Research Report

Distractor Analysis for Multiple-Choice Tests: An Empirical Study With International Language Assessment Data

ETS RR–19-39

Shelby J. Haberman
Yang Liu
Yi-Hsuan Lee

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Distractor Analysis for Multiple-Choice Tests: An Empirical Study With International Language Assessment Data

Shelby J. Haberman,[1] Yang Liu,[2] & Yi-Hsuan Lee[3]

1 Consultant, Jerusalem, Israel
2 University of Maryland. College Park, College Park, MD
3 Educational Testing Service, Princeton, NJ

Distractor analyses are routinely conducted in educational assessments with multiple-choice items. In this research report, we focus on three item response models for distractors: (a) the traditional nominal response (NR) model, (b) a combination of a two-parameter logistic model for item scores and a NR model for selections of incorrect distractors, and (c) a model in which the item score satisfies a two-parameter logistic model and distractor selection and proficiency are conditionally independent, given that an incorrect response is selected. Model comparisons involve generalized residuals, information measures, scale scores, and reliability estimates. To illustrate the methodology, a study of an international assessment of proficiency of nonnative speakers of a single target language used to make high-stakes decisions compares the models under study.

Keywords Item response theory; nominal response model; model fit; test scoring

doi:10.1002/ets2.12275

Known for its effectiveness and economy, the multiple-choice (MC) item format has been widely used in educational assessments across a variety of content domains (for a recent review, see Gierl, Bulut, Guo, & Zhang, 2017). MC items are often dichotomously scored by whether the correct answer is selected; however, it is generally believed that distractors, or incorrect response options, play an important role in determining the quality of MC items and providing diagnostic information about test performance (e.g., Briggs, Alonzo, Schwab, & Wilson, 2006; Haladyna, 2016). Besides benefiting the practice of item writing and test development, distractor analysis via modeling the propensity of selecting distracting options potentially improves measurement precision (Levine & Drasgow, 1983; Thissen & Steinberg, 1984) as well as the detection of unusual response similarity in the context of test security (e.g., Haberman & Lee, 2017; Wollack, 1997).

Distractor analysis can be performed based on either descriptive statistics or item response theory (IRT) models. Examination of marginal distributions of distractor selection and the association between distractor selection and estimated proficiency can be used to identify nondiscriminating or nonfunctioning distractors (Levine & Drasgow, 1983; Wainer, 1989), which in turn guides item revision by content experts. Alternatively, an IRT-based distractor analysis typically relies on fitting a polytomous IRT model to the raw response data in lieu of the dichotomously scored data. The relationship among distractor selection, item score, and ability can be assessed using corresponding model parameters in place of observed statistics. Moreover, because polytomous IRT models take into account the additional discriminative power of incorrect options, scale scores estimated thereof are anticipated to be more precise than those obtained from conventional dichotomous IRT models fitted to the scored data.

A number of IRT models for distractors have been developed in the literature (Bock, 1972; Briggs et al., 2006; Haberman & Lee, 2017; Samejima, 1979; Suh & Bolt, 2010; Thissen & Steinberg, 1984; Thissen, Steinberg, & Fitzpatrick, 1989; Wilson, 1992); many of them are extensions of or modifications to Bock's (1972) nominal response (NR) model, which assumes a log-linear parameterization. There are also finite mixture models for heterogeneous distractor selection styles (e.g., Bolt, Cohen, & Wollack, 2001) and diagnostic classification models for MC items with qualitative latent traits (e.g., de la Torre, 2009). For simplicity, we only focus on distractor models that are readily described in terms of parameters in log-linear models; readers who are interested in other distractor models are referred to the original references. When multiple candidate models are present, it is often desired to select the model of best fit to the observed data. In the present work, model–data fit is gauged by the estimated log-penalty function (Gilula & Haberman, 1994, 1995, 2001; Haberman,

Corresponding author: Y.-H. Lee, E-mail: ylee@ets.org

2005) and generalized residuals (Haberman & Sinharay, 2013; Haberman, Sinharay, & Chon, 2013). Other assessment procedures of model–data fit, such as quadratic-form statistics based on marginal residuals (e.g., Joe & Maydeu-Olivares, 2010; Reiser, 1996), are available but not further discussed.

Empirical evidence suggests that IRT models for raw responses to MC items lead to more reliable estimation of scale scores, since individual preferences of certain distracting options often carry additional information about proficiency. In an analysis of the Raven Progressive Matrices test data, Thissen (1976) observed that, compared to a two-parameter logistic (2PL) model fitted to the dichotomously scored data, the NR model fitted to the raw responses yielded substantially higher test information in the lower half of the latent trait scale. Similar findings were obtained by Lukhele, Thissen, and Wainer (1994) in the context of achievement testing: When the test items are difficult, the lack of measurement precision can be ameliorated by utilizing the additional information carried by the distractors.

The goal of the current study is to conduct an empirical distractor analysis for an international assessment of language proficiency designed for nonnative speakers of a single target language and used in making high-stakes decisions. We are interested in (a) identifying informative distractors, (b) selecting the best-fitting distractor model, and (c) evaluating the added value of using distractor models for scoring. The rest of the report is organized as follows. We first introduce in the Methods section the basic notations, distractor models, procedures to assess model–data fit, and IRT scoring. The data set under study and a road map of our analysis are then described in the Data section. Next, the techniques developed in the Methods section are applied to the data, and the main findings are summarized in the Results section. Implications and future extensions of the present work are discussed in the Discussion and Conclusion section.

Methods

Notation and Setup

Consider a MC test with m items administered to n examinees. Suppose that item j, $1 \le j \le m$, has $r_j > 1$ possible responses indexed by integers from 0 to $r_j - 1$ and that category K_j, $0 \le K_j \le r_j - 1$, is the unique answer key. Let Z_{ij} be examinee i's raw response to item j, $0 \le Z_{ij} \le r_j - 1$, and let $Y_{ij} = k_j(Z_{ij})$ be the dichotomous item score, $Y_{ij} = 0$ or 1, where, for nonnegative integers $z < r_j$, the mapping k_j satisfies $k_j(z) = 1$ if $z = K_j$ and $k_j(z) = 0$ otherwise. The set inverse of k_j, that is, $\mathcal{K}_j\left(y\right) = \left\{z : k_j\left(z\right) = y\right\}$, gives the collection of item responses to item j that correspond to item score y: $\mathcal{K}_j\left(1\right)$ has the single element K_j, while $\mathcal{K}_j\left(0\right)$ is the set of nonnegative integers that are less than r_j and not equal to K_j. Write $\mathbf{Z}_i$ and $\mathbf{Y}_i = \mathbf{k}(\mathbf{Z}_i)$ as the response and score vectors for examinee i with elements Z_{ij} and $Y_{ij} = k_j(Z_{ij})$, $1 \le j \le m$, respectively. The raw responses $\mathbf{Z}_i$, $1 \le i \le n$, are assumed to be independent and identically distributed (i.i.d.). The sample spaces of $\mathbf{Z}_i$ and $\mathbf{Y}_i$ are denoted $\mathcal{Z}$ and $\mathcal{Y}$, respectively.

