The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.1.081
26(1), 81–92

A New Statistical Index for Detecting Cheaters on Multiple Choice Tests

Eun Su Han[a] · Johan Lim[b] · Kyeong Eun Lee[c,1]

[a]Division of Planning and Research, Korea National Institute of Health
[b]Department of Statistics, Seoul National University
[c]Department of Statistics, Kyungpook National University

(Received November 7, 2012; Revised December 3, 2012; Accepted December 28, 2012)

Abstract

It is important to construct a firm basis for accusing potential violators of academic integrity in order to avoid spurious accusations and false convictions. Educational researchers have developed many statistical methods that can either uncover or confirm cases of cheating on tests. However, most of them rely on simple correlation-based measures, and often fail to account for patterns in responses or answers. In this paper, we propose a new statistical index denoted by a Standardized Signed Entropy Similarity Score to resolve this difficulty. In addition, we apply the proposed method to analyze a real data set and compare the results to other existing methods.

Keywords: Detecting Cheaters, Angoff's B, Crawford's method, Error Similarity Analysis, K index, Standardized Signed Entropy Similarity Score.

1. 서론

본 연구가 시작된 동기는 미국의 ETS는 2000년대 초반 한국 수험자 62명의 GRE성적을 수험자들의 답안지를 바탕으로 이들을 부정행위자로 판단하고 시험 성적을 취소하는 사건으로 부터이다. 이러한 과정에서 수험자의 다중선택문항의 답을 바탕으로 부정행위를 결정하는 문제는 대립가설 "수험자가 부정행위를 하였다"에 대한 통계적 가설 검정의 문제를 이해 할 수 있다. 당시 통계학 전공자로서 기존의 접근방식이 궁금하였고 보다 정확하게는 ETS가 내부적으로 시행하는 절차가 궁금하여 이와 관련된 기존의 연구를 살펴볼 기회가 있었다.

이러한 부정행위의 발견과 예방은 학문적 진실성(academic integrity) 측면에서 매우 중요한 논점으로 교육측정(educational measurement)분야의 오랜 연구 주제중 하나이다. 특히 대부분 국가시험의 표준적 형태인 다항선택시험에서의 부정행위의 판단을 위한 다양한 측도, 구체적으로는 임의의 두 수험자의 답안의 유사성에 관한 측도에 대해 많은 연구가 있어왔다.

This research was supported by Kyungpook National University Research Fund 2011.
[1]Corresponding author: Assistant Professor, Department of Statistics, Kyungpook National University, 80 Daehakro, Buk-Gu, Daegu 706-701, Korea. E-mail: artlee@knu.ac.kr

Crawford (1930)는 두 수험자의 전체 오답수(각 수험자의 오답수의 합)에 대한 두수험자의 일치하는 오답수의 상대적인 비율을 구한 후 부정행위가 없는 모든 쌍들의 상대적인 오답비율의 평균과 표준편차를 이용해 표준화 시킨 지표를 제안하였고, Angoff (1974)는 두 수험자의 오답일치수(같은 오답을 선택한 문항수)를 표준화 시킨 지표를 개발하였다. Bellezza와 Bellezza (1989)은 두 수험자의 공통 오답수(오답의 일치와는 상관없이 같이 틀린 문항수) 중 두 수험자의 오답일치수의 분포를 이용하여 p-값을 계산하였고, 1979년 Frederick Kling이 처음 K-지표를 고안했으며 (발표된 논문이 없었음) Holland (1996)가 K-지표와 관련한 이론적인 것들을 발표하였다. K-지표는 관측된 두 수험자의 오답일치수보다 더 많은 오답일치수를 얻을 확률을 구하는 데, 이항 분포를 이용해서 이 확률을 구했다. 이러한 기존의 측도들은 난이도와 같은 문항의 특성이 반영되어 있지 못한 단점이 있다. 보다 구체적으로 이야기하면 A문항은 난이도가 높아 수험자이 모두 틀리는 문항이고 B문항은 난이도가 낮아 모두가 맞추는 문항이라 가정하자. 만일 두 수험자가 A문항을 같이 맞추고 B문항을 같이 다르게 답하면서 틀리는 경우가 그 반대의 경우 (B문항을 같이 맞추고 A문항을 다르게 답하며 두 수험자 모두 오답을 하는 경우)보다 유사성 측도가 높아야 한다. 이러한 기대와는 달리 기존의 연구들은 모든 문항들을 같은 비중을 두고 평균을 취하는 측도를 제한함으로써 문항의 난이도를 전혀 고려하지 못하게 된다. 당연히 위의 두 경우 기존의 측도들은 같은 유사성 측도 값을 제공하게 된다.

이러한 문제를 해결하기 위하여 본 논문에서는 부호화된 엔트로피(signed entropy)를 정의, 이용하여 문항의 특성을 고려한 새로운 유사성 측도를 제안한다. 각 문항의 부호화된 엔트로피는 해당 문항의 각 선택항목에 대하여 수험자들의 응답의 통일성을 나타내는 측도로 문항당 선택항목들에 대하여 비중을 정의한다. 예를 들어 문항 A의 선택항목 (가) 대하여 많은 수험자들이 답으로 표기한 경우 해당 선택항목에 적은 비중을 정의하고 적은 수험자들이 답으로 표기한 경우는 높은 비중을 배정한다. 유사성 측도를 정의 함에 있어서 두 수험자가 같은 항목에 답을 한 경우는 “+”의 점수를 다른 항목에 답을 한 경우는 “−”의 점수를 주고 모든 문항들의 점수의 합으로 유사성 측도를 정의 한다.

