International Journal of Assessment Tools in Education
2021, Vol. 8, No. 1, 156–166

https://doi.org/10.21449/ijate.864078

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **_Research Article_**

# Defining Cut Point for Kullback-Leibler Divergence to Detect Answer Copying

**Arzu Ucar** [1,*], **Celal Deha Dogan** [2]

[1]Hakkari University, Faculty of Education, Department of Educational Sciences, Hakkari, Turkey
[2]Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Ankara, Turkey

**ARTICLE HISTORY**

*Received: Sep. 06, 2020*
*Revised: Dec. 25, 2020*
*Accepted: Jan. 16, 2021*

**Keywords:**
Copy detection,
Cut point,
ROC analysis,
Answer copying,
Kullback-Leibler divergence.

**Abstract:** Distance learning has become a popular phenomenon across the world during the COVID-19 pandemic. This led to answer copying behavior among individuals. The cut point of the Kullback-Leibler Divergence (KL) method, one of the copy detecting methods, was calculated using the Youden Index, Cost-Benefit, and Min Score p-value approaches. Using the cut point obtained, individuals were classified as a copier or not, and the KL method was examined for cases where the determination power of the KL method was 1000, and 3000 sample size, 40 test length, copiers' rate was 0.05 and 0.15, and copying percentage was 0.1, 0.3 and 0.6. As a result, when the cut point was obtained with the Min Score p-value approach, one of the cutting methods approaches, it was seen that the power of the KL index to detect copier was high under all conditions. Similarly, under all conditions, it was observed that the second method, in which the detection power of the KL method was high, was the Youden Index approach. When the sample size and the copiers' rate increased, it was observed that the power of the KL method decreased when the cut point with the cost-benefit approach was used.

## 1. INTRODUCTION

Due to the COVID 19 pandemic period, some exams are required to be administered online, and this situation may increase the examinees' motivation for cheating. Therefore, examinees who cheat and do not cheat should be distinguished to minimize the measurement error that may arise from the copying. Cheating behavior risks the validity of the inferences about students' competence and skills. Cheaters should be detected to minimize the systematic error that may be caused by cheating behavior. Cheating is one of the aberrant behaviors of examinees. Numerous statistical techniques have been developed to detect aberrant response patterns and test scores of examinees. Those techniques detect anomalies associated with different cheating types. There are two main types of cheating behavior. Individual cheating occurs when the student cheats during the exam from a source (other examinees, books, notes, smartphones, etc.). On the other hand, group cheating occurs when at least two examinees cheat in cooperation during or before the exam. Group cheating usually happens when some of the test items are revealed, and a group of examinees shares test items with each other before or

CONTACT: Arzu Uçar ✉ arzukapcik@gmail.com Hakkari University, Faculty of Education, Department of Educational Siences, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

during the exam. Although research on methods used to detect cheating has primarily focused on individual cheating, some methods are used to identify group cheating recently (Belov, 2013; Wollack & Maynes, 2017).

Many studies involve the use of multiple statistical methods to detect individual and group cheating (Karabatson, 2003; Meijer & Sijtsma, 2001; Meijer & Tendeiro, 2014; Wollack, 2006). The methods used to detect individual cheating can be classified as answer copying and similarity analysis, person-fit statistics, relationships between scores on subsets of items within the test, and model approaches (IRT models embedding aberrant behaviors (He, Meadows & Black, 2018). Answer copying and similarity analysis include numerous methods such as Angoff's B and H indices, K index, g2 index, ω index, S1, and S2 indices, VMIs (Variable Match) indices ξ and ξ ∗ indices (Belov, 2011), Wesolowsky's Z similarity index (Wesolowsky, 2000), Generalized Binomial Test (GBT) index (Shu, Henson and Luecht, 2013) and M4 (Maynes, 2014). Person fit statistics differ according to the type of items (dichotomous and polytomous) and the type of model (parametric, non-parametric). Kullback-Leibler Divergence, MPI (Matched Percentile) index, IRI (Irregularity) index, Z-test statistics are the methods used to detect copiers based on the relationships between the scores on subsets of items within the test. DG (Deterministic, Gated IRT Model) model is a commonly used technique in the model approach.

