DOI: 10.55643/ser.1.55.2025.596

Artem Artyukhov
D.Sc. in Economics, Associate Professor of the Research Institute of Trade and Sustainable Business, University of Economics in Bratislava, Bratislava, Slovakia;
ORCID: 0000-0003-1112-6891

Olena Churikanova
D.Sc. in Economics, Professor of the TUBAF Digital, TU Bergakademie Freiberg, Freiberg, Germany;
ORCID: 0000-0001-5703-2271

Oleksandr Dluhopolskyi
D.Sc. in Economics, Professor of the Faculty of Management and Social Sciences, Lubelska Akademia WSEI, Lublin, Poland;
e-mail: dlugopolsky77@gmail.com
ORCID: 0000-0002-2040-8762
(Corresponding author)

Kseniia Bliumska-Danko
Candidate of Economy Sciences, Associate Professor of the Research Center in Economics and Management (RITM), Paris-Saclay Université, Paris, France;
ORCID: 0000-0003-0200-2280

Received: 10/01/2025
Accepted: 20/03/2025
Published: 31/03/2025

THE RELATIONSHIP BETWEEN EDUCATION AND TEXT SIMILARITY CHECKING PROCESS ACROSS COUNTRIES: A QUANTITATIVE ANALYSIS

ABSTRACT

The relationship between education and societal similarity is crucial in understanding global academic and cultural trends. This study aims to examine the correlation between the Education Index (a subindex of the Human Development Index) and the Similarity Percentage across countries, assessing whether higher education levels influence similarity patterns. A quantitative research approach was employed, utilising descriptive statistics, correlation analyses (Spearman, Kendall, and Distance Correlation), and regression models (polynomial and Generalized Additive Models - GAM) based on data from 142 countries. The findings indicate a weak but statistically significant negative correlation between the Education Index and Similarity Percentage, suggesting that higher education levels are slightly associated with lower similarity. However, the Distance Correlation test (dCor = 0.24) implies a more complex, non-linear relationship between the variables. Regression models failed to significantly improve the justification of results, highlighting that the Education index alone does not strongly predict similarity percentages. These results suggest that other socio-economic, cultural, and geopolitical factors play a more substantial role in determining similarity patterns between nations.

Keywords: education index, academic integrity, plagiarism, similarity percentage, correlation analysis, statistical modelling, global trends

JEL Classification: I25, O57, C31

INTRODUCTION

Education plays a fundamental role in shaping societal structures, influencing economic development, cultural exchange, and international cooperation. As nations strive for progress, education levels often serve as indicators of human capital development and innovation potential. However, the extent to which education contributes to similarities or divergences between countries remains an open question. Understanding this relationship is crucial for policymakers, educators, and researchers seeking to assess the broader impact of education on global integration and differentiation.

The concept of similarity between nations encompasses various aspects, including cultural, linguistic, economic, and academic dimensions. One measurable proxy for similarity is the Similarity Percentage, which reflects the degree of resemblance in written and published content across countries. This metric, derived from global plagiarism statistics, provides insights into how closely nations align in terms of academic and intellectual output. Meanwhile, the Education Index, a subindex of the Human Development Index, is a robust indicator of educational attainment and literacy levels worldwide.

Despite the intuitive assumption that higher education levels promote shared knowledge and convergence between nations, the existing literature presents mixed findings. Some studies suggest that education fosters globalisation and intellectual alignment, while others argue that it strengthens national identities and cultural uniqueness, leading to more significant divergence. Furthermore, the relationship between education and similarity may not be strictly linear, requiring advanced statistical approaches to uncover potential patterns.

Given these considerations, this study aims to investigate the correlation between the Education Index and Similarity Percentage across 142 countries. Specifically, it seeks to determine whether higher education levels are associated with increased or decreased similarity between nations. Using a quantitative research approach, the study employs correlation analyses (Spearman, Kendall, and Distance Correlation) and regression models (polynomial and Generalized Additive Models - GAM) to explore the nature of this relationship.

By analysing cross-country data, this research contributes to the ongoing discourse on the global impact of education. The findings will help clarify whether education is a unifying force that fosters international similarity or a differentiating factor that enhances national uniqueness. Additionally, the study provides valuable insights for policymakers, academics, and international organisations seeking to optimise education policies in the context of global development.

LITERATURE REVIEW

LITERATURE REVIEW

The relationship between educational development and plagiarism has been extensively studied, particularly in higher education. Researchers have explored how factors such as students' awareness, institutional policies, and technological interventions affect plagiarism levels.

Understanding plagiarism and its ethical implications is critical in addressing academic dishonesty. Clarke et al. (2022) assessed knowledge and attitudes toward plagiarism among university students in Rwanda, highlighting gaps in awareness that contribute to high plagiarism levels. Similarly, Colella (2018) examined perceptions of plagiarism education and responsibilities, emphasising the role of faculty in shaping student behaviours. Romanowski (2021) further explored preservice teachers' perceptions of plagiarism, noting discrepancies between self-reported understanding and actual academic integrity practices.

Ayon (2017) investigated students' and instructors' perceptions of Turnitin, finding that while many viewed it as an effective deterrent, concerns regarding its limitations persisted. Batane (2010) also evaluated Turnitin's role in combating plagiarism, revealing that students often resorted to superficial changes rather than fully internalising academic integrity principles. McCulloch and Indrarathne (2022) examined plagiarism in English-medium instruction higher education settings, noting that linguistic barriers contribute to unintentional plagiarism. Evering and Moorman (2012) explored plagiarism in the digital age, arguing that traditional definitions of plagiarism must evolve to account for the complexities of online information sharing and digital authorship.

