IRT Goodness-of-Fit Using Approaches from Logistic Regression

Patrick Mair
Wirtschaftuniversität Wien

Steven P. Reise
University of California, Los Angeles

Peter M. Bentler
University of California, Los Angeles

Abstract

We present an IRT goodness-of-fit framework based on approaches from logistic regression. We briefly elaborate the formal relation of IRT models and logistic regression modeling. Subsequently, we examine which model tests and goodness-of-fit indices from logistic regression can be meaningfully used for IRT. The performance of the casewise deviance, a collapsed deviance, and the Hosmer-Lemeshow test is studied by means of a simulation that compares their power to well known IRT model tests. Next, various R^2 measures are discussed in terms of interpretability and appropriateness within an IRT context. By treating IRT models as classifiers, several additional indices such as hit rate, sensitivity, specificity, and area under the ROC curve are defined. Data stemming from a social discomfort scale are used to demonstrate the application of these statistics.

Keywords: Item response theory, logistic regression, model fit.

1. Introduction

It is frequently argued that the benefits of IRT models, including Rasch models, are only realized to the degree that the data meet the appropriate assumptions, and the degree of model to data fit. Nevertheless, the evaluation of fit in IRT modeling has been challenging indeed (see Embretson and Reise 2000). For example although there are a variety of approaches to computing a chi-square test of item or model fit, Orlando and Thissen (2000, 2003) demonstrated that all chi-square approaches are problematic. Beyond the problem of these indices being too powerful, they argue that it is often not clear what the appropriate degree of freedom should be. Even if a statistically correct chi-square index were to be derived, such an index may not yield substantively useful information regarding a lack of fit.

We propose logistic regression as an alternative framework to study IRT goodness-of-fit. We argue for their relative merits compared to traditional approaches. So far logistic regression procedures in IRT were mainly used for detecting differential item functioning (DIF). Further elaborations can be found in Swaminathan and Rogers (1990), Miller and Spray (1993), and Zumbo (1999).

2. Logistic Regression and IRT Modeling

Typically, in dichotomous IRT the data are organized in a binary person × item matrix $\mathbf{Y}$ of dimension $N \times k$ with $v = 1, \ldots, N$ and $i = 1, \ldots, k$. To connect IRT and logistic regression, we represent the matrix $\mathbf{Y}$ as a vector $\mathbf{y}$ of length $l = 1, \ldots, L$ where $L = N \times k$. The vector is established by slicing $\mathbf{Y}$ row-wise which implies that the first $1, \ldots, k$ elements in $\mathbf{y}$ correspond to the k item responses of person 1, the next $k + 1, ..., 2k$ elements to the k item responses of person 2, etc.

In a world without Rasch (or IRT in general) the most intuitive way of analyzing such a $(0, 1)$-response vector would be to predict $\mathbf{y}$, respectively the log-odds of success, from observed covariate patterns. Let us write the log-odds expression as

$$\eta_l = \log\left(\frac{p_l}{1-p_l}\right) \tag{1}$$

for the vector representation $\mathbf{y}$, and as

$$\eta_{vi} = \log\left(\frac{p_{vi}}{1-p_{vi}}\right) \tag{2}$$

for the matrix representation $\mathbf{Y}$.

Note that $\mathbf{y}$ reflects a hierarchical design: item responses are nested within persons. Thus, we could think of covariates on the items and covariates on the persons. Without having covariates we can try to predict the item responses in terms of direct characteristics of persons ("ability") and direct features of items ("easiness"). This leads to following linear model:

$$\begin{aligned}\eta_l &= \beta_0 + \beta_1^{(P)}x_1^{(P)} + \beta_2^{(P)}x_2^{(P)} + \ldots + \beta_v^{(P)}x_v^{(P)} + \ldots + \beta_N^{(P)}x_N^{(P)} + \\ &\quad \beta_1^{(I)}x_1^{(I)} + \beta_2^{(I)}x_2^{(I)} + \ldots + \beta_i^{(I)}x_i^{(I)} + \ldots + \beta_N^{(I)}x_N^{(I)}.\end{aligned} \tag{3}$$

The superscript P refers to the person part in this model, the superscript I to the item part. All the x-variables are binary indicators which reflect the hierarchical structure in $\mathbf{y}$: For $l = 1$ only $x_1^{(P)} = 1$ and $x_1^{(I)} = 1$, the other indicators are 0; for $l = 2$ only $x_1^{(P)} = 1$ and $x_2^{(I)} = 1$, whereas all other indicators are 0, and so on. To express (3) in terms of matrices we need I_k and I_N as identity matrices of order k and N, as well as $\mathbf{1}_N$ and $\mathbf{1}_k$ containing N and k ones, respectively. Let us denote the design matrix for the person part by $\mathbf{X}^{(P)} = I_N \otimes \mathbf{1}_k$ and the design matrix for the item part by $\mathbf{X}^{(I)} = \mathbf{1}_N \otimes I_k$. Due the identifiability issues the first column in $\mathbf{X}^{(I)}$ and the first column in $\mathbf{X}^{(P)}$ are deleted. We will proceed with these reduced matrices but we keep the $\mathbf{X}^{(I)}$ and $\mathbf{X}^{(P)}$ notation. As a consequence, Equation 3 becomes

$$\boldsymbol{\eta} = \beta_0 + \mathbf{X}^{(P)}\boldsymbol{\beta}^{(P)} + \mathbf{X}^{(I)}\boldsymbol{\beta}^{(I)}. \tag{4}$$

$\mathbf{X}^{(P)}$ is of dimension $L \times (N-1)$ and $\mathbf{X}^{(I)}$ of dimension $L \times (k-1)$. It follows that the person parameter vector $\boldsymbol{\beta}^{(P)}$ is of length $N-1$ and the item parameter vector $\boldsymbol{\beta}^{(\boldsymbol{I})}$ of length $k-1$.