Kullback-Leibler Divergence (KL) is a measure of information used in psychometric practice. It is used as an item selection method in Computerized Adaptive Testing (Chang & Ying, 1996). However, it is also used to detect individuals who cheat (Belov, 2014a, 2014b). KL gives the difference between the two distributions. For instance, we used a test to obtain ability distributions before and after manipulation. Hence previous exam results indicated that the examinees do not cheat. The posterior ability distribution of the person who cheats was compared with the posterior ability distribution of the person who did not cheat. Then, we obtained a different value. The greater value of the difference is the greater difference between the individuals' performance in both tests (Belov & Armstrong, 2010). There are many reasons for the difference in distributions. However, what we are interested in is the differentiation that occurs due to cheating.

KL has been a commonly used technique to detect individual copiers because it can be practically used when we have preknowledge about the examinees' ability. KL is one of the methods to detect the copiers. To implement the KL method, we need to find out the cut point used during the individuals' classification under various conditions. However, no standard cut point can be used to classify students with KL values, which interprets KL results vague. Also, no study focuses on defining cut points for KL. Therefore, in this study, it was aimed to obtain the cut point of KL values with two different approaches (Min score P-Value, ROC) and compare the performances (power) of those approaches under various conditions.

### 1.1. Purpose of the Study

In this study, we aim to define the cut point for Kullback-Leibler Divergence in different conditions. Following are the research questions:

1. What is the cut point for Kullback-Leibler Divergence based on
   a) Youden Index
   b) Cost-Benefit approach

in different sample sizes (N=1000, N=30000), copiers' rate (5%, 15%) and copying percentages (10%, 30%, 60%)?

2. What is the cut point for Kullback-Leibler Divergence based on the Min Score p-value approach in different sample sizes (N=1000, N=30000), copiers' rate (5%, 15%), and copying percentages (10%, 30%, 60%)?

3. What is the power of Kullback-Leibler Divergence based on
   a) Youden Index
   b) Cost-Benefit approach
   c) Min Score p-value approach

in different sample sizes (N=1000, N=30000), copiers' rate (5%, 15%) and copying percentages (10%, 30%, 60%)?

## 2. METHOD

### 2.1. Research Design

This research is a simulation study in which some variables are manipulated. We design the levels of the variables considering the previous similar studies and real-life conditions. While in previous studies, test difficulty was defined into three levels (easy, medium, and difficult), we decided to fix this variable as the medium because it reflects real-life conditions (Sunbul & Yormaz, 2018; Zopluoglu, 2016).

The copier's ability and the source is another variable that might affect the power of the copy index (Sotaridona & Meijer, 2002; Steinkamp, 2017; Sunbul & Yormaz, 2018). In the study of van der Linden and Sotaridona (2006), the indexes' power was found high for the cases when low ability individual copies the responses from the high ability one. High ability individuals rely on their knowledge in the tests and answer the items on their own. On the other hand, low ability individuals are more likely to copy someone else's answers (Voncken, 2014). Therefore, during the exams, they copy the answers from their peers. In the light of this information, in this study, we decided to fix the ability of the copier as low and the source of the copier as high because of the real-world scenario that we are high likely to experience.

The copier's ability and the source is another variable that might affect the power of the copy index (Sotaridona & Meijer, 2002; Steinkamp, 2017; Sunbul & Yormaz, 2018). We fixed the copier's ability as lower and that of the source as upper because in real-world generally lower ability individuals copy from the individuals who have the upper ability.

In previous studies, the test length was commonly defined as 40 and 80 items. Because in the real-world, large-scale tests often include approximately 40 items in a sub-test, we decided to fix the test length as 40 (Sotaridona & Meijer, 2002, 2003, Sunbul & Yormaz, 2018; Yormaz & Sunbul, 2017; Wollack, 1997, 2003; Zopluoglu, 2016).

Regarding the related literature, the copier ratio is manipulated as 5% and %15 (Steinkamp, 2017). In the previous studies comparing the power and type 1 error of the copy index, both small and large sample sizes were utilized (from 50 to 10000). However, to be prevented from biased estimations about the item and person parameters, we manipulated sample size as 1000 and 3000 (Hurtz & Weiner, 2018; Sunbul & Yormaz, 2018; van der Linden & Sotaridona, 2006; Yormaz & Sunbul, 2017; Wollack, 2003; Zopluoglu, 2016; Zopluoglu & Davenport, 2012). Based on the relevant literature, in this study, we manipulated copying percentage as lower (10%), medium (30%), and upper (60%). Considering the manipulated variables, we tested 12 conditions (sample size-2 x copiers' percentage-2 x copying percentage -3 = 12). Table 1 presents the manipulated and fixed conditions in the study.