Technological solutions and institutional policies play a significant role in detecting and preventing plagiarism. Holt (2012) found that educating students on plagiarism improved their ability to detect and avoid academic dishonesty. Perkins et al. (2020) similarly emphasised the importance of academic misconduct education in reducing plagiarism incidents. Stapleton (2012) conducted an empirical study on the effectiveness of anti-plagiarism software, noting mixed results depending on students' language proficiency and understanding of citation practices.

Jocoy and DiBiase (2006) explored online plagiarism detection and remediation, demonstrating that targeted interventions can reduce repeated offences. Özbek Güven et al. (2024) assessed Turkish students' attitudes toward plagiarism, finding that cultural factors influenced ethical decision-making. Rumanovská et al. (2024) examined plagiarism in academic environments, underscoring the need for institutional commitment to integrity policies.

The relationship between educational development and plagiarism prevalence has been studied in multiple contexts. Dee and Jacob (2010) examined the concept of rational ignorance in education, revealing that students often engage in plagiarism when they perceive the risks as low. Darmansyah and Darman (2022) analysed plagiarism levels in doctoral programs, finding a correlation between assessment methods and academic integrity breaches. Torres-Diaz et al. (2018) explored the connection between internet accessibility and academic success, highlighting that while digital resources enhance learning, they also facilitate plagiarism.

Kauffman and Young (2015) conducted an experimental study on digital plagiarism, showing that instructional design and copy-paste affordances significantly influence students' likelihood of plagiarising. Ma et al. (2008) examined digital cheating in schools, highlighting the challenges of maintaining academic integrity in an increasingly digitalised learning environment. Grose (2023) discussed the disruptive influence of technology on academic integrity, warning that advancements in digital tools may require new approaches to prevent plagiarism effectively.

Integrating artificial intelligence in education presents opportunities and challenges regarding plagiarism detection and prevention. Crawford et al. (2023) discussed ethical concerns about AI-generated content, emphasising the need for clear academic policies. Kasneci et al. (2023) analysed the potential of large language models like ChatGPT in education, suggesting that while they enhance learning, they also pose risks in terms of academic dishonesty. Yan et al. (2023) conducted a systematic scoping review on the ethical challenges of AI in education, highlighting concerns about its impact on academic integrity.

Academic institutions worldwide have adopted policies to combat plagiarism. Bretag (2013) outlined challenges in addressing plagiarism, including inconsistencies in institutional responses. Kadam (2018) examined academic integrity regulations in India, discussing new policies to curb plagiarism. Yavich and Davidovitch (2024) analysed plagiarism trends in higher education, calling for more vigorous enforcement of academic integrity guidelines.

The literature reveals a complex relationship between educational development and plagiarism. While technological tools and academic policies play crucial roles in detection and prevention, educational institutions must prioritise awareness and ethical training. As AI continues to evolve, further research is needed to address its implications for academic integrity. Higher education institutions can mitigate plagiarism and promote genuine learning by fostering a culture of honesty and accountability.

AIMS AND OBJECTIVES

This research aims to investigate the relationship between the Education Index (a subindex of the Human Development Index) and the Similarity Percentage across countries. Specifically, the study seeks to determine whether higher education levels correlate with increased or decreased similarity in various aspects, such as cultural, economic, or linguistic attributes.

The objectives of this research are:

1. To assess whether higher education levels influence similarity patterns, particularly in terms of plagiarism and academic integrity.
2. To examine the correlation between the Education Index (a subindex of the Human Development Index) and the Similarity Percentage across countries.
3. To use regression models to analyse the relationship between education levels and similar it Education Index (a subindex of the Human Development Index) and the Similarity Percentage across countries.

METHODS

Research design

This study employs a quantitative approach to examine the relationship between the Education Index (a subindex of the Human Development Index from UNDP. (n.d.)) and the Similarity Percentage between countries. Plagiarism Checker X provides a Global Plagiarism Live Statistics (Plagiarism Checker X, n.d.) that displays daily global plagiarism statistics, highlighting the percentage of text similarity worldwide. This tool promotes academic honesty by offering insights into current plagiarism trends and encouraging efforts toward achieving zero plagiarism. The analysis relies on descriptive and inferential statistical methods to identify the two variables' patterns, relationships, and dependencies.

Data collection and sources

The dataset used in this study consists of 142 observations, where each observation represents a country with recorded values for the Education Index and Similarity Percentage (Appendix A). The data were sourced from reliable international databases, ensuring accuracy and relevance. All calculations and visualisations were performed using R Studio, a statistical computing environment.

Variables

1. Dependent variable – the Similarity percentage in checked documents (Plagiarism Checker X, n.d.).
2. The Independent variable is the Education Index (HDI subindex) (UNDP, n.d.).

The Similarity Percentage measures the extent to which a country shares common characteristics with others, while the Education Index assesses the overall educational attainment and literacy levels within a country.

DOI: 10.55643/ser.1.55.2025.596

Descriptive statistics

Descriptive statistics such as the mean, median, standard deviation, skewness, and kurtosis were calculated for both variables to summarise the dataset. These statistics provided insights into the distribution and spread of the data, which were further visualised through histograms and boxplots.

Normality testing

To determine whether the data followed a normal distribution, the Shapiro-Wilk test was performed. Due to the non-normal distribution, non-parametric statistical tests were preferred for further analysis.

Correlation analysis

Three correlation tests were conducted to explore the relationship between the Education Index and the Similarity percentage:

1. Spearman's rank correlation.
2. Kendall's Tau correlation.
3. Distance correlation (dCor) test.

Data visualisation and trend analysis

A scatterplot with a LOESS (Locally Estimated Scatterplot Smoothing) curve was created to visualise the relationship between the Education Index and Similarity Percentage. The trend suggested a non-linear relationship, with fluctuations at different levels of the Education Index.

