Mutual Information Reliability for Latent Class Analysis

Applied Psychological Measurement
2018, Vol. 42(6) 460–477
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621617748324
journals.sagepub.com/home/apm

Yunxiao Chen¹, Yang Liu², and Shuangshuang Xu²

Abstract

Latent class models are powerful tools in psychological and educational measurement. These models classify individuals into subgroups based on a set of manifest variables, assisting decision making in a diagnostic system. In this article, based on information theory, the authors propose a mutual information reliability (MIR) coefficient that summarizes the measurement quality of latent class models, where the latent variables being measured are categorical. The proposed coefficient is analogous to a version of reliability coefficient for item response theory models and meets the general concept of measurement reliability in the Standards for Educational and Psychological Testing. The proposed coefficient can also be viewed as an extension of the McFadden's pseudo *R*-square coefficient, which evaluates the goodness-of-fit of logistic regression model, to latent class models. Thanks to several information-theoretic inequalities, the MIR coefficient is unitless, lies between 0 and 1, and receives good interpretation from a measurement point of view. The coefficient can be applied to both fixed and computerized adaptive testing designs. The performance of the MIR coefficient is demonstrated by simulated examples.

Keywords

reliability, pseudo *R*-square, latent class analysis, cognitive diagnosis model, mastery test, item response theory, computerized adaptive testing, mutual information, Shannon entropy

Introduction

Latent class models (LCMs; Goodman, 1974a, 1974b; Lazarsfeld & Henry, 1968) are widely used in social, behavioral, and health sciences for analyzing data obtained from tests and questionnaires, in which a large number of manifest variables are designed to measure latent attributes of the individuals. LCMs, as opposed to latent trait models that assume quantitative latent attributes, assume the latent attributes to be qualitative and aim at mapping individuals onto distinct subgroups, types, or categories. Readers are referred to Clogg (1995) and Collins and Lanza (2013) for an introduction of modern developments of LCMs.

In this article, the authors focus on latent class analysis in the context of psychological and educational testing, where the manifest variables are discrete responses to items. An LCM is regarded as a measurement model under a latent variable model framework, illustrated in

¹Emory University, Atlanta, GA, USA

²University of Maryland, College Park, USA

Corresponding Author:

Yang Liu, Department of Human Development and Quantitative Methodology, University of Maryland, 1230B Benjamin Building, 3942 Campus Drive, College Park, MD 20742, USA.

Email: yliu87@umd.edu

Figure 1. A graphical representation of latent-class type models.

Figure 1. That is, each individual is assumed to belong to a latent class, represented by a random variable α , and the responses to J items, namely X_1 to X_J , are only driven by the latent class membership α , up to certain measurement error quantified by the deviation of item response probabilities from 0 or 1. Thus, upon observing the responses, inference can be made on α . Given a test and an LCM that describe the underlying mechanism, an important question is how the quality of this measurement can be gauged, that is, the degree to which the latent classes are distinguishable.

Several summary measures of measurement accuracy have been proposed for LCMs. Templin and Bradshaw (2013) proposed a reliability-like measure for the test–retest consistency of a class of restricted LCMs. Cui, Gierl, and Chang (2012) and Wang, Song, Chen, Meng, and Ding (2015) proposed classification consistency and accuracy indices for restricted LCMs under test–retest and single-test administration scenarios, respectively. Another related measure is provided in the software package *Mplus* (Muthén & Muthén, 2010) named “entropy measure”, which describes the distinguishability among latent classes based on the measurement model (Muthén, 2010).

In this article, the authors propose an alternative measure of classification accuracy for LCMs, called the mutual information reliability (MIR) coefficient, with justifications from information theory. It is called a reliability coefficient due to an analogy among the proposed coefficient ‘a version of reliability coefficient for item response theory (IRT) models (Kim, 2012; Nicewander & Thomasson, 1999; Sireci, Thissen, & Wainer, 1991)’ and the general concept of measurement reliability in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, Chapter 2). According to the standards, the general concept of reliability refers to “the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported.” The proposed MIR coefficient can be viewed as an indicator of reliability/precision for latent class analysis wherein the basic measurement problem is to determine the class membership for each respondent.

The proposed coefficient can also be viewed as an extension of the McFadden’s (1974) pseudo R -square coefficient, one of the most popular pseudo R -square measures for logistic regression that evaluates model goodness-of-fit, to LCMs. Thanks to several information-theoretic inequalities, the proposed coefficient is unitless, lies between 0 and 1, and has an intuitive interpretation. The proposed measure is flexible, in the sense that it is applicable to both unrestricted and restricted LCMs, and under both fixed and computerized adaptive testing designs. As the MIR coefficient is proposed under a single test administration setting, the

authors do not compare it with the consistency indices in Templin and Bradshaw (2013), Cui et al. (2012), and Wang et al. (2015). However, the relationship between the MIR coefficient and the classification accuracy index in Cui et al. (2012) and Wang et al. (2015) is discussed. The proposed MIR coefficient improves upon the entropy measure in *Mplus*. The estimation of MIR from sample data is also discussed.

The rest of the article is organized as follows. First, the MIR coefficient is defined as a functional of the latent class measurement model and then its estimation is described. Then, simulation studies are conducted. The article ends with a discussion.

MIR

For ease of exposition, the authors start with a fixed test-length design with J items. However, they point out that the MIR coefficient can be easily generalized to LCM-based computerized adaptive testing designs.