We see from Equation (4) that an increase of, for example $\beta_1^{(P)}$ leads to an increase in the logits of the first k elements of $\boldsymbol{\eta}$ (i.e. the logits of person 1). Thus, we can denote the parameter vector $\boldsymbol{\beta}^{(P)}$ as "person abilities". On item-side it holds that an increase of a particular $\beta_i^{(I)}$ leads to an increase in the i-th, the $2i$-th, etc. up to the Ni-th logits in $\boldsymbol{\eta}$. We therefore refer to $\boldsymbol{\beta}^{(I)}$ as "item easiness".

If we are not interested in person specific estimates but we want to account for a within person measurement error, we can put the intercept β_0 and the person part $\mathbf{X}^{(P)}\boldsymbol{\beta}^{(\boldsymbol{P})}$ into an error component $\boldsymbol{\epsilon}^{(\boldsymbol{P})}$. Consequently, Equation (4) becomes

$$\boldsymbol{\eta} = \mathbf{X}^{(I)}\boldsymbol{\beta}^{(I)} + \boldsymbol{\epsilon}^{(P)}. \tag{5}$$

Looking closer at the decomposition of $\boldsymbol{\epsilon}^{(P)}$ we see that there is the intercept and some (person-related) random component. Thus, in classical multilevel notation, this component is called *random intercept*. Within $\boldsymbol{\epsilon}^{(P)}$ each person v has such a random intercept component which we denote as θ_v. If we want to express the logit a correct response of person v on a particular item i the indicator matrices are not needed, since they only determine the position in $\mathbf{y}$. Therefore, for this person-item combination we have to take into account θ_v and $\beta_i^{(I)}$ only. If we set $\beta_i^{(I)} = -\sigma_i$ the equations above reduce to

$$\eta_{vi} = \theta_v - \sigma_i \tag{6}$$

which corresponds to the well-known Rasch model (Rasch 1960) with σ_i as the item difficulties.

If the parameters would be estimated jointly as represented in Equation (4) it is well known that they are not consistent: As sample size increases, the number of person parameters increases as well. If we use the random intercept representation from Equation (5), this fits into the framework of *generalized linear mixed* *models* (GLMM; De Boeck and Wilson 2004). The person parameters are regarded as random effects and the item parameters as fixed effects. Within the GLMM framework the parameters can be estimated by posing a distribution assumption on the random effects (typically normal). Within the classical IRT framework the most common ways to estimate item parameters in IRT are either *marginal maximum likelihood* (MML; distributional assumption on $\boldsymbol{\theta}$ as in GLMM) or *conditional maximum likelihood* (CML; using the sufficiency property of the margins, only for models of the Rasch family). Person parameters are estimated afterwards by using either ordinary ML or Bayesian approaches.

For the likelihood based fit approaches (mostly model tests) we present in this paper, we provide advice on which type of likelihood is appropriate. For non-likelihood based fit approaches (mostly fit indices) we can use any reliable method of parameter estimation as long as we get an $N \times k$ matrix of expected probabilities denoted by $\widehat{\boldsymbol{\Pi}}$ or the equivalent vector representation $\hat{\boldsymbol{\pi}}$ of length L.

3. Goodness-of-Fit Tests

3.1. Likelihood Ratio Tests

For IRT models that belong to the Rasch family, CML estimation can be applied. For this type of IRT model sample invariance of the parameter estimates holds. Based on this property, several test statistics have been proposed. A taxonomy and detailed elaborations can be found in Glas and Verhelst (1995). The classical approach of CML based model testing was proposed by Andersen (1973). To implement this method a researcher needs to split the sample into $g = 1, \ldots, G$ subgroups in order to compute the corresponding likelihoods $L_C^{(g)}$. The resulting likelihood-ratio test is given by

$$LR = 2\left(\sum_{g=1}^{G} \log L_C^{(g)} - \log L_C\right). \tag{7}$$

This test statistic is asymptotically χ^2-distributed with *df* equal to the number of parameters estimated in the subgroups minus the number of parameters in the total data set. Suárez-Falcón and Glas (2003) have studied the performance of this test statistic in detail by means of simulation. Their results show that LR has high power to detect specific violations. Moreover, the test maintains an acceptable Type I error rate and data meet the assumptions of the Rasch model. Therefore we will use this well-established test as a benchmark for our logistic regression statistics developed below.

A drawback of Andersen's test is that it only works for CML estimated Rasch-type models. Furthermore, the decisions based on the test results can be ambiguous: Different raw score splittings can lead to different decisions. Another practical problem is that for small sample sizes, some item parameters may not be estimable within subsamples due to 0/full item raw scores (within these subsamples). In the case of polytomous models some response categories may not occur within subsamples. Such items can not be taken into account when computing the test statistic.

Outside the Rasch family the question of model fit is mostly judged by the evaluation of aggregated item fit measures in order to obtain a χ^2-distributed model test (see Embretson and Reise 2000, Chapter 9). Logistic regression approaches described in the following sections also allow for the evaluation of fit outside the Rasch family. A crucial problem in our regression-IRT setting is that we have categorical predictors only. For such specifications there is no distinct way of defining a saturated model. Simonoff (1998) elaborates a taxonomy in terms of casewise, collapsing, and contingency table approaches for defining a saturated model. We will use this terminology in the elaborations below.

An additional challenge in the IRT formulation is that the observations can not be grouped according to common predictors since according to Equation (4) they define unique contrasts for items and persons.

The first goodness-of-fit approach is based on the observed 0-1 responses y_{vi} and the estimated model (solving) probabilities $\hat{\pi}_{vi}$. The resulting deviance is

$$D^{(0)} = -2\sum_{v=1}^{N}\sum_{i=1}^{k}\left[y_{vi}\log\left(\frac{\hat{\pi}_{vi}}{y_{vi}}\right) + (1-y_{vi})\log\left(\frac{1-\hat{\pi}_{vi}}{1-y_{vi}}\right)\right]. \tag{8}$$

Due to possible 0 values in the denominators, $D^{(0)}$ has to computed using the deviance residuals

$$d_l = -\sqrt{2|\log(1-\hat{\pi}_l)|} \qquad \forall y_l = 0 \tag{9a}$$
$$d_l = \sqrt{2|\log(\hat{\pi}_l)|} \qquad \forall y_l = 1. \tag{9b}$$

and by taking the sum $D^{(0)} = \sum_{l=1}^{L} d_l^2$.