To specify probability models for the common distribution of the item responses $\mathbf{Z}_i$, $1 \le i \le n$, let $p_{\mathbf{Z}} : \mathcal{Z} \to (0, 1)$ be the response-pattern probability function such that $\sum_{\mathbf{z}\in\mathcal{Z}} p_{\mathbf{Z}}\left(\mathbf{z}\right) = 1$. IRT models for distractors considered in this report amount to restrictive parameterizations of $p_{\mathbf{Z}}$. Let θ_i denote examinee i's one-dimensional proficiency level; θ_i, $1 \le i \le n$, are i.i.d. following a standard normal distribution $\mathcal{N}\left(0, 1\right)$ with density function ϕ. Under the conventional assumption of local independence (e.g., McDonald, 1981), the item responses Z_{ij}, $1 \le j \le m$, are conditionally independent given θ_i for each i, that is, the conditional probability that $\mathbf{Z}_i = \mathbf{z}$ given $\theta_i = \theta$ is $p_{\mathbf{Z}}\left(\mathbf{z}|\theta\right) = \prod_{j=1}^{m} p_{Zj}\left(z_j|\theta\right)$, where z_j is the element j of $\mathbf{z}$ and $p_{Zj}(z|\theta)$ is the probability that $Z_{ij} = z$ given $\theta_i = \theta$. It follows that the marginal probability that $\mathbf{Z}_i = \mathbf{z}$ is

$$p_{\mathbf{Z}}\left(\mathbf{z}\right) = \int_{-\infty}^{\infty} p_{\mathbf{Z}}\left(\mathbf{z}|\theta\right) \phi\left(\theta\right) d\theta. \tag{1}$$

Similar notations are defined for the generating model of the dichotomously scored data $\mathbf{Y}_i$: The marginal probability that $\mathbf{Y}_i = \mathbf{y}$ is

$$p_{\mathbf{Y}}\left(\mathbf{y}\right) = \int_{-\infty}^{\infty} p_{\mathbf{Y}}\left(\mathbf{y}|\theta\right) \phi\left(\theta\right) d\theta, \tag{2}$$

where $p_{\mathbf{Y}}\left(\mathbf{y}|\theta\right) = \prod_{j=1}^{m} p_{Yj}\left(y_j|\theta\right)$ is the probability that $\mathbf{Y}_i = \mathbf{y}$ given $\theta_i = \theta$, y_j is the element j of $\mathbf{y}$, and

$$p_{Yj}\left(y|\theta\right) = \sum_{z\in\mathcal{K}_j(y)} p_{Zj}\left(z|\theta\right) \tag{3}$$

is the probability of receiving item score $0 \le y \le 1$ given $\theta_i = \theta$.

Models for Distractors

This report examines three polytomous IRT models for MC tests.

The Nominal Response Model

In the NR model (Bock, 1972), the conditional probability that $Z_{ij} = z$, $0 \leq z \leq r_j - 1$, given $\theta_i = \theta$ satisfies the log-linear model is

$$p_{Zj}\left(z|\theta\right) = \frac{\exp\left(\alpha_{jz}\theta + \tau_{jz}\right)}{\sum_{k=0}^{r_j-1}\exp\left(\alpha_{jk}\theta + \tau_{jk}\right)}, \tag{4}$$

where α_{jz} and τ_{jz}, $0 \leq z \leq r_j - 1$, are category slopes and intercepts, respectively. Identification constraints are needed to ensure the estimability of the NR model parameters; in particular, we set $\alpha_{jKj} = \tau_{jKj} = 0$.

The remaining two models are based on the following hierarchical representation for the conditional probability that $Z_{ij} = z$ given $\theta_i = \theta$:

$$p_{Zj}\left(z|\theta\right) = p_{Yj}\left(y|\theta\right) p_{Z|Yj}\left(z|\theta\right), \tag{5}$$

where $y = k_j(z)$. In Equation 5, the term

$$p_{Yj}\left(y|\theta\right) = \frac{\exp\left[y\left(\alpha_j\theta + \tau_j\right)\right]}{1 + \exp\left(\alpha_j\theta + \tau_j\right)} \tag{6}$$

specifies the probability of receiving item score $Y_{ij} = y$ under a 2PL model, where α_j and τ_j are the respective item slope and intercept, and $p_{Z|Yj}(z|\theta)$ is the conditional probability that $Z_{ij} = z$ given $Y_{ij} = k_j(z)$ and $\theta_i = \theta$. In this report, we fix $p_{Z|Yj}(K_j|\theta) = 1$ for all θ and further model $p_{Z|Yj}(z|\theta)$ for distracting options $z \in \mathcal{K}_j(0)$.

The Hybrid Model

One possibility is to further express $p_{Z|Yj}(z|\theta)$ using a NR model restricted to the distractors (Suh & Bolt, 2010). More specifically, for $z \in \mathcal{K}_j(0)$,

$$p_{Z|Yj}\left(z|\theta\right) = \frac{\exp\left(\alpha_{jz}^{\star}\theta + \tau_{jz}^{\star}\right)}{\sum_{k\in\mathcal{K}_j(0)}\exp\left(\alpha_{jk}^{\star}\theta + \tau_{jk}^{\star}\right)}, \tag{7}$$

where $\alpha_{jz}^{\star}$ and $\tau_{jz}^{\star}$ are slopes and intercepts for distractors, respectively. Combining Equations 6 and 7 yields a hybrid model, which is hitherto referred to as the 2PLNR model.

The Two-Parameter Logistic Model With Noninformative Distributions

The other possibility is to further assume that the conditional distribution of Z_{ij} given Y_{ij} does not depend on θ, which leads to the 2PL model with noninformative distributions (2PLND). In particular, let $p_{Z|Yj}(z|\theta) = \pi_{jz} > 0$ for noninformative distractors $z \in \mathcal{K}_j(0)$, so that $\sum_{z\in\mathcal{K}_j(0)}\pi_{jz} = 1$, and let $p_{Z|Yj}(K_j|\theta) = \pi_{jKj} = 1$. Consequently, the marginal probability of $\mathbf{Z}_i = \mathbf{z}$ is the product

$$p_{\mathrm{Z}}\left(\mathbf{z}\right) = p_{\mathrm{Y}}\left(\mathbf{k}\left(\mathbf{z}\right)\right)\prod_{j=1}^{m}\pi_{z_j} \tag{8}$$

for all $\mathbf{z} \in \mathcal{Z}$. The 2PLND model is a special case of the 2PLNR model with $\alpha_{jz}^{\star} = 0$ for all $z \in \mathcal{K}_j(0)$; the conditional probability of $Z_{ij} = z$ given $Y_{ij} = 0$ is then reparameterized as $\pi_{jz} = \exp\left(\tau_{jz}^{\star}\right) / \left[\sum_{k\in\mathcal{K}_j(0)}\exp\left(\tau_{jk}^{\star}\right)\right]$. In addition, the 2PLND model can be deduced from the NR model if we set $\tau_{jKj} = 0$ and $\alpha_{jKj} = 0$ for the correct response K_j and $\tau_{jz} = \log\pi_{jz} - \tau_j$ and $\alpha_{jz} = -\alpha_j$ for z in $\mathcal{K}_j(0)$.

Estimation of Model Parameters

Let $\boldsymbol{\gamma}$ be a d-dimensional vector of model parameters defined on some parameter space Γ. To highlight the dependency of p_{Z} on $\boldsymbol{\gamma}$, we now write $p_{\mathrm{Z}}(\mathbf{z};\boldsymbol{\gamma}) = p_{\mathrm{Z}}(\mathbf{z})$. Model examination depends on maximum likelihood (ML) estimation of $\boldsymbol{\gamma}$. Let

$$\ell(\boldsymbol{\gamma}) = (nm)^{-1}\sum_{i=1}^{n}\log p_{\mathrm{Z}}\left(\mathbf{Z}_i;\boldsymbol{\gamma}\right) \tag{9}$$

be the scaled sample log-likelihood and $\widehat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}\in\Gamma}\ell(\boldsymbol{\gamma})$ be the ML estimator of $\boldsymbol{\gamma}$. Conditions for consistency and asymptotic normality of $\widehat{\boldsymbol{\gamma}}$ are provided in the appendix. These conditions are based on Birch (1964) and Berk (1972). For a discussion of applications to IRT, see Haberman (2016, Section 4.4).

Analysis of Distractor Behavior

The three models under study are compared using three types of evaluation criteria. The first type entails model-free summary statistics that quantify distractor behaviors. The second type measures model – data fit, which includes generalized residuals (Haberman et al., 2013; Haberman & Sinharay, 2013) and estimates of the log-penalty function (Gilula & Haberman, 1994, 1995, 2001; Haberman, 2005, 2013). The final type pertains to scale scores and measurement precision.

