate the extent to which the five statistics can detect this alteration with regard to the Rasch model, four sample sizes were selected: 100, 250, 500, and 1000 subjects. Each sample size was simulated under two types of distribution: uniform and normal, with mean 0 and standard deviation 1.In order to increase random guessing, we subtracted a unit of each value from the original samples, resulting in two new distributions (uniform and normal) in each sample size, with mean –1 and standard deviation 1.

Two test lengths were selected: 15 and 30 items, to observe whether any differential effect was produced as a function of text length. The discrimination parameters (*a*) of the items in all tests were 1.00 (expected in the Rasch model), whereas for the difficulty parameters (*b*), two uniform distributions were employed: ±1 logits and ±2 logits. Finally, the pseudo-guessing parameter for all items was fixed at 0.25. In summary, a total of 64 conditions— 4 (Sample Size) × 4 (Type of Distribution) × 2 (Test Length) × 2 (Distribution of b)—were examined. Each experimental condition was replicated 50 times.

Generation of Item-Response Data

The item-response data were generated under the 3-p model with all the discrimination parameters set at 1. The item responses were generated with a computer program that works as follows: Using the examinees' true ability parameters and the true discrimination, difficulty, and pseudo-guessing parameters of the items, it calculates the likelihood of responding correctly to the item (P_{ij}). The program subsequently generates a random number R in the range [0, 1], and it compares it with P_{ij}. If $P_{ij} > R$ the response is 0; if $P_{ij} < R$, then the response is 1.

Fit Evaluation

To examine the behavior of the fit statistics, the ability and item parameters were estimated in each replication using the incorrect model; that is, the Rasch model. Parameter estimation was performed with the ConQuest program (Wu, Adams, & Wilson, 1998). These parameters, together with the original response matrixes, were the basis of a new computer program to calculate the five fit statistics: T-outfit, T-infit, Lz, ECI2z, and ECI4z. Subsequently, with SYSTAT 10.0, the basic statistics were determined (means and standard deviations) and the power of the items that did not fit the model. To determine the power, the cutting-point score ±2 was employed, which corresponds roughly to the nominal rate $\alpha = .05$.

Results

Basic Statistics

In Tables 1 and 2 are displayed the means and standard deviations obtained for each of the fit statistics studied and in each of the manipulated conditions. As expected, the fit statistics revealed considerable differences in their standardization when calculated from the estimated parameters. Thus, in this study, Lz means were higher than 0, indicating that the response patterns obtained for the items were more consistent than those expected by Rasch's model and they increased systematically with sample size, group ability level, and amplitude of test-difficulty interval. Thus, in a 15-item test, the Lz mean changed from .075 (*N* = 100) to .199 (*N* = 1000) when the distribution of group ability was normal and equal to the test difficulty mean, and the test-difficulty interval was logits (see Table 1). If the difficulty interval was increased to logits, then the Lz mean was .218 for *N* = 100, increasing to .699 for *N* = 1000. If the fit statistics in a lower ability group —*N* (–1, 1) — were calculated, they would also be affected by the same conditions. Thus, if the test-difficulty interval was ±1 logits, the mean was .112 for *N* = 100, which increased to .304 for *N* = 1000; and if the difficulty interval was ±2 logits, then the mean changed from .232 for *N* = 100 to .715 for *N* =1000. This pattern occurred regardless of whether the ability distribution was normal or uniform.

If the test length was increased to 30 items, a similar pattern was observed, although ±1 in the logit difficulty interval, the mean of the Lz statistics was approximately 0 when the ability distribution was normal or uniform. Thus, the Lz mean was between .044 (*N* = 100) and –.002 (*N* = 500) when the ability distribution was normal, and between –.011 (*N* = 1000) and .008 (*N* = 500) when the distribution was uniform. However, in a lower ability group, the Lz mean was between .061 (*N* = 100) and .181 (*N* = 1000) when the ability distribution was normal, and it increased from .051 (*N* = 100) to .169 (*N* = 1000) when the ability distribution was uniform. When the difficulty interval was increased to ±2 logits, an increase in the Lz mean was observed in all experimental conditions, from a small sample size (*N* = 100) to a large one (*N* = 1000).