Hosmer and Lemeshow (2000, p.146-147) point out that if the observations are not grouped (e.g. due to common predictor values) the χ^2-approximation is not feasible since the number of parameters increases at the same rate as the sample size. As mentioned above, we have no common predictor values. Therefore we have to find a grouping mechanism for y_l into $g = 1, \ldots, G$ groups with $G \ll L$. A somewhat natural approach within a Rasch framework (and one we will use in the simulation study) is to group the subjects according to their raw scores r_v. Other grouping strategies such as pattern-wise grouping can be considered as well. Simonoff (1998) calls such grouping mechanisms "collapsing".

Let us denote the number of positive responses in raw score group g by y_g. The deviance residual expressions from (9) become

$$d_g = -\sqrt{2n_g|\log(1-\hat{\pi}_g)|} \qquad \forall y_g = 0, \tag{10a}$$
$$d_g = \sqrt{2n_g|\log(\hat{\pi}_g)|} \qquad \forall y_g = n_g, \tag{10b}$$
$$d_g = \operatorname{sgn}(y_g - n_g\hat{\pi}_g)\sqrt{2\left[y_g \log\frac{y_g}{n_g\hat{\pi}_g} + (n_g - y_g)\log\frac{n_g - y_g}{n_g(1-\hat{\pi}_g)}\right]} \qquad \forall 0 < y_g < n_g. \tag{10c}$$

Again, by taking the sum we get the *collapsed deviance*

$$D_g^{(0)} = \sum_{g=1}^{G} d_g^2 \tag{11}$$

which is asymptotically χ^2 with $df = G$. We see that if the sample size increases, the number of deviance components does not increase: The additional observations affect n_g only. As a consequence, the asymptotic χ^2 approximation holds.

The next deviance strategy is based on what Simonoff (1998) refers to as *casewise approach.* Having dichotomous outcomes, that is realizations of L independent Bernoulli trials, the maximized log-likelihood is

$$L = \prod_{l=1}^{L} \hat{\pi}_l^{y_l}(1-\hat{\pi}_l)^{(1-y_l)}. \tag{12}$$

Under the saturated model, $\pi_l := y_l$ and thus Equation (12) becomes

$$L_S^{(1)} = \prod_{l=1}^{L} y_l^{y_l}(1-y_l)^{(1-y_l)} = 1. \tag{13}$$

By means of the full or joint likelihood L_F we can establish the *casewise deviance*

$$D^{(1)} = -2\log L_S. \tag{14}$$

Caution is recommended when using this approach since the χ^2-approximation ($df = L$) is degenerate for the same reason as $D^{(0)}$ (Simonoff 1998; McCullagh and Nelder 1989). Nevertheless we will include the casewise deviance due to the fact that within a logistic regression context this statistic is printed out by most statistical software packages.

The formulation of a saturated model is an essential issue in deviance approaches. Thus we now discuss some further strategies of specifying saturated IRT models. Rost (2004) presents a version of a saturated model which explains the observed data perfectly and has as many parameters as there are independent data points. This model can be used as a general reference model to test various nested IRT submodels against each other. Let Ω_y be the set composed of unique person response patters $\tilde{\mathbf{y}}$. It follows that the saturated likelihood can be expressed as

$$\log L_S^{(2)} = \sum_{\tilde{\mathbf{y}} \in \Omega_y} n_{\tilde{\mathbf{y}}} \log p_{\tilde{\mathbf{y}}} \tag{15}$$

where $p_{\tilde{\mathbf{y}}}$ is the relative frequency of the corresponding unique pattern and $n_{\tilde{\mathbf{y}}}$ the number of persons having $\tilde{\mathbf{y}}$. Let L_M denotes the IRT likelihood estimated by MML. Note that if the parameter estimation is carried out by CML with resulting likelihood L_C, L_M can be computed by means of

$$L_M = \prod_{\tilde{r} \in \Omega_r} p_{\tilde{r}}^{n_{\tilde{r}}} L_C \tag{16}$$

where Ω_r denotes the set of unique person raw scores $\tilde{r}$, $p_{\tilde{r}}$ their relative frequency, and $n_{\tilde{r}}$ their absolute frequency. The corresponding deviance can be expressed by

$$D^{(2)} = -2(\log L_M - \log L_S^{(2)}) \tag{17}$$

and is χ^2 with $df = df_{S^{(2)}} - df_M$ where $df_{S^{(2)}} = 2^k - 1$ and $df_M = k - 1$. Note that this test statistic is included in the software WINMIRA (von Davier 1997)

Another approach to establish a saturated model is to formulate IRT models within a log-linear model framework (Kelderman 1984; Hatzinger 2008). For binary items the corresponding table is of dimension 2^k and by fitting a log-linear model including the k-fold and all lower interactions, the model is saturated and reproduces the observed table perfectly. This provides a flexible framework for parameter and model testing. The drawback of such a log-linear approach is that it is computationally fairly exhaustive (already for $k > 20$) and the 2^k contingency table is usually very sparse such that χ^2 approximations do not hold. For recent developments in this area in terms of model testing see Maydeu-Olivares and Joe (2005), Joe and Maydeu-Olivares (2006), and Cai, Maydeu-Olivares, Coffman, and Thissen (2006).