본 논문의 구성을 소개하면 다음과 같다. 2절에서는 기존의 여러 유사성 측도들을 간략하게 설명하고, 3절에서는 새로운 통계적 유사성 측도를 제안한다. 4절에서는 실제 자료에 대한 분석으로, 새로 제안한 유사성 측도와 다른 기존의 측도들을 이용하여 잠정적 부정행위자들을 찾으려고 한다. 마지막으로 5절에서는 결론 및 본 연구의 성과에 대해 논하려고 한다.

2. 기존의 유사성측도

이번 절에서는 Angoff (1974), Bellezza와 Bellezza (1989), Crawford (1930), 그리고 Kling (Holland, 1996)에 의하여 제안된 네 가지의 유사성 측도에 대하여 간단히 소개하려고 한다.

일반성을 잃지 않고 각 문항이 J개의 선택지가 있고 정답이 모두 1번인 K개의 문항으로 구성된 시험이라고 가정하자. 수험자 A와 B의 답안을 각각

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{J1} & a_{J2} & \cdots & a_{JK} \end{pmatrix} \text{ 와 } \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1K} \\ b_{21} & b_{22} & \cdots & b_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ b_{J1} & b_{J2} & \cdots & b_{JK} \end{pmatrix}$$

로 표기하자. 여기서 a_{jk}는 이항변수로 수험자 A가 k번째 문항에 대하여 j라 응답한 경우 1, 그렇지 않으면 0으로 정의하고 b_{jk}도 같은 방식으로 정의하자.

네가지 유사성 측도들을 정의하는데 사용되는 변수들을 소개하면 다음과 같다:

$$
\begin{aligned}
W_A &= \text{수험자 A의 오답수} = \sum_{k=1}^{K} \mathrm{I}\left(a_{1k} = 0\right), \\
W_B &= \text{수험자 B의 오답수} = \sum_{k=1}^{K} \mathrm{I}\left(b_{1k} = 0\right), \\
W_{AB} &= \text{두 수험자가 공통적으로 틀린 오답수(선택한 선택지에 상관없음)(공통오답수)} \\
&= \sum_{k=1}^{K} \mathrm{I}\left(a_{1k} = 0,\ b_{1k} = 0\right), \\
Q_{AB} &= \text{두 수험자가 같은 오답을 선택한 문항수 (오답일치수)} \\
&= \sum_{k=1}^{K} \sum_{j=2}^{J} \mathrm{I}\left(a_{jk} = 1,\ b_{jk} = 1\right).
\end{aligned}
$$

Crawford (1930)는 두 수험자의 오답일치수(Q)의 절대적인 값(absolute number)을 사용하는 것이 아니라 두 수험자들 ($W_A + W_B$)에 오답수에 대한 상대적 비율을 사용하였다. 부정행위를 하지 않은 수험자들(근처 자리가 아닌 수험자들)의 랜덤표본을 추출하여 각 쌍들의 오답비율의 평균과 두 수험자의 오답비율 차이 검정통계량을 이용한 유사성 측도를 제안 하였다.

$$
\begin{aligned}
z &= \frac{P - \bar{P}}{S_{P-\bar{P}}}, \\
P &= \text{주어진 두 수험자의 } \frac{Q_{AB}}{W_A + W_B}, \\
\bar{P} &= \text{부정행위의 혐의가 없는 모든 수험자 쌍들의 평균 } P, \\
S_{P-\bar{P}} &= \text{두 비율 차이의 표준오차,}
\end{aligned}
$$

여기서 검정 통계량 z는 정규분포를 따르므로, 표준정규분포표를 이용하여 p-값을 얻을 수 있다.

Angoff (1974)는 앞에서 정의된 변수들을 포함하여 12개의 변수들을 이용하여 "A"에서 "H"까지 8개의 지표들을 소개하고 비교 분석하여 최종적으로 B-지표와 H-지표를 선택하여 제안하였다. 이 두 지표는 1970년부터 ETS(Educational Testing Service)에서 오랫동안 성공적으로 사용되어졌다. 특히, Haney (1993)와 Buss와 Novick (1980)에 의해 높이 평가되었던 B-지표는 일종의 오답일치수(Q)를 표준화한 값이다. 오답일치수(Q)들의 분포들은 종종 오른쪽으로 꼬리가 길어서 정규분포가 아닌 경우가 많다. 그럼에도 불구하고 전체평균과 표준편차를 이용해서 표준화하게 되면 정규분포를 이용해서 p-값을 구할 수 없게 된다. 그래서, 오답일치수와 상관관계가 높은 오답의 곱 W_iW_j을 이용, 비슷한 W_iW_j들을 그룹화 한 후 각 그룹에서 Q의 분포를 보면 훨씬 정규분포와 가깝게 된다.