Unrestricted LCMs

The LCM assumes that the responses are governed by the underlying characteristics of the individual, denoted by α , where α may represent, for example, a subtype of social phobia or post-traumatic stress disorder, or mastery/nonmastery of a certain skill. Specifically, α is assumed to be a categorical variable, taking values in $\mathcal{A}$, where each element $a \in \mathcal{A}$ represents the label of a latent class.¹ Let $\mathbf{p} = \{p_a : a \in \mathcal{A}\}$ be the probabilities of the latent classes in the population, satisfying $\sum_{a \in \mathcal{A}} p_a = 1$. That is, if an individual from the population is randomly sampled, the probability that this individual is from latent class α is p_a .

$$P(\alpha = a) = p_a, \quad a \in \mathcal{A}. \quad (1)$$

In addition, it is assumed that the individuals from the same latent class have the same item response distribution

$$P(\mathbf{X} = (x_1, \dots, x_J) | \alpha = a) \triangleq \psi_a(x_1, \dots, x_J).$$

For people from different classes, for example, $a \neq a'$, $\psi_a \neq \psi_{a'}$ in general. It is assumed that the item responses are independent given the latent class membership α , that is, local independence:

$$\psi_a(x_1, \dots, x_J) = \prod_{j=1}^J \psi_{j,a}(x_j), \quad (2)$$

where $\psi_{j,a}(x_j) \triangleq P(X_j = x_j | \alpha = a)$. Unrestricted LCMs have no constraint on $\psi_{j,a}$. Specifically, when x_j s are binary, that is, $x_j \in \{0, 1\}$, the unrestricted LCM assumes that X_j given α is a Bernoulli random variable

$$\psi_{j,a}(x_j) = c_{j,a}^{x_j} (1 - c_{j,a})^{1-x_j},$$

where $c_{j,a} = P(X_j = 1 | \alpha = a)$ is the success probability.

Restricted LCMs

LCMs with constraints are also widely used in psychological and educational measurement. Some examples are provided below.

Diagnostic models. Diagnostic models (Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010; von Davier, 2008) are a class of restricted LCMs for a fine-grained diagnosis of individuals' latent attributes. Specifically, each latent class is assumed to be a K -dimensional vector, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$. Each entry α_k represents ordinal categories of the k th latent attribute, taking values in $\{0, 1, \dots, s_k\}$. s_k is a prespecified highest level of the attribute k . In the context of educational testing, when $s_k = 1$, α_k is binary, indicating the mastery/nonmastery of skill k . In addition, the relationship between the items and the attributes is prespecified by the test designers through the so-called Q-matrix. In particular, $Q = (q_{jk})_{J \times K}$ is a $J \times K$ matrix with binary entries. For each j and k , $q_{jk} = 1$ means that the response to item j is associated with the presence of attribute k and $q_{jk} = 0$ otherwise. More precisely, given the attributes that are measured by the item, the response and the irrelevant attributes are conditionally independent, that is

$$P(X_j | \alpha) = P(X_j | \{\alpha_k : q_{jk} = 1\}).$$

Finally, the form of $\psi_{j,a}(x_j)$ is model specific. For example, when x_j and α_k are binary, a general-form diagnostic model is the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), assuming

$$\begin{aligned} & \text{logit}(\psi_{j,a}(1)) \\ &= \beta_{j,0} + \sum_{k=1}^K \beta_{j,k} q_{jk} a_k + \sum_{1 \leq k_1 < k_2 \leq K} \beta_{j,k_1 k_2} q_{jk_1} q_{jk_2} a_{k_1} a_{k_2} + \dots + \beta_{j,12\dots K} \prod_{k=1}^K q_{jk} a_k, \end{aligned} \quad (3)$$

where $\psi_{j,a}$ is defined in Equation 2. This model can be viewed as a special case of the general diagnostic model proposed in von Davier (2008); see von Davier (2014). This model is considered in the simulation study.

IRT-based mastery test. The statistical model of an IRT-based mastery test can be formulated as a restricted LCM. Consider a mastery test based on the two-parameter logistic (2PL) model (Birnbaum, 1968). Let θ be a continuous random variable, representing the latent trait level of an individual. Then, the 2PL model assumes,

$$P(X_j = 1 | \theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}},$$

where a_j and b_j are known as the discrimination and difficulty parameters, respectively. Moreover, θ is assumed to follow a marginal distribution F , typically a Gaussian distribution. A mastery test focuses on the classification $\theta \leq \tau$ and $\theta > \tau$, where τ is a prespecified mastery threshold. If $\theta \leq \tau$ ($\theta > \tau$), the individual is classified as nonmastery (mastery). This task can be formulated into an LCM. Let $\alpha = I_{\{\theta > \tau\}}$ be the latent class variable, indicating the mastery of the skill. Then,

$$P(\alpha = 1) = P(\theta > \tau) = 1 - F(\tau).$$

$$\psi_0(x_1, \dots, x_J) = \int_{-\infty}^{\tau} \prod_{j=1}^J \frac{e^{a_j(\theta - b_j)x_j}}{1 + e^{a_j(\theta - b_j)}} dF(\theta),$$

and

$$\psi_1(x_1, \dots, x_J) = \int_{\tau} \prod_{j=1}^J \frac{e^{a_j(\theta-b_j)x_j}}{1 + e^{a_j(\theta-b_j)}} dF(\theta).$$

Thus, the classification in a mastery test boils down to making inference on the ordinal latent class α . The formulation can be extended to classifying θ into $P+1$ ordinal categories according to thresholds $\tau_1 < \dots < \tau_P$. A simulation study based on this model is conducted in the sequel.

Information Theory

Three questions are considered under an LCM. At this moment, all the model parameters are assumed to be known.

- A. If an individual is randomly sampled from the population, what is the uncertainty in his or her latent class membership α ?
- B. To make better inference about α , the individual is asked to take the test. After the test, what is the uncertainty in his or her latent class membership α ?
- C. How much uncertainty is reduced by administrating the test?