3.2. Hosmer-Lemeshow Test

Hosmer and Lemeshow (1980) proposed a Pearson χ^2-statistic for logistic regression based on a grouping of the estimated probabilities (see also Hosmer and Lemeshow 2000). Let G denote the number of groups for splitting the corresponding vector of estimated probabilities $\hat{\boldsymbol{\pi}}_l$. One possibility (and the one that we consider) is to split the probabilities by means of the $g = 1, \ldots, G$ percentiles. The corresponding group-wise mean of probabilities is $\overline{\pi}_g$ and o_g are the number of positive responses within group g. The Hosmer-Lemeshow statistic can be expressed as

$$C_g = \sum_{g=1}^{G} \frac{o_g - n_g \overline{\pi}_g}{n_g \overline{\pi}_g (1 - \overline{\pi}_g)} \tag{18}$$

which is asymptotically χ^2 distributed with $df = G-2$. We will use this statistic as an IRT goodness-of-fit test. As a rule of thumb Hosmer and Lemeshow (2000, p. 149) propose to use $G = 10$ percentile groups. We will use this grouping strategy in the following simulation. However note that different percentiles or more natural grouping mechanisms (e.g. person raw score grouping) could also be used. Of course, the grouping strategy affects the performance of the test.

3.3. Simulation Study for Goodness-of-Fit Tests

For the three model tests from logistic regression elaborated above, that is, the collapsed deviance $D_g^{(0)}$ (raw score collapsing), the casewise deviance $D^{(1)}$ and the Hosmer-Lemeshow test C_g ($G = 10$ decile splitting), we study their performance with respect to different IRT scenarios. In addition, as mentioned above, we include also Andersen's LR-test as a benchmark test. Furthermore, Rost's deviance $D^{(2)}$ is included as well because it is commonly used. In what follows, data simulation is carried out by means of the simulation module in the `eRm` package (Mair and Hatzinger 2007, 2008) in `R` (R Developmend Core Team 2008).

Items	Persons	LR	$D_g^{(0)}$	$D^{(1)}$	$D^{(2)}$	C_g
10	200	.03	.04	.16	.00	.04
	500	.05	.06	.33	.00	.28
	1000	.08	.19	.51	.00	.48
	2000	.05	.56	.49	.03	.83
20	200	.08	.01	.52	.00	.05
	500	.02	.00	.71	.00	.08
	1000	.04	.01	.74	.00	.22
	2000	.04	.03	.70	.00	.57
30	200	.06	.01	.73	.00	.13
	500	.04	.00	.91	.00	.05
	1000	.07	.00	.91	.00	.14
	2000	.03	.03	.88	.00	.31

Table 1: Type I Error Rates (Rasch homogeneous data)

In the first scenario in Table 1 we simulate Rasch homogeneous data matrices and count how often each test (wrongly) rejects the H_0 of Rasch homogeneity. Here, H_1 is not specified. As usual when simulating data which are consistent with a particular model of interest, the tests should hold the α-level of .05.

The results in Table 1 indicate that LR and $D_g^{(0)}$ maintain the appropriate α-level in most conditions. For $k = 10$ and $N = 2000$ $D_g^{(0)}$ does not perform satisfactorily. The $D^{(1)}$ statistic appears to rejects too often. This could be caused by the degenerate χ^2-approximation. Suspiciously, Rost's $D^{(2)}$ never rejects H_0. Finally, Hosmer-Lemeshow's C_g performs better as the number of items increases.

As a second scenario we use the strategy proposed by Suárez-Falcón and Glas (2003) for simulating Rasch violations towards a 2-PL, that is, non-parallel item characteristic curves (ICC). We draw item discrimination parameters from a log-normal distribution with $\log \sigma = .25$ which corresponds to a medium violation of the Rasch assumption. Note that H_0 is still Rasch homogeneity whereas H_1 corresponds to a 2-PL. Hence, we count how often the test (rightly) rejects H_0 and decides in favor of the correct H_1. This corresponds to the power of the test.

Items	Persons	LR	$D_g^{(0)}$	$D^{(1)}$	$D^{(2)}$	C_g
10	200	.26	.05	.25	.00	.13
	500	.57	.40	.34	.00	.29
	1000	.84	.86	.43	.00	.65
	2000	.98	.99	.46	.10	.91
20	200	.57	.06	.52	.00	.10
	500	.88	.32	.73	.00	.14
	1000	1.00	.82	.68	.00	.46
	2000	1.00	1.00	.62	.00	.79
30	200	.67	.06	.73	.00	.11
	500	.98	.23	.81	.00	.17
	1000	.99	.87	.85	.00	.37
	2000	1.00	.98	.89	.00	.77

Table 2: Test power for detecting medium Rasch violations ($\log \sigma = .25$)

The results in Table 2 suggest that LR and $C_g^{(0)}$ perform well and detect model violation. For the latter, the sample size must not be too small. The power values of $D^{(1)}$ appear not to be affected by the sample size. C_g shows an increasing power as N increases but in order to achieve values close to 1, a larger N would be required. However, for this setting its performance is inferior to LR and $C_g^{(0)}$. We can also see that $D^{(1)}$ never rejects H_0 except in the case of $N = 2000$ and $k = 10$.

By increasing the standard deviation of the log-normal distribution for drawing the item discrimination parameters ($\log \sigma = .50$) we can simulate strong violations of Rasch homogeneous data. Thus, compared to the former scenario, there is even stronger evidence for a decision of H_1 and the rejection rate should reflect this.

Items	Persons	LR	$D_g^{(0)}$	$D^{(1)}$	$D^{(2)}$	C_g
10	200	.66	.36	.32	.00	.16
	500	.90	.88	.34	.00	.58
	1000	1.00	1.00	.48	.05	.80
	2000	1.00	1.00	.53	.43	.92
20	200	.98	.51	.46	.00	.12
	500	1.00	.99	.55	.00	.43
	1000	1.00	1.00	.60	.00	.66
	2000	1.00	1.00	.61	.00	.91
30	200	1.00	.57	.68	.00	.17
	500	1.00	.99	.69	.00	.47
	1000	1.00	1.00	.73	.00	.63
	2000	1.00	1.00	.76	.00	.92

Table 3: Test power for detecting strong Rasch violations ($\log \sigma = .50$)

From Table 3 we see that $D_g^{(0)}$ is just as good as LR (except for small N). $D^{(1)}$ and $D^{(2)}$ show a similar behavior as in the former scenario. Compared to Table 2, the power of C_g has improved.