Summary Statistics

Let $S(\mathbf{z}) = \sum_{j=1}^{m} k_j\left(z_j\right)$ be the summed score of a response pattern $\mathbf{z}\in\mathcal{Z}$. Define indicator function $\delta_a(b) = 1$ if $a = b$ and $\delta_a(b) = 0$ if $a \neq b$, where a, $b \in \mathbb{R}$. For each distracting option $z \in \mathcal{K}_j(0)$, let σ_{zj}^2, σ_S^2, and let $\sigma_{zj,S}$ be the conditional variance of $\delta_z(Z_{ij})$, the conditional variance of $S(\mathbf{Z}_i)$, and the conditional covariance between $\delta_z(Z_{ij})$ and $S(\mathbf{Z}_i)$ given $Y_{ij} = 0$, respectively. It follows that the conditional point-biserial correlation between $\delta_z(Z_{ij})$ and $S(\mathbf{Z}_i)$ given $Y_{ij} = 0$ is $\rho_{zj,\,S} = \sigma_{zj,\,S}/(\sigma_{zj}\sigma_S)$. Estimating $\rho_{zj,S}$ is straightforward when $\sum_{i=1}^{n}\delta_0\left(Y_{ij}\right) > 0$, an event with a probability that approaches 1 as the sample size n increases. Denote by

$$\widehat{\rho}_{zj,S} = \widehat{\sigma}_{zj,S}/\left(\widehat{\sigma}_{zj}\widehat{\sigma}_S\right) \tag{10}$$

the sample estimate of $\rho_{zj,S}$, where the variance/covariance components are estimated based on the subsample $\{i\colon Y_{ij} = 0\}$. Equation 10 converges to $\rho_{zj,S}$ with probability 1 as the sample size $n \to \infty$. A large $\widehat{\rho}_{zj,S}$ implies that z is an informative distractor for item j.

For each distractor option $z \in \mathcal{K}_j(0)$, we also denote by $\pi_{zj}(s)$ the probability of $Z_{ij} = z$ conditional on the summed score $S(\mathbf{Z}_i) = s$ and an incorrect answer $Y_{ij} = 0$. Provided $\sum_{i=1}^{n}\delta_0\left(Y_{ij}\right)\delta_s\left(S\left(\mathbf{Z}_i\right)\right) > 0$, which again happens with probability arbitrarily close to 1 as $n \to \infty$, $\pi_{zj}(s)$ can be consistently estimated by the observed proportion

$$\widehat{\pi}_{zj}(s) = \frac{\sum_{i=1}^{n}\delta_z\left(Z_{ij}\right)\delta_s\left(S\left(\mathbf{Z}_i\right)\right)}{\sum_{i=1}^{n}\delta_0\left(Y_{ij}\right)\delta_s\left(S\left(\mathbf{Z}_i\right)\right)}. \tag{11}$$

A nearly constant $\pi_{zj}(s)$ in s indicates that z is a noninformative distractor. Such a situation can be identified by a plot (see Figure 2).

Generalized Residuals

Generalized residuals based explicitly on IRT models may be used to analyze distractor behavior under all models considered in this report. For details concerning the large-sample theory of this discussion, see Haberman and Sinharay (2013). Two types of generalized residuals are considered here.

First, consider the following residual statistic for an item j, $1 \le j \le m$, and $\mathbf{z} \in \mathcal{Z}$:

$$e_{zj,S}(\mathbf{z}) = \delta_0\left(k_j\left(z_j\right)\right)S(\mathbf{z})\left[\delta_z\left(z_j\right) - \widehat{p}_{Z|Yj}\left(z_j|\mathbf{z}\right)\right], \tag{12}$$

where $\hat{p}_{Z|Yj}(z|\mathbf{z})$ is the MLE of

$$p_{Z|Yj}(z|\mathbf{z}) = \int_{-\infty}^{\infty} p_{Z|Yj}(z|\theta)\, p_{\theta}(\theta|\mathbf{z})\, d\theta$$

and $p_{\theta}(\theta|\,\mathbf{z}) = p_{\mathbf{Z}}(\mathbf{z}|\,\theta)\phi(\theta)/p_{\mathbf{Z}}(\mathbf{z})$ is the conditional probability density of θ_i given $\mathbf{Z}_i = \mathbf{z}$. If the model under consideration holds, then the statistic

$$D_{zj,S} = \left[\sum_{i=1}^{n} \delta_0\left(Y_{ij}\right)\right]^{-1/2} \sum_{i=1}^{n} e_{zj,S}\left(\mathbf{Z}_i\right)$$

converges in distribution to a normal random variable with mean 0 and variance $\omega^2_{zj,S} > 0$ when the sample size n tends to infinity. A consistent estimate of $\omega^2_{zj,S}$, denoted $\hat{\omega}^2_{zj,S}$, may be based on the Louis approach (Louis, 1982; see Haberman & Sinharay, 2013, for more details). This approach facilitates the estimation of $\omega^2_{zj,S}$ to the problem of computation of root mean squared error in regression analysis, and it has the computational advantage that neither the information matrix nor the Hessian matrix need be used. Let

$$t_{zj,S} = D_{zj,S}/\hat{\omega}_{zj,S} \tag{13}$$

be the generalized residual based on Equation 12, which converges in distribution to $\mathcal{N}(0, 1)$ if the model is correctly specified.

The second generalized residual, often referred to as residuals for item fit (Haberman et al., 2013), compares two estimates of the conditional probability $p_{Z|Yj}(z|\theta)$ for an item j, $1 \le j \le m$, an item response $z \in \mathcal{K}_j(0)$, and a fixed value $\theta \in \mathbb{R}$: the "empirical" estimate

$$\bar{p}_{Z|Yj}(z|\theta) = \frac{\sum_{i=1}^{n} \delta_z\left(Z_{ij}\right) \hat{p}_{\mathbf{Z}}\left(\mathbf{Z}_i|\theta\right)/\hat{p}_{\mathbf{Z}}\left(\mathbf{Z}_i\right)}{\sum_{i=1}^{n} \delta_0\left(Y_{ij}\right) \hat{p}_{\mathbf{Z}}\left(\mathbf{Z}_i|\theta\right)/\hat{p}_{\mathbf{Z}}\left(\mathbf{Z}_i\right)}$$

and the model-based estimate $\hat{p}_{Z|Yj}(z|\theta)$. Let

$$\hat{\Delta}_{zj}(\theta) = \bar{p}_{Z|Yj}(z|\theta) - \hat{p}_{Z|Yj}(z|\theta).$$

If the model holds, $\left[\sum_{i=1}^{n} \delta_0\left(Y_{ij}\right)\right]^{1/2} \hat{\Delta}_{zj}(\theta)$ converges in distribution to a normal random variable with mean 0 and variance $\omega^2_{zj}(\theta) > 0$. The Louis approach can be applied to obtain a consistent estimate of $\omega^2_{zj}(\theta)$, denoted $\hat{\omega}^2_{zj}(\theta)$. When the model is correct, the generalized residual

$$t_{zj}(\theta) = \left[\sum_{i=1}^{n} \delta_0\left(Y_{ij}\right)\right]^{1/2} \hat{\Delta}_{zj}(\theta)/\hat{\omega}_{zj}(\theta) \tag{14}$$

converges in distribution to $\mathcal{N}(0, 1)$ as the sample size n approaches ∞.

Information Analysis

For an arbitrary response-pattern probability function $p_{\mathbf{Z}}$, consider a log-penalty function $-\log p_{\mathbf{Z}}(\mathbf{z};\boldsymbol{\gamma})$ if $\mathbf{Z} = \mathbf{z}$, where $\mathbf{Z}$ is a random vector with the same distribution as $\mathbf{Z}_i$, $1 \le i \le n$ (Gilula & Haberman, 1994, 1995, 2001; Savage, 1971). Then the per-item expected log-penalty is

$$H(\boldsymbol{\gamma}) = -m^{-1} E \log p_{\mathbf{Z}}(\mathbf{Z};\boldsymbol{\gamma}), \tag{15}$$

where E denotes the expectation with respect to the true data-generating model. Let $H_0 = \inf_{\gamma \in \Gamma} H(\boldsymbol{\gamma})$. In particular, the infimum H_0 is attained if the true model is indeed $p_{\mathbf{Z}}(\cdot;\boldsymbol{\gamma}_0)$ for some $\gamma_0 \in \Gamma$, in which case $H(\gamma_0)$ is the average entropy per item. The natural consistent estimate of H_0 is $\hat{H} = -\ell\left(\hat{\gamma}\right)$ (Gilula & Haberman, 1994), so that this estimate may be used to compare models. In samples of intermediate sizes, Akaike or Gilula–Haberman adjustment of $\hat{H}$ of order n^{-1} can be made to reduce bias issues; however, these adjustments are negligible for the data studied in this report, so only $\hat{H}$ is reported.