To answer the above questions, the authors introduce information measures from information theory, a modern discipline studying the quantification, storage, and communication of information (e.g., Cover & Thomas, 2012). In particular, the Shannon (1948) entropy quantifies the amount of uncertainty in a random variable. The Shannon entropy of α satisfying Equation 1 is defined as

$$\text{SHE}(\alpha) = - \sum_{a \in \mathcal{A}} p_a \log p_a. \quad (4)$$

It has the following properties:

1. When $p_a = 1$ and $p_{a'} = 0$ for $a' \neq a$, $\text{SHE}(\alpha) = 0$, meaning that there is no uncertainty in α .
2. $\text{SHE}(\alpha)$ as a function of the class membership probabilities $\mathbf{p}$ is maximized when $p_a = 1/|\mathcal{A}|$ for all $a \in \mathcal{A}$, where $|\mathcal{A}|$ is the total number of latent classes. Figure 2 illustrates the relationship between $\mathbf{p}$ and $\text{SHE}(\alpha)$ when $\mathcal{A} = \{0, 1\}$ and thus $|\mathcal{A}| = 2$ and $\mathbf{p} = (p_0, p_1)$. The x -axis represents p_1 and the y -axis represents $\text{SHE}(\alpha)$. Specifically, according to this graph, the maxima of $\text{SHE}(\alpha)$ is achieved when $p_0 = p_1 = 0.5$.

Generally, the more concentrated or unbalanced the distribution of α is ‘the smaller the entropy value becomes’; thus, it is a reasonable measure of uncertainty. For example, let the latent classes $\mathcal{A} = \{0, 1, 2, 3, 4\}$. Then,

$$\text{SHE}(\alpha) = \begin{cases} 0, & \text{when } p_0 = 1 \text{ and } p_1 = p_2 = p_3 = p_4 = 0; \\ 0.69, & \text{when } p_0 = p_1 = 0.5 \text{ and } p_2 = p_3 = p_4 = 0; \\ 1.10, & \text{when } p_0 = p_1 = p_2 = 1/3 \text{ and } p_3 = p_4 = 0; \\ 1.39, & \text{when } p_0 = p_1 = p_2 = p_3 = 1/4 \text{ and } p_4 = 0; \\ 1.61, & \text{when } p_0 = p_1 = p_2 = p_3 = p_4 = 1/5. \end{cases}$$

Figure 2. A graphical illustration of SHE when $\mathcal{A} = \{0, 1\}$, showing the SHE as a function of the probability of Latent Class I (p_1).

Note. SHE = Shannon entropy.

Thus, $\text{SHE}(\alpha)$ provides a reasonable answer to the Question A.

After taking the test, the uncertainty of α is characterized by the conditional distribution of α given $\mathbf{X}$, that is, the posterior distribution. Thus, for a particular response pattern $\mathbf{x} = (x_1, \dots, x_J)$, the Shannon entropy of α given $\mathbf{X} = \mathbf{x}$ is

$$\text{SHE}(\alpha|\mathbf{X}=\mathbf{x}) = - \sum_{a \in \mathcal{A}} P(\alpha = a|\mathbf{X}=\mathbf{x}) \log P(\alpha = a|\mathbf{X}=\mathbf{x}),$$

where the posterior distribution is calculated as

$$P(\alpha = a|\mathbf{X}=\mathbf{x}) = \frac{\psi_a(x_1, \dots, x_J) p_a}{\sum_{a' \in \mathcal{A}} \psi_{a'}(x_1, \dots, x_J) p_{a'}}. \quad (5)$$

Taking the randomness of $\mathbf{X}$ into account, the uncertainty in α can be described by the conditional entropy, defined as

$$\text{SHE}(\alpha|\mathbf{X}) = \sum_{\mathbf{x}} \text{SHE}(\alpha|\mathbf{X}=\mathbf{x}) p(\mathbf{X}=\mathbf{x}),$$

which quantifies the amount of information in α unexplained by $\mathbf{X}$ (Cover & Thomas, 2012). Here $p(\mathbf{X}=\mathbf{x})$ is the marginal distribution of $\mathbf{X}$,

$$p(\mathbf{X}=\mathbf{x}) = \sum_{a \in \mathcal{A}} \psi_a(x_1, \dots, x_J) p_a.$$

The conditional entropy $\text{SHE}(\alpha|\mathbf{X})$ provides a reasonable answer to Question B.

Moreover, the difference between the entropy of α and the conditional entropy $\text{SHE}(\alpha|\mathbf{X})$ is always nonnegative, and is known as the mutual information between α and $\mathbf{X}$. More precisely, the mutual information is defined as

Figure 3. A Venn diagram representation of the information measures, including the entropy of latent class membership $SHE(\alpha)$, the entropy of responses $SHE(X)$, the conditional entropy of latent class membership given responses $SHE(\alpha|X)$, and the mutual information $MI(\alpha, X)$; Specifically, $SHE(\alpha) = SHE(\alpha|X) + MI(\alpha, X)$.

Note. SHE = Shannon entropy.

$$MI(\alpha, X) = SHE(\alpha) - SHE(\alpha|X).$$

The mutual information quantifies the amount of information that X contains about α . It is non-negative and is strictly positive when X and α are not independent. It means that by taking a test, the uncertainty in α is reduced, as long as the measurement is informative (i.e., X and α are not independent). Therefore, the mutual information $MI(\alpha, X)$ can be used to answer Question C as a summary measure.

Figure 3 provides a Venn diagram representation of the information measures above. Specifically, the entire circle on the left represents the value of $SHE(\alpha)$, summarizing the total amount of uncertainty in α when the individual is sampled from the population. The right circle represents the entropy of the responses X defined similarly as in Equation 4, which is the total amount of information in X . The overlapping part of the two circles is the mutual information $MI(\alpha, X)$, the amount of shared information between α and X . The nonoverlapping part of the left circle is the conditional entropy $SHE(\alpha|X)$, quantifying the uncertainty in α unexplained by X .

