JOURNAL OF APPLIED MEASUREMENT, *23*(3/4), 113-116

COMMENTARIES

Measurement and Data Information

Michelle Liou
Institute of Statistical Science
Academia Sinica

Engelhard's (2022) "Pillars of Measurement Wisdom" provides an effective bridge between statistics and measurement, conveyed in a language that is understandable to both scholars and practitioners. Recent advances in computation and unprecedented access to data have raised some questions pertaining to the validity of our deductions and underlying ethical issues. It might be important to streamline Engelhard's pillars as well as to ensure that advanced technologies in statistics are appropriately utilized for solving measurement problems. *Data information* is a key concept that could create a new trend of integrated pillars for both statistics and measurement. Differential item functioning (DIF), as widely discussed in education and social science, is a good example to illustrate this idea. Two methods that have been applied to assess DIF in dichotomously scored items are logistic regression (Rogers & Swaminathan, 1993) and the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1986). By assessing data information, it is shown below that different pillars of statistics/measurement can be monitored at the same time. The residuals of fitting a DIF model are actually higher-order interactions among variables and can be interpreted scientifically.

DIF is a criterion for evaluating the fairness of measurement; that is, test items should measure the same ability across takers independent of factors irrelevant to the test supposed to measure. Let X be a binary item response with X = 0 or 1 for incorrect and correct answers, respectively, on a test. Let Y be the matching variable which can be a taker's ability level estimated by an item response theory model or his/her total raw score, and Z be a grouping variable such as ethnicity or gender. The cross-classified (X, Y, Z) table is a three-way 2 × J × K contingency table with the joint probability density function f_{ijk}, for i = 1 or 2, j = 1, …, J and k = 1, …, K. The mutual information (MI) or relative entropy among categorical variables (Kullback & Leibler, 1951) defines the expectation of the logarithmic ratio between the joint likelihood and the product of marginal likelihoods under the independence condition. Denote $\hat{I}$(X, Y, Z) as the sample

Requests for reprints should be sent to Michelle Liou, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan; email: mliou@stat.sinica.edu.tw

estimate of MI of the three variables, which can be decomposed into the following terms (Cheng, Liou, & Aston, 2010; Cheng, Liou, Aston, & Tsai, 2008):

$$\hat{I}(X; Y; Z) = \hat{I}(Y; Z) + \hat{I}(X; Y) + \hat{I}(X; Z|Y), \quad (1)$$

where the right-hand side of Equation 1 gives two 2-way MI terms; for example, $\hat{I}(X; Y)$ is the MI of the 2-way (X, Y) table, which is an analogy of $\hat{I}(X; Y; Z)$ in the 3-way table. The conditional MI (CMI) term $\hat{I}(X; Z|Y)$ is the expected deviance of population data from the table assuming conditional independence between X and Z across the levels of Y. Likelihood ratio (LR) statistics corresponding to the three terms are orthogonal to each other and can be tested for statistical significance using Chi-Square distributions with, respectively, (J-1)(K-1), (I-1)(J-1), and (I-1)(K-1)J degrees of freedom. The information identity in Equation 1 also indicates that the 3rd 2-way (X, Z) does not exist when (Y, Z) and (X, Y) are already in the model. If $\hat{I}(X; Z|Y)$ is replaced by $\hat{I}(X; Z)$ in Equation 1, this additional 2-way effect measures the partial association between X and Z given Y. The CMI term in Equation 1 can be further decomposed into two orthogonal terms; that is,

$$\hat{I}(X; Z|Y) = \hat{Int}(X; Y; Z) + \hat{Par}(X; Z|Y), \quad (2)$$

where $\hat{Int}(X; Y; Z)$ denotes the interaction between the three variables, and $\hat{Par}(X; Z|Y)$ denotes the partial association between X and Z given Y.