The statistics based on residuals (T-outfit and T-infit) presented a pattern similar to that obtained by the Lz statistic. In some conditions, the means were negative, indicating that the simulated response patterns have less variation than expected by the model, approaching the Guttman pattern. Thus, in short tests (*n* = 15) with normal ability distribution and difficulty interval of ±1 logits, the T-outfit mean was –.050 (*N* = 100), which decreased to –.200 (*N* = 1000), and the T-infit mean was –.068 (*N* = 100), which decreased to –.128 (*N* = 1000). Again, the same effect was observed when the ability distribution was uniform, that is, the T-outfit mean was –.063 (*N* = 100), decreasing to –.220 (*N* = 1000), and the T-infit mean was –.047 (*N* = 100), decreasing to –.165 (*N* = 1000). When the difficulty interval increased, the T-outfit and T-infit means in all experimental conditions decreased. The same occurred when the fit statistics were calculated in a lower ability group (see Table 1).

When test length was increased to 30 items, the distribution was centered, and the test difficulty interval was between ±1 logits, a similar effect was obtained to that observed in Lz. The T-outfit mean was between .001 (*N* = 100) and –.037 (*N* = 500) when test length was 30 items, whereas the T-infit mean was –.025 (*N* = 500), which changed to .023 (*N* = 1000). The test was subsequently compared to a lower ability group, resulting in a T-outfit mean between –.069 (*N* = 100) and –.150 (*N* = 1000), whereas the T-infit mean was between –.062 (*N* = 100) and –.199 (*N* = 1000). If the distribution was uniform, the T-outfit mean was between –.027 (*N* = 100) and –.164 (*N* = 1000), and the T-infit mean was between –.058 (*N* = 100) and –.203 (*N* = 1000). If the test difficulty interval was increased to logits, all T-outfit and T-infit means were considerably reduced.

However, the means of the statistics Eci2z and Eci4z were similar in the experimental conditions manipulated in this study. Thus, when using a short test (*n* = 15), difficulty interval of logits, and normal ability distribution, the Eci2z mean was between –.002 (*N* = 100) and –.007 (*N* = 1000)—see Table 1—and the Eci4z mean was between –.024 (*N* = 100) and –.046 (*N* = 1000). The Eci4z mean was an unexpected value only in a few conditions. Thus, with a 30-item test, difficulty interval of ±2 logits, and normal ability distribution (see Table 2), the Eci4z mean was –.122 (*N* = 500) and –.163 (*N* = 1000). In the same experimental conditions, but with uniform distribution, the Eci4z mean was –.141 (*N* = 1000).

Regarding the standard deviations, a more or less common pattern in all five fit statistics was observed. Thus, the standard deviations only maintained their expected value of 1 when the sample size was relatively small (100 and 250), but they increased considerably when the sample size was 500 or higher. That is, with a short test (*n* = 15), difficulty interval of ±1 logits, and normal and centered ability distribution (see Table 1), the Lz standard deviation was .978 (*N* = 100) and .988 (*N* = 250), but it increased to 1.442 (*N* = 1000). In the lower ability group, the standard deviation was between .959 (*N* = 100) and 1.573 (*N* = 1000).

The variability of the fit statistics increased considerably when the test difficulty interval was logits, especially when the sample size was 500 or more. Thus, with a short test (*n* = 15) and normal ability distribution, the standard deviation of Lz was 1.235 (*N* = 500) and 1.605 (*N* = 1000), slightly higher than those obtained in the same sample sizes at the difficulty interval of ±1 logits (see Table 1). If the ability distribution was uniform, then a small increase in the standard deviations of all statistics was observed in the various experimental conditions. This was systematically repeated in all fit statistics, observing high standard deviations in the Eci2z and Eci4z statistics when the sample size was 500 or more persons. A similar effect was observed in the standard deviations of the Lz, T-outfit, and T-infit statistics when increasing test length to 30 items, but very few relevant differences were observed in short tests (*n* = 15). That is, the standard deviations of these statistics changed as a function of the type of ability distribution (normal vs. uniform) of the

Table 1

Means and Standard Deviations (in brackets) of the Five Fit Statistics in the 15-Item Test with Normal and Uniform Distributions, Two Difficulty Intervals, and Four Sample Sizes