To summarize, we conclude that the collapsing deviance $D_g^{(0)}$ performs similarly to Andersen's LR and therefore we can consider $D_g^{(0)}$ as a goodness-of-fit test for IRT models. Note that unlike the Andersen's LR, $D_g^{(0)}$ is not limited to Rasch models. The problem with Rost's $D^{(2)}$ is that, similar to log-linear approaches, it is defined over 2^k response patterns. Thus, it works only if the number of items is very low and at the same time the sample size large. Due to a poor χ^2 approximation the casewise deviance $D^{(1)}$ should not be considered for IRT or for logistic regression in general. The main problem for the Hosmer-Lemeshow C_g is the performance under the true H_0 in Table 1. Based on the simulation results we can conclude that C_g needs larger N and larger k in order to achieve acceptable performance values.

4. Logistic Regression Fit Indices for IRT

Sometimes it is desirable to model acceptability by means of (standardized) goodness-of-fit indices. Especially when n becomes very large, as for instance in large scale assessment, test statistics tend to become significant. A common approach that is used primarily for model comparisons, are information criteria (IC) such as Akaike's AIC, its consistent version $cAIC$, or BIC within a Bayesian context. For (non-nested) competing models, the model which minimizes an IC of choice is typically selected. A striking drawback of IC-based approaches is that the value of the IC can not be interpreted in a substantive manner. Therefore, in this section we present alternative goodness-of-fit measures which are bounded between 0 and 1, interpretable, and can be used for model selection.

4.1. Coefficients of Determination

In linear regression, the coefficient of determination R^2 is a popular measure for a descriptive evaluation of models in terms of explained variance. Having categorical responses as in logistic regression, many different R^2 proposals have been made. Mittlböck and Schemper (1996) examined 12 R^2 measures within a logistic regression context in terms of their interpretability, bounds (preferably [0; 1]), and influence of linear transformations on the predictors. Only two measures fulfilled their criteria which can be applied in a straightforward manner on the full $\mathbf{y}$ vector (see also Hosmer and Lemeshow 2000, p. 164). The first is the squared Pearson correlation coefficient

$$R_P^2 = \frac{\left[\sum_{l=1}^{L}(y_l - \overline{y})(\hat{\pi}_l - \overline{\pi})\right]^2}{\left[\sum_{l=1}^{L}(y_l - \overline{y})^2\right]\left[\sum_{l=1}^{L}(\hat{\pi}_l - \overline{\pi})^2\right]} \tag{19}$$

with $\overline{y}$ as the mean of the response patterns and $\overline{\pi}$ as the mean of the model probabilities. A second, regression sum-of-squares-like R^2-measure, is

$$R^2_{SS} = 1 - \frac{\sum_{l=1}^{L}(y_l - \hat{\pi}_l)^2}{\sum_{l=1}^{L}(y_l - \overline{y})^2} \tag{20}$$

which is based on a proportional reduction in dispersion of the response. Both, R^2_P and R^2_{SS} relate the variance of the predicted values to the variance of the response vector. As a consequence, they are bounded between $[0;1]$ and they can be interpreted in a straightforward manner.

Commonly, statistical software packages report various Pseudo-R^2 measures based on the likelihood L_0 of a baseline model and the likelihood of the estimated model (e.g. McFadden 1974; Cox and Snell 1989; Nagelkerke 1991). For IRT models the easiest way to compute these measures is by means of Equation (4). Let us denote the likelihood of any GLMM specified IRT model by L_G and the likelihood of the baseline only model, that is $\boldsymbol{\eta} = \beta_0$, by L_0. This intercept-only model implies that there are neither item effects nor person effects. McFadden's R^2 (McFadden 1974), for instance, can be expressed as

$$R^2_{MF} = \frac{\log L_0 - \log L_G}{\log L_G} \tag{21}$$

and can be interpreted in two ways: As proportional reduction in the deviance statistic (Menard 2000) or, within an information theoretic context, as the ratio of the estimated information gain when using GLMM-IRT model in comparison with the baseline model to the estimate of the information potentially recoverable by including all possible explanatory variables (Shtatland, Kleinman, and Cain 2002). However, this interpretation is not as intuitive as for R^2_P or R^2_{SS}. For a discussion of the range of R^2 values in logistic regression see Walsh (1989).

The crucial question at this point is how can we define a parsimoniuos null model in order to get L_0? Within the classical IRT context we can fit parsimonious IRT models by means of linear logistic test models (LLTM; Scheiblechner 1972; Fischer 1973). These are based on a linear decomposition of the item parameters in terms of $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\lambda}$. In this expression, $\boldsymbol{\beta}$ is the item parameter vector of length k, $\boldsymbol{\lambda}$ a vector of length $k' < k$ containing the basic parameters, and $\mathbf{W}$ a design matrix of dimension $k \times k'$. The most parsimonious model we can fit is the one with one design vector $\mathbf{w}$ and thus one basic parameter λ only. Note that due to identifiability issues we can not use $\mathbf{w} = \mathbf{1}$. We have to fix the metric by for example setting the first design vector element to 0 and the remaining elements to 1 (which results in $\beta_1 = 0$). Therefore a one-parameter-only model (using e.g. CML) has to be estimated. Pertaining to Equation (21), the resulting L_0' replaces L_0 and L_C from CML estimation replaces L_G.

Another possibility of defining a null model can be established by taking into account equations (12) and (13). For defining a null model we set $\hat{\pi}_l := \overline{y}$ (Agresti 2002, Section 6.2.5). Specifically, for the null model, the likelihood becomes

$$L_0'' = \prod_{l=1}^{L} \overline{y}^{y_l}(1-\overline{y})^{(1-y_l)}. \tag{22}$$

Hence, in Equation (21) L_0'' replaces L_0 and the full likelihood L from Equation (12) replaces L_G. In an analogous manner, other Pseudo-R^2 can be easily established.