Recall that the 2PLND model assumes that the distractor selection is unrelated to the latent variable being measured and is nested within the other two candidate models. Thus it is expected to produce larger generalized residuals and estimated log-penalty functions. This expectation is empirically verified in this study, with more focus on examining whether such a difference is substantial enough to justify the use of more complex distractor models.

Scale Scores

For each response pattern $\mathbf{z} \in \mathcal{Z}$, define the expected a posteriori (EAP) score

$$\mu_T(\mathbf{z}) = \int_{-\infty}^{\infty} T(\theta)\, p_\theta(\theta|\mathbf{z})\, d\theta \tag{16}$$

of the test characteristic curve, also known as the expected score function,

$$T(\theta) = E[S(\mathbf{Z})|\theta] = \sum_{j=1}^{m} p_{Yj}(1|\theta). \tag{17}$$

Using the EAP score of $T(\theta)$ rather than that of θ ensures that scores obtained from different distractor models are compared on the same scale. To quantify the precision of the EAP score, we also compute the associated posterior variances of $T(\theta)$:

$$\sigma^2_T(\mathbf{z}) = \int_{-\infty}^{\infty} \left[T(\theta) - \mu_T(\mathbf{z})\right]^2 p_\theta(\theta|\mathbf{z})\, d\theta. \tag{18}$$

As Equations 16 and 18 depend on item parameters $\boldsymbol{\gamma}$, plugging in $\widehat{\boldsymbol{\gamma}}$ yields the respective empirical estimates $\widehat{\mu}_T(\mathbf{z})$ and $\widehat{\sigma}^2_T(\mathbf{z})$ for $\mu_T(\mathbf{z})$ and $\sigma^2_T(\mathbf{z})$. The overall measurement precision is gauged by the following reliability measure (Haberman & Sinharay, 2010):

$$\rho^2 = \frac{\mathrm{Var}\left[\mu_T(\mathbf{Z})\right]}{\mathrm{Var}\left[\mu_T(\mathbf{Z})\right] + E\left[\sigma^2_T(\mathbf{Z})\right]}. \tag{19}$$

In Equation 19, the expectation and variance are taken with respect to the generating model of $\mathbf{Z}$, which can be further estimated by the corresponding sample statistics. Let $\overline{\mu}_T = n^{-1}\sum_{i=1}^{n} \widehat{\mu}_T\left(\mathbf{Z}_i\right)$, $\overline{\sigma}^2_T = n^{-1}\sum_{i=1}^{n} \widehat{\sigma}^2_T\left(\mathbf{Z}_i\right)$, and $s^2\left(\widehat{\mu}_T\right) = n^{-1}\sum_{i=1}^{n}\left[\widehat{\mu}_T\left(\mathbf{Z}_i\right) - \overline{\mu}_T\right]^2$. Then ρ^2 can be estimated by

$$\widehat{\rho}^2 = \frac{s^2\left(\widehat{\mu}_T\right)}{s^2\left(\widehat{\mu}_T\right) + \overline{\sigma}^2_T}. \tag{20}$$

One basic question to be addressed is whether IRT models incorporating distractor information improve over those ignoring such information in terms of generating more reliable estimates of latent variable scores. It is worth noting that the 2PLND model yields the same scores and reliability estimate as what would be obtained from the 2PL model based solely on dichotomized item scores. The additional computational work involved in fitting the NR and 2PLNR models is justified, if a substantial enough increase in the precision of individual scores and overall reliability is observed, if problems with sparse data are not serious, and if issues of public policy discussed in the "Discussion and Conclusion" section can be adequately addressed.

Data

The data were responses of $n = 12{,}123$ examinees from a single administration of a large-scale international assessment of language proficiency in a target language that is not the native language of the examinees. Owing to confidentiality requirements, little background information of the testing program itself can be disclosed. For simplicity, analysis was confined to items with two item scores and four possible item responses, that is, $\mathcal{K}_j(0) = \{0, 1, 2\}$ and $\mathcal{K}_j(1) = \{3\}$ for all j. In all, 29 listening items and 39 reading items were studied. Separate analyses were conducted for listening and reading items. For each section, only examinees who responded to all items were considered, so that 11,383 examinees were used for listening and 10,232 examinees were used for reading.

Results

Summary Statistics

The conditional point-biserial correlations between distractor choices and summed scores given incorrect answers, that is, $\widehat{\rho}_{zj,S}$ (Equation 10), have means close to 0 and standard deviations about .15 for listening and about .14 for reading; the

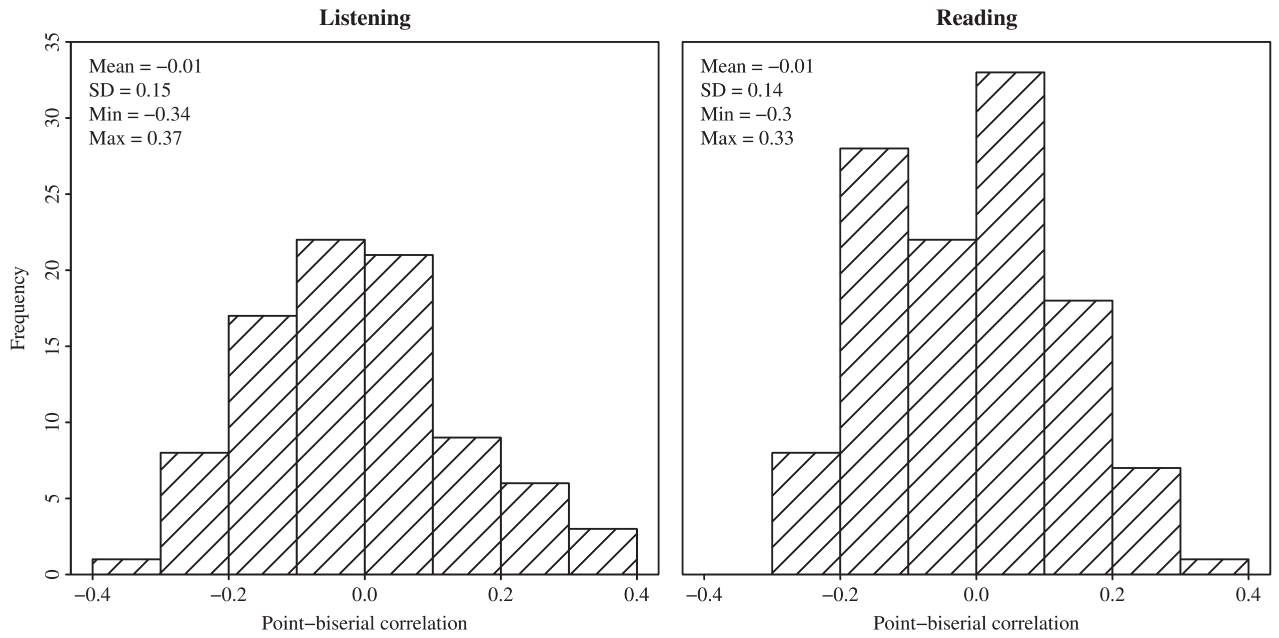

Figure 1 Histograms for the conditional point-biserial correlations between distractor selections and summed scores given incorrect answers, that is, $\hat{\rho}_{zj,S}$. SD = standard deviation. Min = minimum. Max = maximum.

