A Q_3 -Based Permutation Test for Assessing Local Independence

Applied Psychological Measurement

2016, Vol. 40(2) 157–160

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621615622635

apm.sagepub.com

W. Holmes Finch¹ and Heather Jeffers¹

Abstract

Item response theory (IRT) is a powerful statistical methodology used in the analysis of psychological and educational assessments. IRT rests on three fundamental assumptions about the data, including local independence, which means that after accounting for the latent trait(s) being measured, the item responses are independent of one another. Traditionally, this assumption is assessed using Yen's Q_3 statistic. However, Q_3 does not have a known sampling distribution, and thus, it is typically used in a descriptive fashion, such that values larger than an arbitrary cut-value (e.g., 0.2) indicate the presence of local dependence. The current study introduces a formal test of the null hypothesis that for a given item pair Q_3 is 0, based on permutation test methodology. A small simulation study carried out to assess the Type I error and power rates of the Q_3 permutation test found that this new statistic maintains good Type I error control, while also yielding power for detecting local dependence at a rate higher than that associated with the use of the 0.2 cut-value.

Keywords

local independence, Yen's Q_3 , item response theory

All of the commonly used unidimensional item response theory (IRT) models rest on three primary assumptions about the data: (a) the relationship between an item response and the latent trait being measured is monotonic, (b) the latent trait being measured by the scale is unidimensional, and (c) the item responses are locally independent. Local independence (LI), which is the focus of this study, means that item responses are uncorrelated when conditioning on the measured latent trait. A number of statistics have been proposed for assessing LI, with Yen's Q_3 (Yen, 1984) being perhaps the most popular and accurate for this purpose (Kim, De Ayala, Ferdous, & Nering, 2011). For an item pair, Q_3 is the correlation between item residuals, where the residual is the difference between the observed item responses and the responses predicted for each item by an appropriate IRT model (e.g., three-parameter logistic). A variety of approaches for testing the null hypothesis of LI using Q_3 have been suggested, but none has been found effective (Chen & Thissen, 1997; Glas & Suarez Falcon, 2003; Kim et al., 2011). Thus, researchers typically use an ad hoc rule, concluding that LI is violated when Q_3 exceeds 0.2, as suggested by Yen (1993).

¹Ball State University, Muncie, IN, USA

Corresponding Author:

W. Holmes Finch, Department of Educational Psychology, Ball State University, Muncie, IN 47306, USA.

Email: whfinch@bsu.edu

The purpose of this study is to introduce a permutation test for Q_3 ($Q3PT$) and to investigate its performance with a simulation study. Permutation tests allow for testing a null hypothesis without making distributional assumptions about the variable in question or the test statistic itself (Pesarin & Salmaso, 2010). Such tests have in common the creation of a sampling distribution of the statistic of interest (e.g., Q_3) under the null hypothesis, by calculating its value for all possible permutations of the data, or in the case of very large samples, a set of random permutations of the data. This permutation approach can be applied to testing the null hypothesis that $Q_3 = 0$; that is, an item pair is LI. $Q3PT$ is calculated using the following steps:

1. Calculate Q_3 for a pair of items with the observed data, using an appropriate IRT model.
2. For each pair of items, j, k , randomly sort residuals for one item in the pair.
3. Calculate Q_3 for the item pair j, k using the randomly sorted residuals.
4. Repeat Steps 2 and 3 for all data permutations (or a large number of random permutations) to develop a distribution of Q_3 values under the null hypothesis of LI.
5. Compare Q_3 calculated in Step 1 for the original data with the distribution of permuted Q_3 values. If it is greater than the 95th percentile of the permutation distribution, reject the null hypothesis of LI and conclude that the items are locally dependent.
6. Repeat Steps 1 through 5 for each pair of items for which the LI assumption is to be assessed.

The goal of this study is to introduce the $Q3PT$ for assessing LI. Given that there is not a dependable hypothesis test for this statistic, researchers must rely on a rule of thumb ($Q_3 > 0.2$) to determine whether pairs of items are LI. Chen and Thissen (1997) noted that this approach tends to identify locally dependent item pairs less often than it should (low power). Despite this weakness, Q_3 remains perhaps the most effective tool for assessing LI. Therefore, it would be useful if an accurate method for testing the null hypothesis of LI using Q_3 could be developed. It is with this purpose in mind that the current study was conducted.

Method

A small simulation study was conducted using SAS (SAS Institute, 2010) to examine the performance of $Q3PT$. Data were simulated using the three-parameter logistic testlet model, with item parameter values for data generation were taken from the calibration sample ($N = 6,452$) of a large national standardized reading assessment. Testlets were used to induce local dependence among items. Two test lengths were included, 20 and 40 items, and the number of examinees simulated was 500, 1,000, or 2,000. Local item dependence was simulated to be small (testlet variance of 0.25) or large (testlet variance of 1.0). Testlet length (i.e., percentage of locally dependent items) was simulated to be 25% or 50% of the total test. $Q3PT$ ($\alpha = .05$) and Q_3 with the 0.2 cutoff value ($Q3_{.0.2}$) used to assess LI. The outcome variables were Type I error (rejecting LI when no testlet was present), false positive rates for LI when one item in the pair belonged to the testlet and the other did not (False Positive Type 1 (FP1)), false positive rates for LI when neither item was in the testlet but a testlet was present (False Positive Type 2 (FP2)), and power rates for assessing LI.