**Table 1.** *Simulation Design Conditions and Levels.*

| Condition | Number of Levels | Level Values |
|---|---|---|
| Sample Size | 2 | 1000, 3000 |
| Copiers' Percentage | 2 | 5%, 15% |
| Copying Percentage | 3 | 10%, 30%, 60% |
| Test Difficulty* | 1 | Medium |
| Person Parameter of Source/Copier* | 1/1 | Upper-Lower |
| Test Length* | 1 | 40 |

**fixed variable*

## 2.2. Simulation Data

The Rasch model is one of the Item Response Theory models. It has some advantages, such as being mathematically less complex and easy to apply. Moreover, it is the most frequently used model in the exam programs because encountering parameter estimation problem is less. Therefore, we used the Rasch model in this study. The ability of 10000 participants and the difficulty parameter of 40 items was produced under the standard normal distribution $N$ (0,1). Considering the population's abilities and the difficulty parameters of the test items, dichotomous (1-0) response matrices were simulated based on the Rasch model. For the simulations, we utilized the "mirt" package (Chalmers, 2019) in the R program.

Sunbul and Yormaz (2018) denoted the ability level of the copiers as (-3.00, -1.50), (-1.50, 0.00), and the ability of the source as (0.00, 1.50), (1.50, 3.00). We denoted the ability of copiers in a wider range. In this way, we reduced the interference with the ability level of the copier. In addition to this, since the performance of similarity indices in identifying copiers increases with the increase of the difference between the ability levels of the copier and the source, we selected the ability of the source individuals (1.51, 3) from the individuals with high ability in order to ensure that the difference between the abilities of the copier individuals and the source individuals is greater (van der Linden & Sotaridona (2006)). Therefore, the individuals with low (-3, 0), medium (0.01, 1.50), and high (1.51, 3) abilities were randomly selected from the population (Sunbul & Yormaz, 2018). Low, medium, and high ability levels respectively include 20%, 60%, and 20% of the sample.

Copiers in the sample were randomly assigned among low ability individuals. The sources, who are individuals that the copiers copied their answers from, were randomly assigned among high ability individuals. In this study, we assigned only one copier for each source. Responses of the individuals, who are assigned as copiers, were manipulated so that their responses become similar to the sources' responses. Data simulation is repeated 100 times per each condition.

## 2.3. Analysis

The Kullback-Leibler divergence, one of the common methods, was utilized to detect copiers (Kullback & Leibler, 1951). KL reveals the difference between the two distributions, is calculated with the expression in the equation:

$$D(g||h) = \int_{-\infty}^{+\infty} g(x) \ln \frac{g(x)}{h(x)} dx \tag{1}$$

KL values were obtained by estimating the individuals' abilities twice before and after manipulation and comparing those two distributions. For the analysis, the 'irtoys' (Partchev, 2017) and 'LaplaceDemon' packages (Singmann, 2020) were used in the R program.

We used two methods to find the cut point for KL values. Firstly, to find the cut point, for every 100 iterations, the lowest KL values among the copiers were selected and created a new distribution of the lowest KL values. We repeated this process for each condition separately, and in the end, we obtained 12 distributions. We defined the cut point separately for each distribution based on the 0.05 alpha value (We call this approach as Min Score p-value). Secondly, ROC analysis (Swets & Pickett, 1982; Swets & Swets, 1979) was utilized for all KL values to define the cut point. ROC analysis can classify the data as binary or multi-category. In this study, data were classified as copier and non-copier based on the ROC curves. These curves are used to determine the relationship between sensitivity (Se) and specificity (Sp). The ROC curve is obtained by coordinates (1-Sp (c); Se (c)) for all possible cut points c; where Se (c) and Sp (c);

