Routledge
Taylor & Francis Group

Student plagiarism in higher education in Vietnam: an empirical study

Khang Do Ba, Khai Do Ba, Quoc Dung Lam, Dao Thanh Binh An Le, Phuong Lien Nguyen, Phuong Quynh Nguyen and Quoc Loc Pham

Faculty of Economics and Commerce, Hoa Sen University, Ho Chi Minh City, Vietnam

ABSTRACT

This paper assesses and compares the prevalence of plagiarism across different student and assignment characteristics at a university in Vietnam, using the similarity index reported by the text-matching software Turnitin as a proxy measure of plagiarism on a sample of 681 student papers. The findings present a level of match higher than reported in earlier studies at universities outside of Vietnam, with an average similarity index of 29.06%. Controlling for the gender and major of the students, the extent of plagiarism is implied to be negatively correlated with the students' academic performances and with the likelihood of being caught, and positively correlated with the length of the assignments. Thus, this study, relying on actual text-matching data rather than self-reported surveys, provides the first empirical test of two theoretical plagiarism models proposed in the literature. The explanatory factors confirmed by this study illustrate the potential benefits of the use of software tools to detect plagiarism and can help refine academic integrity policy formulations for universities.

Introduction

Plagiarism has been an issue in institutions of higher learning since at least the mid-eighteenth century (Quah, Stewart, & Lee, 2012). Current research literature indicates a high prevalence of plagiarism, cheating and other forms of academic misconduct in colleges and universities (McCabe, Treviño, & Butterfield, 2001; Stone, Jawahar, & Kisamore, 2009). The easy access to information via the Internet is perceived to have led to a higher level of plagiarism among students (Batane, 2010). Although this has not been confirmed by empirical researches, Scanlon and Neumann (2002) warned that the amount of online plagiarism 'should be a matter of concern' (p. 383).

Numerous studies in the literature have explored the nature and level of plagiarism practiced by students. However, the different methods of measuring plagiarism used in these studies make it difficult to conduct empirical comparative or trend studies of the problem. Many authors (Comas-Forgas & Sureda-Negre, 2010; Martin, Rao, & Sloan, 2011; McCabe, Treviño, & Butterfield, 2002; Rettinger & Kramer, 2009; Walker, 2010)

CONTACT Khang Do Ba   ✉ khang.doba@hoasen.edu.vn

have also attempted to empirically investigate the explanatory factors of the phenomenon, mostly with the reliance on surveying students. The responses from self-reported surveys, while appropriate for an attitudinal study of plagiarism, might be biased and inaccurate in measuring the actual practice of plagiarism. The relatively recent introduction of software tools to help detect plagiarism thus has opened new opportunities to investigate the phenomenon with more objective data. Despite the overwhelming amount of empirical studies on the problem, relatively few authors have ventured to develop conceptual models to explain the motivations for people to plagiarize. Notable among these were the models proposed by Collins, Judge, and Rickman (2007) on the economics of general plagiarism, and by Quandt (2012) on student plagiarism. The authors, however, presented no empirical test for their models.

In Vietnam, plagiarism in higher education has been mentioned mostly in public media and through informal social-network anecdotes of gross violations. Ngo (n.d.) claimed that cheating is currently getting worse at Vietnamese colleges and universities where students often plagiarize from articles on the Internet without a single citation. Phan (2006) argued that the plagiarism problem might be aggravated due to the educational traditions promoting reproduction of learned texts. Yet, on the whole, this problem in the country has hardly been investigated formally with empirical studies.

In such a context, the first objective of this research is to analyze the extent of plagiarism by students, using data reported by Turnitin, a text-matching tool adopted at a university in Vietnam. In addition, we explored and measured the effects of the students' and reports' characteristics, as well as the implementation of new academic integrity policies and assignment-submission protocols, on the similarity index provided by Turnitin. Thus, the study provides the first known empirical test of the cost-benefit models of plagiarism developed by Collins et al. (2007) and Quandt (2012).