4.2. IRT Models as Classifiers

In the field of data mining and machine learning, logistic regression is typically considered as a classification method (see e.g. Witten and Frank 2005). By assuming a threshold τ (e.g., $\tau = .5$) the estimated response probabilities $\hat{\pi}_{vi}$ are dichotomized into $\hat{y}_{vi} = 1$ if $\hat{\pi}_{vi} \geq .5$ and $\hat{y}_{vi} = 0$ otherwise. The corresponding frequencies can be arranged in an "predicted vs. observed" 2×2 classification table which in classification terminology is called *confusion matrix* (see Table 4).

TP denotes the *true positives*, TN the *true negatives*, FP the *false positives*, and FN the *false negatives*. Based on this confusion matrix, various measures can be computed. The most important for our IRT context are the following:

- Accuracy: $ACC = (TP + TN)/L$ which denotes how many observations are correctly classified.

	observed		
predicted	1	0	Σ
1	n_{11} (TP)	n_{12} (FP)	n_{1+} (P_{pred})
0	n_{21} (FN)	n_{22} (TN)	n_{2+} (N_{pred})
Σ	n_{+1} (P_{obs})	n_{+2} (N_{obs})	n_{++} (L)

Table 4: IRT Confusion Matrix.

- True positive rate (hit rate or sensitivity): $TPR = TP/P_{obs}$; that is among the responses observed as 1, how many were correctly classified as 1.
- False positive rate (false alarm rate): $FPR = FP/N_{obs}$; that is among the responses observed as 0, how many were incorrectly classified as 1.
- Specificity: $SP = TN/N_{obs} = 1 - FPR$; i.e among the responses observed as 0, how many were correctly classified as 0.

A more sophisticated measure of the classification accuracy can be achieved by means of *receiver operation characteristic* curve, short ROC curve. The ROC curve is established by plotting the FPR on the x-axis against the TPR on the y-axis.

There are several important coordinates in the ROC space (Fawcett 2006): The lower left point $(0,0)$ refers to the fact that the classifier commits no FP but also gains no TP. On the upper right point $(1,1)$ the classifier returns TP with probability .5. The upper left point $(0,1)$ represents (unrealistic) perfect classification, that is no FP ($FPR = 0$) and only TP ($TPR = 1$).

If the classifier point lies on the upper left-hand side of the ROC graph, it is said to be "conservative": it classifies 1 only with an underlying strong evidence (i.e. high solving probability) which can result in a low TPR, since in practice more than only the most evident responses are observed as 1. Simultaneously such classifiers make few FP errors. If the classifier point is on the upper right-hand side it is thought of as "liberal": it classifies 1 also with weak evidence which results in a high TPR. At the same time, the FPR can be high as well since also 0 responses are classified as 1. If a classifier is on the diagonal it is as good (or better bad) as a coin toss. This is one point of view of the ROC space and it can be used for the comparison of different IRT models. In this case τ should be the same across the models.

Another way to perform ROC analysis regards the determination of the goodness-of-classification of a particular classifier, or in our case the performance of an IRT model. The starting point of this type of ROC analysis is the question what happens if the cutpoint τ is varied within $[0,1]$? For each $\tau \in [0,1]$ we compute FPR and TPR, based on the underlying confusion matrix. We start with $\tau = 1$ which results in point $(0,0)$. By lowering τ stepwise we move from left to right on the resulting ROC curve. Finally, for $\tau = 0$ we reach point $(1,1)$. To evaluate goodness-of-classification we can evaluate the area under the curve (AUC). Basically, AUC varies in the interval $[.5,1]$ where $AUC = 0.5$ denotes random classification (coin toss) and $AUC = 1$ perfect classification. Hosmer and Lemeshow (2000, p. 162) refer to "acceptable discrimination" if $.7 \leq AUC < .8$, "excellent discrimination" if $.8 \leq AUC < .9$, and "outstanding discrimination" if $AUC \geq .9$. Based on AUC we can compute the Gini-coefficient (Gini 1955) by means of $2 \times AUC - 1$.

As an additional result, we can plot the cutoff values τ on the x-axis and the corresponding sensitivities/specificities on the y-axis. The point where the two curves intersect is the optimal cutpoint where the sensitivity and specificity of classification are maximized. We provide an example in the following section.

5. Example: Social Discomfort Scale

In this section we apply the test statistics from Section 3 and the fit indices from Section 4 on a social discomfort dataset (SOD). This dataset is also analyzed in Reise and Waller (2003) and Waller and Reise (2009). To form a Rasch homogenous scale, a subset of items from a social discomcort scale was selected.

Test statistic	Value	df	p-value
Andersen LR	18.711	12	.096
Collapsed Deviance $D_g^{(0)}$	189.690	156	.034
Casewise Deviance $D^{(1)}$	22222.105	22297	.638
Rost Deviance $D^{(2)}$	3835.788	8179	1.000
Hosmer-Lemeshow C_g	6.284	8	.615

Table 5: Test results for social discomfort data

Specifically, $k = 13$ items were identified ($N = 2000$), based on the item-wise Wald criterion (Glas and Verhelst 1995, Section 5.4).

Because the `eRm` package uses ML estimation of the person parameters, subjects with complete 1 or complete 0 response patterns have to be eliminated. Our sample ultimately consisted of $N = 1717$ subjects and the length of the response vector $\mathbf{y}$ is $L = 1717 \times 13 = 22321$.

In Table 5 we see that the p-value of Andersen's LR (median split) is .096 which suggests that the Rasch model fits (even though closely). The collapsed deviance $D^{(0)}$ would reject (again closely) the Rasch H_0 with $p = .034$. The three remaining tests statistics make a decision in favor of H_0.

When we applied the proposed R^2 measures to this dataset, we obtained values of $R_P^2 = .296$, $R_{SS}^2 = .296$, and $R^2 = .327$. Hence, approximately 30% of the variance in the observed data can be explained by the fitted model (R_{SS}^2). Or, alternatively, the squared correlation between between the observed and predicted values is approximately .30 (R_P^2; $r = .544$).

	observed		
predicted	1	0	Σ
1	4958	2043	7001
0	3361	11959	15320
Σ	8319	14002	22321

Table 6: Confusion matrix for social discomfort data.