$$
\begin{aligned}
t &= \frac{\left(Q_{AB} - \bar{Q}_{W_iW_j}\right)}{S_{Q.W_iW_j}}, \\
\bar{Q}_{W_iW_j} &= W_AW_B\text{가 속한 } W_iW_j\text{그룹에서 } Q_{AB}\text{의 평균,} \\
S_{Q.W_iW_j} &= W_AW_B\text{가 속한 } W_iW_j\text{그룹에서 } Q_{AB}\text{의 표준오차.}
\end{aligned}
$$

Bellezza와 Bellezza (1989)의 오류 유사성 분석(Error Similarity Analysis; ESA)으로 더 잘 아려진 Bellezza와 Bellezza의 지표는 오답일치수 Q가 근사적으로 이항분포 (W_{AB}, P_{ESA})를 따르는 것을 이용

하여, 두 수험자가 우연히, 관측된 오답 일치수보다 같거나 더 많은 오답일치수를 가질 확률, 즉, $P(Q \geq Q_{AB})$이다. 여기서, ESA에서 사용된 P_{ESA}는 Crawford의 B-지표에서 사용된 $\bar{P}$와는 다른 것으로 부정행위가 없는 수험자들이 우연히 같은 오답을 선택할 확률의 추정치이다.

$$p = \sum_{i=Q_{AB}}^{W_{AB}} \frac{W_{AB}!}{i!(W_{AB}-i)!} P_{ESA}^{i}(1-P_{ESA})^{W_{AB}-i},$$

$p =$ 두 수험자가 우연히, 관측된 오답 일치수보다 같거나 더 많은 오답일치수를 가질 확률,

$P_{ESA} =$ 부정행위가 없는 두 수험자가 같은 오답을 선택할 확률의 추정치이다.

1979년 Frederick Kling이 발표된 논문이 없이 K-지표를 고안하였고 Holland (1996)가 K-지표와 관련한 이론들을 발표하였고 1980년대 초반에 ETS에 의해 많이 사용되어졌다. 이 지표는 아마도 Bellezzas와 Bellezzas (1989)에게 ESA의 아이디어를 제공한것으로 보여진다.

ESA나 K-지표는 P를 어떤 값을 사용했느냐에 따라 차이가 많이 난다. 작은 값의 P를 사용하면 작은 값의 p-값을 가지게 되므로 더 많은 쌍들이 유의하게 되어 위양성(false positive)을 증가시킬 수 있다. K-지표는 Bellezzas의 방법과 거의 유사하지만 W_{AB}대신 W_A를 사용했다는 점이 다르다.

$$p = \sum_{i=Q_{AB}}^{W_A} \frac{W_A!}{i!(W_A-i)!} P_K^{i}(1-P_K)^{W_A-i},$$

$P_K =$ 부정행위가 없는 두 수험자가 같은 오답을 선택할 확률의 추정치이다.

정의에 의해, 수험자 A의 오답수는 두 수험자의 공통오답수보다 크므로, 두 지표 모두 같은 P를 사용하게 되면, K-지표는 더 많은 항의 합이 p-값에 포함되어 ESA의 방법에 의한 p-값보다 커지게 된다.

3. 새로운 통계적 측도

N명 수험자는 K개 문제로 이루어진 시험을 치는 것을 가정한다. 게다가 각 문제는 J 선택지를 가지는 것을 가정한다. 그들의 응답들은 $\{\mathbf{X}_i\}$, 여기서 $i = 1, 2, \ldots, N$ 그리고 $\mathbf{X}_i$는 앞절에서 정의한 행렬이며 아래의 모수

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{J1} & p_{J2} & \cdots & p_{JK} \end{pmatrix}$$

를 가지는 함께 곱 다항분포(Product Multinomial distribution), 수험자들 사이의 독립성, 그리고 각 수험자들의 응답사이의 독립성을 가정한다. 실제로, 수험자들 사이의 독립성과 각 수험자들의 응답사이의 독립성에 대한 가정은 어느 정도 이상적일 수 있다. 몇 몇 수험자는 주요 문항에 대한 유사한 교육을 받았거나 혹은 시험을 준비하는 동안 함께 공부하였기 때문에 그들 사이의 종속은 나타날지도 모른다. 또한, 만약 시험에서 같은 주제와 관련하여 여러 문항들이 있는 경우, 즉, 문항들의 군집(clustering)이 있다면 문항들 사이에 종속이 있을 것이다. 문항들간의 종속성은 유사한 문제들을 그룹화 해서 그 그룹들을 단위화해서 다루어 질 수 있다. 또한 수험자들 중에서 종속은 단순곱다항분포(Simple Product Multinomial Distribution) 대신에 혼합곱다항분포(Mixture Product Multinomial Distribution) 사용하여 다루어 질 수 있다.