$$Se(c) = P(T_+|D = 1) = P(T \geq c|D = 1), \tag{2}$$

$$Sp(c) = P(T_-|D = 0) = P(T < c|D = 0). \tag{3}$$

In the formulas, the T values higher than the cut points mean that the individual is a copier. Sensitivity is the degree of defining a copier correctly. On the other hand, specificity is the degree of identifying a non-copier correctly. The ROC analysis presents a graph showing the specificity and the sensitivity (1-specificity) values in the x and y-axis and a curve regarding those values. The graph makes the interpretation easier. In the end, ROC analysis gives us the area under the ROC curve (AUC), which shows the correctness of cut points and the mean of all possible cut points. Thus, it is much more beneficial to evaluate all cut points considering AUC (Bamber, 1975; Swets, 1979). AUC values vary between 0.5 (non-informative) and 1 (excellent). However, ROC analysis offers several cut points criteria using assumptions based on sensitivity and specificity measures or functions defined as a linear combination of both measures. Besides, ROC curve criteria allow the selection of optimum cut points based on the risks and benefits of right and wrong decisions due to the classification outcome. We used some of these several cut points criteria. One of them is Youden Index, and the other is the Cost-Benefit method.

The Youden index (Youden, 1950) is one of the most common indicators used to evaluate the ROC curve. Youden index is the maximum difference between true positive and false positive rates (Krzanowski & Hand, 2009).

$$YI(c) = Se(c) + Sp(c) - 1 \tag{4}$$

The benefits and risks of each type of decision are combined with the prevalence of classification to find Se and 1-Sp values in the ROC curve; this provides the minimum average risk (maximum average benefit) in a given diagnosis (McNeill, Keeler, & Adelstein, 1975; Metz, 1978; Metz, Starr, Lusted & Rossmann, 1975; Swets & Swets, 1979). In a situation where there are two possible alternative decisions, the expected risk of classification use C can be expressed as follows:

$$\begin{aligned} C(c) = C_0 + C_{TP}p\, S_e(c) + C_{TN}(1-p)\, S_p(c) + C_{FP}(1-p)\left(1 - S_p(c)\right) \\ + C_{FN}p\left(1 - S_e(c)\right) \end{aligned} \tag{5}$$

$C_{TP}$, $C_{TN}$, $C_{FP}$, $C_{FN}$ represent the average risks of the results from the decision type, and $C_0$ represents the overhead risk. We used the 'OptimalCutpoints' package (Raton-Lopez & Rodriquez-Alvarez, 2019) in R to compute cut points for KL values. In the end, we compute the power ratios of the cut points obtained.

## 3. RESULT / FINDINGS

### 3.1. Results

The cut point of KL values calculated under various conditions was calculated for the ROC analysis (Youden Index and Cost-Benefit) and the p-value of the minimum score (Min Score *p*-value). Table 2 shows the calculated cut points for different conditions.

It is observed that the cut points based on the Min Score p-value approach ranged from 0.00000000059 to 0.00000545898. For the Youden Index, the cut point obtained were in the range of 0.00000926385-0.00009678113. On the other hand, the cut points obtained with the Cost-Benefit approach varied between 0.00001011724 and 0.00035431080. The lowest cut point was obtained as 0.00000000059 in the Min Score p-value approach. (Sample size was 1000, copiers' rate 0.05, and copying percentage was 0.6. Table 3 presents the Power of KL method to detect copiers for the cut points obtained by Youden Index, Cost-Benefit, and Min Score p-value approaches

**Table 2.** *Cut point of KL values of the table.*

| Sample Size | Copiers' Rate | Copying Percentage | Min Score p-value | Youden Index | Cost-Benefit |
|---|---|---|---|---|---|
| 1000 | 0.05 | 0.1 | 0.00000305008 | 0.00002678292 | 0.00002918371 |
| | | 0.3 | 0.00000545898 | 0.00002854412 | 0.00003208120 |
| | | 0.6 | 0.00000000059 | 0.00003000205 | 0.00003379222 |
| | 0.15 | 0.1 | 0.00000121844 | 0.00009678113 | 0.00034437004 |
| | | 0.3 | 0.00000000188 | 0.00006357387 | 0.00035431080 |
| | | 0.6 | 0.00000000380 | 0.00008453689 | 0.00034498847 |
| 3000 | 0.05 | 0.1 | 0.00000044474 | 0.00000986877 | 0.00001011724 |
| | | 0.3 | 0.00000073166 | 0.00000926385 | 0.00001039373 |
| | | 0.6 | 0.00000070757 | 0.00000973595 | 0.00001085132 |
| | 0.15 | 0.1 | 0.00000049923 | 0.00002917059 | 0.00012223349 |
| | | 0.3 | 0.00000042948 | 0.00003221728 | 0.00011512144 |
| | | 0.6 | 0.00000037981 | 0.00002460582 | 0.00012614019 |

