Measuring the Power of Learning.™

Research Memorandum
ETS RM–20-01

A Program to Group Test Takers Identified by Multiple Security Analyses

Shelby J. Haberman

Yi-Hsuan Lee

March 2020

ETS Research Memorandum Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Senior Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Director

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Consultant

Priya Kannan
Managing Research Scientist

Sooyeon Kim
Principal Psychometrician

Anastassia Loukina
Research Scientist

John Mazzeo
Distinguished Presidential Appointee

Donald Powers
Principal Research Scientist

Gautam Puhan
Principal Psychometrician

Jesse Sparks
Research Scientist

Elizabeth Stone
Research Scientist

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ariela Katz
Proofreader

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

A Program to Group Test Takers Identified by Multiple Security Analyses

Shelby J. Haberman
Consultant, Jerusalem, Israel

Yi-Hsuan Lee
Educational Testing Service, Princeton, New Jersey

March 2020

Corresponding author: Y.-H. Lee, E-mail: ylee@ets.org

Suggested citation: Haberman, S. J., & Lee, Y.-H. (2020). *A program to group test takers identified by multiple security analyses* (Research Memorandum No. RM-20-01). Educational Testing Service.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/

To obtain a copy of an ETS research report, please visit http://www.ets.org/research/contact.html

Action Editor: Sooyeon Kim

Reviewers: Carol Eckerly and Shuhong Li

Copyright © 2020 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and MEASURING THE POWER OF LEARNING. are registered trademarks of Educational Testing Service (ETS). All other trademarks are the property of their respective owners.

Abstract

The program CONNECT is constructed to treat a problem in security analysis, although other applications relevant to educational testing are also possible. In the basic application of the program, data are available from separate analyses applied to different test sections to identify incorrect answer keys for the section employed exactly or nearly exactly by a group of test takers. It is often true that for different test sections, the groups of test takers identified are not identical but do overlap. The problem is to employ all the available data to identify groups of test takers who appear to have interacted on some sections.

Key words: security analysis, linking items

Acknowledgments

The authors thank Carol Eckerly, Sooyeon Kim, and Shuhong Li for their suggestions on the report. The work was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D170026. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

Background

In matching analysis used for several testing programs (Haberman & Lee, 2017), incorrect answer keys for test sections are identified with an unusual number of test takers who follow the key exactly or nearly exactly. Such keys may be obtained via Internet or via one test taker looking over another's shoulder in a testing center. The analysis leads to identification of disjoint groups of test takers each associated with a specific key. When multiple test sections exist, it is common for such keys to be identified for more than one section. A natural issue is the relationship between groups identified for each section. For example, for one section, Test Takers 1, 2, and 3 might be found associated with the same answer key; however, for the other section, a common key is associated with Test Takers 3, 4, and 5. Given that Test Taker 3 is common to keys for both groups, there is obvious reason to suspect that Test Takers 1, 2, 3, 4, and 5 have some common connection. In practice, far more complex situations may be encountered for which identifications of test takers connected to each other via answer keys for different sections is relatively challenging without a suitable computer program.

Example 1

For a relatively simple illustrative example, consider example1.csv.[1] For privacy reasons, test takers are identified only by person numbers that do not reflect their actual identities and the testing program and test sections are not identified. The 101 rows have labels in the set V of 101 persons found associated with different keys for one or more test sections. The leftmost column for Row v in V provides the row label v. Thus the first row in the leftmost column has the entry person1. The remaining set of columns is $W = \{B, C, D\}$. These columns correspond to three test sections. In the row for test taker v in V, if an entry is present in the column for section w in W, then that entry is a member $s(v, w)$ of V that corresponds to a test taker who represents the group of test takers associated with a key that v is suspected to have used. The test taker $s(v, w)$ may be described as a source, although that person may not have actually provided the answers to the test takers in the group. For example, in the first row, person1 appears to have employed the same answer key as person72 on Section D. One test taker may have multiple sources for different sections. For example, the row for person7 lists that individual as the source for Section B, has source person48 for Section C, and has source person51 for Section D. The CONNECT program[2] identifies test takers with apparent connections to each other via common sources across Sections

B, C, and D.