우리는 $\boldsymbol{X}$와 $\boldsymbol{Y}$ 사이의 유사성 정도를 엔트로피를 이용해서 정의하려고 한다. 두 변수 X와 Y의 엔트로피는 다음과 같이 정의된다.

$$\begin{aligned} H(X,Y) &= \mathrm{E}\left(-\log P(X,Y)\right) \\ &= \sum_{x,y}\{-\log P(X=x,Y=y)\}P(X=x,Y=y) \\ &= \sum_{x=y}\{-\log P(X=x,Y=x)\}P(X=x,Y=x)+\sum_{x\neq y}\{-\log P(X=x,Y=y)\}P(X=x,Y=y). \end{aligned}$$

우리는 여기서 $x \neq y$인 경우 $\log P(X=x,Y=y)$의 부호를 $-$ 대신 $+$ 부호를 바꾸어 부호가 있는 엔트로피(Signed Entropy)라고 정의하자.

$$\mathrm{SH}(X,Y) = \sum_{x=y}\{-\log P(X=x,Y=x)\}P(X=x,Y=x) + \sum_{x\neq y}\log P(X=x,Y=y)P(X=x,Y=y).$$

이것을 이용, 두 수험자 간의 응답, $\boldsymbol{X}$와 $\boldsymbol{Y}$의 유사성 척도를 부호가 있는 엔트로피 유사성 점수(Signed Entropy Similarity Score)라고 하고 다음과 같이 정의하자.

$$S(\mathbf{X},\mathbf{Y}) = \sum_{k=1}^{K}\sum_{j=1}^{J}(-2)\cdot\log(p_{jk})\cdot x_{jk}\cdot y_{jk} + \sum_{k=1}^{K}\sum_{j_1\neq j_2}\log(p_{j_1,k}\cdot p_{j_2,k})\cdot x_{jk_1}\cdot y_{jk_2}.$$

제안된 유사성 측도는 단순히 오답의 유사성을 측정하는 것이 아니라 대다수 응답으로 부터 벗어나는 문항들에 큰 가중치를 배정하게 되며 그러한 문항들에서 두 수험자의 일치하는 정도가 높게 되면 더 큰 값을 가지게 된다.

$S(\mathbf{X},\mathbf{Y})$를 벡터와 행렬을 이용해 다음의 식으로

$$S(\mathbf{X},\mathbf{Y}) = \sum_{k=1}^{K}\mathbf{x}_k'\mathbf{W}_k\mathbf{y}_k$$

표기 할 수 있으며, 여기서 $\mathbf{x}_k = (x_{1k},\ldots,x_{Jk})'$, $\mathbf{y}_k = (y_{1k},\ldots,y_{Jk})'$이며 $\mathbf{W}_k$는 $J\times J$ 행렬이며 (j_1,j_2)번째 원소는 다음과 같다:

$$[\mathbf{W}_k]_{j_1,j_2} = \begin{cases} -2\log p_{jk}, & \text{if } j_1 = j_2 = j, \\ \log p_{j_1k} + \log p_{j_2k}, & \text{if } j_1 \neq j_2. \end{cases}$$

$S(\mathbf{X},\mathbf{Y})$를 표준화하기 위해, $S(\mathbf{X},\mathbf{Y})$의 평균과 분산을 살펴보자. 먼저, $\mathbf{x}_k'\mathbf{W}_k\mathbf{y}_k$의 평균과 분산을 각각 μ_k와 σ_k^2라 하자. 그러면 다항분포의 성질과 두 변수의 독립성 때문에

$$\begin{aligned} \mu_k &= E\left(\mathbf{x}_k'\mathbf{W}_k\mathbf{y}_k\right) \\ &= E\left(\mathbf{x}_k\right)'\mathbf{W}_k\mathrm{E}\left(\mathbf{y}_\mathrm{k}\right) \\ &= -2\sum_{j=1}^{J}\log p_{jk}\cdot p_{jk}^2 + \sum_{j_1\neq j_2}\left(\log p_{j_1k}+\log p_{j_2k}\right)\cdot p_{j_1k}\cdot p_{j_2k}, \\ \sigma_k^2 &= \mathrm{Var}\left(\mathbf{x}_\mathrm{k}'\mathbf{W}_\mathrm{k}\mathbf{y}_\mathrm{k}\right) \\ &= E\left(\mathbf{x}_k'\mathbf{W}_k\mathbf{y}_k\right)^2 - \mu_k^2 \\ &= \sum_{j=1}^{J}4\left(\log p_{jk}\right)^2 p_{jk}^2 + \sum_{j_1\neq j_2}\left(\log p_{j_1k}+\log p_{j_2k}\right)^2\cdot p_{j_2k}\cdot p_{j_2k} - \mu_k^2 \end{aligned}$$

Table 4.1. Examinee seating plans

Room I							Room II						
1		8		14		21	28		35		42		48
2		9		15		22	29		36		43		49
3		10		16		23	30		37		44		50
4		11		17		24	31		38		45		51
5		12		18		25	32		39		46		52
6		13		19		26	33		40		47		
7				20		27	34		41				

이 된다. 그러면 $S(\mathbf{X}, \mathbf{Y})$의 평균과 분산은 문항들의 독립성 때문에 $\mu = \sum_{k=1}^{K} \mu_k$ 그리고 $\sigma^2 = \sum_{k=1}^{K} \sigma_k^2$이 된다.