When using the cut point obtained with the Youden Index, the power of detecting the copiers of the KL method was observed as the lowest 0.6311 under 1000 sample size, 0.15 copiers' rate, and 0.6 copying percentage conditions. On the other hand, the highest power (0.8328) was obtained under a 1000 sample size, 0.05 copiers' rate, and 0.3 copying percentage conditions.

**Table 3.** *Power of KL Methods Based on Cut Points Method.*

| Sample Size | Copiers' Rate | Copying Percentage | Youden Index | Min Score p-value | Cost-Benefit |
|---|---|---|---|---|---|
| 1000 | 0.05 | 0.1 | 0.8084 | 0.9221 | 0.7866 |
| | | 0.3 | 0.8328 | 0.8980 | 0.8120 |
| | | 0.6 | 0.8097 | 1.0000 | 0.7868 |
| | 0.15 | 0.1 | 0.6959 | 0.9441 | 0.3975 |
| | | 0.3 | 0.7028 | 0.9964 | 0.3479 |
| | | 0.6 | 0.6311 | 1.0000 | 0.2994 |
| 3000 | 0.05 | 0.1 | 0.8168 | 0.9547 | 0.8116 |
| | | 0.3 | 0.8079 | 0.9325 | 0.7884 |
| | | 0.6 | 0.8108 | 0.9306 | 0.7910 |
| | 0.15 | 0.1 | 0.7058 | 0.9496 | 0.3830 |
| | | 0.3 | 0.7191 | 0.9533 | 0.4331 |
| | | 0.6 | 0.7513 | 0.9588 | 0.4126 |

When using the cut points based on the Min Score p-value approach, the power of detecting the copiers of the KL method was observed as the lowest 0.8980 under 1000 sample size, 0.05 copiers rate, and 0.3 copying percentage conditions. On the other hand, the highest power (1.000) was obtained under a 1000 sample size and 0.6 copying percentage conditions. For the Cost-Benefit approach power of detecting the copiers of the KL method was observed as the lowest 0.2994 under 1000 sample size, 0.15 copiers' rate, and 0.6 copying percentage conditions. On the other hand, the highest power (0.81) was obtained under 3000 sample size, 0.05 copiers' rate, and 0.1 copying percentage conditions. Moreover, comparing three methods to define cut points regarding all conditions Min Score p-value approach has the highest power rates while the Cost-Benefit approach has the lowest rates.

