Detecting Answer Similarity Using Nonparametric Item Response Models

Xi Wang, Wonsuk Kim, Louis Roussos

Measured Progress
Dover, NH

Presentation at Conference of Test Security 2017

Introduction

Answer Similarity Analysis

- Pairwise comparison

  Source responses: ABDCADBCA
  Copier responses: ABCDCABCD
- Detection of answer copying, teacher/school administrator intervention
- Two classes of detection statistics:
  - Known response model for regular response processes
  - Model-based vs Non-model based

Introduction

Answer Similarity Analysis

- Two classes of detection statistics:

  Non model-based: K-index (*Holland, 1996*) and variations (*Sotaridona & Meijer, 2002; Sotaridona & Meijer, 2003*), Kappa(*Sotaridona et al., 2006*)

  Model-based: ω-index (*Wollack, 1997*), g_2 statistic (*Frary et al., 1977*), S-Check statistic (*Wesolowsky, 2000*), generalized binomial test (GBT; *van der Linden & Sotaridona, 2006*), and the M_4 statistic (*Maynes, 2014*).

Introduction

Answer Similarity Analysis

- Nominal response model
  - Unstable estimation: non-convergence, large parameter estimates for low-discriminating items
  - Poor model fit
- Non-parametric response model
  - Fewer assumptions, more flexible
  - Under mild assumptions, the curved smoothed “ICC estimates and ordinal ability estimates simultaneously converge to their true values” (Douglas, 1997)

Introduction

Research Purpose

- Evaluate the statistical properties of two well-known model-based statistics, ω and GBT, when the nonparametric estimation is used.
- Detection level: pair-level and group level

Detection Statistics

ω-index

- M_{CS}: total # of matched responses; $M_{cs} = \sum_{i=1}^{n} I_{csi}$, I_{csi} is an indicator function whether *c* and *s* have a matching response on item *i*.
- $E(M_{CS}|U_S) = \sum_i P(U_{iC} = u_{iS}|\theta_c, \boldsymbol{u_S})$; sum of the probability that a copier chooses the same answer as the source given the copier's ability; $P(U_{iC} = u_{iS}|\theta_c, \boldsymbol{u_S})$ estimated nonparametrically.
- $\sigma_{M_{CS}|U_S} = \sum_i P(U_{iC} = u_{is}|\theta_C, \boldsymbol{u_S})(1 - P(U_{iC} = u_{is}|\theta_C, \boldsymbol{u_S}))$
- $\omega = \frac{M_{CS} - E(M_{CS}|U_S)}{\sigma_{M_{CS}|U_S}}$~*N*(0,1) asymptotically

Detection Statistics

Generalized Binomial Test (GBT)

- $M_{CS} = \sum_{i=1}^{n} I_{csi}$;
- Under H_0,

$$P_i(I_{cs} = 1) = \sum_{k=1}^{K} P_i(U_C = U_S = k|\theta_C, \theta_S)$$

$$= \sum_{k=1}^{K} P_i(U_S = k|\theta_S) P_i(U_c = k|\theta_C)$$

- M_{cs} is the sum of independent Bernoulli random variables. It follows generalized binomial distribution
- Probability Density Function: Lord-Wingersky recursive formula

measured progress

Group-level Detection

Step 1: Pair-level detection

- Detection on each possible pair of examinees in a group

Step 2: Group-level detection

- Compute the # of detected pairs in each group (N_F); N_F~Binom(N_P,α), where N_P is the total number of pairs analyzed in the group, ignoring dependence among pairs.
- Find the critical value corresponding to right-tailed *p*-value of 0.05, and compare it to N_F. If N_F exceeds the critical value, the group is flagged.

Nonparametric Estimation

- Kernel Smoothing

$$\hat{p}(U = k|\theta) = \sum_{j=1}^{n} w_j(\theta) I(U_j = k)$$

- Bandwidth: fixed bandwidth $h = 1.06\sigma_\theta n^{-1/5}$, rule of thumb of Silverman(1986).
- Kernel function: standard normal distribution
- Ability estimation: rank-order of total test scores, rank-order is transformed to the quantile under *N*(0,1).