Linderberg-Feller 정리에 의해, 우리는 어떤 $\delta > 0$대하여 $\sigma^{2*} = \lim_{K\to\infty} \sigma^2/K$ 그리고 $\min_{j,k} p_{jk} > \delta$일때, $(S(\mathbf{X}, \mathbf{Y}) - \mu)/\sqrt{K}$가 정규분포(Gaussian distribution)로 약수렴(Weakly converge)한다는 것을 알 수 있다. 따라서 잠재적 부정행위자들은 아래에 정의된 표준화된 부호가 있는 엔트로피 유사성 점수(Standardized Signed Entropy Similarity Score; SSESS)를 이용해 찾아질수 있다.

$$\mathrm{SS}(\mathbf{X}, \mathbf{Y}) = \frac{S(\mathbf{X}, \mathbf{Y}) - \mu}{\sigma}\Big|_{P=\hat{P}}$$

위의 통계량은 근사적으로 표준정규분포이다. $\hat{\mathbf{P}}$은 사전확률 Dirichlet$(1, 1, \ldots, 1)$과 PMD(Product Multinomial Distribution)에서 모수 $\mathbf{p}_k = (p_{1k}, \ldots, p_{Jk})'$의 베이즈 추정치이다.

$$\hat{p}_{jk} = \frac{n_{jk} + 1}{N + J},$$

여기서 $n_{jk} = k$번째 문제에 j를 답한 수험자들의 수이다. $\hat{\mathbf{P}}$은 아래의 행렬을 가진다.

$$\hat{\mathbf{P}} = \begin{pmatrix} \hat{p}_{11} & \hat{p}_{1,2} & \cdots & \hat{p}_{1K} \\ \hat{p}_{21} & \hat{p}_{2,2} & \cdots & \hat{p}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{p}_{J1} & \hat{p}_{J2} & \cdots & \hat{p}_{JK} \end{pmatrix}.$$

4. 사례

이번 절에서는 실제 시험 자료에 우리가 제안한 SSESS와 이전에 사용된 네가지 방법들을 적용하려고 한다. 이 자료는 2010년 1학기 때 경북대학교 통계학과 학생들의 전공기초과목의 기말고사 자료이다. 52명의 수험자들(모든 수험자들은 1∼52까지 숫자로 표시)이 강의실 두 곳(Table 4.1 참조)에서 나누어서 시험을 쳤다. 수험자들의 자리는 수험자들이 선택하였다.

시험은 총 80문항으로 이루어졌으며 이 중, 46문항이 사지선다형 문제이며, 34문항이 참(True), 거짓(False)문항이다. 전체 표본은 $m = (52 \times 51)/2 = 1326$ 쌍의 수험자들로 이루어져 있다. 평균점수는 약 45개정도이고 표준편차는 약 10개정도이다. 최고점수는 66개이고 최저점수는 22개이다. 점수의 분포를 보면 Figure 4.1에서 알 수 있듯이 중심부에 모여있음을 알 수 있다.

80문항들에 대한 수험자들의 답을 선택한 유형을 살펴보기 위하여 각 문항의 정답을 1이 되도록 변경

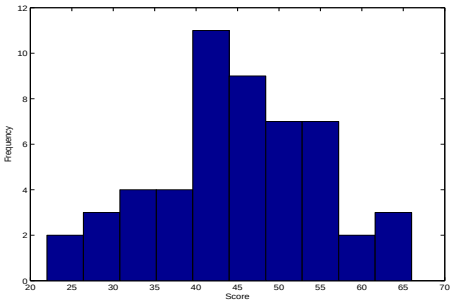

Figure 4.1. Histogram of test scores

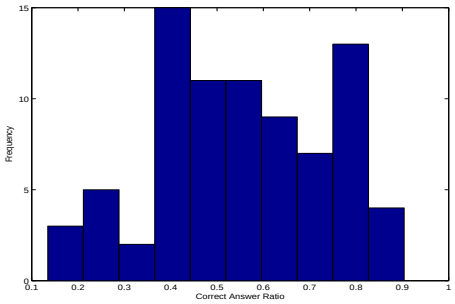

Figure 4.2. Histogram of examinees correctly answered ratios

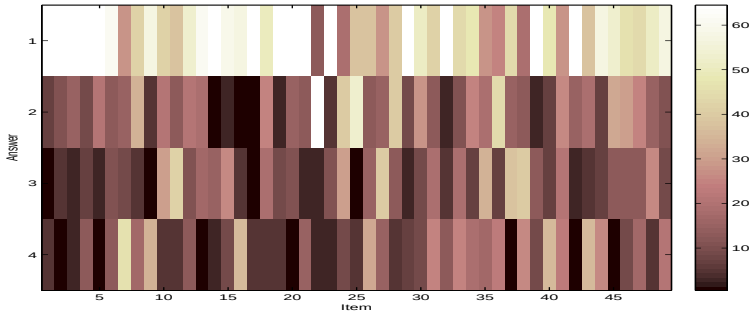

Figure 4.3. Heatmap of Four-Choice items

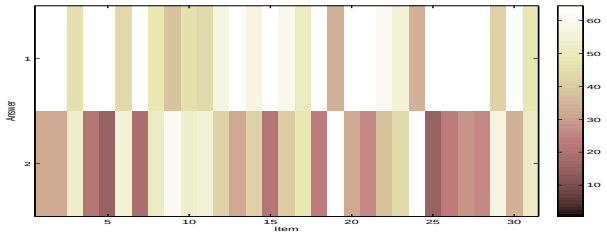

Figure 4.4. Heatmap of True-False items

하여 답에 대한 수험자들의 상대적 빈도수(×100)를 heatmap(밝을수록 빈도수가 높음)을 이용하였다 (Figure 4.3과 Figure 4.4 참조). 수험자들이 정답을 선택한 비율의 히스토그램 (Figure 4.2)에서 보듯이 3문항은 20% 미만의 수험자들이 맞추었으며 17문항은 75% 이상의 수험자들이 맞추기도 했다.