Regression models

To further investigate the nature of the relationship, different regression models were applied:

1. Polynomial Regression (Quadratic and Cubic).
2. Generalised Additive Model (GAM).

RESULTS

Table 1 presents descriptive statistics for two variables Similarity percentage and Education index (HDI subindex).

Table 1. Summary statistics for variables. *(Source: authors' calculation in R studio)*

	Similarity percentage	Education index (HDI subindex)
Min.	0.0400	0.2400
1st Qu.	0.1200	0.5700
Median	0.1400	0.7200
Mean	0.1405	0.6977
3rd Qu.	0.1600	0.8400
Max.	0.2600	1.0100

The Similarity percentage values range between 4% and 26%, with most values clustering between 12% and 16%. The Education index varies significantly, from 0.24 (very low education levels) to 1.01 (close to perfect education index). The median and mean values for the Education index (~0.70) indicate that, on average, the education levels are relatively high, with some countries dragging the lower end.

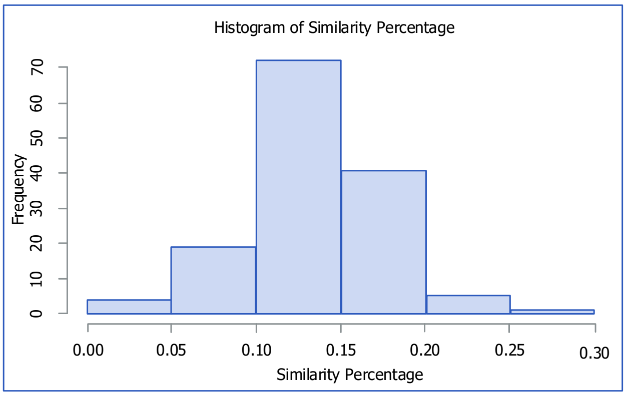
Table 3. Descriptive statistics. *(Source: authors' calculation in R studio)*

Statistical indicators	Similarity percentage	Education index (HDI subindex)
Number of observations	142	142
The arithmetic mean (average) of the variable	0.14	0.7
Standard deviation	0.04	0.18
Median	0.14	0.72
Mean calculated after removing outliers	0.14	0.71
Median Absolute Deviation	0.03	0.19
Minimum	0.04	0.24
Maximum	0.26	1.01
Difference between max and min	0.22	0.77
Skewness	-0.13	-0.53
Kurtosis	0.89	-0.53
Standard error	0	0.02

For the Similarity percentage, the data is usually distributed with very low skewness. The mean and median are equal, reinforcing symmetry. Low standard deviation (0.04) means the values are tightly packed around the mean. The range is small (0.22), suggesting a limited spread.

For the Education index (HDI subindex), the distribution is slightly left-skewed (-0.53), meaning there are more values at higher education levels. The mean (0.7) is lower than the median (0.72), confirming the negative skew. The standard deviation (0.18) is relatively high, meaning the education index varies more across observations. The kurtosis (-0.53) suggests a relatively flat distribution, with fewer extreme values than a normal distribution.

Tohistogram, presented in Figure 1, were built.



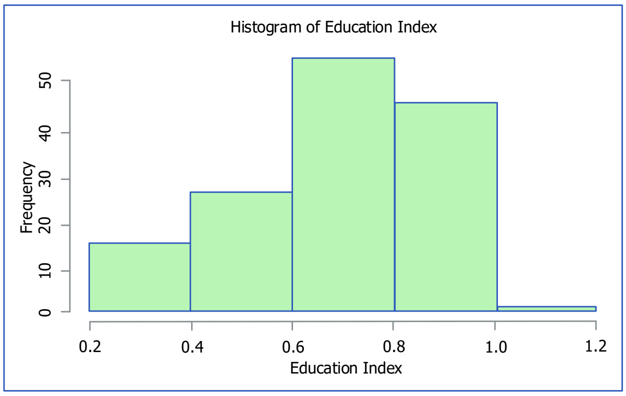


Figure 1. Histograms for visualisation visualise the spread and skewness.

The Similarity Percentage histogram shows a bell-shaped distribution, suggesting it is close to a normal distribution. The left and right tails are relatively symmetrical, supporting the near-zero skewness (-0.13) found earlier. The spread is narrow, confirming the low standard deviation (0.04).

The distribution of the Education Index appears left-skewed (negative skew), meaning more observations are concentrated towards higher education index values. The highest frequency is around 0.6–0.8, indicating that most regions have an education index in this range. There are fewer observations in the lower range (0.2–0.4) and upper range (1.0+), meaning extreme values are less common. This aligns with the skewness value of -0.53.

A boxplot (Figure 2) helps detect outliers.

DOI: 10.55643/ser.1.55.2025.596

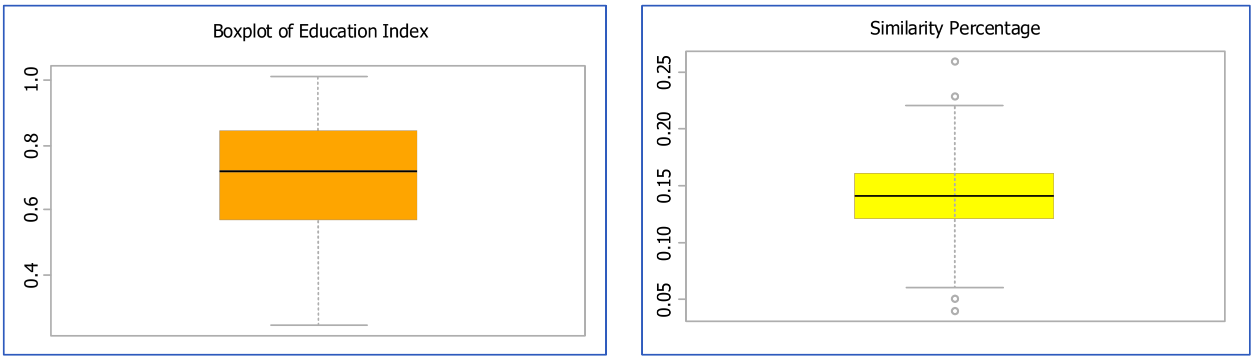


Figure 2. Boxplot of Education Index and Similarity percentage.