Simulation Design

Pair-level Type-I error

- Generating model: nominal response model.
- Generating item parameters: 40 items from a large-scale state assessment.
- Generating ability pairs: 36 pairs.
- Data generation replicated for 500 times

	Source θ				
		-1.5	-1	…	2
Copier θ	-2				
	-1.5				
	⋮				
	1.5				

measured progress

Simulation Design

Pair-level Power

1) Source provides all item responses to the copier, and the copier accepts all without thinking.
   - S and C have matching responses on ALL items (both incorrect and correct)
2) Source and copier collaborate so that the copier can get help with any item that the copier has difficulty with.
   - S and C have matching responses on ALL Incorrect items

measured progress

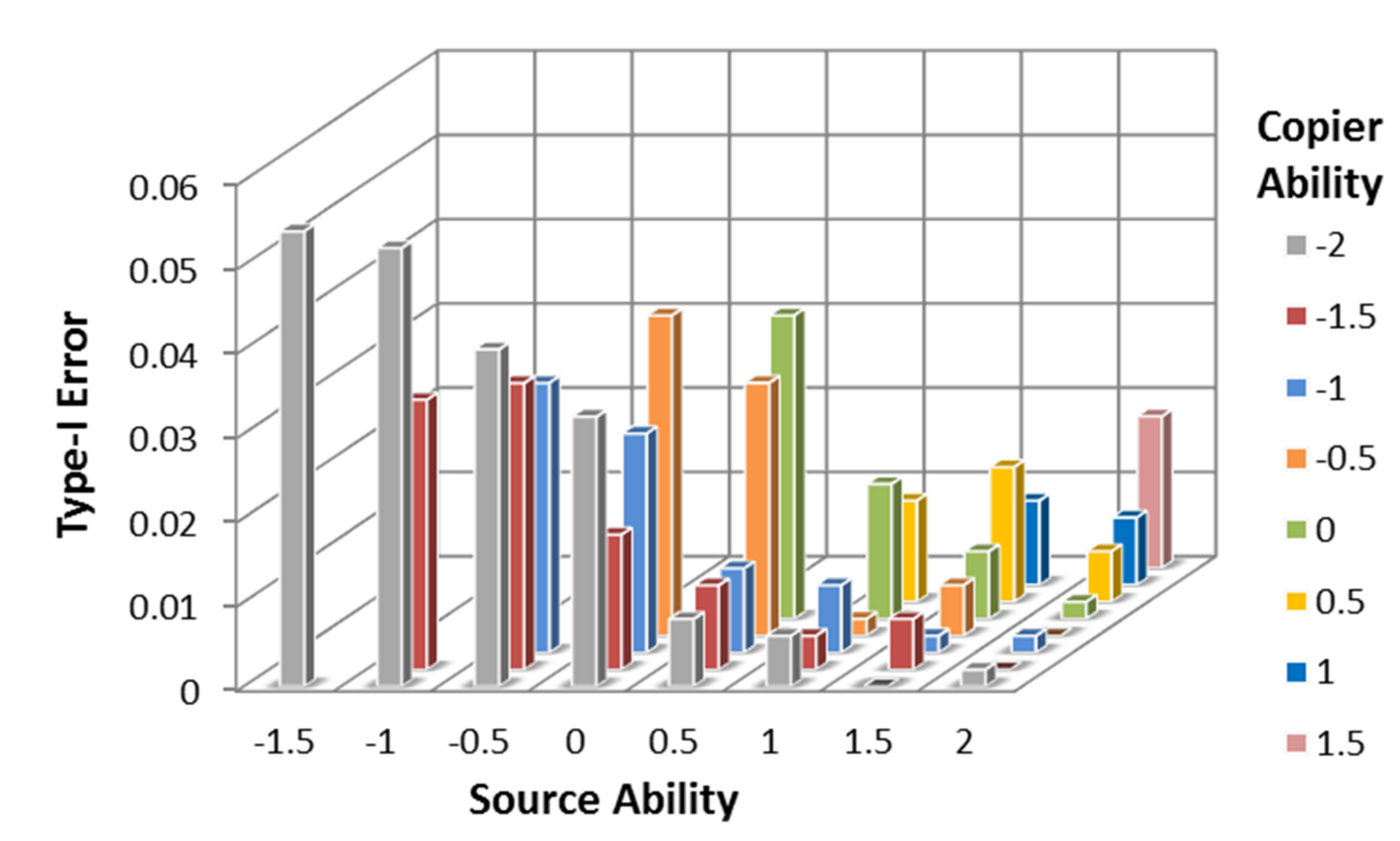
Simulation Design

Pair-level Power

3) Copier gets answers for items s/he has difficulty with by looking at the source's answer whenever possible.
   - S and C have matching responses on X% Incorrect items (X=20,40)