**Figure 1.** *The Conditions' Interaction Effects for Power of KL Methods Based on Cut Points Methods.*

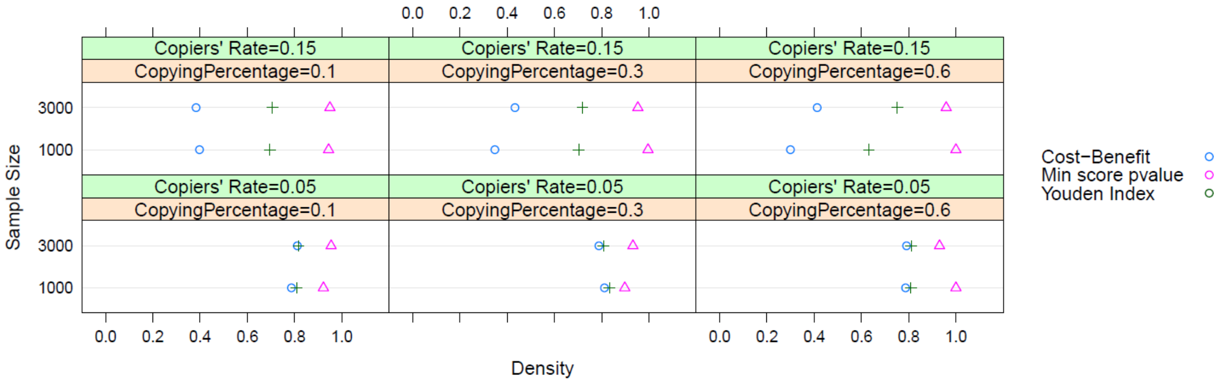

Figure 1 shows the interaction effect plot for the power of the KL method to detect the copier. Regarding the cut point obtained with the Min Score p-value approach, the KL method performed better than other approaches under all conditions. Youden Index method produced the second-best values, and the Cost-Benefit approach produced the worst values regarding the power of copy detection.

## 4. DISCUSSION and CONCLUSION

The Kullback-Leibler Divergence method was used to detect the copiers under various sample sizes, copiers' rates, and copying percentages. Cut points for the KL method were obtained using three approaches (Min Score p-value approach, Youden Index, Cost-Benefit approach). The power of the KL method was computed for the cut points obtained by three approaches. The findings were compared under the manipulated conditions (sample size, copiers' rate, and copying percentage).

Findings showed that the KL method's performance to detect copiers was higher under all conditions when the Min Score p-value approach was used. Especially in cases where the sample size was 1000, and the copying percentage was 0.60, the KL method correctly detected all the copiers. On the other hand, in the cases where the copiers' rate was 0.05 Youden Index and Cost-benefit approaches produced similar values.

Individuals are classified using the Cost-Benefit approach in clinical practice. There is a procedure to be performed for individuals diagnosed after classification. The Cost-Benefit approach determines cut points for this procedure to be both more useful and less cost outcome (Metz, 1978; Zou, et al., 2013). Because the procedure will be performed for each individual to be classified as false positive, otherwise it increases the cost. However, for the individual classified as a false negative, the procedure should not be applied because it will not provide a significant result. The study results revealed that the cut points obtained as a result of the analysis were higher than the cut points in other approaches to minimize the cost.  When the difference between cheating individuals' distributions is less than the cut points, these individuals could not be identified as cheating individuals. Therefore, when the Cost-Benefit approach was used to define the cut point, negative relation was obtained between the copiers' rate and the KL method's power. In other words, the more copiers we had in the sample, the less power the KL method had to detect copiers. However, the copiers' rate did not affect KL methods' power when we used the cut point obtained by the Min Score p-value approach. When the Min Score p-value approach was used to define the cut point, the KL method performed better in detecting the copiers.

When the difference between posterior ability distributions of individuals is high, the KL method with Min Score p-value approach performs better since it uses the minimum KL score of copiers in the computation process. On the other hand, in cases where there are no copiers in the sample, the Min Score p-value approach may detect individuals as copiers, although they are not (false positive). In other words, Min Score p-value Approach might inflate the type 1 error. The Youden index might perform better than the Min Score P-value approach to control the type 1 error.

In contrast to the Cost-Benefit approach's criteria, such as misclassification-cost and the minimum difference value as in the Min Score P-value approach, the Youden index displays a balancing approach. As can be seen from the findings, the cut points obtained according to the other two methods are located between both methods' cut points. In other words, the Youden index makes the classification in a balanced way by maximizing/minimizing a particular combination of sensitivity and specificity. Therefore, the cut points obtained with the Youden index is higher than the cut points obtained with the Min Score p-value approach (Raton-Lopez, et al., 2014). So, when we use the Min Score p-value approach to define the cut point, the KL method's power increases. The cost-Benefit approach decreases the type one error more than other methods do. In order to decide the cut point methods to be used, the researcher should consider the benefits and risks they will take after the decision (Lindahl & Danell, 2016).

Findings showed that the KL method's performance to detect copiers was higher under all conditions wthe hen Min Score p-value approach was used. To detect copiers with the KL method, cut scores are;

- minimum 0.00000000059 maximum 0.00000545898 based on Min Score p-value approach.
- minimum 0.00000926385 maximum 0.00009678113 based on Youden Index.
- minimum 0.00001011724 maximum 0.00035431080 based on Cost-Benefit approach.