Table 6 provides the confusion matrix for the social discomfort data. From this table, based on $\tau = .5$, an accuracy $ACC = .758$ results, that is, about 75% of the responses are correctly classified. The sensitivity or true positive rate is $TPR = .560$ which suggests that among all observed 1 responses, 56% were correctly classified. The false positive rate is $FPR = .146$ and, consequently, the specificity $SP = .854$. This implies that among all observed 0 responses, 85% were correctly classified. Thus, the specificity is a considerably large. The weakness of this IRT model lies in the ability to detect correctly 1 responses and we can conclude that this IRT classfier behaves "conservative": The overall accuracy is good but is classifies 1 responses only if there is strong evidence. The TPR is considerably low as well as the FPR.

A more detailed performance examination is provided the ROC curve on the left hand side of Figure 1. The $AUC = .819$ which, according the rules of thumb from Section 4.2 suggests an "excellent discrimination". Correspondingly, the Gini coefficient is $GC = .637$. The dotted lines show the FPR and TPR for $\tau = 0.5$. In order to examine the cutpoint behavior we can look at the right plot in Figure 1. This plot suggests that the optimal cutoff value, that is where sensitivity and specificity are maximized, is $\tau_{opt} = .379$. A reproduction of the confusion matrix using this cutpoint results in a slightly lower overall accuracy of .739 and specificity of .742, but we achieve a noticeable gain in sensitivity to .736.

6. Discussion

We have adapted several logistic regression approaches in order to devise new methods for judging goodness-of-fit in IRT. For overall model tests, several approaches such as the casewise and collapsed deviance, Rost's LR-test, and the Hosmer-Lemeshow test were developed and reviewed, respectively. Within this context we discussed the definition of a saturated model in IRT. These tests were subject of a simulation study whereas Andersen's LR-test acted as benchmark test. It resulted that the collapsed

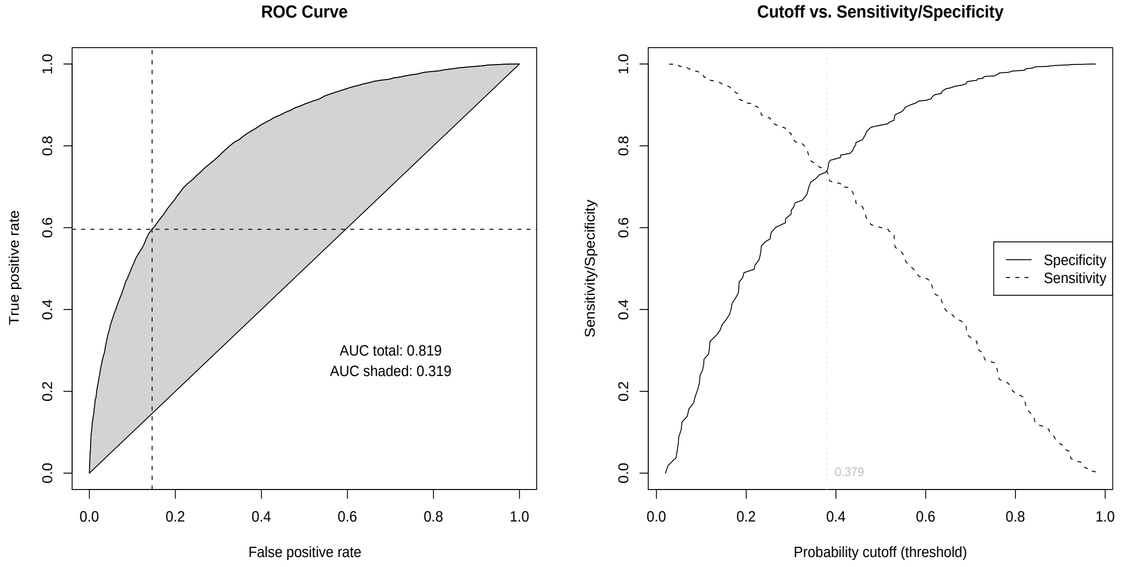

Figure 1: ROC plot and cutoff plot

deviance was powerful in detecting Rasch violations towards a 2-PL and was able to hold the Type I error rate.

In the second part we focused on fit indices. First, we adapted three R^2-measures from logistic regression with special emphasis on their interpretability. Within this context we discussed the definition of null models in IRT. Second, by considering IRT models as classifiers, the confusion matrix was the starting point of several additional fit indices such as various (mis)classification rates, the area under die ROC curve, and the Gini coefficient. The classifier approach offers numerous additional diagnostic tools that can be found in data mining literature. An extensive computational implementation is given by means of the R package ROCR (Sing, Sander, Beerenwinkel, and Hefgauer 2005). The practical use of these fit indices was demonstrated on a real data example.

Future research needs to explore different grouping mechanisms for Hosmer-Lemeshow tests and how they impact the testing power and Type I error rate. For example, non-percentile based grouping (e.g. raw scores, patterns) as well as dynamic percentile grouping (depending on the number of items/persons) should be considered. Another important topic for future research is to extend these approaches to polytomous IRT models. Pertaining to the GLMM framework, De Boeck and Wilson (2004) show how polytomous models can be formulated as multinomial logit models. For likelihood-based goodness-of-fit approaches the extensions appear to be straightforward. Concerning approaches based on model probabilities, corresponding modifications of the fit indices have to be found. For instance, Hand and Till (2001) generalize the ROC analysis for multiple class classifiers. Such generalization would provide a comprehensive framework for examining goodness-of-fit of rating scale models, partial credit models, nominal response models, and graded response models.

References

Agresti A (2002). *Categorical Data Analysis*. Wiley, New York, 2nd edition.