Pair-level Results

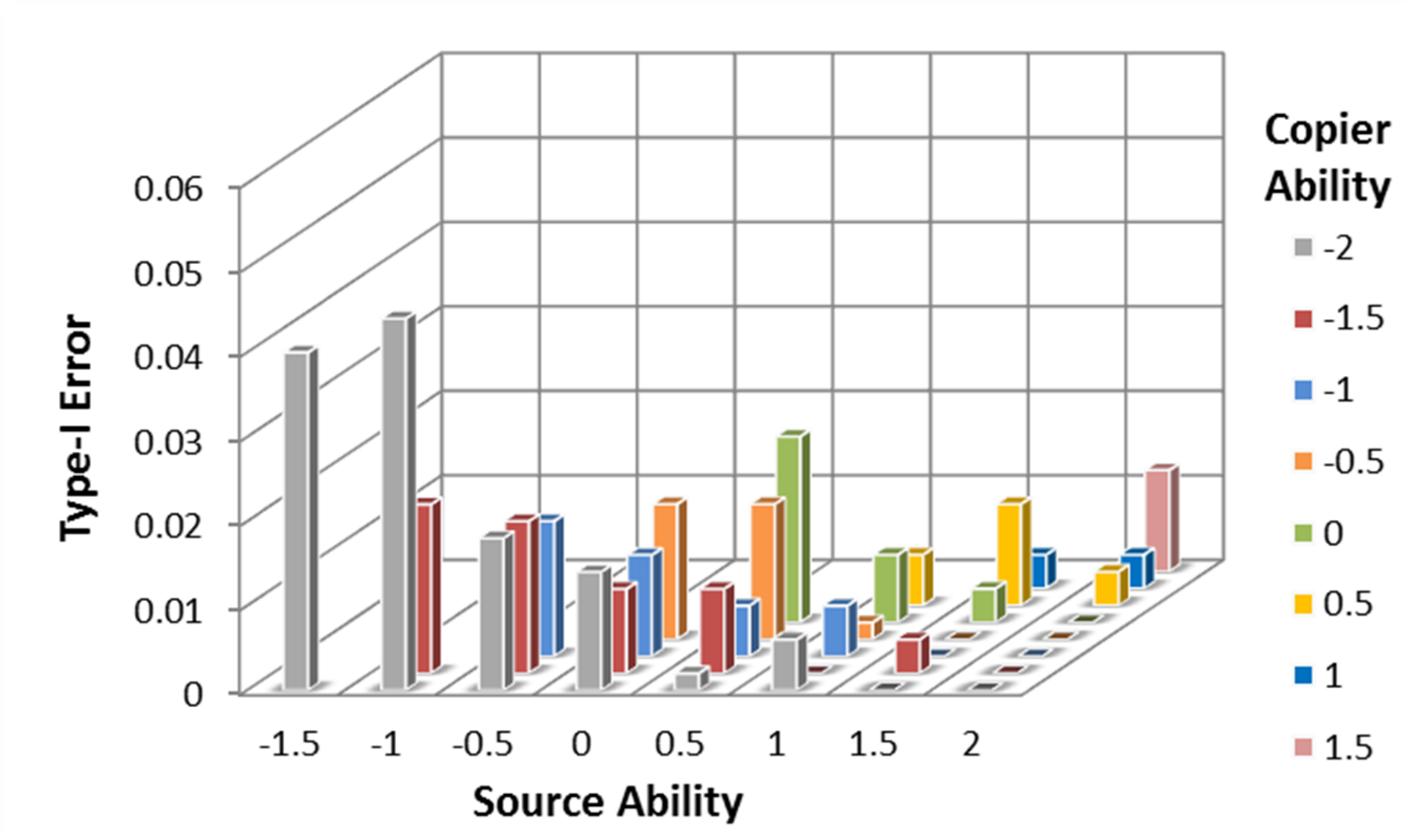
- Type-I error for Omega



measured progress.

Pair-level Results

- Type-I error for GBT



measured progress.

Pair-level Results

- Pair-level power for detecting exact matching: Omega

		Source θ							
		-1.5	-1	-0.5	0	0.5	1	1.5	2
Copier θ	-2	1	1	1	1	1	1	1	0.99
	-1.5		1	1	1	1	1	1	0.99
	-1			1	1	1	1	1	0.99
	-0.5				1	1	1	1	0.99
	0					1	1	1	0.99
	0.5						1	1	0.99
	1							1	0.99
	1.5								0.99