In this study, we manipulated and the sample size, copiers' rate, copy percentage. Item difficulty parameters, sources, and copiers' abilities indexed are fixed. So different findings might be obtained when conditions are adjusted in different ways. The standard cut points of KL used by researchers are essential to detect copiers in tests developed in accordance with various measurement theories. Thus, by using various measurement theories, standard cut points of KL can be obtained from different simulation studies. In addition to various measurement theories, results for the type one error and power of KL are needed, when using standard cut points calculated for different values of α under various conditions (sample size, test length, measurement theories, ability distribution, etc.). When using the standard cut points calculated for different α values, new studies investigating the type one error and power of KL can be planned.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## Authorship contribution statement

**Arzu Uçar:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Celal Deha Doğan:** Methodology, Writing the original draft, Supervision, and Validation.

## ORCID

Arzu Uçar https://orcid.org/0000-0002-0099-1348
Celal Deha Doğan https://orcid.org/0000-0003-0683-1334

## 5. REFERENCES

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology, 12*, 387-415. https://doi.org/10.1016/0022-2496(75)90001-2

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and K-index. *Applied Psychological Measurement*, *34*, 379–392. https://doi.org/10.1177/0146621610370453

Belov, D. (2011). Detection of Answer Copying Based on the Structure of a High-Stakes Test. *Applied Psychological Measurement, 35*(7), 495-517. https://doi.org/10.1177/0146621611420705

Belov, D. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement,50*, 141-163. https://doi.org/10.1111/jedm.12008

Belov, D. (2014a). *Detection of Aberrant Answer Changes via Kullback– Leibler Divergence* (Report No. RR 14-04). Law School Admission Council.

Belov, D. I. (2014b). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing, 2*, 37-58. http://dx.doi.org/10.7333%2Fjcat.v2i0.36

Chalmers, P. (2020). Multidimensional item response model (mirt) [Computer software manual]. https://cran.r-project.org/web/packages/mirt/mirt.pdf

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229. https://doi.org/10.1177/014662169602000303

He, Q., Meadows, M., & Black, B. (2018). *Statistical techniques for studying anomaly in test results: a review of literature* (Report No: Ofqual 6355-5). Office of Qualifications and Examinations Regulation.

Hurtz, G., & Weiner, J. (2019). Analysis of test-taker profiles across a suite of statistical indices for detecting the presence and impact of cheating. *Journal of Applied Testing Technology, 20*(1), 1-15. http://www.jattjournal.com/index.php/atp/article/view/140828

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298. https://doi.org/10.1207/S15324818AME1604_2

Krzanowski, W., & Hand, D. (2009). *ROC curves for continuous data.* Chapman and Hall/CRC Press.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86. https://www.jstor.org/stable/2236703

Lindahl, J., & Danell, R. (2016). The information value of early career productivity in mathematics: a ROC analysis of prediction errors in bibliometricly informed decision making. *Scientometrics, 109*, 2241-2262. https://doi.org/10.1007/s11192-016-2097-9

Maynes, D. (2014). Detection of non-independent test taking by similarity analysis. In N.M. Kingston & A.K. Clark (Eds.), Test Fraud: Statistical Detection and Methodology (pp. 52-80). Routledge Research in Education.

McNeill, B., Keeler, E., & Adelstein, S. (1975). Primer on Certain Elements of Medical Decision Making, with Comments on Analysis ROC. *The New England Journal of Medicine, 293*, 211-215.https://www.researchgate.net/publication/22346698_Primer_on_Certain_Elements_of_Medical_Decision_Making

Meijer, R., & Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement, 25*, 107-135. https://doi.org/10.1177/01466210122031957

Meijer, R., & Tendeiro, J. (2014). *The use of person-fit scores in high stakes educational testing: How to use them and what they tell us.* (Report No. RR 14-03). Law School Admission Council.

Metz, C. (1978). Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*, *8*, https://doi.org/10.1016/S0001-2998(78)80014-2

Metz, C., Starr, S., Lusted, L., & Rossmann, K. (1975). Progress in Evaluation of Human Observer Visual Detection Performance Using the ROC Curve Approach. In C. Raynaud & A. E. Todd-Pokropek (Eds.), Information processing in scintigraphy (pp. 420-436). Orsay.

