Evaluating Local Dependence (LD) Using Pairwise LD Indices in Psychological Data

Saifa Pirani

PSYC 859 Item Response Theory

Fall 2019

Introduction

The concept of local independence in Item Response Theory (IRT) models has been a topic of interest in the past two decades. Local independence, in a unidimensional or correctly specified multidimensional IRT model, suggests that item responses on two or more items are unrelated, holding the underlying latent variable constant. More formally, local independence is the assumption that the probability of observing similar item responses is a product of the probabilities of observing the item responses independently (see Chen & Thissen, 1997; McDonald, 1994; Stout, 2002 for details). For psychological measures, this means that item responses on all items in a scale measuring psychological constructs such as depression are assumed to be independent of each other. This is a fundamental assumption of IRT models. If the assumption is not satisfied then Local Dependence (LD) is present. Violating this assumption can lead to inaccurate and biased estimates of item parameters including the ability parameter and can produce overestimation of item parameters in mispecified models (Chen & Thissen, 1997; Steinberg & Thissen, 1996; Wainer & Wang, 2000; Thissen, Steinberg, & Mooney, 1989). In psychological measurement, where responses to items measuring psychological constructs are related, this assumption is often violated (Houts & Edwards, 2013).

**Existing Local Dependence Indices and Statistics**

In the past two decades, researchers have proposed a plethora of indices and statistics to address the issue of LD. Examples of a few existing LD indices and statistics include: Pearson’s χ2 (P- χ2 ; Chen and Thissen, 1997), LR *G2* , Wald test, LR test (Wilks, 1935), Transformed Q3 (Yen, 1984), Mantel–Haenszel test with multiple testing corrections (Ip, 2001), Jackknife Slope Index (JSI; Edwards, Houts, & Cai, 2018), and confirmatory factor analysis modification indices (MI). The current paper focuses on three LD indices: P - χ2, JSI, and MI.

**P - χ2 (Chen and Thissen, 1997).** The P- χ2 test is a statistical test to evaluate the difference between expected and observed frequencies. More formally, this is a pairwise null hypothesis test of LD with an approximate χ2 distribution. For IRT, P- χ2 test can be used to detect LD between two categorical variables from a 2 x 2 contingency table. When the differences between the observed and expected frequencies exceed 3.84 (5%) and/or 6.63 (1%; Chen & Thissen, 1997), LD is present.

**Confirmatory Factor Analysis Modification Indices.** MIs obtained from confirmatory factor analysis (CFA) output are comparable to LD indices in detecting LD (e.g., Kim, 2007; Steinberg & Thissen, 1996; Swygert, McLeod, & Thissen, 2001). In traditional CFA settings, MIs are used as a diagnostic statistic to evaluate improvement in overall model fit if a select pathway usually the error covariance of two items is freed. Similar to the P- χ2 test, MIs have an underlying χ2 distribution. If the MI values of two items exceeds the nominal levels of 3.84 (5%) and/or 6.63 (1%), the error covariance between those two items is considered significantly high.

**JSI Jackknife Slope Index (JSI; Edwards, Houts, & Cai, 2018).** The JSI was formally introduced as new diagnostic procedure to detect LD by Edwards, Houts, and Cai (2018). Prior to the explicit introduction, the JSI procedure has been used in previous studies (Houts & Edwards, 2013). This procedure utilizes the inflation of slope parameters, caused by LD items in the model to detect the source of LD. This is an *ad hoc* procedure performed once parameter estimates for the full model are obtained. One item is removed iteratively to obtain revised slope parameter estimates, then the item is replaced. The JSI value for an item is computed by subtracting the a-parameter when the item is removed from the a-parameter in the full model over the standard error of the full model a- parameter. The threshold for detecting LD using the JSI procedure is computed as the mean plus two standard deviations of the summed JSI values.

**Summary of Recent Studies**

Starting from Chen & Thissen (1997) there have many studies that have progressively compared and contrasted LD indices using various conditions (e.g., varying sample sizes, etc.). The studies are discussed below.

**Chen & Thissen (1997)** examined LD with two models: surface local dependence (SLD) and underlying local dependence (ULD) using four LD indices (i.e., P - χ2). SLD occurs when item responses between two or more items are nearly identical. ULD occurs when there is a common latent trait between a pair and more items in a model. To simulate these models, Chen and Thissen varied the number of items and varied the values of the joint probability of participants responding to a pair of items. Their study results indicated that while P - χ2 test was a good detector of LD in SLD models, it was not sensitive to LD in the ULD model.