For the Education Index, the median (bold horizontal line) is slightly above 0.7, which aligns with the median value in the descriptive statistics. The interquartile range (IQR) (box width) spans approximately from 0.6 to 0.85, meaning 50% of the data falls within this range. The whiskers extend to approximately 0.4 (minimum) and 1.0 (maximum), suggesting a moderate spread. No extreme outliers are present, confirming that the distribution is well-contained within its expected limits. This supports the previous histogram's left-skewed shape, meaning there are more observations at higher education index values.

For the Similarity percentage, the median is around 0.14, which aligns with the mean and median from the descriptive statistics (mean = 0.14, median = 0.14). The IQR spans from approximately 0.11 to 0.16, meaning most values fall within this narrow range. The whiskers extend from around 0.04 to 0.21, indicating the full spread of the data. Outliers are present above 0.21 and below 0.06, shown as small circles. These outliers represent values significantly different from the rest of the data but not extreme. The presence of outliers but minimal skewness (-0.13) suggests that the distribution is primarily normal but has a few distant values.

The Shapiro-Wilk test checks whether a dataset follows a normal distribution. Shapiro-Wilk test for the Education Index has the result – W = 0.95948, p-value = 0.0003379. That means the null hypothesis should be rejected. The Education Index is not normally distributed. The Shapiro-Wilk test for the Similarity Percentage (W = 0.97187, p-value = 0.005019) means that the Similarity Percentage is not normally distributed. Since the data is not normally distributed, using non-parametric statistical tests (Spearman correlation instead of Pearson correlation) is better.

Spearman's Rank Correlation proves a weak but statistically significant (p-value = 0.02163) negative correlation (-0.1926351) between the Education Index and Similarity Percentage. As the Education Index increases, the Similarity Percentage tends to decrease slightly. The correlation is weak so that other factors may be influencing the relationship. However, some values are repeated (ties exist). Since Spearman's correlation is rank-based, ties can affect exact p-value calculation/

Kendall’s Tau correlation is used when the data has many tied ranks (duplicate values). It is more robust than Spearman when dealing with small datasets or many ties. The τ (Kendall's rank correlation coefficient) = -0.1388, p-value = 0.01942. Both Spearman and Kendall show a weak negative correlation, meaning higher education levels are slightly associated with lower similarity percentages. Spearman's coefficient (-0.19) is slightly stronger than Kendall's (-0.13). Kendall is more robust to tie in the data, so its correlation is slightly lower.

The Distance Correlation (dCor) test measures both linear and non-linear dependencies between variables. It provides a more flexible measure of association than Pearson, Spearman, or Kendall.

Table 4. The Distance Correlation (dCor) outputs.

dCor independence test (permutation test) **data: index 1, replicates 1000**			
dCor	0.24009	p-value	0.00999
sample estimates:			
dCov	dCor	dVar(X)	dVar(Y)
0.01227264	0.24008985	0.12320206	0.02120854

DOI: 10.55643/ser.1.55.2025.596

The dCor = 0.24 suggests a moderate relationship between the Education Index and Similarity Percentage. Unlike Pearson/Spearman/Kendall, dCor detects both linear and non-linear dependencies, meaning there might be a hidden pattern in the data. The p-value < 0.05 means a statistically significant relationship. This means the observed correlation is unlikely to be due to random chance. Unlike Spearman (-0.19) and Kendall (-0.13), Distance Correlation (0.24) is more vigorous, suggesting a more complex (possibly non-linear) relationship.

The scatterplot (Figure 3) visualises the relationship between **the Education Index and Similarity Percentage** using a **LOESS (Locally Estimated Scatterplot Smoothing) curve** (red line) with a confidence interval (shaded area).