MIR

The MIR coefficient for LCM is defined as

$$MIR(\alpha; X) = 1 - \frac{SHE(\alpha|X)}{SHE(\alpha)} = \frac{MI(\alpha, X)}{SHE(\alpha)}. \quad (6)$$

This MIR coefficient lies between 0 and 1. It can be interpreted as the proportion of uncertainty in the latent class α reduced by taking the test. It provides a measure of the classification quality.

The above MIR coefficient is defined at the latent class level. For diagnostic models, the attribute-level reliability can be defined as

$$MIR(\alpha_k; X) = 1 - \frac{SHE(\alpha_k|X)}{SHE(\alpha_k)} = \frac{MI(\alpha_k, X)}{SHE(\alpha_k)}, \quad (7)$$

where α_k is the k th attribute.

In what follows, some theoretical properties of the MIR coefficients in Equations 6 and 7 are discussed.

1. When $\mathbf{X}$ and α are independent, that is, the items are irrelevant to the latent classes being measured, $P(\alpha = a | \mathbf{X} = \mathbf{x}) = P(\alpha = a)$ for all a and $\mathbf{x}$, so $\text{SHE}(\alpha | \mathbf{X}) = \text{SHE}(\alpha)$ and hence $\text{MIR}(\alpha, \mathbf{X}) = 0$.
2. Consider two sets of responses, $\mathbf{X} = (X_1, \dots, X_J)$ and $\tilde{\mathbf{X}} = (X_1, \dots, X_J, X_{J+1})$. Then

$$\text{MIR}(\alpha; \mathbf{X}) \leq \text{MIR}(\alpha; \tilde{\mathbf{X}}). \quad (8)$$

The above inequality becomes equality when X_{J+1} and α are conditionally independent given the previous responses $\mathbf{X}$, that is, the additional item does not provide additional information about the latent classes. This inequality implies that the quality of the measurement always increases when an informative item is added. Specifically, if $X_{J+1} = X_j$ for some $j \leq J$, which might happen when two identical items are administered, then X_{J+1} and α are conditionally independent given $\mathbf{X}$ and the quality of measurement does not improve. Generally, when severe local dependence (LD) exists between the new item and the previous ones, quality of measurement improves little under the proposed framework.

A similar inequality applies to the MIR at the attribute level for diagnostic models, that is,

$$\text{MIR}(\alpha_k; \mathbf{X}) \leq \text{MIR}(\alpha_k; \tilde{\mathbf{X}}).$$

These inequalities imply that the more repeated measurement, the better the quality of measurement, which, however, means higher time cost to the respondents.

3. Another common practice in diagnostic modeling is to dichotomize the item responses when they are polytomous, to analyze them using existing diagnostic models for binary responses. A typical argument in the development of polytomous diagnostic models is the information loss when dichotomizing the items (e.g., Ma & de la Torre, 2016). Our MIR provides a theoretical justification to that. Let $\mathbf{X} = (X_1, \dots, X_J)$ be the responses to the original polytomous items and $\mathbf{X}^* = (X_1^*, \dots, X_J^*)$ be the dichotomized responses. The distribution of $\mathbf{X}^*$ is induced from that of $\mathbf{X}$. Then the following inequalities are true:

$$\begin{aligned} \text{MIR}(\alpha; \mathbf{X}^*) &\leq \text{MIR}(\alpha; \mathbf{X}), \\ \text{MIR}(\alpha_k; \mathbf{X}^*) &\leq \text{MIR}(\alpha_k; \mathbf{X}). \end{aligned} \quad (9)$$

That is, dichotomizing the items reduces the measurement quality of the test.

The proof of the inequalities in Equations 8 and 9 relies on the data processing inequality of mutual information; see Chapter 2, Cover and Thomas (2012) for further details.

Further Justifications

Further justifications are provided for the proposed coefficient as a reliability measure.

MIR coefficient and McFadden's pseudo R-square. The MIR coefficient can be understood from a regression point of view. Specifically, the proposed MIR coefficient is linked to McFadden's

pseudo R -square measure (McFadden, 1974) for logistic regression. This pseudo R -square measures the amount of information in binary response Y in logistic regression captured by explanatory variables $\mathbf{Z} = (Z_1, \dots, Z_p)$.

Consider a logistic regression model M that specifies the conditional distribution of Y given $\mathbf{Z}$. It is further assumed that explanatory variables $\mathbf{Z}$ are also random with a finite state space. Furthermore, let M_0 be the null model without any explanatory variable, that is, the unconditional Bernoulli model. The McFadden's pseudo R -square measure for model M is defined as

$$R^2 = 1 - \frac{l_n(M)}{l_n(M_0)},$$

where $l_n(M)$ and $l_n(M_0)$ are the maximized log-likelihood functions under models M and M_0 , respectively, and n is the sample size. It is not difficult to show that $-l_n(M)/n$ converges to the conditional entropy $\text{SHE}(Y|\mathbf{Z})$ and $-l_n(M_0)/n$ converges to the Shannon entropy $\text{SHE}(Y)$, when the sample size n grows to infinity. Consequently, with a large sample size n , $R^2 \approx 1 - \text{SHE}(Y|\mathbf{Z})/\text{SHE}(Y)$. This asymptotic result builds a connection between McFadden's pseudo R -square and the proposed MIR coefficient. The difference is that, instead of having binary observable response Y , we have latent variable α that is unobservable and can take more than two values. The responses $\mathbf{X}$ in the LCM play the same role as the explanatory variables $\mathbf{Z}$ in the logistic regression model.

