Classification Reliability of Psychological and Educational Tests: Methods for Evaluating Classification Consistency

Authors: Chen Jingyi, Song Lihong, Wang Wenyi, Song Lihong

Date: 2025-04-11T19:36:27+00:00

Abstract

Psychological, educational, and medical tests are widely used for classifying examinees; however, reliability coefficients such as internal consistency and α cannot directly evaluate classification reliability. How to assess the classification reliability of criterion-referenced tests has become an important issue of concern for researchers and practitioners. This study, from the perspective of classification consistency methods, explores classification consistency estimation models for single-administration tests, analyzes the developmental trajectories and core ideas of various representative methods, and examines real data from personality tests, achievement tests, diagnostic tests, etc., in conjunction with relevant software packages and programs for each method. Combining theoretical analysis and data analysis, this research summarizes the advantages, disadvantages, and influencing factors of various methods, proposes recommendations for selecting various methods, discusses issues such as interval estimation of classification consistency, and promotes research, application, and reporting of classification consistency for classification tests.

Full Text

the test dataset are used to simulate two parallel test response patterns for each individual in the test set. The trained MLM then predicts categories for these two simulated response patterns, and classification consistency is calculated. This method can also be combined with bootstrap sampling to obtain interval estimates.

3. Real Data Analysis

To demonstrate the application of various classification consistency estimation methods, this study analyzes three types of real data: personality test data,

academic achievement test data, and cognitive diagnostic test data.

3.1 Personality Test Data

The data come from the Mindful Attention Awareness Scale (MAAS; Chen et al., 2012), which consists of 15 items using a 6-point Likert scale (1-6). The sample size is $N = 836$. The total score ranges from 15 to 90. Following the standard practice, a cut score of 60 is used to classify individuals into "high mindfulness" and "low mindfulness" groups.

We applied the LL method, Lee method (CTT), and W method to estimate classification consistency. For the LL method, the effective test length was calculated first. For the W method, the standard error of measurement was estimated as $\sigma_e = \sigma_x\sqrt{1-\alpha}$, where α is the Cronbach's alpha coefficient. The results are shown in Table 1.

3.2 Academic Achievement Test Data

The data come from a city-wide Grade 8 Mathematics Achievement Test, consisting of 20 multiple-choice items (0-1 scoring) and 5 constructed-response items (0-4 scoring). The sample size is $N = 2000$. Three cut scores ($\lambda_1 = 40, \lambda_2 = 60, \lambda_3 = 80$) were used to classify students into four proficiency levels: "Below Basic", "Basic", "Proficient", and "Advanced".

Since the test contains mixed item formats, we used the IRT-based Lee method and the MIRT-based method. A bi-factor model (Bi-MIRT) was fitted to the data. The classification consistency was estimated using the individual approach:

$$\phi_i = \sum_{h=1}^{4} P_{ih}^2$$

where P_{ih} is the probability of examinee i being classified into level h. The test-level consistency is $\phi = \frac{1}{N}\sum_{i=1}^{N}\phi_i$.

3.3 Cognitive Diagnostic Test Data

The data come from the Fraction Subtraction Data (Tatsuoka, 1984), which includes 20 items measuring 8 attributes. The sample size is $N = 536$. The DINA model was used for parameter estimation. We calculated pattern-level and attribute-level classification consistency using the Wang method and the Johnson method.

For attribute k, the individual consistency is:

$$\phi_{ik} = p_{ik}^2 + (1 - p_{ik})^2$$

where p_{ik} is the posterior probability of mastering attribute k:

$$p_{ik} = \frac{\sum_{\alpha\in\Omega: I_k(\alpha)=1} L(x_i|\alpha)P(\alpha)}{\sum_{\alpha\in\Omega} L(x_i|\alpha)P(\alpha)}$$

The results for the first three attributes are presented in Table 2.

4. Discussion and Recommendations

Based on the theoretical review and data analysis, we summarize the following recommendations for selecting classification consistency estimation methods:

1. **Model Fit**: The choice of method should first depend on the measurement model that best fits the data. For tests following CTT assumptions, the LL or Lee methods are appropriate. For large-scale assessments using IRT, the IRT-based Lee or Rudner methods are preferred.
2. **Decision Rules**: When complex decision rules (conjunctive or compensatory) are used in multidimensional assessments, MIRT-based methods or MLM-based methods should be employed.
3. **Test Length**: For short tests, the W method or IRT-based methods (Guo method) tend to be more robust than the LL method.
4. **Reporting**: Researchers should report not only the classification consistency index (ϕ) but also the kappa coefficient (κ) to account for chance agreement, and provide interval estimates (e.g., 95% confidence intervals) using bootstrap or Bayesian methods.

In conclusion, evaluating classification consistency is vital for ensuring the reliability of criterion-referenced decisions. Future research should further explore the interval estimation of these indices and their application in computerized adaptive testing.

Note: Figure translations are in progress. See original paper for figures.