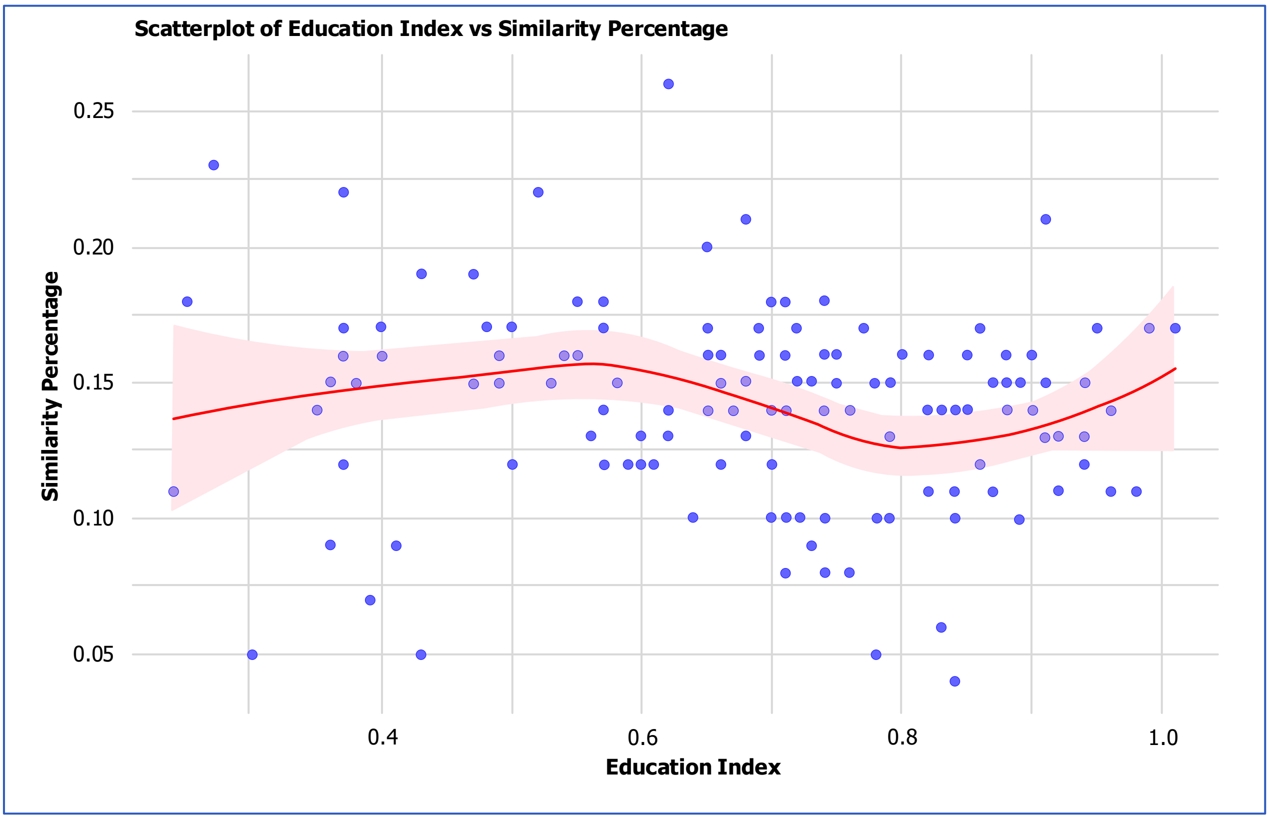


Figure 3. Visualisation of the relationship between Education Index and Similarity Percentage.

Figure 3 demonstrates that at lower values of the Education Index, around 0.3 to 0.5, the Similarity Percentage exhibits a slight upward movement. The trend declines slightly as the Education Index moves into the mid-range, approximately 0.6 to 0.8. However, at the higher end of the Education Index spectrum, from 0.9 to 1.0, the Similarity Percentage appears to rise again.

The confidence interval, represented by the shaded pink region, estimates the trend's uncertainty. The confidence interval is more expansive at low and high extreme values, indicating higher uncertainty. In contrast, the confidence interval narrows for mid-range Education Index values, suggesting more excellent reliability in the observed trend. This visualisation reinforces the idea that the relationship is not strictly linear, which aligns with the Distance Correlation result (dCor = 0.24), suggesting the presence of a moderate non-linear association.

Comparing these findings to the earlier statistical analyses, the Spearman (-0.19) and Kendall (-0.13) correlations indicated a weak negative association, implying a slight downward trend. However, the LOESS curve reveals a more complex, non-monotonic relationship that is not captured by rank-based correlations. The Distance Correlation analysis detected a moderate association, which is now evident in the scatterplot, as the trend does not follow a straightforward linear decline.

This visualisation suggests that the relationship between the Education Index and Similarity Percentage is influenced by additional factors, causing fluctuations in the trend rather than a simple increasing or decreasing pattern. A purely linear model would not adequately explain this relationship and a polynomial or non-linear regression approach might be necessary to quantify the observed trends better. If a deeper analysis is required, fitting a polynomial regression model, such as a quadratic or cubic function, could help better capture the variations in the data.

DOI: 10.55643/ser.1.55.2025.596

Since the LOESS curve suggested a non-linear relationship, a polynomial regression model (e.g., quadratic or cubic) can help quantify the trend. However, the quadratic (Table 5) and cubic (Table 6) polynomial models do not significantly improve the explanation of the Similarity Percentage (y) based on the Education Index (x). Both models have low Adjusted R^2 values, meaning they fail to explain much variance. The cubic model performs slightly better than the quadratic model, but the improvement is minimal.

Although the LOESS plot suggested non-linearity, the polynomial regression does not find strong statistical support for a clear quadratic or cubic relationship. The weak p-values for x^2 and x^3 indicate that adding polynomial terms does not significantly enhance the model.

Table 5. The outputs of the quadratic polynomial models.

Call:
lm(formula = y ~ poly(x, 2), data = df)

Residuals:

Min	IQ	Median	3Q	Max
-0.099193	-0.021605	0.003846	0.020190	0.116426

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(intercept)	0.140493	0.003012	46.640	<2e-16***
poly(x,2)1	-0.060207	0.035896	-1.677	0.0957.
poly(x,2)2	-0.011744	0.035896	-0.327	0.7440

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.0359 on 139 degrees of freedom
Multiple R-squared: 0.02058, Adjusted R-squared: 0.006484
F-statistic: 1.46 on 2 and 139 DF, p-value: 0.2357

Table 6. The outputs of the cubic polynomial models.

Call:
lm(formula = y ~ poly(x, 3), data = df)

Residuals:

Min	IQ	Median	3Q	Max
-0.103710	-0.022518	0.006223	0.018991	0.114221

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(intercept)	0.140493	0.002992	46.963	<2e-16***
poly(x,2)1	-0.060207	0.035648	-1.689	0.0935.
poly(x,2)2	-0.011744	0.035648	-0.329	0.7423
poly(x, 3)3	0.061088	0.035648	1. 714	0.0888.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.03565 on 138 degrees of freedom
Multiple R-squared: 0.04098, Adjusted R-squared: 0.02014
F-statistic: 1.966 on 3 and 138 DF, p-value: 0.122

Although the LOESS plot suggested non-linearity, the polynomial regression does not find strong statistical support for a clear quadratic or cubic relationship. The weak p-values for x^2 and x^3 indicate that adding polynomial terms does not significantly enhance the model. A Generalized Additive Model (GAM) could better capture the non-linearity.