The two orthogonal LR statistics for testing significance in Equation 2 are analogy of the Breslow–Day (Breslow et al., 1980) and MH tests (Mantel & Haenszel, 1959), respectively, which are not orthogonal to each other, however. We may consider terms in Equation 2 as the MI counterparts of the Breslow-Day and MH tests. All MI and CMI terms in Equations 1 and 2 can be estimated by the maximum likelihood estimation method under independence and conditional independence assumptions. The right-hand side of Equation 1 can further be expressed as

$$\hat{I}(X; Y; Z) = \hat{I}(Y; Z) + \hat{I}[X; (Y, Z)]$$
$$= \hat{I}(Y; Z) + \hat{I}(X; Y) + \hat{I}(X; Z|Y). \quad (3)$$

The term $\hat{I}[X; (Y, Z)]$ in Equation 3 corresponds to the total information to be fitted by a logistic regression model, in which X is the target variable to be predicted by Y and Z.

Rogers and Swaminathan (1993) proposed the logistic regression analysis for detecting uniform and nonuniform DIF in dichotomous items. The binary logistic model in DIF analysis considers the logarithmic odds between the probability of receiving score 1 relative to that of score 0 on the target item as follows:

$$\log\left(\frac{f_{1,j,k}}{f_{0,j,k}}\right) = \beta_0 + \beta_j^Y + \beta_k^Z + \beta_{jk}^{YZ}, \quad (4)$$

for j = 1, …, J and k = 1, …, K. The variable Y denotes the total score after deleting the score on the target item, and Z is the grouping variable (focal versus reference groups). An item displays uniform DIF if $\hat{\beta}_k^Z$ is significant and $\hat{\beta}_{jk}^{YZ}$ is insignificant. If $\hat{\beta}_{jk}^{YZ}$ is significant, then the item is declared to have nonuniform DIF. When Y is categorical, the model in Equation 4 is identical to $\hat{I}[X; (Y, Z)] = \hat{I}(X; Y) + \hat{I}(X; Z|Y)$. As indicated in Equation 2, $\hat{I}(X; Z|Y)$ can be decomposed into the partial association and interaction terms corresponding to uniform ($\hat{\beta}_k^Z$) and nonuniform ($\hat{\beta}_{jk}^{YZ}$) DIF, respectively. An empirical example may illustrate the correspondence between models in Equations 3 and 4.

Empirical data contain item responses on the Basic Competence Test for Junior High School Students (BCtest) developed for measuring the ability of resolving practical problems using 9th grade algebra and geometry knowledge by the Research Center for Psychological and Educational Testing in 2009. For the illustrative purpose, we selected

11 geometry items that were homogeneous in content based on factor analysis on item scores, and a random sample of 29,710 takers after deleting cases with missing responses. We considered genders as the grouping variable and total raw scores as the matching variable (without counting scores on the item under the DIF analysis). An interested reader may refer to Smith (2004) for a comparison between the MH procedure and Rasch model for assessing DIF. Among the 11 items, LR tests based on information decomposition in Equation 2 suggested that 6 out of 11 items were classified as non-DIF items at $\alpha = 0.05$, and the other 5 items were classified as nonuniform DIF. Logistic regression in Equation 4 treating Y as a continuous variable suggested similar results except for one out of these 6 non-DIF items that showed significant $\hat{\beta}_{jk}^{YZ}$. Additionally, one nonuniform DIF item in the LR test was classified as uniform DIF with significant $\hat{\beta}_k^Z$ in logistic regression analysis. The cross-classifying (X, Z) according to the levels of Y suggested that the item with significant $\hat{\beta}_{jk}^{YZ}$ in logistic regression showed significant contingency coefficients with p values equal to 0.02 and 0.01 for Y = 2 and 3, respectively, (i.e., genders and item scores were correlated when Y = 2 and 3) which could explain the insignificant $\hat{Int}(X; Y; Z)$ and $\hat{Par}(X; Z|Y)$ in the LR test. In the BCtest, which has been used as an entrance examination for senior high school, bias in low-level scores is not necessarily of concern in selecting cut-off scores at the upper-tail of the score distribution.