			Difficulty Interval							
			[–1, +1]				[–2, +2]			
			Sample size				Sample size			
Distribution		Statistic	100	250	500	1000	100	250	500	1000
Normal Distribution	(0, 1)	Lz	.075	.108	.141	.199	.218	.359	.481	.699
			(.978)	(.988)	(1.184)	(1.442)	(.860)	(.988)	(1.235)	(1.605)
		T-outfit	–.050	–.097	–.152	–.200	–.145	–.331	–.457	–.650
			(.958)	(1.035)	(1.191)	(1.435)	(.961)	(1.086)	(1.378)	(1.769)
		T-infit	–.068	–.099	–.128	–.190	–.208	–.333	–.474	–.708
			(.900)	(.966)	(1.165)	(1.412)	(.829)	(.945)	(1.165)	(1.519)
		Eci2z	–.002	–.003	–.005	–.007	–.004	–.016	–.019	–.019
			(1.034)	(1.057)	(1.241)	(1.458)	(1.039)	(1.208)	(1.486)	(1.959)
		Eci4z	–.024	–.022	–.031	–.046	–.052	–.066	–.095	–.130
			(1.036)	(1.055)	(1.235)	(1.444)	(1.045)	(1.197)	(1.442)	(1.872)
	(–1, 1)	Lz	.112	.161	.223	.304	.232	.354	.490	.715
			(.959)	(1.104)	(1.263)	(1.573)	(.936)	(1.058)	(1.284)	(1.638)
		T-outfit	–.087	–.151	–.195	–.283	–.190	–.323	–.413	–.638
			(.983)	(1.100)	(1.262)	(1.552)	(.995)	(1.135)	(1.415)	(1.819)
		T-infit	–.113	–.168	–.235	–.319	–.237	–.366	–.524	–.761
			(.958)	(1.101)	(1.253)	(1.560)	(.910)	(1.027)	(1.230)	(1.551)
		Eci2z	–.002	–.002	–.003	–.003	–.006	–.002	–.011	–.009
			(1.090)	(1.237)	(1.438)	(1.802)	(1.194)	(1.382)	(1.701)	(2.227)
		Eci4z	–.008	–.008	–.014	–.012	–.032	–.023	–.063	–.067
			(1.091)	(1.236)	(1.426)	(1.788)	(1.189)	(1.373)	(1.674)	(2.159)
Uniform Distribution	(0, 1)	Lz	.061	.098	.135	.183	.203	.329	.460	.642
			(1.014)	(1.187)	(1.316)	(1.728)	(.990)	(1.204)	(1.483)	(1.995)
		T-outfit	–.063	–.110	.148	–.220	–.186	–.340	–.494	–.693
			(1.008)	(1.164)	(1.304)	(1.677)	(1.084)	(1.249)	(1.561)	(2.113)
		T-infit	–.047	–.089	–.123	–.165	–.184	–.311	–.439	–.625
			(1.001)	(1.169)	(1.295)	(1.703)	(.944)	(1.163)	(1.417)	(1.893)
		Eci2z	.002	.001	–.001	–.002	–.003	–.004	–.011	–.018
			.(1.105)	(1.232)	(1.340)	(1.723)	(1.153)	(1.406)	(1.718)	(2.206)
		Eci4z	–.012	–.019	–.025	–.030	–.044	–.042	–.061	–.098
			(1.111)	(1.230)	(1.327)	(1.704)	(1.151)	(1.392)	(1.663)	(2.202)
	(–1, 1)	Lz	.098	.139	.200	.280	.214	.335	.456	.654
			(1.030)	(1.235)	(1.516)	(1.950)	(.989)	(1.279)	(1.613)	(2.168)
		T-outfit	–.067	–.141	–.190	–.281	–.172	–.309	–.442	–.614
			(1.029)	(1.200)	(1.472)	(1.864)	(1.097)	(1.369)	(1.715)	(2.314)
		T-infit	–.111	–.146	–.218	–.307	–.221	–.358	–.489	–.720
			(1.028)	(1.241)	(1.515)	(1.948)	(.957)	(1.232)	(1.541)	(2.062)
		Eci2z	.000	–.003	–.005	–.005	.000	–.004	–.005	–.011
			(1.133)	(1.357)	(1.677)	(2.178)	(1.226)	(1.616)	(2.080)	(2.816)
		Eci4z	–.017	–.006	–.014	–.016	–.023	–.034	–.045	–.074
			(1.132)	(1.352)	(1.663)	(2.150)	(1.216)	(1.579)	(2.008)	(2.720)