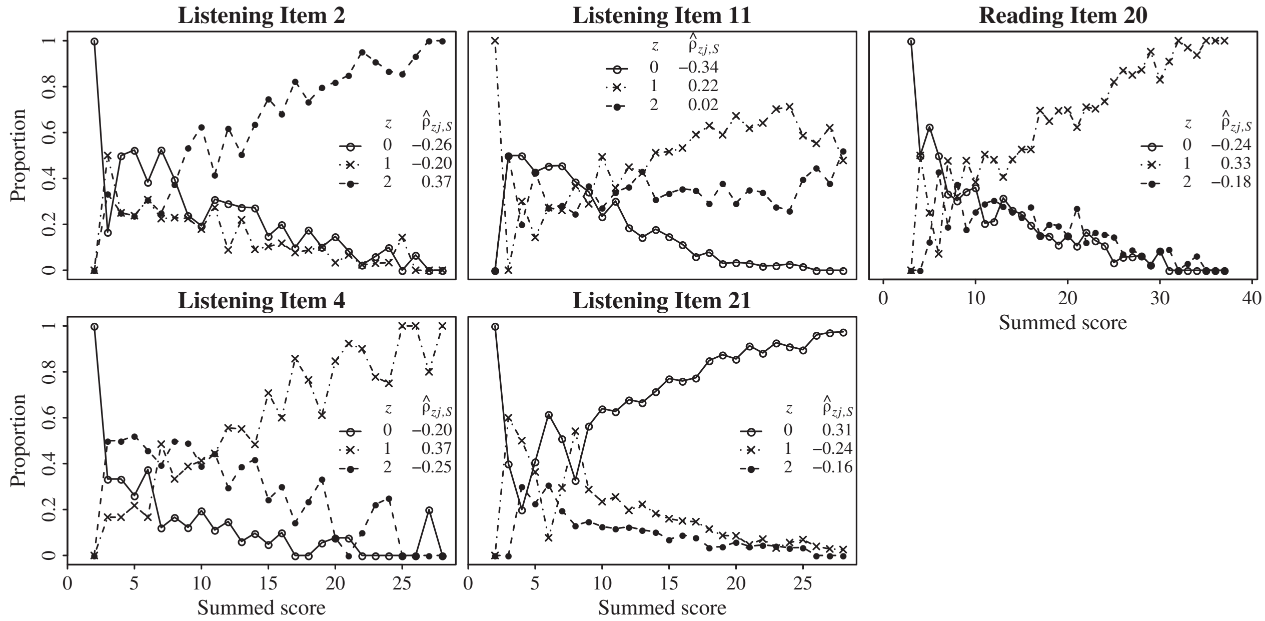

Figure 2 Observed proportions of distractor choices given incorrect answers, that is, $\hat{\pi}_{zj}(s)$, plotted against summed score levels. Plots are only created for the five items (four listening items and one reading item) with $|\hat{\rho}_{zj,S}| > 0.3$ for $z = 0,1,2$. Numbers for different distracting options are shown in distinct line types and symbols.

maximum magnitudes of those correlations are .37 for listening and .33 for reading. The empirical distributions of the conditional point-biserial correlations are displayed in Figure 1. Four listening items and one reading item have $|\hat{\rho}_{zj,S}| >$ .3; for those items, we further calculate the observed proportions of distractor selections conditional on incorrect answers at each summed score level, $\hat{\pi}_{zj}(s)$, given in Equation 11 for $z = 0, 1, 2$ (see Figure 2).

We observe from Figure 2 that distracting options with strong positive (negative) conditional point-biserial correlations are selected more (less) often as the proficiency level (measured by summed scores) increases. In contrast, the conditional

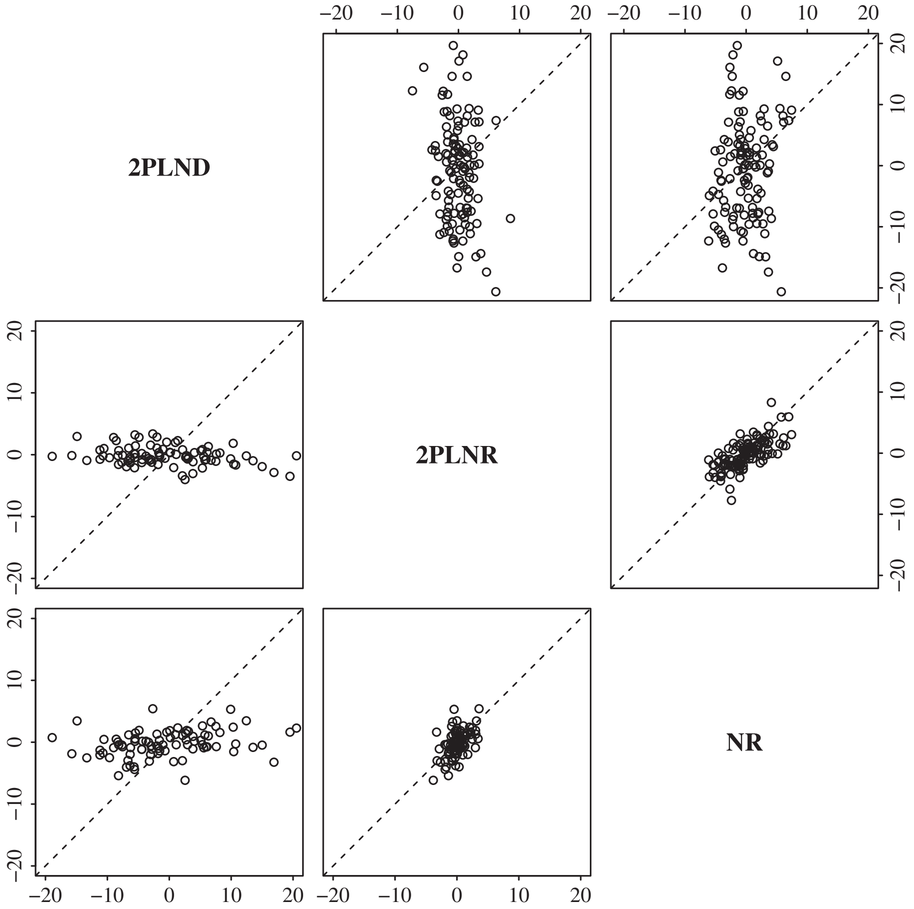

Figure 3 Generalized residuals $t_{zj,S}$ for each distractor response $z \in \mathcal{K}_j(0)$ and each item j. The statistics were plotted for each pair of the fitted models. The lower-triangular panels in the scatterplot matrix correspond to listening items, and the upper-triangular panels correspond to reading items. 2PLND = two-parameter logistic model with noninformative distributions. 2PLNR = hybrid model. NR = nominal response model.

proportions of selection for noninformative distractors, such as $z = 2$ for Listening Item 11, tend to be stable across a wide range of summed score levels. In examination of graphs, note that the estimated proportions of estimates are less precise at the extreme ends of the scale because (a) relatively fewer examinees have low scores in the test and (b) many high-score examinees are excluded because they correctly answered the item (i.e., $z = 3$).

Generalized Residuals

For the three models under study, $t_{zj,S}$ were obtained for each distractor response $z \in \mathcal{K}_j(0)$ and each item j, $1 \leq j \leq m$. To facilitate comparisons among the three fitted models, residuals are displayed as a scatterplot matrix in Figure 3: Results for listening items were plotted in lower-triangular panels, and those for reading items were plotted in upper-triangular panels.

In no case were data fully compatible with the models considered; nonetheless, there are some notable differences among the three models under study. For the listening test, the average values of $|t_{zj,S}|$ are 1.64 for the NR model and 1.12 for the 2PLNR model. In contrast, the 2PLND model is less successful: The average value of $|t_{zj,S}|$ is 6.34. A similar pattern is observed for the reading scale: The average values of $|t_{zj,S}|$ are 2.44 for the NR model, 1.77 for the 2PLNR model, and 6.63 for the 2PLND model. Owing to the large sample sizes, averages of $|D_{zj,S}|$ were also calculated. For listening, the average is .015 for the NR model, .018 for the 2PLNR model, and .254 for the 2PLND model. For reading, the averages are .027, .017, and .332, respectively. Both the criteria of average $|t_{zj,\,S}|$ and average $|D_{zj,S}|$ as well as the scatterplots (Figure 3) suggest that the NR and 2PLNR models are comparable and fit the data appreciably better than the 2PLND model.