Table 7. The GAM results.

Family: gaussian **Link function: identity** **Formula: y ~ s(x)**				
Parametric coefficients:				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.140493	0.002962	47.43	<2e-16***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Approximate significance of smooth terms:				
	EDF	**Ref.df**	**F**	**p-value**
s(x)	3.407	4.262	1.743	0.154
R-sq.(adj) = 0.0395 Deviance explained = 6.27% GCV = 0.0012856 Scale est. = 0.0012457, n = 142				

The GAM model allows for a flexible non-linear relationship but does not significantly improve model performance. The smooth term (s(x)) is not statistically significant ($p = 0.154$), meaning that the curvature detected in the LOESS plot does not have strong predictive power. Additionally, the Adjusted R^2 remains low (0.0395), indicating that the model explains minimal variance.

The results confirm that Education Index alone is not a strong predictor of Similarity Percentage. Other factors may influence the outcome, or a more complex interaction of variables might be at play.

The plot (Figure 4) visualises the smooth function (s(x)) fitted in the GAM model for the relationship between Education Index (x) and Similarity Percentage (y). The red line represents the estimated smooth function, while the grey shaded area is the confidence interval (95%).

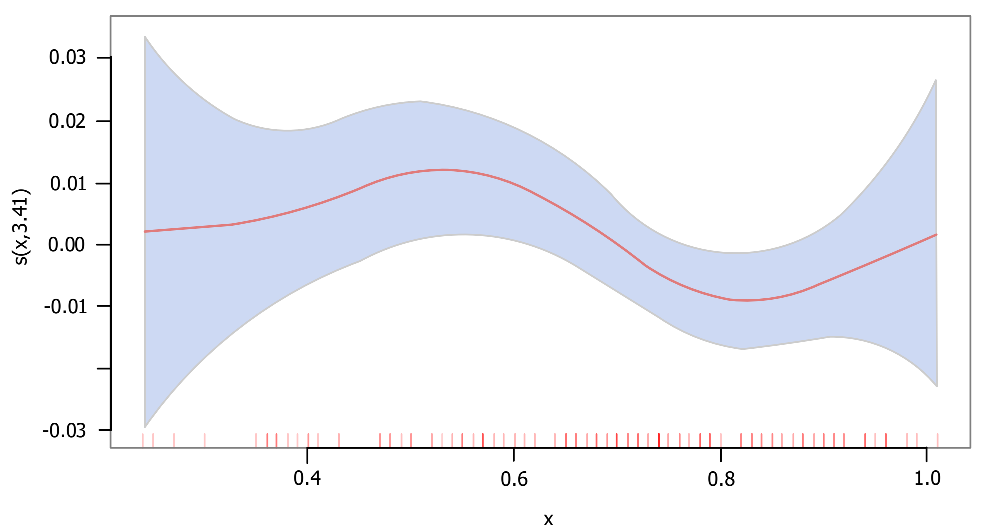


Figure 4. The visualisation of the smooth function (s(x)) is fitted in the GAM model.

The estimated smooth function (red line) is nearly flat, indicating that the Education Index has little to no effect on the Similarity Percentage. A slight dip between 0.6 and 0.8 suggests a non-linear weak impact, but it is not strong enough to be statistically significant. The confidence interval (grey shading) is wide at the ends, indicating higher uncertainty at extreme values of x (low and high Education Index values).

DOI: 10.55643/ser.1.55.2025.596

The wide grey bands at both ends mean that predictions at the lowest and highest Education Index values are less reliable. The estimated effect is insignificant since the entire confidence interval includes zero. This confirms the p-value from the GAM model ($p = 0.154$), suggesting no firm evidence of a relationship.

The GAM model suggests no strong relationship between the Education Index and the Similarity Percentage. While there is a slight non-linear pattern, the confidence intervals indicate high uncertainty, and the effect is not statistically significant. The adjusted R^2 remains low (0.0395), meaning the model explains only ~4% of the variance.

This result aligns with previous findings from polynomial regression, reinforcing the conclusion that the Education Index alone does not explain much variation in Similarity Percentage.

DISCUSSION

The findings of this study reveal a weak negative correlation between the Education Index and the Similarity Percentage, as indicated by Spearman's (-0.19) and Kendall's (-0.13) correlation tests. This suggests that higher education levels are associated with lower similarity percentages, consistent with some literature studies but diverging from others. The Distance Correlation (dCor = 0.24) suggests a moderate, possibly non-linear relationship, implying that the influence of education on similarity may be more complex than a simple linear trend.

Several studies from the literature review provide insights into the relationship between educational levels and plagiarism, helping contextualise these results.

Prior research has established that education improves plagiarism awareness and understanding of academic integrity (Holt, 2012; Perkins et al., 2020). The weak negative correlation found in this study aligns with Bretag (2013), who identified that while education enhances awareness of plagiarism, external factors such as institutional policies and technological resources play a crucial role in plagiarism prevention. This is further supported by Clarke et al. (2022), who noted significant gaps in students' knowledge of plagiarism despite high educational exposure. Similarly, Evering and Moorman (2012) argued that traditional plagiarism detection methods may not effectively address the complexities of plagiarism in digital environments, aligning with the non-linear pattern observed in this study.

Batane (2010) and Ayon (2017) examined Turnitin's effectiveness, showing that while anti-plagiarism software reduces similarity levels, it does not necessarily eliminate plagiarism tendencies, particularly in students with lower educational backgrounds. This finding is reflected in this study's results, as the non-linear relationship suggests that education alone does not entirely determine plagiarism levels.

Additionally, Grose (2023) and Crawford et al. (2023) highlighted the disruptive role of AI in plagiarism, which could contribute to anomalies in the data. If AI-generated content becomes more prevalent, the correlation between education and similarity percentages may weaken further due to the increasing ease of text modification and paraphrasing.