Table 2
Means and Standard Deviations (in brackets) of the Five Fit Statistics in the 30-Item Test with Normal and Uniform Distributions, Two Difficulty Intervals, and Four Sample Sizes

			Difficulty Interval							
			[–1, +1]				[–2, +2]			
			Sample size				Sample size			
Distribution		Statistic	100	250	500	1000	100	250	500	1000
Normal Distribution	(0, 1)	Lz	044	.000	–.002	–.001	.090	.126	.182	.249
			(.948)	(1.021)	(1.204)	(1.434)	(.858)	(1.052)	(1.278)	(1.654)
		T-outfit	.001	–.025	–.037	–.035	–.076	–.155	–.216	–.306
			(.968)	(1.040)	(1.211)	(1.431)	(.933)	(1.148)	(1.432)	(1.860)
		T-infit	–.011	.021	–.025	.023	–.056	–.094	–.154	–.216
			(.919)	(.993)	(1.169)	(1.393)	(.816)	(.992)	(1.190)	(1.531)
		Eci2z	–.005	–.011	–.017	–.026	–.021	–.041	–.063	–.082
			(1.095)	(1.169)	(1.364)	(1.612)	(1.134)	(1.384)	(1.699)	(2.216)
		Eci4z	–.019	–.024	–.036	–.062	–.050	–.077	–.122	–.163
			(1.094)	(1.168)	(1.354)	(1.593)	(1.135)	(1.364)	(1.643)	(2.133)
	(–1, 1)	Lz	.061	.090	.132	.181	.104	.175	.249	.343
			(1.968)	(1.094)	(1.306)	(1.585)	(.879)	(1.064)	(1.322)	(1.706)
		T-outfit	–.039	–.064	–.116	–.150	–.085	–.155	–.249	–.341
			(.975)	(1.098)	(1.309)	(1.574)	(.975)	(1.196)	(1.485)	(1.945)
		T-infit	–.062	–.100	–.143	–.199	–.096	–.181	–.261	–.365
			(.963)	(1.085)	(1.293)	(1.567)	(.837)	(1.000)	(1.240)	(1.582))
		Eci2z	–.005	–.007	–.013	–.013	–.011	–.014	–.029	–.038
			(1.131)	(1.297)	(1.555)	(1.927)	(1.240)	(1.565)	(1.981)	(2.635)
		Eci4z	–.012	–.019	–.021	–.026	–.028	–.039	–.059	–.081
			(1.133)	(1.292)	(1.546)	(1.910)	(1.223)	(1.534)	(1.937)	(2.563)
Uniform Distribution	(0, 1)	Lz	.001	.012	.008	–.011	.063	.107	.162	.209
			(.990)	(1.100)	(1.317)	(1.719)	(.970)	(1.273)	(1.612)	(2.143)
		T-outfit	–.004	–.037	–.041	–.066	–.076	–.185	–.262	–.365
			(.979)	(1.095)	(1.287)	(1.653)	(1.032)	(1.357)	(1.710)	(2.290)
		T-infit	.016	.009	.013	–.015	–.027	–.066	–.124	–.154
			(.962)	(1.078)	(1.290)	(1.683)	(.916)	(1.201)	(1.527)	(2.008)
		Eci2z	–.003	–.007	–.012	–.013	–.024	–.040	–.053	–.083
			(1.115)	(1.201)	(1.437)	(1.850)	(1.211)	(1.556)	(1.980)	(2.653)
		Eci4z	–.018	–.018	–.034	–.040	–.052	–.066	–.090	–.141
			(1.117)	(1.200)	(1.425)	(1.827)	(1.199)	(1.521)	(1.917)	(2.545)
	(–1, 1)	Lz	.051	.078	.116	.169	.088	.156	.222	.325
			(1.046)	(1.288)	(1.629)	(2.031)	(.977)	(1.317)	(1.682)	(2.283)
		T-outfit	–.037	–.081	–.118	–.164	–.069	–.185	–.267	–.369
			(1.048)	(1.286)	(1.594)	(1.988)	(1.063)	(1.442)	(1.682)	(2.498)
		T-infit	–.058	–.090	–.140	–.203	–.089	–.163	–.236	–.366
			(1.039)	(1.277)	(1.626)	(2.022)	(.931)	(1.243)	(1.585)	(2.135)
		Eci2z	–.003	–.010	–.012	–.015	–.007	–.018	–.026	–.032
			(1.192)	(1.447)	(1.834)	(2.334)	(1.235)	(1.823)	(2.372)	(3.238)
		Eci4z	–.013	–.016	–.023	–.031	–.030	–.034	–.046	–.074
			(1.190)	(1.436)	(1.840)	(2.313)	(1.322)	(1.782)	(2.312)	(3.142)