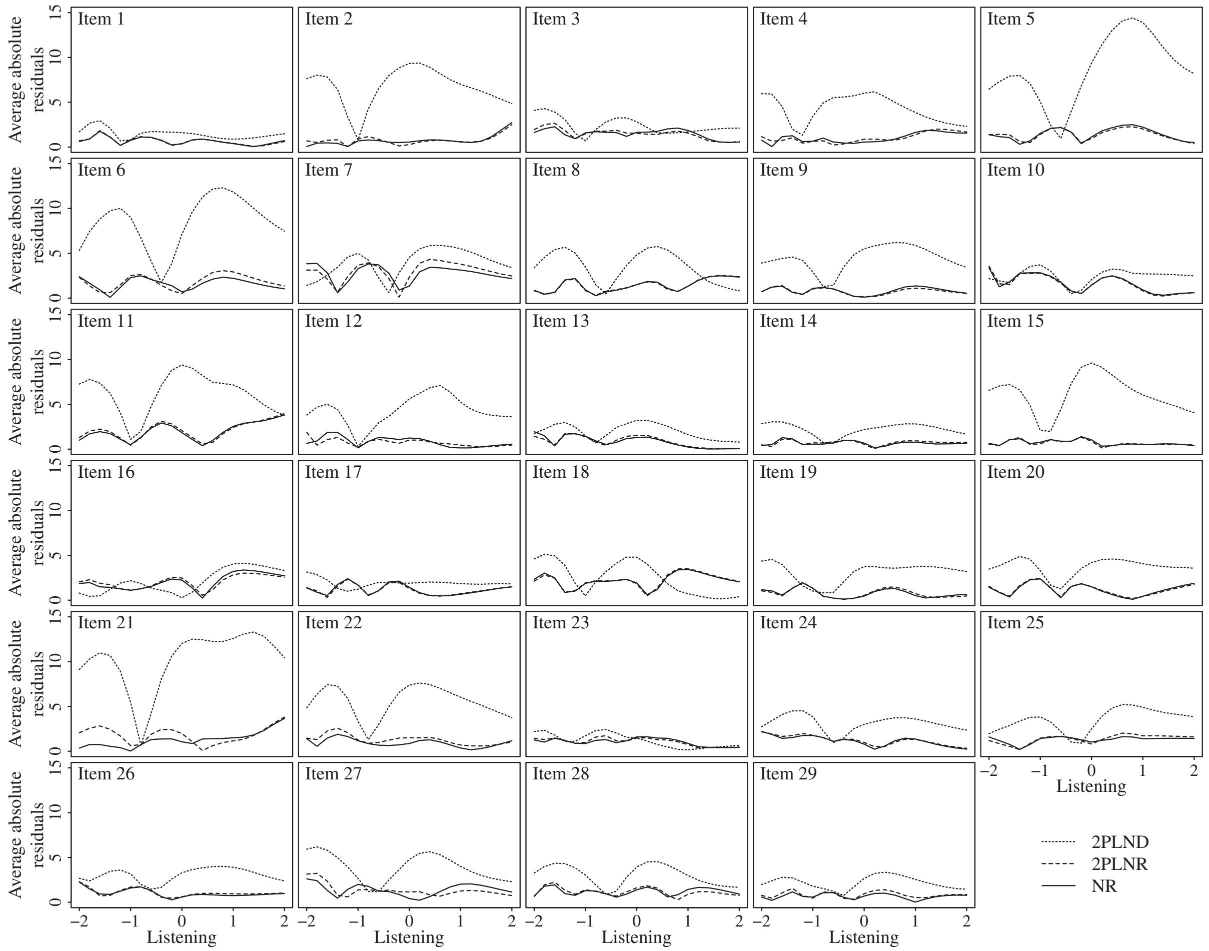

Figure 4 Average absolute residuals, that is, $\frac{1}{3}\sum_{z\in\mathcal{K}_j(0)} \mid t_{zj}(\theta) \mid$, as each item has three distractors, for listening items. Results for the three fitted models are shown in different line types. 2PLND = two-parameter logistic model with noninformative distributions. 2PLNR = hybrid model. NR = nominal response model.

Residuals $t_{zj}(\theta)$ were evaluated for each model for θ equal to $h/5$, where hs are integers between -10 and 10. In Figures 4 and 5, the average $|t_{zj}(\theta)|$ values across $z \in \mathcal{K}_j(0)$ were plotted for each item j at various θ levels for listening and reading, respectively.

Again, we observe that none of the three models agrees with the data perfectly, but the fit of the 2PLND model appears to be noticeably worse than the fits of the NR and 2PLNR models. To summarize the performance of the three fitted models, we further averaged $|t_{zj}(\theta)|$ over distractors $z \in \mathcal{K}_j(0)$, $\theta = h/5$, with integers $-10 \leq h \leq 10$ and items $1 \leq j \leq m$. For listening, the averages were 1.29 for the NR model, 1.24 for the 2PLNR model, and 4.00 for the 2PLND model. For reading, the averages were 1.75 for the NR model, 1.71 for the 2PLNR model, and 4.16 for the 2PLND model. The averages of the $\mid \widehat{\Delta}_{zj}(\theta) \mid$ across z, θ, and j were also examined to check on sizes of discrepancies. For listening, the averages were .018 for the NR model, .017 for the 2PLNR model, and .066 for the 2PLND model. For reading, averages were .025 for the NR model, .025 for the 2PLNR model, and .072 for the 2PLND model. The 2PLND model appears to be substantially less successful than are the NR and 2PLNR models, and the latter two models have roughly comparable performance. In summary, the results of generalized residuals further strengthen the conclusion that both the listening and reading tests contain items with informative distracting options.

Information Analysis

In terms of overall model–data fit, the estimated log-penalty functions also indicate the resemblance between the NR and 2PLNR models as well as their superiority over the 2PLND model. As noted earlier, because of the large sample sizes, different estimates of the log-penalty function are essentially the same, so only $-\ell\left(\hat{\boldsymbol{\gamma}}\right)$ is reported.