The BCtest test data also include 5 geographic locations (i.e., rural and urban areas) of takers in Taiwan, and the grouping variables can be genders and areas (A):

$$\hat{I}(X; Y; Z; A) = \hat{I}(Y; Z; A) + \hat{I}[X; (Y, Z, A)]$$
$$= \hat{I}(Y; A) + \hat{I}(Z; A) + \hat{I}(Y; Z|A)$$
$$+ \hat{I}(X; Y) + \hat{I}(X; Z|Y) + \hat{I}[X; A|(Y, Z)]. \quad (5)$$

Equation 5 satisfies the rule of a valid information identity; that is, with 4 variables, there are at most three 2-way, two 3-way, and one 4-way MI terms (Liou et al., 2023). If one computes the MI values for the 6 terms in Equation 5 separately, the sum of these values should be equal to the $\hat{I}(X; Y; Z; A)$ value. The last two terms in Equation 5 can each be decomposed into partial and interaction effects. The logistic regression can be expressed as

$$\log\left(\frac{f_{1,j,k}}{f_{0,j,k}}\right) = \beta_0 + \beta_j^Y + \beta_k^Z + \beta_{jk}^{YZ} + \beta_1^A. \quad (6)$$

When β_j^Y, β_k^Z, and β_{jk}^{YZ} are already in the model, the last term β_l^A for $l = 1, \ldots, 5$ estimates $\hat{Par}[X; A|(Y, Z)]$. The deviance value after fitting Model 6 to data estimates $\hat{Int}[X; A; (Y, Z)]$. Among the 6 non-DIF items, three of them showed significant partial associations with geographic areas. Logistic regression treating Y as a continuous variable also suggested similar results.

Our DIF analysis has made it clear that when assessing DIF in complex scenarios, the MI approach advocates the integration of various methods (e.g., MH and logistic regression) under a single data-information framework. In this example, the data per se supervise various pillars of measurement, including likelihood, measurement invariance, and regression. In Equations 3 and 5, residuals are presented as high-order interactions among variables, which can be interpreted scientifically regardless of their statistical significance.

Finally, it is important to remember that educational measurement involves ranking takers according to two dimensions: (1) relative competence on items designed according to *psychometric* properties (i.e., individual differences) or (2) differential competence before and after an intervention designed according to specific *edumetric* properties (i.e., individual gain; Carver, 1974). The pillars of statistical analysis, including aggregation, likelihood, information, inter-comparisons,

regression, design, and residuals, rely largely on the principle of variation within a sample. In the absence of variation, all such statistical analysis grinds to a halt. In other words, these pillars reflect the value and applicability of the psychometric dimension to measurement-related issues including the MI approach. Ideally, an educational program should seek to maximize the gains achieved between pre- and post-assessments, rather than maximizing between-person variation. In situations where assessment focuses on the gains made by the individual before and after an intervention, it is important to divorce research and practice from the principle of variation among test takers. In other words, researchers must develop advanced theories and technologies that are independent of this principle by tailoring measurement to edumetric properties.

References

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research. Volume I: The analysis of case-control studies*. IARC Scientific Publications.

Carver, R. P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, *29*(7), 512–518.

Cheng, P. E., Liou, M., & Aston, J. A. D. (2010). Likelihood ratio tests with three-way tables. *Journal of the American Statistical Association*, *105*(490), 740–749.

Cheng, P. E., Liou, M., Aston, J. A. D., & Tsai, A. C. (2008). Information identities and testing hypotheses: Power analysis for contingency tables. *Statistica Sinica*, *18*(2), 535–558.

Engelhard, G., Jr. (2022). The pillars of measurement wisdom. *Journal of Applied Measurement*, *23*(3/4), 80–95.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, *1986*(2), 1–24.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Liou, J.-W., Liou, M., & Cheng, P. E. (2023). Modeling categorical variables by mutual information decomposition. *Entropy*, *25*, 750. https://doi.org/10.3390/e25050750

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116.

Smith, R. M. (2004). Detecting item bias with the Rasch Model. *Journal of Applied Measurement*, *5*(4), 430–449.