다중비교시 실험별 오류(Family Wise Error Rate)를 관리하기 위하여 본페로니 방법(Bonferroni's method)으로 교정된 $\alpha_{PC} = \alpha/m = 0.05/1326 = 0.000038$를 이용하였으며 Storey (2003)의 positive-False Discovery Rate(pFDR)인 q-값(q-value)도 참고로 명시하였다. 결과표는 총 1326쌍의 결과를 p-값의 오름차순으로 정렬한다음 상위 10쌍만 혹은 작은 p-값 가지는 쌍들을 본 논문에 게재 하였다.

이전에 연구된 4가지 방법의 결과는 Table 4.2, Table 4.3, Table 4.4, Table 4.5에 제시되어 있으며, 이것은 또한 3장에서 정의한 변수들의 값들과 p-값과 q-값의 결과이며, Table 4.6은 우리가 새롭게 제안한 Signed Entropy Similarity Score 방법의 결과인 p-값과 q-값을 제시하였다.

Table 4.2. Crawfold's test of proportions

Examinee Pair	W_A	W_B	W_{AB}	Q_{AB}	P	z	p-value	q-value
(13, 24)	37	35	24	21	0.292	3.38	0.00072	0.782
(11, 13)	28	37	20	18	0.277	3.02	0.00251	0.782
(20, 49)	39	49	25	24	0.273	2.92	0.00350	0.782
(5, 17)	36	14	6	2	0.040	−2.77	0.00557	0.782
(39, 47)	50	56	41	28	0.264	2.71	0.00673	0.782
(24, 25)	35	37	21	19	0.264	2.70	0.00686	0.782
(12, 17)	31	14	5	2	0.044	−2.66	0.00774	0.782
(3, 39)	16	50	10	3	0.045	−2.64	0.00833	0.782
(10, 28)	47	16	9	3	0.048	−2.59	0.00972	0.782
(10, 48)	47	38	26	22	0.259	2.58	0.00989	0.782

Table 4.3. Angoff's B Index

	Examinee Pair	W_A	W_B	W_{AB}	Q_{AB}	B Index	p-value	q-value
$G = 1$	(40, 47)	58	56	45	29	4.470	0.00001	0.003
	(39, 47)	50	56	41	28	4.223	0.00002	0.005
	(36, 47)	51	56	38	26	3.730	0.00019	0.020
	(40, 46)	58	48	42	26	3.730	0.00019	0.020
	(39, 40)	50	58	42	25	3.483	0.00050	0.034
	(40, 49)	58	49	35	25	3.483	0.00050	0.034
$G = 2$	(40, 47)	58	56	45	29	4.438	0.00001	0.005
	(39, 47)	50	56	41	28	4.153	0.00003	0.009
	(36, 47)	51	56	38	26	3.584	0.00034	0.047
	(40, 46)	58	48	42	26	3.584	0.00034	0.047
$G = 3$	(40, 47)	58	56	45	29	4.071	0.00005	0.038

사지 선다형 문항들과 달리, True/False 문항들의 답은 정답 아니면 오답이므로 수험자들의 오답 일치수를 과장시키는 경향이 있다. $\bar{P}$의 값을 사전에 고정시킨 경우, 수험자들의 오답일치수 Q_{AB}를 직접적으로 사용하는 Bellezza 방법이나 Kling 방법의 p-값을 더 작게 만들어서 오판단 할 가능성을 높이고 있다. Bellezza 방법에서, 오답일치수인 Q의 값이 바로 공통오답수(W_{AB})가 되므로 p-value ($P(X \geq Q_{AB})$)는 굉장히 작게 된다. 그래서 True/False 문제를 제외한 사지 선다형 문항들만 고려하여 계산하는 것이 더 합리적이다. 반면, 우리가 제안한 방법은 단순히 오답에 대한 일치도를 고려하는 것이 아니라, 응답들의 상대적인 일치도를 고려하고 있기 때문에, True/False 문항들을 제외시킬 이유가 없으며 단순히 확률 추정치 계산에서 분모의 K-값이 2로 바뀔 뿐이다.

먼저, Crawford의 지표는 부정행위를 하지 않은 수험자들(근처 자리가 아닌 수험자들)의 랜덤표본을 추출하여 각 쌍들의 오답일치수의 상대적비율들의 평균과 표준편차를 이용하여 특정수험자들의 오답일치수의 상대적 비율이 평균비율과 차이나는 지 검정한 것이다. 시험이 강의실 두 곳에서 이루어졌기 때문에, 강의실 다른 수험자들은 부정행위 혐의가 전혀 없으므로 그러한 수험자들의 쌍들, 675($= 27 \times 25$) 쌍들의 오답일치수의 상대적 비율들을 이용하였다. 평균은 0.392이고 표준편차는 0.0409이다. Table 4.2에서 보듯이 유의한 쌍은 하나도 없음을 알 수 있고, 상위에 있는 수험자들도 근처 자리가 아닌 것으로 보아 이 지표로서는 부정행위 혐의가 있는 수험자을 발견할 수 없다.