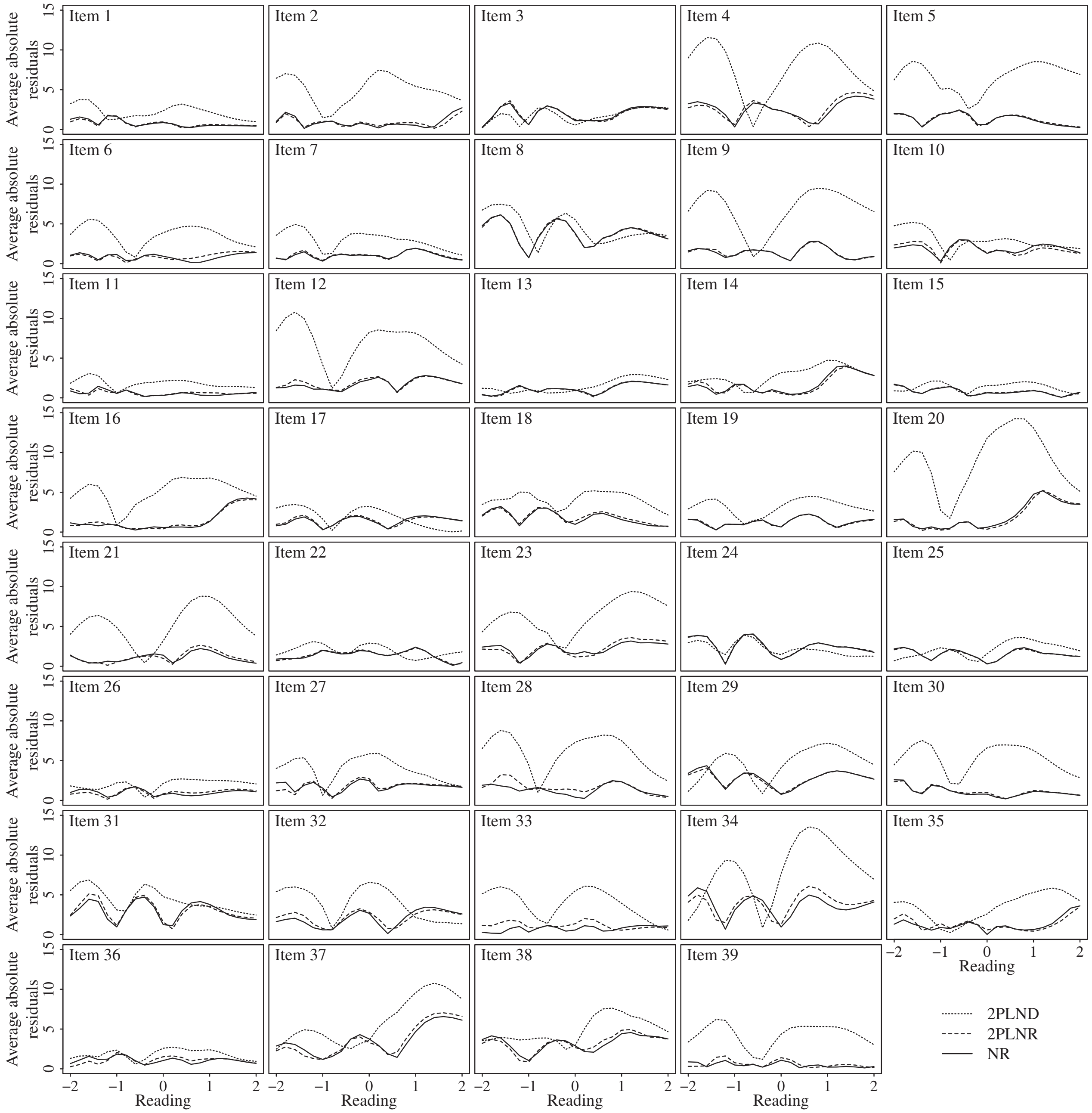

Figure 5 Average absolute residuals, that is, $\frac{1}{3}\sum_{z\in\mathcal{K}_j(0)}|t_{zj}(\theta)|$, as each item has three distractors, for reading items. Results for the three fitted models are shown in different line types. 2PLND = two-parameter logistic model with noninformative distributions. 2PLNR = hybrid model. NR = nominal response model.

For listening, the values of $-\ell\left(\hat{\boldsymbol{\gamma}}\right)$ are .5816 for the NR model, .5815 for the 2PLNR model, and .5854 for the 2PLND model. For reading, the values of $-\ell\left(\hat{\boldsymbol{\gamma}}\right)$ are .6449 for the NR model, .6448 for the 2PLNR model, and .6497 for the 2PLND model. The information criteria for the NR and 2PLNR models are the same up to the first three decimal places, and they are .004–.005 smaller than those of the 2PLND model. To help make sense of the magnitude of the information criteria, we fit the 1PL model with noninformative distributions (the 1PLND model), which further constrains $\alpha_j = \alpha$ for all j in a 2PLND model. Given the range of the slope estimates in the 2PLND models, the 1PLND models are expected to fit the data much less well. The information criteria for the 1PLND model are .5876 for the listening scale and .6523 for the reading scale—they are only about .002 larger than the values for the 2PLND model. Consequently, we infer that

Table 1 Correlations of Expected A Posteriori Scores Among the Three Fitted Models in Four Summed Score Groups

			Score group[a]			
Scale	Correction	Model	1	2	3	4
Listening	No	2PLNR vs. 2PLND	.886	.918	.955	.983
		2PLNR vs. NR	.995	.999	.999	.999
		2PLND vs. NR	.901	.925	.957	.982
	Yes	2PLNR vs. 2PLND	.979	.987	.995	.999
		2PLNR vs. NR	.999	1.000	1.000	1.000
		2PLND vs. NR	.982	.988	.995	.999
Reading	No	2PLNR vs. 2PLND	.891	.925	.963	.984
		2PLNR vs. NR	.997	.999	1.000	.999
		2PLND vs. NR	.905	.932	.964	.982
	Yes	2PLNR vs. 2PLND	.986	.991	.996	.999
		2PLNR vs. NR	1.000	1.000	1.000	1.000
		2PLND vs. NR	.987	.992	.996	.999

Note. Correlations before and after the correction of range restriction are shown separately. 2PLND = two-parameter logistic model with noninformative distributions. 2PLNR = hybrid model. NR = nominal response model.

[a] 1 = lowest; 4 = highest.

a .004–.005 decrease in $-\ell\left(\hat{\gamma}\right)$ that the NR and 2PLNR models achieve, in comparison with the 2PLND model, in fact indicates a substantial improvement in model fit.

Scale Scores

For both the listening and reading tests, the correlations of the estimated EAP scores $\hat{\mu}_T\left(\mathbf{Z}_i\right)$, $i = 1, \ldots, n$, obtained from the three models are always above .99. It seems to suggest that modeling informative distractors does not have a large impact on the estimated scale scores in general. We proceed to partition the samples into four subgroups based on their summed scores $S(\mathbf{Z}_i)$[1] and compute correlations within each group. For listening, the four groups have respective summed score ranges [0,11], [12,17], [18,23], and [24,29]; for reading, the four groups have respective summed score ranges [0,15], [16,23], [24,31], and [32,39]. As the ranges of EAP scores are restricted within groups, we apply the standard correction known as Thorndike's (1949) Case 2. Both the corrected and uncorrected results are reported in Table 1. In lower summed score groups, we observe weaker correlations of EAP scores between the models assuming informative distractors (NR and 2PLNR) and the model assuming noninformative distractors (2PLND); once the correction is applied, the attenuation remains but becomes much less salient. In the meantime, the correlation between the NR and 2PLNR models is nearly perfect across all the score groups before and after the correction.

Next, the estimated posterior variances $\hat{\sigma}^2_T\left(\mathbf{Z}_i\right)$, $i = 1, \ldots, n$, were contrasted between models assuming informative distractors (NR and 2PLNR) and the 2PLND model. Figure 6 displays the distribution of posterior variance ratios by the four summed score groups. An improvement in measurement precision by modeling informative distractors is reflected by a ratio less than 1. For both pairs of models under comparison, the average posterior variance ratios across all examinees are .94 for the listening test and .92 for the reading test, which suggests slightly improved measurement precision. Within the lowest score groups in both tests, more than 10% reductions are observed for more than half of the examinees; the 2PLNR model tends to yield slightly smaller posterior variances compared to the NR model. In addition, the within-group average of posterior variance ratios approaches 1 as the summed score level increases, which implies diminished utility of modeling informative distractors.

As for the estimated reliability of the scale $\hat{\rho}^2$ (Equation 20), exploitation of distractors has very little gain (<.01), similar to the other overall criteria of measurement precisions we have discussed. For listening, the reliability estimates are .878 for both the NR and 2PLNR models and .868 for the 2PLND model. For reading, the reliability estimates are .909 for the NR and 2PLNR models and .900 for the 2PLND model.

Discussion and Conclusion

The methodology introduced in this report provides a relatively simple yet theoretically grounded framework for distractor analysis. It includes examinations of conditional correlations/proportions, generalized residuals, information criteria,

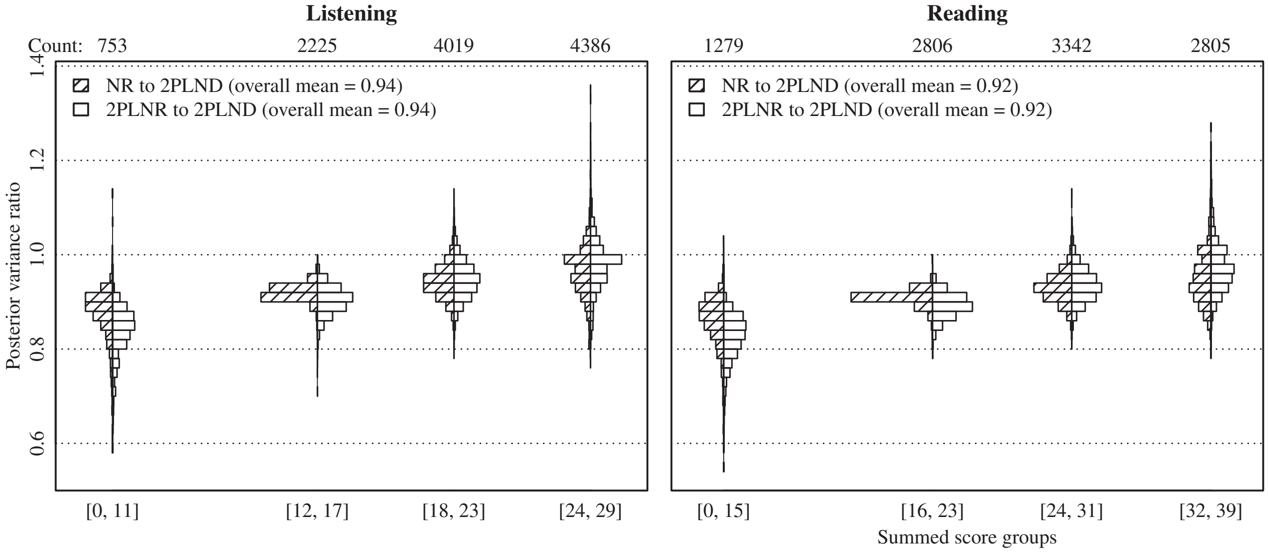

Figure 6 Histograms of posterior variance ratios between the NR model and the 2PLND model (shaded bars) and between the 2PLNR and the 2PLND model (unfilled bars) in four summed score groups. The counts of examinees within each summed score group are shown at the top of each panel. Results are displayed for the (left) listening test and (right) reading test, respectively. 2PLND = two-parameter logistic model with noninformative distributions. 2PLNR = hybrid model. NR = nominal response model.