mean group ability, and the sample size, but no appreciable changes were observed due to increase in test length. Thus, if n = 15 items, ability distribution is normal, and the difficulty interval is ±1 logits (see Table 1), the standard deviation of T-outfit was 1.435 (N = 1000), and in the same conditions but with a 30-item test (see Table 2), the standard deviation of T-outfit was 1.431 (N = 1000). Only the standard deviations of Eci2z and Eci4z increased slightly because of the increase in test length. Thus, with a 15-item test, normal ability distribution, and difficulty interval of logits (see Table 1), the standard deviation of Eci2z was 1.458 (N = 1000), but if the test length was increased to 30 items, the standard deviation of Eci2z increased to 1.612 (N = 1000).

Power of Fit Statistics

In Tables 3 and 4 is showed the power of each of the fit statistics examined, in each of the manipulated conditions. As all the items were simulated under the modified 3-p model with the probability of randomly correct guessing at c = .25, it was expected that none the items would be detected as fitting the model, assuming that, for the Rasch model, the probability of random correct guessing is 0. However, Table 3 (n = 15) and Table 4 (n = 30) provide very different results from those expected. Generally, the power was low or very low, especially in the Lz, T-outfit, and T-infit statistics, and somewhat higher in Eci2z and Eci4z. When the ability distribution was normal, regardless of whether the test length was short (n = 15) or longer (n = 30), the power of the fit statistics was very low when sample size was 500 or more. For example, when the test difficulty interval was ±1 logits, at the sample size of 500 and normal ability distribution, the power of the five fit statistics was between 9% (Lz and T-infit) and 11% (Eci2z and Eci4z). If the fit statistics were calculated in the lower ability group, the power was between 12% (T-infit) an 18% (Eci2z and Eci4z).

As expected, the power increased with sample size in all experimental conditions, but in sample size N = 1000, using a lower ability group than the mean test difficulty, uniform distribution, and item difficulty interval of ±2 logits, the power of the fit statistics was close to 70%. Thus, when n = 30, test difficulty interval was ±1 logits, the power of Eci2z and Eci4z was 50% and 49%, respectively, which increased to 71% and 70% when the test difficulty interval

Table 3
Power of the Five Fit Statistics in the 15-Item Test with Normal and Uniform Distributions, Two Difficulty Intervals, and Four Sample Sizes

			Difficulty Interval							
			[–1, +1]				[–2, +2]			
			Sample size				Sample size			
Distribution		Statistic	100	250	500	1000	100	250	500	1000
Normal Distribution	(0, 1)	Lz	.03	.05	.09	.14	.03	.04	.10	.30
		T-outfit	.04	.05	.10	.15	.04	.05	.18	.38
		T-infit	.03	.04	.09	.14	.03	.04	.09	.23
		Eci2z	.06	.06	.11	.16	.06	.10	.20	.36
		Eci4z	.06	.07	.11	.16	.06	.09	.17	.34
	(–1, 1)	Lz	.04	.08	.13	.26	.03	.06	.15	.35
		T-outfit	.05	.08	.13	.24	.04	.07	.21	.42
		T-infit	.05	.07	.12	.25	.03	.06	.14	.32
		Eci2z	.08	.13	.18	.32	.10	.15	.27	.44
		Eci4z	.08	.12	.18	.31	.09	.15	.25	.44
Uniform Distribution	(0, 1)	Lz	.06	.09	.11	.30	.04	.10	.18	.53
		T-outfit	.05	.09	.13	.29	.06	.12	.24	.57
		T-infit	.06	.09	.11	.29	.04	.09	.15	.43
		Eci2z	.08	.12	.13	.29	.08	.16	.26	.46
		Eci4z	.08	.11	.13	.27	.07	.16	.25	.44
	(–1, 1)	Lz	.05	.12	.22	.40	.04	.11	.24	.55
		T-outfit	.05	.09	.20	.37	.05	.16	.34	.56
		T-infit	.05	.13	.22	.39	.04	.11	.27	.52
		Eci2z	.07	.16	.27	.45	.11	.24	.43	.63
		Eci4z	.08	.16	.27	.44	.11	.24	.40	.61