MIR coefficient and IRT marginal reliability. The concept of reliability for IRT models is well established, whereas it is less well defined for LCMs. In what follows, a connection is established between the MIR coefficient and a version of IRT reliability.

There are two different ways of defining a reliability coefficient for IRT models, as (a) the correlation between ability estimates on two parallel forms of a test and (b) the squared correlation between the true abilities and estimates (Kim, 2012). The authors focus on the latter and use expected a posteriori (EAP) for ability estimation, for which, the IRT reliability is defined as

$$R_{\text{IRT}} = 1 - \frac{E(\text{Var}(\theta|\mathbf{X}))}{\text{Var}(\theta)}. \quad (10)$$

where θ is the trait level of an individual measured in this IRT model. For more justifications and theoretical properties of Equation 10, see Kim (2012) and the references therein. In what follows, Equation 10 is interpreted from an information theoretic point of view and then similarities between this IRT reliability and the proposed MIR coefficient are discussed.

1. If an individual is randomly sampled from the population, the uncertainty in his or her latent trait level θ is captured by the marginal distribution of θ . As latent trait is a continuous random variable, the variance of θ provides a reasonable summary of its uncertainty (Zidek & van Eeden, 2003).
2. After taking a test and observing $\mathbf{X} = \mathbf{x}$, the uncertainty of θ is characterized by the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$, for which $\text{Var}(\theta|\mathbf{X} = \mathbf{x})$ provides a summary measure. Taking all possible response patterns into account, this uncertainty is summarized by

$$E(\text{Var}(\theta|\mathbf{X})) = \sum_{\mathbf{x}} \text{Var}(\theta|\mathbf{X} = \mathbf{x})P(\mathbf{X} = \mathbf{x}),$$

where $p(\mathbf{X} = \mathbf{x})$ is the marginal distribution of $\mathbf{X}$ under the IRT model.

Figure 4. A Venn diagram representation of the uncertainty measures under an IRT model, including the variance of latent trait $\text{Var}(\theta)$, the expected conditional variance $E(\text{Var}(\theta|\mathbf{X}))$, and the variance of EAP estimate $\text{Var}(E(\theta|\mathbf{X}))$ or equivalently, the covariance between the latent trait θ and the EAP estimate $\text{Cov}(\theta, E(\theta|\mathbf{X}))$; specifically, $\text{Var}(\theta) = E(\text{Var}(\theta|\mathbf{X})) + \text{Var}(E(\theta|\mathbf{X}))$.

Note. IRT = item response theory; EAP = expected a posteriori.

3. Consequently, $\text{Var}(\theta) - E(\text{Var}(\theta|\mathbf{X} = \mathbf{x}))$ measures the uncertainty of θ reduced by the test, and R_{IRT} can be interpreted as the proportion of uncertainty reduction after taking the test. It follows from the law of total variance that

$$\text{Var}(\theta) - E(\text{Var}(\theta|\mathbf{X})) = \text{Var}(E(\theta|\mathbf{X})),$$

which is guaranteed to be nonnegative. Moreover, $\text{Var}(E(\theta|\mathbf{X})) = \text{Cov}(\theta, E(\theta|\mathbf{X}))$, which can be further interpreted as the amount of information that θ and $E(\theta|\mathbf{X})$ share.

Following the discussion above, a Venn diagram for the IRT reliability can be drawn (Figure 4), similar to the one for MIR (Figure 3): The relationship among $\text{Var}(\theta)$, $E(\text{Var}(\theta|\mathbf{X}))$, and $\text{Var}(E(\theta|\mathbf{X}))$ in the IRT reliability closely resembles that among the Shannon entropy $\text{SHE}(\alpha)$, conditional entropy $\text{SHE}(\alpha|\mathbf{X})$, and mutual information $\text{MI}(\alpha, \mathbf{X})$ in MIR, respectively.

MIR coefficient and classification error (CE)/accuracy. Another measure of the classification quality is the CE or, equivalently, the classification accuracy. More precisely, the CE is defined as

$$P(\hat{\alpha}(\mathbf{X}) \neq \alpha),$$

where $\hat{\alpha}(\mathbf{X})$ is the Bayesian classifier, defined as

$$\hat{\alpha}(\mathbf{X}) = \underset{a \in \mathcal{A}}{\text{argmax}} P(\alpha = a|\mathbf{X}).$$

The classification accuracy is defined as $P(\hat{\alpha}(\mathbf{X}) = \alpha) = 1 - P(\hat{\alpha}(\mathbf{X}) \neq \alpha)$, which is first proposed in Cui et al. (2012) as a measure of reliability for diagnostic models. In Cui et al. (2012), however, $\hat{\alpha}(\mathbf{X})$ can be other model-based classifier, not restricted to the Bayesian classifier above.

The MIR coefficient and CE are closely related, in the sense that when items are added, the MIR coefficient increases and the CE decreases. In what follows, an illustrative example is

Figure 5. A latent class model example showing the relationship between the test length J and the CE, and that between the test length J and the MIR coefficient.
 Note. CE = classification error; MIR = mutual information reliability.

provided under an unrestricted LCM, with $\mathcal{A} = \{0, 1\}$, $P(\alpha = 0) = 1/3$, and $P(\alpha = 1) = 2/3$, and all the items are dichotomous with the same item parameters, that is, $P(X_j = 1 | \alpha = 1) = 0.8$ and $P(X_j = 1 | \alpha = 0) = 0.2$. Figure 5 presents the relationship between the test length J and the CE, and that between the test length J and the MIR coefficient.