Example 2

The same kind of data can appear in a very different context. Consider example2.csv. Here linkage of test items is considered for multiple test administrations. To preserve test security, item codes are changed, administrations are not identified, and not all items to be linked are included in the table. The table has 25 rows. The set V of row labels are item codes, so that the leftmost column of the row for Item1 is Item1. The remaining 15 columns correspond to a set W of test administrations B to P to which the item possibly can be linked. If not blank, the row for the item v in V and the column for the administration w in W corresponds to a representative item $s(v, w)$ in V that appears in Administration w. For example, in the row for Item1 and column for Administration E, Item1 appears. Here CONNECT shows what items can be linked to each other across Administrations B to P.

In general, the data used form a table with n rows associated with a set V of $n > 1$ members, a column that displays v for the corresponding row, and a set W of m other columns. In the language of graph theory (Harary, 1973), V is the set of vertices. The set S is the set of pairs (v, w) such that the row associated with v in V and the column associated with w in W has an entry $s(v, w)$ in V. In graph theory, the set $\mathcal{E}$ of edges consists of the sets $\{v, s(v, w)\}$, (v, w) in S, and the corresponding unordered graph is defined by V and $\mathcal{E}$. This memorandum exploits the notion of a connected graph.

Definition 1

Two members v_1 and v_2 of V are not connected if disjoint subsets V_1 and V_2 of V satisfy the following conditions:

1. The union of V_1 and V_2 is V.

2. v_1 is in V_1 and v_2 is in V_2.

3. If v is in V_1, w is in W, and (v, w) is in S, then $s(v, w)$ is in V_1.

4. If v is in V_2, w is in W, and (v, w) is in S, then $s(v, w)$ is in V_2.

Otherwise, v_1 and v_2 are connected.

Connectivity is an equivalence relationship. The relationship is reflexive in the sense that any member v of V is connected to itself, for v cannot belong to two disjoint sets. The relationship is symmetric due to the symmetry of the definition. Thus v_1 and v_2 in V are connected if, and only if, v_2 and v_1 are connected. The relationship is transitive. Let v_1 and v_3 in V be connected, and let v_3 and v_2 in V be connected. If v_1 and v_2 are not connected, then a contradiction results. Define V_1 and V_2 as in Definition . Either v_3 is in V_1 and therefore v_3 and v_2 are not connected or v_3 is in V_2 and v_3 and v_1 are not connected. Thus v_1 and v_2 are connected. As is true for any equivalence relationship, a class $\mathcal{C}$ of disjoint subsets of V is defined so that every member v in V is in a member $C(v)$ of $\mathcal{C}$, and v_1 and v_2 in V are connected if, and only if, they are in the same member $C(v_1) = C(v_2)$ of $\mathcal{C}$. Each member C of $\mathcal{C}$ is said to be connected, and V itself is connected if $\mathcal{C}$ has the single member V. The program CONNECT determines the class $\mathcal{C}$.

Definition 1 also implies that v in V and $s(v, w)$ in V are connected whenever w is in W and the pair (v, w) is in S. For example, in example1.csv person52 and person68 are connected. In this case, the set $\{\text{person52}, \text{person68}\}$ is in $\mathcal{C}$, for person52 is the only test taker connected to person68.

Although the CONNECT program has terminology and data files designed for its use for test security, the program can be used for very different applications. The data in example2.csv provides one case; however, it should be noted that connectivity also arise in the study of log-linear models for incomplete contingency tables (Fienberg, 1972; Goodman, 1968). The next section provides instructions for use of the program. The last section provides a detailed discussion of the analysis of the data in example1.csv.

Program Instructions

CONNECT is a program written in Fortran 95 that is invoked via a command-line prompt. The program is used together with an instruction file. If the instruction file is example2.txt then the command is

```
connect<example2.txt
```

The instruction file contains a single namelist statement. The statement begins with an ampersand followed immediately by the group name parameters and then followed by a space. The remainder of the statement includes pairs of namelist variable names and variable values. The order of the namelist variables does not matter. An example is provided by example2.txt.