**Kim (2007)** ran a Monte Carlo simulation study to examine the impact of test length, LD level, and locally dependent item percentage on the performance of ten LD indices (e.g., P - χ2 , LR G2, MI, etc.) using Type I error, power, and false positive rates. While no LD index had the “best” performance in terms of type I error, power, and false positive rate, Kim (2007) recommended the use of select indices (i.e., Q3 , residual correlation, MI, etc.). Kim also found that as the number of items increased, power and Type I error rates of the MIs to detect LD improved. However, in the presence of strong LD (75% items are LD), MI displayed the highest false positive rates.

**Liu & Maydeu – Olivares (2012)** examined the performances of bivariate and triplet- wise (three items are locally dependent) LD indices (e.g., P - χ2 , LR G2, *M*3, etc.) in simulated data. In the first simulation containing no LD, they found that the bivariate and trivarite P - χ2 were not only inconsistent in detecting LD, but also their power and Type I error rate did not improve with an increase in sample size. The authors choose not to test the power of the bivariate and trivarite P - χ2 test in the second simulation. In their second simulation study, they generated data to follow a bifactor LD triplet model and a 2-factor independent clusters model. The bifactor LD triplet model is similar to the SLD model, as defined by Chen & Thissen (1997), where a set of three items are locally dependent. The 2-factor independent cluster model is similar to the ULD model, where there are two independent underlying latent traits. This model generates LD between items of each latent trait when items are assumed to have a unidimensional model structure. After simulating data under these two model conditions, Liu and Maydeu – Olivares evaluated the power and Type I error rates of select LD indices.

**Houts & Edwards (2013)** compared the performance of Q3 (Yen, 1984), Fisher’s r-to-z transformed Q3, likelihood ratio test statistic G2 (Chen & Thissen, 1997), JSI, NOHARM-based χ2G/D (McDonald, 1997), and confirmatory factor analysis modification indices in a Monte Carlo study. Uniquely, this was the first LD index comparison study to simulate conditions that mimic psychological assessment conditions. Meaning that, population parameters to simulate data were drawn from the distributional properties commonly found in psychological data. Additionally, conditions including small sample sizes, small number of items, and multiple ordered responses were simulated. The JSI performed well independently and relative to other LD indices. Results suggested that JSI and likelihood ratio test *G2* demonstrated good relative performance. The authors suggested that given the limitations of JSI, researchers should use JSI with the supplementary index, LR *G2*. In the 2PL model, MIs showed both high power and high Type I error rates in majority of the conditions indicating that MIs have liberal statistical criterion.

These comparative studies have made huge contributions to our understanding of LD indices. Overall, researchers have demonstrated power and Type I error rate of various indices in simulated random and simulated psychological data (Houts & Edwards, 2013). However, thus far, there has not been evidence of a “best” index to detect LD (Kim, 2007). Like previous studies, the current study compares LD indices in various conditions. Particularly, the newest LD procedure, the JSI, is closely examined.

While two recent studies have presented results that compare the new JSI procedure with other LD procedures (Houts & Edwards, 2013; Edwards et al., 2018), the performance of the JSI procedure has not been examined in bifactor structure models and multidimensional structure model containing three items with LD (Liu & Maydeu – Olivares, 2012). Houts & Edwards (2013) explained that the JSI experiences a loss of power in the presence of more than one pair of LD. In an applied example, Edwards et al., (2018) discovered that when the source of LD is not known *a priori*, JSI can effectively detect and indicate pairs containing LD. In both papers, the authors call for further investigation of the JSI independently and relative to other LD indices in diverse model conditions. As such, the current study attempts to answer the following questions: How does the JSI perform in a bifactor structure model and multidimensional structure model with three locally dependent items relative to select LD indices and independently in a small number of items? Does the JSI experience a loss of power as Houts & Edwards (2013) noted in conditions where there are more than one pair of locally dependent items? Finally, does the JSI procedure provide adequate evidence in simulated psychological data to detect LD when the source of LD is not known *a priori*?