scale scores, and reliability. In the example considered in this report, the simple 2PLND model, which assumes that distractor selections are not informative given incorrect answers, appears to fit the data worse than the NR and 2PLNR models do in terms of generalized residuals and information criteria. On the other hand, the latter two models assume informative distractors and yield almost identical fit. In addition, the relative weakness of the 2PLND model appears to have a small overall effect on the estimation of scale scores and reliability. The EAP scores obtained from the NR and 2PLNR models are less correlated with those from the 2PLND models among examinees with low proficiency levels; however, the difference becomes negligible after a correction of range restriction. In the meantime, smaller posterior variances are also observed for the NR and 2PLNR models in the low-proficiency groups, which implies better measurement precision. Other data may lead to different conclusions: For example, the real data example in Bock (1972) did lead to a larger effect on reliability than what has been observed in the current work.

Although this report focuses on item scores that are 0 or 1, the methodology developed is readily applied to polytomously scored items and other possible models for distractors. However, challenges arise when test items have a large number of distracting options, as it is difficult to estimate option-specific parameters precisely when the frequencies of a response option are low. In addition, although the analysis in the report assumes that all examinees receive the same items, many of the methods developed can be extended to the scenarios in which examinees in the same administration encounter different items as a result of, for example, adaptive testing and item bank rotation.

It is also questionable whether the general public would accept assigning different credits for different wrong answers, especially when the overall impact in measurement precision is minimal. For the language assessment example, scores in the lowest performance group may be too low to be of practical interest: In particular, they are not much above the expected score for an examinee who randomly guesses. An additional analysis reveals that distractor selection indeed appears more dispersed as the performance decreases, approaching a random guessing pattern. It calls for future policy research to determine the added value of distractor analysis to score-based decision-making.

Finally, there is an appreciable cost of maintenance for a testing program if distractor models are used operationally in place of standard IRT models. Because the distractor models involve more free parameters, more quality-control efforts are required to assure that the parameters are precisely estimated and remain stable across various test administrations and subpopulations of examinees.

Acknowledgments

The authors thank Rebecca Zwick, John Donoghue, and Peter van Rijn for their advice on an earlier version.

Note

[1] Originally, five equal-width subgroups were created for each test. Owing to the small sample sizes, we merged the first two groups so that the correlation coefficients can be estimated more precisely.

References

Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Annals of Mathematical Statistics*, *43*, 193–204. https://doi.org/10.1214/aoms/1177692713

Birch, M. (1964). A new proof of the Pearson–Fisher theorem. *Annals of Mathematical Statistics*, *35*, 817–824. https://doi.org/10.1214/aoms/1177703581

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. https://doi.org/10.1007/bf02291411

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*, 381–409. https://doi.org/10.3102/10769986026004381

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33–63. https://doi.org/10.1207/s15326977ea1101_2

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183. https://doi.org/10.1177/0146621608320523

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*, 1082–1116. https://doi.org/10.3102/0034654317726529

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656. https://doi.org/10.2307/2290867

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *Annals of Statistics*, *23*, 1130–1142. https://doi.org/10.1214/aos/1176324701

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 129–187. https://doi.org/10.1111/0081-1750.00094

Haberman, S. J. (2005). *Latent-class item response models* (Research Report No. RR-05-28). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02005.x

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Haberman, S. J. (2016). Exponential family distributions relevant to IRT. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 2. Statistical tools* (pp. 47–70). Boca Raton, FL: CRC Press.

Haberman, S. J., & Lee, Y.-H. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses* (Research Report No. RR-17-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12150

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227. https://doi.org/10.1007/s11336-010-9158-4

Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of the American Statistical Association*, *108*, 1435–1444. https://doi.org/10.1080/01621459.2013.835660

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*, 417–440. https://doi.org/10.1007/s11336-012-9305-1

Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 392–409). New York, NY: Routledge.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*, 393–419. https://doi.org/10.1007/S11336-010-9165-5

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement*, *43*, 675–685. https://doi.org/10.1177/001316448304300301

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *44*, 226–233. https://doi.org/10.1111/j.2517-6161.1982.tb01203.x

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*, 234–250. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100–117. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

Reiser, M. (1996). Analysis of residuals for the multionmial item response model. *Psychometrika*, *61*, 509–528. https://doi.org/10.1007/BF02294552

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Office of Naval Research Report Nos. 79-4, N00014-77-C-0360). Knoxville, TN: University of Tennessee, Department of Psychology.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801. https://doi.org/10.2307/2284229

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, *75*, 454–473. https://doi.org/10.1007/s11336-010-9163-7

Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, *13*, 201–214. https://doi.org/10.1111/j.1745-3984.1976.tb00011.x

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501–519. https://doi.org/10.1007/bf02302588

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*, 161–176. https://doi.org/10.1111/j.1745-3984.1989.tb00326.x

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York, NY: John Wiley.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*, 191–208. https://doi.org/10.1111/j.1745-3984.1989.tb00328.x

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, *16*, 309–325. https://doi.org/10.1177/014662169201600401

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*, 307–320. https://doi.org/10.1177/01466216970214002

Appendix

Regularity Conditions

Let the parameter space Γ be an open subset of $\mathbb{R}^d$. For each possible observation vector $\mathbf{z} \in \mathcal{Z}$, let $p_{\mathrm{Z}}(\mathbf{z}; \boldsymbol{\gamma})$ be a positive and continuously differentiable function of $\boldsymbol{\gamma}$ such that, for each $\mathbf{z} \in \mathcal{Z}$ and $\boldsymbol{\gamma} \in \Gamma$, the gradients of $\log p_{\mathrm{Z}}(\mathbf{z}; \boldsymbol{\gamma})$ span the space $\mathbb{R}^d$. If the model holds, that is, $p_{\mathrm{Z}}(\mathbf{z}) = p_{\mathrm{Z}}(\mathbf{z}; \boldsymbol{\gamma}_0)$ for some $\boldsymbol{\gamma}_0 \in \Gamma$, then assume that $p_{\mathrm{Z}}(\mathbf{z}; \boldsymbol{\gamma})$ converges to $p_{\mathrm{Z}}(\mathbf{z}; \boldsymbol{\gamma}_0)$ for all $\mathbf{z} \in \mathcal{Z}$ only if $\boldsymbol{\gamma}$ converges to $\boldsymbol{\gamma}_0$. More generally, assume that some $\boldsymbol{\gamma}_0 \in \Gamma$ exists such that $H(\boldsymbol{\gamma}) = -E(\ell(\boldsymbol{\gamma})) \geq H_0 = H(\boldsymbol{\gamma}_0)$ and $H(\boldsymbol{\gamma})$ only approaches H_0 if $\boldsymbol{\gamma}$ approaches $\boldsymbol{\gamma}_0$.

Suggested citation:

Haberman, S. J., Liu, Y., & Li, Y.-H. (2019). *Distractor analysis for multiple-choice tests: An empirical study with international language assessmernt data* (Research Report No. RR-19-39). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12275

Action Editor: Rebecca Zwick

Reviewers: Peter van Rijn and John Donoghue