Table 4
Power of the Five Fit Statistics in the 30-Item Test with Normal and Uniform Distributions, Two Difficulty Intervals, and Four Sample Sizes

			Difficulty Interval							
			[–1, +1]				[–2, +2]			
			Sample size				Sample size			
Distribution		Statistic	100	250	500	1000	100	250	500	1000
Normal Distribution	(0, 1)	Lz	.04	.05	.10	.15	.03	.06	.11	.23
		T-outfit	.05	.05	.09	.16	.04	.08	.15	.35
		T-infit	.04	.05	.09	.14	.02	.05	.09	.19
		Eci2z	.07	.10	.16	.23	.08	.15	.26	.48
		Eci4z	.07	.09	.15	.23	.08	.15	.25	.46
	(–1, 1)	Lz	.04	.07	.13	.27	.03	.05	.10	.33
		T-outfit	.04	.07	.13	.24	.04	.08	.19	.45
		T-infit	.05	.07	.12	.25	.02	.04	.08	.23
		Eci2z	.08	.13	.21	.38	.11	.22	.37	.60
		Eci4z	.08	.13	.21	.37	.10	.21	.35	.58
Uniform Distribution	(0, 1)	Lz	.05	.07	.13	.25	.05	.11	.24	.42
		T-outfit	.05	.07	.12	.25	.05	.13	.28	.56
		T-infit	.05	.07	.12	.25	.05	.10	.22	.37
		Eci2z	.07	.10	.18	.32	.11	.21	.37	.60
		Eci4z	.08	.10	.17	.32	.10	.21	.36	.58
	(–1, 1)	Lz	.06	.12	.26	.44	.04	.12	.29	.61
		T-outfit	.06	.12	.25	.41	.05	.17	.37	.65
		T-infit	.06	.12	.26	.43	.03	.09	.24	.56
		Eci2z	.10	.17	.34	.50	.14	.32	.52	.71
		Eci4z	.10	.17	.33	.49	.13	.30	.49	.70

was increased to ±2 logits. The increase was also significant for the Lz, T-outfit, and T-infit statistics, which, at the difficulty interval of ±1 logits obtained power values of 44%, 41%, and 43%, respectively; and they increased to 61%, 65%, and 56%, when the difficulty interval was logits.

Conclusions

In this simulation study, we examined whether three fit statistics that are habitually used in the area of detection of aberrant response patterns (appropriateness measure) can also be used to detect items that do not fulfill the assumptions of the Rasch model.

In view of the results obtained, it seems that the usefulness of the item fit statistics (T-outfit and T-infit) and of the statistics of appropriateness measure (Lz, Eci2z, and Eci4z) is very limited because when using estimated parameters, low or very low detection rates, usually with sample sizes of less than 500 examinees, were detected.

In any case, all the computer programs of parameter estimation with IRT models include one or more item fit statistics, so that psychologists should decide whether these fit statistics are useful when making decisions to select items. If they decide that they are useful, the following information should be taken into account. First, the Lz, T-outfit, and T-infit statistics do not tend toward the expected values when they are calculated using estimated parameters. These results are in accordance with those reported by Smith (1991) on the evaluation of item fit and by Moleenar and Hoijtink (1990, 1996) on the evaluation of person fit. As occurs in person-fit evaluation (Li & Olejnik, 1997; Noonan et al., 1992), the distributions of these statistics depend on sample size, test-difficulty interval amplitude, group ability level, and test length. Nevertheless, if the item difficulty interval is relatively narrow, as the test length increases, the properties of the distributions of the Lz, T-outfit, and T-infit statistics improve considerably.