The current study empirically evaluates JSI by examining the independent performance and relative performance of the JSI compared to P - χ2, and MI in a simulation study. While there is strong evidence of other LD indices performing better than the traditional procedures, P - χ2 and MI, to detect LD, the traditional procedures were selected based on their accessibility and popularity in psychological measurement literature. These procedures were used as markers for psychological researchers to better understand JSI values.

The current study mimics the methodology of past LD research to obtain the power and Type I error rates for the three LD indices. Simulation conditions are described below. Note, the current study reflects preliminary results. This study can easily be expanded to obtain comprehensive results.

**Data Simulation**

Methodology for the current study was inspired by Chen & Thissen (1997), Edwards & Houts (2018), and Liu & Maydeu-Olivares (2012). The purpose of the current study is to evaluate the LD indices P - χ2, MIs, and the JSI using data that are simulated to retain the statistical properties commonly found in psychological measures. Particularly, with data that contain three locally dependent items and small test length (6 items). Three different LD conditions are examined: no LD, bifactor model structure LD, and multidimensional model structure LD. All conditions were evaluated using two sample sizes (*N* = 250 and *N* = 1,000) with 1,000 replications each.

In the no LD case, parameter conditions were modeled after Edwards & Houts (2018), where the a-parameters were simulated from a N(1.7, 0.3) distribution to mimic distributions found in common psychological measures (Hill et al., 2004). The difficulty/ severity parameters or b- parameters and values of θ were simulated from a N(0,1) distribution similar to Chen and Thissen (1997), Edwards et al. (2018), and Kim (2007).

Like the no LD case, the bifactor model structure LD case parameters were modeled after Edwards & Houts (2018). The slope parameters for all items were drawn from a N(1.7, 0.3) distribution for the first factor and the Items 3 to 6 slope parameters were drawn from N (2.55, 0.15) distribution for the second factor. In essence, Items 3 to 6 loaded the first and second factor inducing a bifactor model structure case (Edwards & Houts, 2018; Houts & Edwards, 2015). The b-parameters and θ were simulated from a N(0,1) distribution. Identical procedures were followed for the multidimensional model structure LD. However, slopes for Items 3 to 6 only loaded on the second factor to simulate a multidimensional model structure, with orthogonal factors.



1. Bifactor model structure



1. Multidimensional model structure

The simulated data for both conditions were analyzed using the packages “mirt” and “lavaan” in R (Chambers, 2012; Rossell, 2012). First, P - χ2 and JSI values were extracted for all items using the “mirt” package (Chambers, 2012). Then, modification indices were extracted for all items using the package lavaan (Rossell, 2012). Only results for item pair 1 & 2, item pair 3 & 4, and item pair 5 & 6 are reported for simplicity. The average values for each pair across all replications and conditions are presented in (Tables 1, 2, & 3). P - χ2, MI, and JSI values that crossed a certain threshold value were considered evidence for local dependence. For, the P - χ2 and MI the nominal threshold value of 3.84 was retained. Any P - χ2 and MI value for a pair above this threshold was flagged (Chen & Thissen, 1997; Kim, 2007). Finally, as suggested by Edwards, Houts, & Cai (2013), the threshold values for the JSI method were calculated *ad hoc* as the mean plus two standard deviations of the pair-wise summed JSI values. For each replication, positive JSI values above the unique threshold were flagged as indicating LD. This is the arbitrary “mean plus 2SD” rule-of-thumb.

Results on Tables 1, 2, & 3 display the ratio of LD recovery for P - χ2, MI, and JSI over 1,000 replications. These results demonstrate two statistical concepts: power and type 1 error. Power, the ability of the LD procedure to correctly identify pairs that are dependent, is computed as the ratio of correctly identified LD over 1,000 replications. Type I error occurs when LD is detected, where none should be present. Similar to power, Type I error is computed as the ratio of misidentified LD over 1,000 replications. Thus, for the no LD condition, consistent with previous studies, we expected all LD procedures to retain low Type I error rates. For the bifactor and the multidimensional model structure LD case, we expected item pairs 1& 2 and/or 5 & 6 to provide evidence of high power to detect local dependence and for item pair 3 & 4 we expected a low Type I error rate. As in Houts & Edwards (2013), rule-of-thumb thresholds for power and Type I error are used to discuss results, where a power of 0.8 or above and Type I error rate below or equal to 0.05 are considered satisfactory.

