Revealing the Structure of Item Dependency with SAS® CLUSTER and TREE Procedures in Educational Testing and Measurement
Yung-chen Hsu, American Council on Education, Washington, DC

ABSTRACT
In the development of large-scale educational or psychological tests, identifying potentially dependent items is one of the essential tasks before further item treatments or additional psychometric procedures can be performed. For fast screening, several indices have been proposed to quantify the pair-wise dependency for each item pair. Items with relative strong dependency are flagged for further examination by content experts. However, managing this practice is still time consuming, especially when many test forms are developed simultaneously. This paper introduces the use of the CLUSTER and TREE procedures to reveal the dependency structures. The graphical tree structure helps test development practitioners to lessen the ambiguity of inconsistency among different indices and to further reduce the number of flagged item pairs. When dealing with item triplets or more than three items, this approach appears to have advantage in providing insight into the underlying dependency. Both simulation and real data are used as examples to illustrate the applicability.

INTRODUCTION
In most testing service and assessment research organizations, identifying locally dependent items is one of the essential tasks in the tests development process. Assessing local item dependence (LID) also plays an important role in the study of evaluating new psychometric models and alternative item treatment strategy. LID presents covariations among subset of items. Without properly accounting for the effect of LID can produce overestimate of item parameters and test precision (Chen & Thissen, 1997; Sireci, Thissen, & Wainer, 1991, Wainer & Thissen, 1996; Yen 1993).

Although the possible cause of LID varies, such as those listed in Yen’s (1993) compendium, LID is considered as an empirical phenomenon in educational and psychological tests and is often observed in applied settings—among items in paper-and-pencil tests as well as in the computer adaptive testing environment. In the practice of test development, a typical procedure that most psychometricians use is to compute LID indices for preliminary screening. Over the years, several LID indices were proposed and were useful for screening dependent items. For example, the $Q_r$ statistics described by Yen (1984, 1993) is the conditional correlation of item scores given model predictions based on ability estimates. Chen and Thissen (1997) proposed four other indices. The four indices all use the expected frequencies from psychometric model as the theoretical frequencies to compute the values. In addition, some indices based on information entropy (Tsai & Hsu, 2005) also used expected frequency from psychometric model. Generally, the screening step does allow psychometricians to effectively exclude most item pairs for further investigating the dependency and reduce the diagnostic effort of content experts. However, there are still issues in real practice. This study illustrates a way to assist psychometricians to smooth out some of the glitches in the test development process by providing a holistic view of the structure of item dependency, not just the item pairs.

LID ASSESSMENT ISSUES
The following are some practical issues which psychometricians often encounter in the LID assessment process:

1. Because each index has its own characteristics, all commonly used indices are often generated to provide better insight in preliminary review. When inconsistency occurs, psychometricians generally keep more items for content experts for further investigation. Since multiple test forms are often developed simultaneously, more flagged items will take more time and cost.

2. Once item dependence is identified, recursive purification processes may be performed to improve the accuracy of estimation because dependent items are part of the estimation process for some LID indices. The circularity of refinement steps is similar to the implementation for Mantel-Haenszel index (Dorans & Holland, 1993) or logistic regression method (French & Maller, 2007) in differential item functioning (DIF) analysis. After re-scaling, LID assessment will be performed again to review the impact.

3. Psychometricians might be more interested in finding association within a cluster, such as the items under a reading passage, not just a pair. Therefore, besides checking the magnitude of an LID index, the
possible links from the other pairs may be considered, too. The uncertainty often requires more time to judge and to diagnose.

4. The number of pairs grows exponentially as the number of items increases. Even, the indices can be sorted first, the process of examining the indices for the pairs may be time consuming and confusion.

TREE DENDROGRAM FOR LID ASSESSMENT

An LID index indicates the association between only two items—a pair. However, procedures that are common to utilize the LID indices are limited due to the lack of clear interpretation of the strength of dependency. Therefore, psychometricians and context experts wish to have a graphical approach to provide a holistic view of the dependency structure for the whole items in a test form, such that a psychometrician can effectively flag the clusters of dependent items. Hsu & Tsai (2009) first suggested that a tree dendrogram could be used to visualize the improvement of LID purification process with methods based on nonparametric item response theory. The approach could be generalized as a standard procedure in test development to illustrate the overall structure of dependency for a test form, not just pair-wise association.

A tree dendrogram is a diagram to illustrate a hierarchy of categories using certain distance metric to represent cluster dissimilarity. Generating a simple tree dendrogram is very straightforward with SAS/STAT® CLUSTER and TREE procedures as long as the data type is arranged as a distance matrix. Since the indices indicate the association of two items, but the distance data set for PROC CLUSTER represents dissimilarity, it is necessary to convert the LID indices into certain quantities to represent the dissimilarity for constructing tree dendrogram.