Angoff의 B-지표는 $W_i \cdot W_j$의 비슷한 값들끼리 묶어 그룹으로 묶어 그 그룹안에서 Q_{AB}를 표준화한것이다. 그룹 수에 따라 B-지표의 분포는 많이 달라질 수 있다. 그룹 수를 1개에서 부터 점차 늘려보았는

Table 4.4. Bellezzas' Error Similarity Analysis

	Examinee pair	W_A	W_B	W_{AB}	Q_{AB}	p-value	q-value
All	(20, 49)	39	49	25	24	0.00017	0.220
	(4, 8)	37	43	22	20	0.00387	1.000
	(24, 25)	35	37	21	19	0.00561	1.000
	(11, 45)	28	28	11	11	0.00650	1.000
	(36, 42)	51	26	16	15	0.00678	1.000
	(11, 13)	28	37	20	18	0.00810	1.000
	(13, 24)	37	35	24	21	0.00853	1.000
	(23, 37)	27	23	10	10	0.01028	1.000
	(14, 33)	33	39	19	17	0.01163	1.000
	(7, 45)	32	28	14	13	0.01503	1.000
Four-Choice	(20, 49)	25	28	13	12	0.00096	0.856
	(4, 8)	22	30	15	13	0.00205	0.856
	(14, 33)	23	28	15	13	0.00205	0.856
	(11, 45)	17	20	8	8	0.00258	0.856
	(14, 24)	23	21	13	11	0.00690	1.000
	(1, 24)	13	21	10	9	0.00702	1.000
	(12, 47)	21	39	6	6	0.01145	1.000
	(23, 37)	16	14	6	6	0.01145	1.000
	(28, 42)	11	15	20	15	0.01176	1.000
	(11, 13)	17	22	12	10	0.01247	1.000

Table 4.5. Kling's Index K

	Examinee Pair	W_A	W_B	W_{AB}	Q_{AB}	p-value	q-value
All	(12, 47)	31	56	26	21	0.00057	0.404
	(24, 40)	35	58	29	23	0.00061	0.404
	(20, 49)	39	49	25	24	0.00183	0.505
	(17, 40)	14	58	14	11	0.00203	0.505
	(30, 40)	22	58	18	15	0.00325	0.505
	(7, 49)	32	49	26	20	0.00335	0.505
	(9, 40)	32	58	25	20	0.00335	0.505
	(11, 13)	28	37	20	18	0.00343	0.505
	(11, 49)	28	49	23	18	0.00343	0.505
	(25, 40)	37	58	28	22	0.00511	0.608
Four-Choice	(12, 47)	21	39	20	15	0.00011	0.151
	(24, 40)	21	39	19	13	0.00254	1.000
	(11, 34)	17	32	15	11	0.00335	1.000
	(1, 24)	13	21	10	9	0.00416	1.000
	(4, 8)	22	30	15	13	0.00450	1.000
	(10, 48)	33	23	17	13	0.00752	1.000
	(14, 33)	23	28	15	13	0.00752	1.000
	(7, 33)	19	28	14	11	0.01089	1.000
	(11, 13)	17	22	12	10	0.01308	1.000
	(11, 49)	17	28	15	10	0.01308	1.000

Table 4.6. Standardized Signed Entropy Simlarity Score

Examinee Pair	W_A	W_B	W_{AB}	Q_{AB}	p-value	q-value
(38, 39)	41	50	12	9	0.00002	0.020
(3, 27)	16	23	9	6	0.00007	0.032
(34, 39)	50	50	11	8	0.00018	0.045
(39, 49)	50	49	12	11	0.00021	0.045
(18, 39)	33	50	10	6	0.00023	0.045
(1, 3)	21	16	15	13	0.00037	0.048
(12, 39)	31	50	14	8	0.00038	0.048
(1, 17)	21	14	12	7	0.00044	0.048
(1, 7)	21	32	12	7	0.00052	0.048
(1, 42)	21	26	13	10	0.00054	0.048

데, 그룹 수에 따라 지표의 변동이 심한 것을 볼 수 있었고, Table 4.3에서 보는 것처럼 그룹 수에 따라 유의한 쌍들의 수도 많이 변함을 알 수 있다. 그룹 수가 4개 이상일 때는 유의한 쌍이 없었다. 상위에 있는 수험자 쌍 중 네 쌍 $(40, 47)$, $(39, 47)$, $(40, 46)$, $(39, 40)$은 자리도 인접하여 있어서 부정행위가 의심스럽기는 하지만 일단 이러한 쌍들이 오답수가 너무 많다. 수험자 47번의 오답수가 제일 많다.

Bellezza와 Bellezza의 ESA 지표 계산에서 사용된 P_{ESA}는 부정행위가 없는 수험자들의 쌍들만 고려하기 위하여 Crawford 지표처럼 다른 강의실에서 치른 수험자들의 쌍만을 고려하여 Q_{AB}/W_{AB}의 평균이며, 전체문항을 모두 고려한 경우는 0.63이며 사지 선다형 문제만 고려한 경우는 0.47이다.