The non-monotonic relationship identified in this study, where similarity percentages fluctuate at different levels of the Education Index, aligns with McCulloch and Indrarathne's (2022) findings. They observed that students in higher education institutions often engage in unintentional plagiarism due to linguistic challenges, which suggests that high education levels alone do not equate to lower plagiarism rates.

Moreover, Özbek Güven et al. (2024) and Yavich and Davidovitch (2024) found that attitudes toward plagiarism vary significantly across cultures. This could explain the non-linear trend observed in this study, where countries with similar education indices may display different plagiarism tendencies due to sociocultural influences.

The failure of polynomial regression models to significantly explain the variance in similarity percentages suggests that additional factors must be considered. This aligns with Dee and Jacob (2010), who posited that rational ignorance in education may lead students to plagiarise when the perceived risk is low. Similarly, the weak predictive power of Generalized Additive Models (GAM) supports the notion that education alone does not fully determine plagiarism levels (Romanowski, 2021).

Limitations

While this research provides insights into the relationship between the Education Index and Similarity Percentage, several limitations must be noted:

1. Data constraints – the study is based on 142 countries, which may not fully represent global diversity. The Education Index and Similarity Percentage may not capture all relevant factors influencing similarity.

2. Methodological limits – non-parametric statistical tests establish correlation but not causation. Confounding variables influencing both education levels and similarity were not included.
3. External influences – factors such as history, politics, and economics, which likely impact similarity, were not analysed.
4. Model constraints – polynomial regression and GAM models showed low explanatory power, suggesting missing influential variables.
5. Temporal scope – this study is cross-sectional and does not examine trends over time.

To enhance findings, future studies should expand datasets, investigate causal relationships, incorporate socio-economic variables, and explore machine learning models for better predictive accuracy.

CONCLUSIONS

This research aimed to investigate the relationship between the Education Index and Similarity Percentage across countries. The study sought to determine whether higher education levels correlate with increased or decreased similarity in various national characteristics.

A quantitative approach was employed using descriptive statistics, correlation analyses (Spearman, Kendall, and Distance Correlation), and regression models (polynomial and Generalized Additive Models - GAM). Data from 142 countries were analysed, and visualisation techniques such as histograms, boxplots, and LOESS smoothing were applied to explore trends.

The results indicate a weak negative correlation between the Education Index and Similarity Percentage, as shown by Spearman's and Kendall's tests (-0.19 and -0.13, respectively). Higher education levels may be associated with slightly lower similarity across countries. The Distance Correlation test (dCor = 0.24, $p < 0.01$) pointed to a moderate association, indicating that the relationship between the two variables may not be strictly linear. Regression models, including polynomial and Generalized Additive Models (GAM), failed to improve explanatory power significantly, implying that education alone does not strongly predict similarity percentages. These findings suggest that other factors, such as geopolitical, cultural, and economic influences, likely play a more substantial role in shaping similarity between nations.

Education policies should be developed with a broader context in mind, as education levels alone do not strongly determine similarity patterns between nations. Policymakers should consider cultural, historical, and economic factors when designing international collaborations or policy frameworks. Since the weak correlation suggests that multiple elements beyond education influence the similarity between nations, policies should integrate aspects such as trade, governance models, and historical ties to foster global integration effectively. Further research is necessary to incorporate additional socio-economic variables and conduct longitudinal analyses that can better explain the role of education in shaping national similarities over time. While education remains a crucial factor in national development, its direct impact on similarity between countries appears limited. A comprehensive and multi-faceted policy approach is required to ensure more informed and effective decision-making.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

All authors have contributed equally.

FUNDING

The research is prepared in the frame of the project "AICE - With Academic Integrity to EU Values: Step by Step to Common Europe" (project number ERASMUS-JMO-2021-HEI-TCH-RSCH-101048055). This project is co-founded by European Union. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

CONFLICT OF INTEREST

The Authors declare that there is no conflict of interest.