First, Pearson’s $X^2$ (Chen & Thissen, 1997) index was transformed into contingency coefficient to interpret measure of strength of association. Specifically, it is

$$\text{Pearson’s contingency coefficient} = \sqrt{\frac{X^2}{X^2 + N}}$$

where $N$ is the total sample size. Then use one minus the value as the distance measure. $Q_3$ index (Yen, 1984, 1993) is the correlation coefficient of the item pair residuals. An optional transformation is to use Fisher’s $z$-transformation from normal theory on $Q_3$ (Yen, 1993):

$$\frac{1}{2} \ln \left( \frac{1 + Q_3}{1 - Q_3} \right)$$

However, this study adopted one minus the absolute of $Q_3$ as the measure. For the difference of observed and expected mutual information entropies (Tsai & Hsu, 2005), one minus the index value was used because they are positive numbers in the interval between zero and one.

The following code is an illustrative example to show the steps. Assume the data set $q4c$ has two columns, $x$ and $y$, as item numbers and the one column, $q3$, the value of LID index. The macro `Mfm` uses SAS/IML® to re-arrange the data into matrix form. After attaching the sequence of item number as `id`, the data set for PROC CLUSTER is created by specifying `type=distance` in the data step. The code example uses average linkage clustering method in the PROC CLUSTER procedure. Generally, other methods will generate similar results in LID assessment applications. The output SAS data set from the PROC CLUSTER procedure is rendered by the following PROC TREE procedure. Using the same statements as in SAS/GRAPH® to specify axis properties, the PROC TREE procedure creates the final tree dendrogram.

```sas
%macro namesx(name=,ni=);
%do i=1 %to &ni;
  &name&i
%end;
%mend namesx;

%macro Mfm(nx=,ny=,nix=,INF=,OUTF=,lix=);
proc means data=&INF noprint;
  var &nx &ny;
  output out=dTmpl max=xmax ymax;
%mend Mfm;
```
DATA SIMULATION

Local independent responses of a 40-item test were generated first. The LID condition was then simulated based on the same data under null condition. The scenario of item chaining was modeled and four pairs of dependent items were simulated. Typical example of this scenario arises with correlated items in some tests where an examinee’s obtains the partial information from the first item will lead to a correct answer of the second item. The scenario is common in psychological or educational assessments due to the use of item blocks or multi-step items.

The two graphs in figure 1 illustrate both LID indices based Person’s $X^2$ and information entropy, respectively. There are 780 pairs for a 40-item test. PROC GPLOT from SAS/GRAPH® was used to create the graphs. Since this is a simulation, the dependency is simulated to be strong and the graphs can reveal the dependent pairs. A tree dendrogram will be able provide a holistic view of the structure.

The left graph in figure 2 is the tree dendrogram of the same data using entropy based LID index with average clustering method. The dependent item pairs can be identified easily. The right graph illustrates the improvement of an LID purification process. Namely, the four pairs were removed in the re-scaling process.
A REAL LID ASSESSMENT EXAMPLE

Figure 3 shows an LID assessment example using a real data set. $Q_3$ was used in the example. The data set was derived from a reading test collected by General Educational Development (GED) Testing Service in 2008 Item Tryout Study. In practice, the real data are often not as clean as the simulation data. After sorting the $Q_3$ indices, it is still not so easy to pick the items. Reading test has several item blocks. When seeing two items with moderate index values but were on two different blocks, people may often wonder and confuse. In fact, the context experts did identify items in Math and Writing tests that a test taker may gain information from another block in the past analysis. The tree dendrogram can help us eliminate the confusion of the first case and alert the potential dependency of the second case. In the right graph, items 45, 46, and 48 are in the same cluster. Indeed, they are
within the same reading passage. Without the dendrogram, it will be rather tricky to find the item triplet, such as to observe a pair first, and to obtain some hints from the other pairs to make the link.

**Figure 3. LID indices and tree dendrogram of a reading test**

**CONCLUSIONS**

Constructing dendrogram plots is an effectively way to illustrate the structure of item dependency for a test form. GED Testing Service started to use tree dendrogram to assist LID assessment for analyzing tryout data collected for developing the next series test. Using the CLUSTER and TREE procedures, dendrograms were generated automatically for psychometricians and content experts to review. Not much programming effort is required, but the time saving is huge especially many test subjects and test forms are being analyzed. The graph provides a holistic view to quickly flag dependent items for further investigation, identify possible clusters that warrant attention, reveal the relationship among anchor items and new items, and visualize the overall dependency in a single view. Further, the unexpected adjustment and re-scaling work can be reduced. It is feasible and promising to integrating the use of tree dendrogram into LID assessment practice or in assessment software application to help testing service organizations to lower operational cost.

**REFERENCES:**


ACKNOWLEDGMENTS
SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are registered trademarks or trademarks of their respective companies.
The authors would like to thank Tsung-hsun Tsai for his advice of this paper.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:

Yung-chen Hsu
General Educational Development Testing Service
American Council on Education
One Dupont Circle NW, Suite 250
Washington, DC 20036-1163
Work Phone: (202) 659-8875
Email: yung-chen_hsu@ace.nche.edu