Kling의 지표 계산에서 사용된 P_K계산에서 사용된 P_K는 다른 강의실에서 치른 수험자들의 쌍만을 고려하여 $Q_{AB}/\min(W_A, W_B)$의 평균이며, 전체문항을 모두 고려한 경우는 0.37이며 사지 선다형 문제만 고려한 경우는 0.30이다. 두 지표 모두 부정행위를 의심할 만한 쌍들을 찾지 못했다.

우리가 제안한 Standardized Signed Entropy Simlarity Score 방법에서 찾은 상위 10개 쌍들을 보면 앞의 방법들에 의해 찾은 쌍들과는 다른 유형들을 보이는 데, W_{AB}이나 Q_{AB}는 상대적으로 작은 값을 가지고 있음을 알 수 있다.

앞 네가지 지표들은 문항의 난이도 고려없이 단순히 오답일치수를 고려하게 됨으로 때로는 너무 많은 잠재적 부정행위자를 찾거나 전혀 발견하지 못함으로 인해 부정행위자를 찾는 지표로서 역활을 하지 못하고 있다. 그와 달리 SSESS로서 찾은 $(38, 39)$는 실제 부정행위를 하였는 지 확인하지는 못 하였지만 실제 수험자들의 자리가 인접하여 (앞/뒤 자리)로 부정행위가 의심스럽다.

5. 결론

본 논문에서, 우리는 선다형 시험에서 부정행위를 확인하기 위한 새로운 통계적 유사성 측도를 제안했다. 이 지표는 수험자들의 선택한 선택지 빈도의 분포에 따라 가중치를 고려한, 즉 각 문항의 난이도를 반영한 응답일치도로서 정답/오답과 무관하게 된다. 실제 시험 자료 분석에서 보여지는 결과와 같이 이전에 사용된 네가지 지표들은 문항의 난이도 고려없이 단순히 오답일치수를 고려하게 됨으로 때로는 너무 많은 잠재적 부정행위자를 찾거나 전혀 발견하지 못함으로 인해 부정행위자를 찾는 지표로서 역활을 하지 못하고 있다. 그와 달리 우리가 제안한 SSESS로서 찾은 한 쌍은 실제 부정행위를 하였는 지 확인할 수 없지만 실제 수험자들의 자리가 인접하여 (앞/뒤 자리)로 부정행위가 의심스러운 쌍을 발견할 수 있었다.

References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters, *Journal of the American Statistical Association*, **69**, 44–49.

Bellezza, F. S. and Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error similarity analysis, *Teaching of Psychology*, **16**, 151–155.

Buss, W. G. and Novick, M. R. (1980). The detection of cheating on standardized tests: Statistical and legal analysis, *Journal of Law and Education*, **9**, 1–64.

Crawford, C. C. (1930). Dishonesty in objective tests, *School Review*, **38**, 776–781.

Haney, W. M. (1993). Cheating and escheating on standardized tests, *Paper presented at the annual meeting of the American Educational Research Association*, Atlanta, GA.

Holland, P. W. (1996). Assessing unusual agreement between the incorrect answers of two examinees using the K-Index: Statistical theory and empirical support, *ETS Research Report No.96-7.* Princeton, NJ: Educational Testing Service.

Storey, J. D. (2003). The positive False Discovery Rate: A Bayesian interpretation and the q-value, *The Annals of Statistics*, **6**, 2013–2035.

다중선택 시험에서 부정행위자 발견을 위한 새로운 통계적 측도

한은수[a] · 임요한[b] · 이경은[c,1]

[a]국립보건연구원 연구기획과, [b]서울대학교 통계학과, [c]경북대학교 통계학과

(2012년 11월 7일 접수, 2012년 12월 3일 수정, 2012년 12월 28일 채택)

요 약

학문적 진실성(academic integrity)을 위반하는 잠재적 부적행위를 판단할 때, 잘못된 결정을 피하기 위해서는 확고한 근거를 마련하는 것이 중요하다. 교육학 연구자들은 부정행위를 발견 혹은 확신 할 수 있는 많은 통계적인 방법들을 발전시켰다. 그러나, 대부분의 방법들은 단순히 상관계수를 기초로한 방법들이어서 종종 응답자들의 패턴을 설명하기가 어렵다. 이 논문에서는, 이런 어려움을 해결하기 해결하기 위하여 표준화된 부호 엔트로피 유사성 점수(Standardized Signed Entropy Similarity Score)라는 새로운 통계적인 측도를 제안한다. 또한, 이 제안한 방법을 실제 시험 자료를 이용 부정행위자를 발견하는 데 적용하였고, 다른 기존의 방법들과 비교하였다.

주요용어: 부정행위 발견, Angoff의 B, Crawford의 방법, Error Similarity Analysis, K-지표, 표준화된 부호 엔트로피 유사성 점수.

이 논문은 2011학년도 경북대학교 학술연구비에 의하여 연구되었음.

[1]교신저자: (706-701) 대구 광역시 북구 대학로 80, 경북대학교 통계학과, 조교수. E-mail: artlee@knu.ac.kr