DOI: 10.55643/ser.1.55.2025.596

REFERENCES

1. Ayon, N. (2017). Students' and Instructors' Perceptions of Turnitin: A Plagiarism Deterrent? *Creative Education*, *8*, 2091-2108. https://doi.org/10.4236/ce.2017.813141
2. Batane, T. (2010). Turning to Turnitin to Fight Plagiarism among University Students. *Educational Technology & Society, 13*(2), 1–12. https://home.hamptonu.edu/wp-content/uploads/sites/67/2021/03/plagiarism.pdf
3. Bretag, T. (2013). Challenges in Addressing Plagiarism in Education. *PLoS Med, 10*(12), e1001574. https://doi.org/10.1371/journal.pmed.1001574
4. Clarke, O., Chan, W. Y., Bukuru, S., Logan, J., & Wong, R. (2022). Assessing knowledge of and attitudes towards plagiarism and ability to recognise plagiaristic writing among university students in Rwanda. *Higher Education*, *85*(2), 247–263. https://doi.org/10.1007/s10734-022-00830-y
5. Colella, J. (2018). Plagiarism education, perceptions, and responsibilities in post-secondary education. Electronic Theses and Dissertations, 7509. https://scholar.uwindsor.ca/etd/7509
6. Crawford, J., Cowling, M., & Allen, K. A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, *20*(3). https://doi.org/10.53761/1.20.3.02
7. Darmansyah, & Darman, R. A. (2022). Plagiarism level analysis of online take-home examination results and its correlation with doctoral program students' learning outcome. *International Journal of Instruction, 15*(2), 243-256. https://doi.org/10.29333/iji.2022.15214a
8. Dee, T. S., & Jacob, B. A. (2010). Rational ignorance in education: a field experiment in student plagiarism. *NBER working paper series*. Working Paper 15672. http://www.nber.org/papers/w15672
9. Evering, L.C., & Moorman, G. (2012). Rethinking Plagiarism in the Digital Age. *Journal of Adolescent & Adult Literacy, 56*(1), 35-44. https://www.jstor.org/stable/23367758
10. Grose, T. K. (2023). Disruptive Influence. *ASEE Prism, 32*(3), 14–17. https://www.jstor.org/stable/10.2307/48734149
11. Holt, E. A. (2012). Education Improves Plagiarism Detection by Biology Undergraduates. *BioScience*, *62* (6), 585–592. https://doi.org/10.1525/bio.2012.62.6.9
12. Jocoy, C., & DiBiase, D. (2006). *Plagiarism by Adult Learners Online: A case study in detection and remediation*. The International Review of Research in Open and Distributed Learning. https://www.irrodl.org/index.php/irrodl/article/download/242/466/0
13. Kadam, D. (2018). Academic integrity and plagiarism: The new regulations in India. *Indian journal of plastic surgery: official publication of the Association of Plastic Surgeons of India*, *51*(2), 109–110. https://doi.org/10.4103/ijps.IJPS_208_18
14. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274
15. Kauffman, Y., & Young, M. F. (2015). Digital plagiarism: An experimental study of the effect of instructional goals and copy-and-paste affordance. *Computers & Education*, *83*, 44–56. https://doi.org/10.1016/j.compedu.2014.12.016
16. Ma, H. J., Wan, G., & Lu, E. Y. (2008). Digital Cheating and Plagiarism in Schools. *Theory Into Practice, 47*(3), 197-203. http://www.jstor.org/stable/40071543
17. McCulloch, S., & Indrarathne, B. (2022). Plagiarism in EMI higher education: conceptual understanding of staff and students in four South Asian countries. *Higher Education Research & Development*, *42*(3), 694–713. https://doi.org/10.1080/07294360.2022.2102589
18. Özbek Güven, G., Inceoğlu, F., Yilmaz, Ş., & Ince, F. (2024). Evaluation of Turkish Academic and Student Attitudes on Plagiarism: Validity and Reliability of the Plagiarism Attitudes Scale. *Journal of Empirical Research on Human Research Ethics, 19*(1-2), 58-70. https://doi.org/10.1177/15562646241231597
19. Perkins, M., Gezgin, U. B., & Roe, J. (2020). Reducing plagiarism through academic misconduct education. *International Journal for Educational Integrity*, *16*(1). https://doi.org/10.1007/s40979-020-00052-8
20. Plagiarism Checker X. (n.d.). *Global plagiarism Live statistics*. Plagiarism Checker X - Text Similarity Detector. https://plagiarismcheckerx.com/global-plagiarism-live-stats
21. Romanowski, M. H. (2021). Preservice Teachers' Perception of Plagiarism: A Case from a College of Education. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-021-09395-4
22. Rumanovská, Ľ., Lazíková, J., Takáč, I., & Stoličná, Z. (2024). Plagiarism in the Academic Environment. *Societies*, *14*(7), 128. https://doi.org/10.3390/soc14070128
23. Stapleton, P. (2012). Gauging the effectiveness of anti-plagiarism software: An empirical study of second language graduate writers. *Journal of English for Academic Purposes*, *11*(2), 125–133. https://doi.org/10.1016/j.jeap.2011.10.003
24. Torres-Diaz, J., Duart, J., & Hinojosa-Becerra, M. (2018). Plagiarism, Internet and Academic Success at the University. *Journal of New Approaches in Educational Research, 7*(2), 98-104. https://doi.org/10.7821/naer.2018.7.324
25. UNDP. (n.d.). *Human Development Index*. UNDP. Human Development Reports. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI
26. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and

ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90-112. https://doi.org/10.1111/bjet.13370

27. Yavich, R., & Davidovitch, N. (2024). Plagiarism among Higher Education Students. *Education Sciences*, *14*(8), 908. https://doi.org/10.3390/educsci14080908

Артюхов А., Чуріканова О., Длугопольський О., Блюмська-Данько К.

ЗВ'ЯЗОК МІЖ ОСВІТОЮ ТА ПРОЦЕСОМ ПЕРЕВІРКИ ПОДІБНОСТІ ТЕКСТУ В РІЗНИХ КРАЇНАХ: КІЛЬКІСНИЙ АНАЛІЗ

Взаємозв'язок між освітою та соціальною подібністю має вирішальне значення для розуміння глобальних академічних і культурних тенденцій. Це дослідження має на меті вивчити кореляцію між Індексом освіти (субіндексом Індексу людського розвитку) та відсотком подібності між країнами, оцінюючи, чи впливає рівень вищої освіти на моделі подібності. У роботі використано кількісний дослідницький підхід із використанням описової статистики, кореляційного аналізу (Spearman, Kendall та Distance Correlation) та регресійних моделей (поліноміальні й узагальнені адитивні моделі – GAM) на основі даних зі 142 країн. Отримані результати вказують на слабку, але статистично значущу негативну кореляцію між індексом освіти й відсотком подібності, що свідчить про те, що рівень вищої освіти дещо пов'язаний із нижчою схожістю. Однак тест кореляції відстаней (dCor = 0,24) має на увазі більш складний, нелінійний зв'язок між змінними. Регресійні моделі не змогли суттєво покращити обґрунтування результатів, підкреслюючи, що освіта сама собою не дає чіткого прогнозу відсотків подібності. Ці результати свідчать про те, що інші соціально-економічні, культурні та геополітичні фактори відіграють більш суттєву роль у визначенні закономірностей подібності між націями.

Ключові слова: індекс освіти, академічна доброчесність, плагіат, відсоток подібності, кореляційний аналіз, статистичне моделювання, світові тренди

JEL Класифікація: I25, O57, C31