A connection to the entropy measure in Mplus. An entropy index is defined on page 34 of B. Muthén (2010) for unrestricted LCMs, which is related to the proposed MIR coefficient. It is estimating the following entropy quantity:

$$E(\alpha; \mathbf{X}) = 1 - \frac{\text{SHE}(\alpha | \mathbf{X})}{-\log\left(\frac{1}{|\mathcal{A}|}\right)}. \quad (11)$$

The difference between this entropy measure and the proposed MIR coefficient is that the denominator $\text{SHE}(\alpha)$ in the second term of Equation 6 is replaced by $-\log(1/|\mathcal{A}|)$ in Equation 11. A few remarks are provided.

1. As $-\log(1/|\mathcal{A}|)$ is the entropy of a uniform distribution over the set $\mathcal{A}$, the entropy measure $E(\alpha; \mathbf{X})$ and MIR coefficient $\text{MIR}(\alpha; \mathbf{X})$ coincide when α is uniformly distributed on $\mathcal{A}$.
2. The authors have discussed previously that a uniform distribution has the maximum entropy among all discrete distributions supported on $\mathcal{A}$. In particular,

$$-\log(1/|\mathcal{A}|) \geq \text{SHE}(\alpha).$$

Thus, $E(\alpha; \mathbf{X})$ is always an upper bound of $\text{MIR}(\alpha; \mathbf{X})$.

3. Comparing with the MIR coefficient, the entropy measure does not take the difficulty of the measurement problem (quantified by $\text{SHE}(\alpha)$) into account. As a result, it may not

be a good measure when comparing different measurements. In addition, the proportion of uncertainty reduction interpretation is no longer valid for $E(\alpha; \mathbf{X})$.

Estimation

When Model Parameters Are Known

When the model parameters are known, they are plugged into Equations 6 and 7 to compute the latent class level and attribute-level MIR coefficients. However, a technical difficulty is that the evaluation of the conditional entropy $\text{SHE}(\alpha|\mathbf{X})$ or $\text{SHE}(\alpha_k|\mathbf{X})$ involves summing over all possible response patterns. There are, in total, 2^J response patterns, and thus the summation is computationally infeasible for a large J . When J is large, the authors approximate $\text{SHE}(\alpha|\mathbf{X})$ by Monte Carlo simulations. The procedure is described as follows.

1. Simulate M independent samples from the known unrestricted or restricted LCM, denoted by $\mathbf{x}_1, \dots, \mathbf{x}_M$.
2. For each individual i , compute the posterior distribution given $\mathbf{X}_i = \mathbf{x}_i$, $P(\alpha_i = a|\mathbf{X}_i = \mathbf{x}_i)$, $a \in \mathcal{A}$ using Equation 5.
3. Then, $\text{SHE}(\alpha|\mathbf{X})$ is approximated by

$$\widehat{\text{SHE}}(\alpha|\mathbf{X}) = \frac{1}{M} \sum_{i=1}^M \text{SHE}(\alpha_i|\mathbf{X}_i = \mathbf{x}_i),$$

and the Monte Carlo standard error of $\widehat{\text{SHE}}(\alpha|\mathbf{X})$ is

$$SE\left(\widehat{\text{SHE}}(\alpha|\mathbf{X})\right) = \sqrt{\frac{\sum_{i=1}^M \left(\text{SHE}(\alpha_i|\mathbf{X}_i = \mathbf{x}_i) - \widehat{\text{SHE}}(\alpha|\mathbf{X})\right)^2}{M(M-1)}}.$$

The authors then approximate $\text{MIR}(\alpha, \mathbf{X})$ by $\widehat{\text{MIR}}(\alpha, \mathbf{X}) = 1 - \widehat{\text{SHE}}(\alpha|\mathbf{X})/\text{SHE}(\alpha)$, which has a standard error $SE(\widehat{\text{SHE}}(\alpha|\mathbf{X}))$. $\text{MIR}(\alpha_k, \mathbf{X})$ can also be approximated in the same manner.

When Model Parameters Are Unknown

When the model parameters are unknown, $\text{MIR}(\alpha, \mathbf{X})$ or $\text{MIR}(\alpha_k, \mathbf{X})$ is estimated based on item response data. In what follows, the estimation of $\text{MIR}(\alpha, \mathbf{X})$ is described and that of $\text{MIR}(\alpha_k, \mathbf{X})$ is similar.

1. Estimate the model using an expectation maximization (EM) algorithm and denote the estimate by $\hat{\psi}_{j,a}(\cdot)$, $j = 1, \dots, J$, $a \in \mathcal{A}$ and $\hat{\mathbf{p}}$. See Collins and Lanza (2013) and Rupp et al. (2010) for details of the corresponding EM algorithms.
2. The estimated model is then plugged into the evaluation of the posterior distribution of α_i s. That is,

$$\hat{P}(\alpha_i = a|\mathbf{X}_i = \mathbf{x}_i) = \frac{\hat{\psi}_a(x_{i1}, \dots, x_{iJ})\hat{p}_a}{\sum_{a' \in \mathcal{A}} \hat{\psi}_{a'}(x_{i1}, \dots, x_{iJ})\hat{p}_{a'}}.$$

3. MIR ($\alpha, \mathbf{X}$) is estimated by

$$\widehat{\text{MIR}}(\alpha, \mathbf{X}) = 1 - \frac{\frac{1}{N} \sum_{i=1}^N \widehat{\text{SHE}}(\alpha_i | \mathbf{X}_i = \mathbf{x}_i)}{-\sum_{a \in \mathcal{A}} \hat{p}_a \log \hat{p}_a},$$

where

$$\widehat{\text{SHE}}(\alpha_i | \mathbf{X}_i = \mathbf{x}_i) = -\sum_{a \in \mathcal{A}} \hat{P}(\alpha_i = a | \mathbf{X}_i = \mathbf{x}_i) \log(\hat{P}(\alpha_i = a | \mathbf{X}_i = \mathbf{x}_i)).$$

Furthermore, the standard deviation for MIR($\alpha, \mathbf{X}$) can be computed using a nonparametric bootstrap method (Efron & Tibshirani, 1993). It proceeds as follows. Let B be the total number of bootstrap data sets to be resampled.