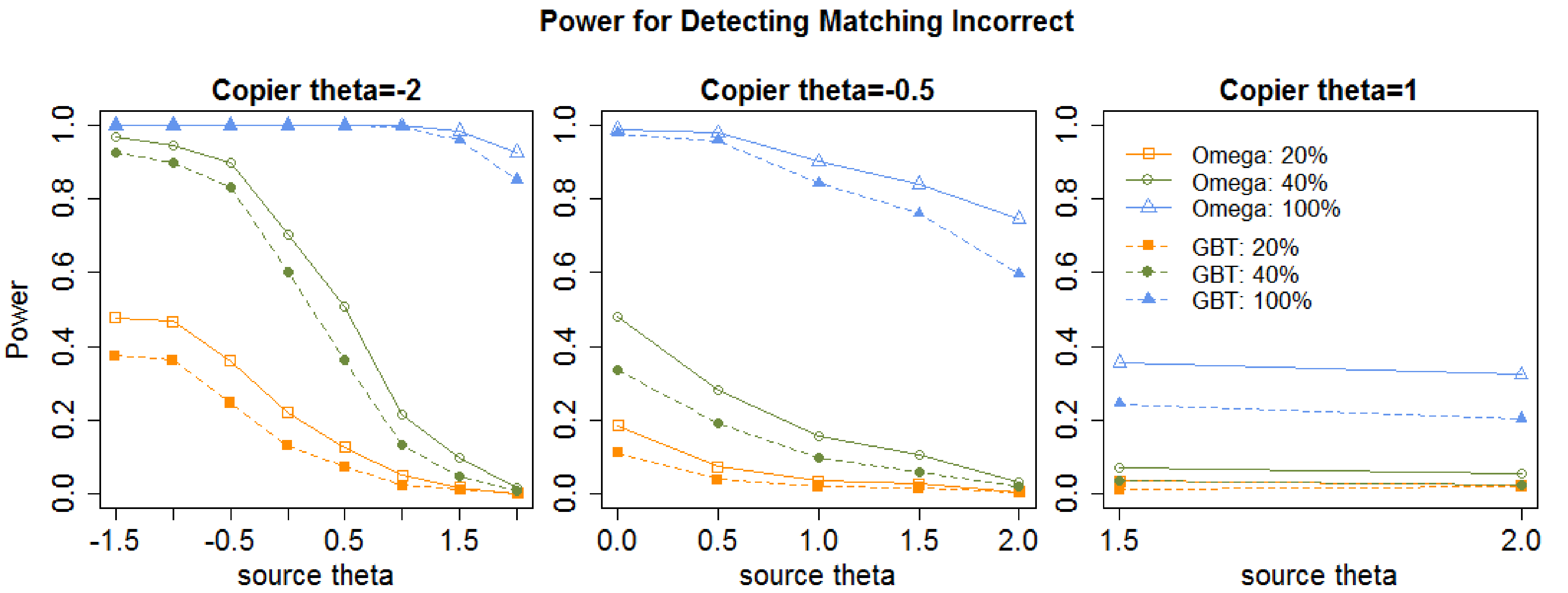
Pair-level Results

- Pair-level power for detecting exact matching: GBT

		Source θ							
		-1.5	-1	-0.5	0	0.5	1	1.5	2
Copier θ	-2	1	1	1	1	1	1	0.99	0.92
	-1.5		1	1	1	1	1	0.99	0.92
	-1			1	1	1	1	0.99	0.92
	-0.5				1	1	1	0.99	0.92
	0					1	1	0.99	0.92
	0.5						1	0.99	0.92
	1							0.99	0.92
	1.5								0.92

Pair-level Results

- Pair-level power for detecting matching incorrect



Simulation Design

Group-level Type-I error

- Group size: 25, 50
  - mimics the size of a class or the number of students associated with one teacher
- $\theta \sim N(-0.4, 1)$.
  - -0.4 is the "below proficient" cut in a real testing program. Centering at -0.4 mimics a low-ability class with 50% students being below proficient.

Simulation Design

Group-level Power

1) Answer change after test administration

   1.1. Lucky cheating: The teacher has sufficient time to change responses of all students.

   Simulation: changes on 25% items, randomly selected from the more difficult half of the test.

   1.2. Smart cheating: teacher only changes answers on 30% students to avoid being caught.

   Simulation: changes on a certain number of incorrect items such that student's raw score is above the raw score cut for "proficient"

Simulation Design

Group-level Power

2) Item Exposure

- The teacher exposes 25% items before the test administration to all students in the class.
- Simulation: Introduce $\Delta\theta$ to each student on the exposed items. $\Delta\theta \sim unif(0.1, 0.77)$, where 0.77 is the difference between two adjacent cut scores.

Group-level Results

- Group-level Type-I error and Power

		Detect Rate		% Detected Pairs	
		Omega	GBT	Omega	GBT
N=25	Type-I error	0	0	0.025	0.015
	Lucky	1	0	0.167	0.040
	Smart	0	0	0.020	0.007
	Exposure	0	0	0.019	0.008
N=50	Type-I error	0	0	0.026	0.015
	Lucky	1	0.99	0.166	0.068
	Smart	0	0	0.026	0.008
	Exposure	0	0	0.024	0.012

Discussions

- Type-I error

  Conservative both at pair-level and at group-level. Conservativeness can be a desired property in practice.

- Power

  (1) High power to detect extreme cheating (exact matching);

  (2) Largely affected by the proportion of matching responses.

  (3) Omega has larger type-I error and slightly larger power than GBT

Thank you.

Wang.xi@measuredprogress.org