The following variables are used in this statement:

- infile
- outfile
- nkeys
- width

infile

The name of the input file. This file is a csv file, so that variable entries are separated by commas. In example2.txt, the file is example2.csv. Names cannot contain more than 32 characters.

outfile

The name of the output file. This file is a csv file. In example2.txt, the name is example2out.csv.

nkeys

The value of m. If nkeys is omitted, then m is 1. Values of m less than 1 are regarded as errors. The label nkeys is based on the program use in security analysis in which the $s(v, w)$ are individuals associated with unauthorized answer keys. In example2.txt, nkeys is 15 because the set W of test administrations has 15 members.

width

The number of characters in a label for a member of V. This value is fixed and should be specified. Values less than 1 are regarded as errors.

In example2.txt, width is 6 because the names of items have 6 characters. Note that use of a text editor for reading example2.csv shows that the entry for Item1 is “Item1 ” rather than “Item1” and “   ” is the entry for a blank.

The input file specified by infile is a comma-separated file with rows that correspond to the members v of V. The leftmost column is the name for v, and the remaining nkeys columns are the names for the $s(v, w)$ for w in W. If (v, w) is not in S, then the name is a blank entry. In example2.csv, the name “Item1 ” is the first column. The next three columns are “   ” because (v, w) is not in S for the corresponding w in W. “Item1 ” is the first column because this item does appear in the corresponding test administration.

The output file specified by outfile is also a comma-separated file. The format is very similar to the format of the file specified by infile; however, an extra column is inserted at the extreme left that provides a representative member $r(C)$ of each member C of $\mathcal{C}$. Rows are rearranged so that members of a set C in $\mathcal{C}$ appear together. For example, in example2out.csv, $\mathcal{C}$ consists of the single set V and $r(V)$ is Item1. In the leftmost column, Item1 then appears in all rows. In the first line in the file, the entries are the same as in the first line of example2.csv, except they are moved right one column. The ordering of rows is changed. For example, the second line of example2out.csv corresponds to Item10 rather than to Item2. For score linking, the important result of the analysis is that all items can be linked by use of the available administrations.

In the case of example1out.csv, which corresponds to example1.csv, $\mathcal{C}$ has 5 members that are represented by person1, person2, person90, person52, and person50. The rearrangement of table rows is therefore somewhat more complex. The next section provides a detailed discussion of this example.

An Example of Security Analysis

Consider the data in example1.csv. The format of this file has been discussed already in the Example 1 section. In this case, the instruction file is example1.txt. Here infile is example1.csv and outfile is example1out.csv. The three test sections and therefore three possible key sources are indicated by setting nkeys equal 3. The name widths are given by width, which is 9. This width is chosen to accommodate person100 and person101. This example has a mixture of cases, for $\mathcal{C}$ has 5 members. The largest member of $\mathcal{C}$ includes 76 test takers (represented by person1), the second largest has 19 test takers (represented by person2), and the remaining three cases just have two test takers. The first two cases likely involve actions somewhat more involved than two people communicating with each other.

References

Fienberg, S. E. (1972). The analysis of incomplete multi-way contingency tables. *Biometrics, 28*(1), 177–202. https://doi.org/10.2307/2528967

Goodman, L. A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association, 63*(324), 1091–1131. https://doi.org/10.1080/01621459.1968.10480916

Haberman, S. J., & Lee, Y.-H. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses* (Research Report No. RR-17-23). Educational Testing Service. https://doi.org/10.1002/ets2.12150

Harary, F. (1973). *Graph theory.* Addison-Wesley.

Notes

[1] Example files discussed in this report can be obtained by downloading them from https://www.ets.org/Media/Research/RM-20-01-examplefiles.zip. You can also obtain a copy of the zipped file by emailing researchreports@ets.org.

[2] The CONNECT program discussed in this report, as well as the documentation, is available by contacting Jeff Wright at ETS at jwright@ets.org.