Andersen EB (1973). "A goodness of fit test for the Rasch model." *Psychometrika*, **38**, 123–140.

Cai L, Maydeu-Olivares A, Coffman DL, Thissen D (2006). "Limited information goodness of fit testing of item response theory models for sparse 2^p tables." *British Journal of Mathematical and Statistical Psychology*, **59**, 173–194.

Cox DR, Snell EJ (1989). *The analysis of binary data.* Chapman & Hall, London, 2nd edition.

De Boeck P, Wilson M (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* Springer, New York.

Embretson S, Reise S (2000). *Item Response Theory for Psychologists.* Erlbaum, Mahwah, NJ.

Fawcett T (2006). "An introduction to ROC analysis." *Pattern Recognition Letters*, **27**, 861–874.

Fischer GH (1973). "The linear logistic test model as an instrument in educational research." *Acta Psychologica*, **37**, 359–374.

Gini C (1955). "Variabilità e mutabilità [Variability and Mutability; reprint of Gini's 1912 paper]." In E Pizetti, T Salvemini (eds.), "Memorie di metodologica statistica," pp. 211– 382. Libreria Eredi Virgilio Veschi, Rome.

Glas CAW, Verhelst ND (1995). "Testing the Rasch model." In GH Fischer, IW Molenaar (eds.), "Rasch Models: Foundations, Recent Developments and Applications," pp. 69–96. Springer, New York.

Hand DJ, Till RJ (2001). "A simple generalization of the area under the ROC curve to multiple class classification problems." *Machine Learning*, pp. 171–186.

Hatzinger R (2008). "A GLM framework for item response theory models: Reissue of 1994 Habilitation thesis." *Research Report Series Department of Statistics and Mathematics 66*, Wirtschaftsuniversität Wien.

Hosmer DG, Lemeshow S (2000). *Applied Logistic Regression.* Wiley, New York, 2nd edition.

Hosmer DW, Lemeshow S (1980). "A goodness-of-fit test for the multiple logistic regression model." *Communications in Statistics*, **A10**, 1043–1069.

Joe H, Maydeu-Olivares A (2006). "On the asymptotic distribution of Pearson's X^2 in cross-validation samples." *Psychometrika*, **71**, 587–592.

Kelderman H (1984). "Loglinear Rasch model tests." *Psychometrika*, **49**, 223–245.

Mair P, Hatzinger B (2007). "Extended Rasch modeling: The eRm package for the application of IRT models in R." *Journal of Statistical Software*, **20**, 1–20.

Mair P, Hatzinger R (2008). *eRm: Extended Rasch Modeling.* R package version 0.10-0.

Maydeu-Olivares A, Joe H (2005). "Limited and full information estimation and testing in 2^n contingency tables: A unified framework." *Journal of the American Statistical Association*, **100**, 1009–1020.

McCullagh P, Nelder JA (1989). *Generalized Linear Models.* Chapman & Hall, London, 2nd edition.

McFadden D (1974). "Conditional logit analysis of Qualitative choice behaviour." In P Zarembka (ed.), "Frontiers in Econometrics," pp. 105–142. Academic Presc, New York.

Menard S (2000). "Coefficients of determination for multiple logistic regression analysis." *The American Statistician*, **54**, 17–24.

Miller TR, Spray JA (1993). "Logistic discriminant function analysis for DIF identification of polytomously scored items." *Journal of Educational Measurement*, **30**, 107–122.

Mittlböck M, Schemper M (1996). "Explained variation for logistic regression." *Statistics in Medicine*, **15**, 1987–1997.

Nagelkerke NJD (1991). "A note on a general definition of the coefficient of determination." *Biometrika*, **78**, 691–692.

Orlando M, Thissen D (2000). "Likelihood-based item-fit indices for dichotomous item response theory models." *Applied Psychological Measurement*, **24**, 50–64.

Orlando M, Thissen D (2003). "Further investigation of the performance of $S-X^2$: An item fit index for use with dichotomous item response theory models." *Applied Psychological Measurement*, **27**, 289–298.

R Developmend Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Rasch G (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Reise SP, Waller NG (2003). "How many IRT parameters does it take to model psychopathology items?" *Psychological Methods*, **8**, 164–184.

Rost J (2004). *Lehrbuch Testtheorie/Testkonstruktion [Textbook test theory/test construction]*. Huber, Bern, 2nd edition.

Scheiblechner H (1972). "Das Lernen und Lösen komplexer Denkaufgaben. [The learning and solving of complex reasoning items.]." *Zeitschrift für Experimentelle und Angewandte Psychologie*, **3**, 456–506.

Shtatland ES, Kleinman K, Cain EM (2002). "One more time on R^2 measures of fit in logistic regression." *NESUG 15 Proceedings*, pp. 222–226.

Simonoff JS (1998). "Logistic regression, categorical predictors, and goodness-of-fit: It depends on who you ask." *The American Statistician*, **10**, 10–14.

Sing T, Sander O, Beerenwinkel N, Hefgauer T (2005). "`ROCR`: Visualizing classifier performance in `R`." *Bioinformatics*, **21**, 3940–3941.

Suárez-Falcón JC, Glas CAW (2003). "Evaluation of global testing procedures for item fit to the Rasch model." *British Jotrnal of Mathematical and Statistical Psychology*, **56**, 127–143.

Swaminathan, Rogers HJ (1990). "Detecting differential item functioning using logistic regression procedures." *Journal of Educational Measurement*, **27**, 361–370.

von Davier M (1997). "WINMIRA: Program description and recent enhancements." *Methods of Psychological Research Online*, **2**.

Waller NG, Reise SP (2009). "Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI." In S Embretson, JS Roberts (eds.), "New Directions in Psychological Measurement with Model-Based Approaches," American Psychological Association, Washington, DC.

Walsh A (1989). "Logistic regression and model fit." *Teaching Sociology*, **17**, 419–420.

Witten IH, Frank E (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman, San Francisco, CA, 2nd edition.

Zumbo BD (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary ramework for binary and likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense, Ottawa, ON.