1. For each $b \in \{1, \dots, B\}$, sample with replacement from the individual set $\{1, 2, \dots, N\}$ and obtain indices $I_1^b, \dots, I_N^b$. The b th bootstrap data set is obtained by $\mathbf{x}_i^{*b} = \mathbf{x}_{I_i^b}$, $i = 1, 2, \dots, N$.
2. For the b th bootstrap data set, fit the LCM and obtain a bootstrap estimate of the MIR $\widehat{\text{MIR}}_b(\alpha, \mathbf{X})$.
3. The standard deviation of $\widehat{\text{MIR}}(\alpha, \mathbf{X})$ is then estimated by the bootstrap standard error

$$SE_B(\widehat{\text{MIR}}(\alpha, \mathbf{X})) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left(\widehat{\text{MIR}}_b(\alpha, \mathbf{X}) - \frac{1}{B} \sum_{r=1}^B \widehat{\text{MIR}}_r(\alpha, \mathbf{X}) \right)^2}$$

Simulation Studies

Study 1

This study demonstrates that the MIR coefficient calculated from a misspecified model could be inflated and does not reflect the actual classification quality. The results suggest that the model-data fit of the calibrated LCM should be established prior to the calculation of any measures of classification quality.

$N = 100,000$ individual response vectors to 60 items were simulated from a two-attribute ($K = 2$) LCDM with LD between every pair of adjacent items. LD is introduced to the LCDM (Equation 3) for item j ($j = 2, \dots, 60$) by adding the response to the previous item x_{j-1} as an additional covariate, that is,

$$\text{logit}(c_{j,a}) = \beta_{j,0} + \beta_{j,1}q_{j1}a_1 + \beta_{j,2}q_{j2}a_2 + \beta_{j,12}q_{j1}q_{j2}a_1a_2 + \gamma x_{j-1},$$

in which $c_{j,a} = P(X_j = 1 | \alpha = a, X_{j-1} = x_{j-1})$ denotes the conditional success probability for item j and γ is the additional coefficient determining the strength of LD. The first item follows a regular LCDM. The authors note that the LD-generating mechanism has been utilized by Glas and Suárez Falcón (2003) and Liu and Thissen (2014) in the context of IRT modeling. In the current simulation study, the γ parameter was fixed at 1.5 for all $j = 2, \dots, 60$. The remaining item parameters were randomly generated: in particular, intercepts $\beta_{j,0} \sim \text{Uniform}(-1.5, 0.5)$, main effects $\beta_{j,k} \sim \text{Uniform}(0.5, 1.5)$, and interactions $\beta_{j,k_1k_2} \sim \text{Uniform}(0.5, 1.5)$. The population

Figure 6. MIR (top panel) and CE (bottom panel) for overall classification under a misspecified LCDM model omitting the local dependence effect.

Note. In the top panel, the MIRs calculated from the misspecified model (dotted line) and true model (solid line) were plotted against the test length $J = 10, 15, \dots, 60$. In the bottom panel, the model-implied CE measures corresponding to misspecified and true models were displayed in dotted and solid lines, respectively. The actual CE for the true attribute profile under the misspecified model was shown as the long-dashed line. MIR = mutual information reliability; CE = classification error; LCDM = log-linear cognitive diagnostic model.

proportions of the $2^K = 4$ attribute patterns (0, 0), (1, 0), (0, 1), and (1, 1) were set to 0.40, 0.26, 0.09, and 0.25, respectively.

The overall MIR and CE measures were calculated for the first J simulated items: the test length $J = 10, 15, \dots, 60$. Under each test-length condition, a two-attribute LCDM without LD, which is thus misspecified, was fitted to the simulated responses using the R package CDM (Robitzsch, Kiefer, George, & Uenlue, 2016). Then, the parameter estimates were used to calculate the MIR and CE as though the model were correctly specified. The actual CE of the Bayes classifier under the misspecified model can be evaluated using the simulated respondents' true attribute profiles. As the LD effect was omitted from the fitted model, it is expected that the actual CE is higher than the model-implied counterpart. As a benchmark for comparison, the MIR and CE from the true data generating model (with LD) were also calculated.

By comparing the solid and dotted lines in Figure 6, it is observed that the model misspecification triggers an inflation of MIR and a deflation of model-implied CE. When $J = 60$, for example, the MIR and CE implied by the fitted LCDM are .93 and 0.03, respectively; however, the true model-implied measures are only .78 and 0.11. The finding is consistent with the intuition that the amount of reduced uncertainty contributed by each item was overestimated due to the omission of LD. Moreover, classification of individuals based on a misspecified measurement model is problematic: Across all test-length conditions, the actual CE under the misspecified LCDM is uniformly higher than the true model-implied value and is more than twice as high as the value obtained by assuming the fitted model is correct. For instance, under the misspecified model, a 50-item test might be believed to yield very accurate classification because

Figure 7. The marginal distribution of the latent trait and the mastery threshold for the mastery tests in simulation Study 1.

Note. The nonmastery and mastery groups are defined as individuals whose latent trait level is below and above τ , respectively.

the model-implied MIR is above .9 and the corresponding CE is less than 0.05. Nevertheless, the actual CE based on the misspecified classifier is in fact 0.14, which is much less desirable.