**Results**

Results for the no LD condition results are presented in Table 1. Overall, all three LD indices across both sample sizes have low Type I error rates. Particularly, for the P - χ2 test Type I error did not exceed 1%. For, MI Type I error rates range from 5% to 8%. Finally for the JSI Type 1 Error rates ranged from 1% to 5%. As per the rule-of-thumb thresholds, the MI did not have a satisfactory Type I error.

The bifactor model structure LD case provided interesting results (Table 2). For the data with a sample size of *N* = 250, all three indices had high power to detect LD for item pair 1& 2. However, all three failed to detect LD for item pair 5 & 6. While, P - χ2 and the JSI had satisfactory Type I error rates, MI had a Type I error rate of 12%. Model results were identical in the *N* = 1,000 sample. However, the JSI no longer retained high power to detect LD for item pair 1& 2.

Finally, for the multidimensional model LD structure for the *N* = 250 sample size, none of the indices adequately recovered LD for item pair 1 & 2 and item pair 5 & 6. Interestingly, the JSI detected LD in item pair 3& 4 at 22% where none was present. Model results improved in the *N* = 1,000 sample. While P - χ2 and the JSI evidenced high power to detect LD for pair 1 & 2, they had low power to detect LD for item pair 5 & 6. MI had the opposite effect. MI recovered LD 98% of the time in item pair 5 & 6. Type I error rates remined high for MI and the JSI and low for P - χ2 (Table 3).

**Conclusion**

Starting from Chen & Thissen (1997), simulation studies to evaluate the performance of various LD indices have been conducted in abundance in the past 20 years (Houts & Edwards, 2013; Kim, 2007; Liu & Maydeu – Olivares, 2012). The current study combines select elements of previous studies’ methodology and uniquely examines that performance of pairwise JSI with P - χ2 and MI procedures in conditions of a) no LD, b) bifactor model structure LD, and c) multidimensional model structure LD. A small simulation is conducted to compare the three LD indices in data that mimic psychological assessment conditions with six items.

Notably, study results indicate that while the JSI detects LD adequately in the multidimensional model structure, it does not perform well in the bifactor model structure for smaller sample sizes. The JSI consistently retains low Type I error rates across sample sizes. In the presence of three locally dependent items, all three indices fail to detect LD in both item pairs 1& 2 and 5& 6 simultaneously. However, all three do detect at least one source of LD.

Consistent with Houts & Edwards (2013), the JSI experienced a loss of power in both the multidimensional and bifactor model structures with three locally dependent items. Finally, when individual JSI matrices are examined across both conditions with LD and sample sizes, the JSI did provide evidence the source of LD for the three locally dependent items *a priori*. This is consistent with Houts & Edwards (2013) and Edwards et al., (2018)’s recommendation to not use fixed or arbitrary threshold values to evaluate LD, but rather examine the JSI matrix individually to locate the source of LD.

It is important to note that results from the current study are only generalizable to psychological assessment data that mirror the simulation conditions. Even though population parameters were randomly drawn from distributional properties commonly seen in real world psychological assessment data, this study is not comprehensive enough to make generalizable conclusions about the behaviors of these three LD indices. This study can be expanded with varying sample sizes, number of items, and with multiple ordered response options to better address the behavior of these indices.

Resources

Chalmers RP (2012). “mirt: A Multidimensional Item Response Theory Package for the R Environment.” Journal of Statistical Software, **48**(6), 1–29.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological methods*, *23*(1), 138-149.

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni, J. W. (2007). Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Journal of Medical Care,* S39-S47.

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement*, *37*(7), 541-562.

Ip, E. H. S. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*(1), 109-132.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In *Handbook of modern item response theory* (pp. 257-269). Springer, New York, NY.

Rosseel Y (2012). “lavaan: An R Package for Structural Equation Modeling.” Journal of Statistical Software, **48**(2), 1–36. <http://www.jstatsoft.org/v48/i02/>.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*(1), 81-97.

Swygert, K. A., McLeod, L. D., & Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. In *Test Scoring* (pp. 229-262). Routledge.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple‐categorical‐response models. *Journal of Educational Measurement*, *26*(3), 247-260.Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220.

Wilks, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica, Journal of the Econometric Society*, 309-326.

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*(2), 125-145.

Tables