Partchev, I. (2017). A collection of functions related to ıtem response theory (irtoys) [Computer software manual]. https://cran.r-project.org/web/packages/irtoys/irtoys.pdf

Raton-Lopez, M. & Rodriquez-Alvarez, X. M. (2019.). Computing optimal cut points in diagnostic tests (OptimalCutpoints) [Computer software manual]. https://cran.r-project.org/web/packages/OptimalCutpoints/OptimalCutpoints.pdf

Raton-Lopez, M., Rodriquez-Alvarez, X. M., Suarez- Cadarso, C., & Sampedro-Gude, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cut points in Diagnostic Tests. *Journal of Statistical Software,61(8)*, 1-36. https://www.jstatsoft.org/v061/i08

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response. *Psychometrika, 78*, 481-497. https://doi.org/10.1007/s11336-012-9311-3

Singmann, H. (2020). Complete Environment for Bayesian Inference (LaplaceDemon) [Computer software manual]. https://cran.r-project.org/web/packages/LaplacesDemon/LaplacesDemon.pdf

Sotaridona, L., & Meijer, R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement, 39*(2), 115-132. https://www.jstor.org/stable/1435251

Sotaridona, L., & Meijer, R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*(1), 53-70. https://www.jstor.org/stable/1435054

Steinkamp, S. (2017). Identifying aberrant responding: Use of multiple measures [Doctoral dissertation]. https://conservancy.umn.edu/bitstream/handle/11299/188885/Steinkamp_umn_0130E_18212.pdf?sequence=1&isAllowed=y

Sunbul, O., & Yormaz, S. (2018). Investigating the performance of omega index according to item parameters and ability levels. *Eurasian Journal of Educational Research, 74*, 207-226. https://ejer.com.tr/public/assets/catalogs/en/11_EJER_SYormaz.pdf

Swets, J. (1979). ROC Analysis Applied to the Evaluation of Medical Imaging Techniques. *Investigative Radiology, 14*(2), 109-121.

Swets, J., & Pickett, R. (1982). *Evaluation of diagnostic systems: methods from signal detection theory.* Academic Press.

Swets, J., & Swets, J. (1976). ROC approach to cost/benefit analysis. In KL. Ripley & A. Murray (Eds.), Proceedings of the Sixth IEEE Conference on Computer Applications in Radiology. IEEE Computer Society Press.

van der Linden, W., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*(3), 283-304. https://www.jstor.org/stable/4122441

Voncken, L. (2014). *Comparison of the Lz* Person-Fit Index and ω Copying-Index in Copying Detection*. (First Year Paper). Universiteit van Tilburg. http://arno.uvt.nl/show.cgi?fid=135361

Wesolowsky, G. (2000). Detecting excessive similarity in answers on multiple choice exams.

*Journal of Applied Statistics, 27*(7), 909-921. https://doi.org/10.1080/02664760050120588

Wollack, J. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21*(4), 307-320. https://doi.org/10.1177/01466216970214002

Wollack, J. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40*(3), 189–205. https://www.jstor.org/stable/1435127

Wollack, J. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education, 19*(4), 265-288. https://doi.org/10.1207/s15324818ame1904_3

Wollack, J., & Maynes, D. (2017). Detection of test collusion using cluster analysis. In G. Cizek & J. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 124-150). Routledge.

Yormaz, S., & Sunbul, O. (2017). Determination of Type I Error Rates and Power of Answer Copying Indices under Various Conditions. *Educational Sciences: Theory & Pracice, 17*(1), 5-26. https://doi.org/10.12738/estp.2017.1.0105

Youden, W. (1950). Index for Rating Diagnostic Tests. *Cancer, 3*, 5-26. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Zopluoglu, C. (2016). Classification performance of answer-copying indices under different types of IRT models. *Applied Psychological Measurement, 40*, 592–607. https://doi.org/10.1177/0146621616664724

Zopluoglu, C., & Davenport, E. (2012). The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, *72*(6), 975-1000. https://doi.org/10.1177/0013164412442941

Zou, K. H., Yu, C.-R., Liu, K., Carlsson, M. O., & Cabrera, J. (2013). Optimal Thresholds by Maximizing or Minimizing Various Metrics via ROC-Type Analysis. *Academic Radiology*, *20(7),* 807–815. https://doi.org/10.1016/j.acra.2013.02.004