Study 2

Two mastery tests are then considered based on the 2PL model, each of which has 15 items and 10,000 examinees. The classification threshold was set to $\tau = 1.5$ for both tests, and the latent trait of the population was assumed to follow a standard normal distribution, as illustrated in Figure 7. The item parameters were chosen so that the test information for Test A spreads out for a sufficient range of latent trait levels and the test information for Test B concentrates around τ , where test information is a function of θ that characterizes the degree of estimation precision at different θ values (Embretson & Reise, 2000). The test information functions for both tests are shown in Figure 8.

In this analysis, the model parameters were assumed to be known. For both tests, the regular IRT reliability that is designed to reflect the overall estimation accuracy, the MIR coefficient that considers the classification nature of mastery tests, the mean square error (MSE) of the posterior mean (EAP) estimates of examinees' latent trait level, and the CE were calculated. The results are shown in Table 1. In particular, it can be seen that Test A has a higher regular IRT reliability but a lower MIR than Test B, which is consistent with the results of MSE of the EAP estimates and the CE. This is because, for mastery tests, the bottleneck is to correctly classify the examinees whose latent trait levels are close to the classification threshold τ . Examinees whose latent trait levels are far away from the threshold can be correctly classified without much difficulty. Therefore, a test with high regular IRT reliability may not perform well, due to wasting a certain number of items to measure high or low ability examinees accurately. Finally, it is observed that even with a 5.0% CE, the MIR coefficient is not high (MIR = 0.54). This is because a 5.0% CE rate is relatively large, considering the size of the mastery group (6.7% of the population).

Figure 8. The test information functions for Test A and Test B in simulation Study I.

Table 1. The Results of Simulation Study I, Including the Regular IRT Reliability, the MIR Coefficient for Mastery Test, the MSE of the EAP Estimates of Examinees' Latent Trait Level, and the CE.

	R_{IRT}	MIR	MSE	CE (%)
Test A	.84	.54	.16	5.0
Test B	.65	.80	.36	2.1

Note. R_{IRT} = regular IRT reliability; IRT = item response theory; MIR = mutual information reliability; MSE = mean square error; EAP = expected a posteriori; CE = classification error.

Discussion

This article proposes a MIR coefficient, a measure of classification accuracy for unrestricted and restricted LCMs. This measure is analogous to a version of reliability measure for IRT models and accommodates the discreteness of the latent variables by borrowing ideas from information theory. The proposed measure can also be viewed as an extension of McFadden's pseudo R -square for logistic regression to latent variable model. Thanks to information theory, the proposed MIR coefficient has desirable theoretical properties. Specifically, it takes value between 0 and 1, increases under repeated measurements, and approaches 1 as the number of repeated measurements grows. For long tests, Monte Carlo simulations must be invoked to approximate the expectation over all possible response patterns in the calculation of the conditional entropy. When the measurement model parameters are unknown, the MIR coefficient can be estimated by plugging in the point estimates, and the corresponding standard errors can be obtained via bootstrapping.

Findings from the simulation studies are summarized and discussed as follows. The first simulation study shows that under model misspecification, the MIR coefficient estimate can be greatly inflated, providing overly optimistic results. It implies that a coefficient of measurement quality is contingent upon the measurement model: Applying the reliability measure derived from one model to data generated from a different model can be grossly misleading. Therefore, the model-data fit should be established prior to the calculation of any measure of classification quality. While the comparative fit of LCMs may be assessed by various information criteria, the absolute fit can be tested via the recently developed limited-information goodness-of-fit

statistics (Hansen, Cai, Monroe, & Li, 2016). The second simulation study shows that the regular IRT reliability is not an ideal benchmark for the assembly of a mastery test, because it favors a combination of items with high measurement precision throughout the latent trait scale while what is needed for good classification is only high measurement precision around the threshold value. This result implies that the measure of measurement quality should be selected to match the goal of the measurement (e.g., estimation/classification).

In future research, the proposed measure will be extended to general latent variable models to provide a unified framework for measuring the model-based measurement quality, retaining the information-theoretic properties of the proposed measure. Based on the proposed framework, it is of interest to investigate the effects of all kinds of model misspecification on the measurement quality of LCMs, especially under the Q-matrix misspecification in diagnostic models and the existence of LD, which could provide quantification of the potential overestimate of test qualities under the misspecification of measurement model. Finally, applications of the proposed method in test assembly and CAT design will also be discussed.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. To clarify, Greek letter " α " is used to denote a random variable and letter " a " to denote elements in $\mathcal{A}$.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-422). Reading, MA: Addison-Wesley.
- Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311-359). New York, NY: Springer.
- Collins, L. M., & Lanza, S. T. (2013). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, NJ: John Wiley.
- Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Hoboken, NJ: John Wiley.
- Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, 49, 19-38.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: CRC Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Glas, C. A., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27, 87-106.
- Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A modified latent structure approach. *American Journal of Sociology*, 79, 1179-1259.

- Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.
- Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, 69, 225-252.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, 77, 153-162.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.
- Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology*, 67, 496-513.
- Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69, 253-275.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105-142). New York, NY: Academic Press.
- Muthén, B. (2010). *Technical appendices*. Available from <http://www.statmodel.com>
- Muthén, L., & Muthén, B. (2010). *Mplus users guide* (Version 5.21) [Computer software and manual]. Los Angeles, CA: Author.
- Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied psychological measurement*, 23, 239-247.
- Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2016). *CDM: Cognitive diagnosis modeling* (R package version 4.8-0). Retrieved from <https://CRAN.R-project.org/package=CDM>
- Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6, 219-262.
- Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, 30, 251-275.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287-307.
- von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*, 2014(2), 1-13.
- Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, 52, 457-476.
- Zidek, J., & van Eeden, C. (2003). Uncertainty, entropy, variance and the effect of partial information. *Lecture Notes-Monograph Series*, 42, 155-167.