Results

Type I error rates (proportion of times that the null hypothesis of LI was rejected when the data were generated from a unidimensional model with no Local Dependence [LD] present) across

Table 1. Type I Error Rate for a Unidimensional Model With No Local Dependence (LD), by Sample Size (*n*) and Number of Items.

<i>n</i>	Items	Q3PT	Q3_0.2
500	20	0.06	0.08
1,000	20	0.05	0.08
2,000	20	0.06	0.08
500	40	0.05	0.08
1,000	40	0.05	0.08
2,000	40	0.05	0.08

Note. Q3PT = permutation test for Q_3 .

Table 2. False Positive Type 1 (FP1), False Positive Type 2 (FP2), and Power Rates by Degree of Local Dependence (LD) by Magnitude of LD, Proportion of LD Items, Sample Size (*n*), and Number of Items.

		FPI				FP2				Power			
		Q3PT		Q3_0.2		Q3PT		Q3_0.2		Q3PT		Q3_0.2	
<i>n</i>	I	25%	50%	25%	50%	25%	50%	25%	50%	25%	50%	25%	50%
Weak LD													
500	20	0.08	0.08	0.11	0.08	0.09	0.09	0.13	0.10	0.88	0.92	0.49	0.53
1,000	20	0.07	0.07	0.12	0.08	0.07	0.08	0.13	0.10	0.94	0.97	0.55	0.60
2,000	20	0.04	0.05	0.12	0.09	0.07	0.07	0.13	0.11	1.00	1.00	0.61	0.66
500	40	0.06	0.07	0.07	0.06	0.07	0.07	0.08	0.09	0.90	0.92	0.79	0.81
1,000	40	0.06	0.06	0.07	0.06	0.06	0.07	0.09	0.09	0.96	1.00	0.86	0.86
2,000	40	0.05	0.05	0.08	0.07	0.06	0.06	0.09	0.09	1.00	1.00	0.88	0.90
Strong LD													
500	20	0.08	0.08	0.16	0.11	0.08	0.08	0.17	0.13	0.93	0.99	0.83	0.84
1,000	20	0.07	0.07	0.17	0.12	0.07	0.07	0.17	0.14	0.99	1.00	0.87	0.88
2,000	20	0.04	0.05	0.17	0.11	0.06	0.06	0.18	0.14	1.00	1.00	0.91	0.93
500	40	0.06	0.06	0.09	0.08	0.07	0.07	0.10	0.10	0.96	1.00	0.86	0.88
1,000	40	0.05	0.06	0.09	0.08	0.06	0.06	0.10	0.10	1.00	1.00	0.93	0.94
2,000	40	0.05	0.06	0.09	0.08	0.06	0.06	0.11	0.11	1.00	1.00	0.98	0.98

Note. Q3PT = permutation test for Q_3 .

item pairs for the methods by sample size and number of items appear in Table 1. *Q3PT* maintained the nominal Type I error rate of 0.05 across sample sizes and numbers of items. However, *Q3_0.2* had somewhat more elevated rejection rates when no LD was simulated in the data.

FP1 rates (rejection of LI for a pair when one item in the pair belonged to the testlet and the other did not) for each method by sample size, number of items, and degree of LD appear in the leftmost columns in Table 2. *Q3_0.2* had the highest FP1 rates across virtually all conditions simulated here, whereas *Q3PT* generally had FP1 rates at or near 0.05 across many of the conditions simulated here. FP2 occurred when LD was simulated in the data, but neither item in the tested pair was part of the testlet. Rejection rates under this condition (Table 2) were highest for *Q3_0.2* across sample sizes, number of items, magnitude of LD, and proportion of items included in the testlet. In addition, *Q3PT* FP2 rates declined with larger

samples, more items, and were largely impervious to LD magnitude and proportion of items in the testlet.

Power rates for the methods by sample size, number of items, magnitude of LD, and proportion of items belonging to the testlet appear in the four rightmost columns in Table 2. Across all conditions, *Q3PT* exhibited power rates for rejecting LI when it was not present of 0.88 or higher. Power for *Q3_0.2* was well below 0.80 for many of the conditions examined here, with its lowest rates when LD was weak, with 20 items.

Discussion

The purpose of this study was to introduce *Q3PT* for assessing the LI assumption in IRT. This new approach was designed to improve upon known problems with the popular *Q3* statistic, namely, the lack of a set criterion for determining when LD is present for an item pair and low power for identifying it when present. Based on the simulation study, *Q3PT* shows promise in terms of both maintaining acceptable Type I error and false positive rates, while also yielding high power for identifying LD items. Therefore, *Q3PT* appears to offer a potentially useful alternative for psychometricians and researchers in educational measurement when the LI assumption must be assessed. Researchers are encouraged to study this method further and consider its possible application in practice.

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.

References

- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.
- Glas, C. A. W., & Suarez Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106.
- Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement, 35*, 447-471.
- Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications, and software*. Chichester, UK: Wiley.
- SAS Institute. (2010). SAS version 9.1. Cary, NC: SAS Institute.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.