Second, despite the fact that the Lz, T-outfit, and T-infit statistics are based on different concepts to evaluate the fit of the items to the model, no important differences were observed in the distributional properties of these statistics. In fact, their behavior was observed to be similar independently of the factors manipulated in this study.

Third, the behavior of the statistics Eci2z and Eci4z (Tatsuoka, 1984) is satisfactory in all experimental conditions, except for some isolated case of Eci4z. Therefore, it seems that their distribution (the mean) is relatively stable, independently of test length, sample size, type of ability distribution (normal vs. uniform), and group ability level. The same cannot be said about their variability, which showed an increase as a function of sample size, test length, difficulty interval amplitude, and type of ability distribution. These results partially contrast with those found when these statistics are applied to evaluate person fit. Thus, Noonan et al. (1992) found that the ECI4z statistic fit a normal distribution better—both the mean and the standard deviation—than the ECI2z statistic and the Lz, T-outfit, and T-infit statistics, showing a less skewed distribution and showing less influence of sample size. However, Noonan et al. (1992) study used true parameters, not estimated ones.

Fourth, a small sample size (*N* = 250 or less) may be sufficient (Lord, 1983) to obtain estimations that are consistent with item difficulty parameters, but it will not help to decide whether or not the items fit the Rasch model, because the power of the five fit statistics was relatively low or very low.

Fifth, the power of the fit statistics drops drastically when the item is a multiple-choice item and can be guessed correctly at random. In this case, not even a large sample size (*N* = 1000) ensures sufficient power of any of the five statistics to guarantee that an item does not follow the assumption of the Rasch model, where the probability of correct guessing at random is 0. In any case, if *N* = 250 or higher, the ECI2z and ECI4z statistics present higher power to detect these items than the statistics based on the likelihood function (Lz) or on T-outfit and T-infit residuals.

Finally, in practically all the experimental conditions manipulated, the power of the Eci2z and Eci4z statistics was between 5% and 10% higher than that Lz, T-outfit, and T-infit, although these differences were even greater when the ability distribution was normal, their mean was the same as the test mean, and the sample size was less than 1000 cases.

In view of these results, perhaps some of the well-known computer programs used to examine the fit of the Rasch model should include the Eci2z and Eci4z statistics as item and person fit statistics.

References

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10,* 167-174.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10,* 59-67.

Drasgow, F., Levine, M. V., & Mclaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Fisher, G. H., & Molenaar, I. W. (Eds.) (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Gulliksen, H. (1950). *Theory of mental test*. New York: Wiley.

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

Harnish, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 104-122). Vancouver, Canada: Educational Research Institute of British Columbia.

Hulin, Ch. L., Drasgow, F., & Parsons, Ch. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow-Jones Irwin.

Levine, M. V., & Rubin, D. F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269-290.

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*, 215-231.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51-61). New York: Academic Press.

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

Meijer, R. R. (1996). The influence of the presence of deviant item score patterns on the power of a person-fit statistic. *Applied Psychological Measurement*, *20*, 141-154.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person-fit indices. *Psychometrika*, *55*, 75-106.

Molenaar, I. W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, *9*, 27-45.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, *19*, 121-129.

Nering, M. L. (1997). The distribution of indexes of person-fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, *21*, 115-127.

Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement*, 16, 345-352.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University Chicago Press.)

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*, 127-137.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*, 213-229.

Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement*, *11*, 47-57.

Smith, R. M. (1991). The distributional properties of Rasch item-fit statistics. *Educational and Psychological Measurement*, *51*, 541-565.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

SYSTAT (v. 10.0) (2000). *The system for statistics*. SPSS Inc.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*, 95-110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, *7*, 81-96.

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *Acer ConQuest: Generalised item response modelling software*. Melbourne, Australia: Australian Council for Educational Research.

Received: November 18, 2004
Revision received: January 1, 2005
Accepted: January 26, 2005