The university under study is a young, medium-sized private university in Ho Chi Minh City, Vietnam, formally certified in 2006 and currently with an enrollment of about 10,000 students. It has four faculties offering a variety of undergraduate programs, six of which were used in our study: five programs from the Faculty of Economics and Commerce (Human Resources Management, Business Administration, Finance and Banking, Accounting and Marketing) and one program from the Faculty of Languages & Cultural Studies (Tourism). Starting from the 2013–2014 academic year, the university introduced the use of Turnitin, becoming the first university using this system for student assignments in the Vietnamese language.

The findings help draw a clearer picture of the level of student plagiarism in Vietnamese higher education, and provide the first quantitative basis for administrators in Vietnamese universities to spur policy changes to address this issue. By evaluating and verifying the factors identified in theoretical plagiarism models, our results may extend beyond the country's borders. At the very least, the insights gained will call for the attention of both educators and students towards necessary policy priorities to counter the issue of plagiarism.

A literature review will be presented in the next section, discussing the prevalence of plagiarism in different parts of the world, as well as factors that lead to academic dishonesty. Following this will be a description of the data and the research framework used in the study. The results from the analyses of the data will then be reported and discussed. The final section will summarize the findings and discuss the implications of the study.

Literature review

Several studies have tried to establish a baseline of how many students cheat and how often they cheat (Josien & Broderick, 2013), but most of these studies on plagiarism in higher education are exploratory in nature. They used self-reported survey data, mostly with simple descriptive statistics, focusing on measuring the extent of plagiarism among different academic populations ranging from high school to college to doctoral students. Some reports attempted to determine individual and contextual indicators of plagiarism such as gender, type of assignment, laziness, years enrolled, parents' education and participation in extracurricular activities (Crown & Spiller, 1998; Josien & Broderick, 2013; McCabe & Treviño, 1997; Walker, 2010; Young, 2013). Others have applied more complex institutional-level analyses by examining students' plagiarism attitudes and practices (Martin et al., 2011; Walker, 2010), student and faculty's perceptions and perspectives on cheating behaviors (Higbee, Schultz, & Sanford, 2011; Pritchett, 2010), peer behaviors (Owunwanne, Rustagi, & Dada, 2010), students' understanding and acceptance of academic integrity policies (O'Dwyer, Risquez, & Ledwith, 2010), the perceived severity of penalties for cheating (Young, 2013), the impact of culture on the understanding and attitudes towards cheating (Tran, 2012), as well as the types of sources students use when doing research (Balingit, 2008).

The results of these studies have indicated that plagiarism has been occurring all over the world, from high-school students (Balingit, 2008) to undergraduates in the United States (Pritchett, 2010; Scanlon & Neumann, 2002), New Zealand (Walker, 1998, 2010), Spain (Comas-Forgas & Sureda-Negre, 2010), Australia (Devlin & Gray, 2007; Yeo, 2007), Austria (Kayaoğlu, Erbay, Flitner, & Saltaş, 2016), South Korea (Ledesma, 2011) and Taiwan (Chien, 2016). Depending on each researcher's definition of academic dishonesty, data collection method and tested variables, prior studies reported that between 13% and 95% of college students engage in some form of academic dishonesty (as reviewed in McCabe & Treviño, 1993). Additionally, a large number of students believe that their peers cheat, suggesting that students regularly judge plagiarism by others to be more prevalent than by themselves (Scanlon & Neumann, 2002). Meanwhile, faculty and instructors view cheating as a more serious infraction than students do (Pritchett, 2010). However, the methods of measurement used in these studies are different, making it difficult to conduct comparative or trend studies. Most of the available literature relies on survey data and is, therefore, only appropriate for attitudinal studies.

Another line of research has aimed to investigate the explanatory factors of plagiarism in higher education. These factors varied between the different studies, ranging from objective data such as sex, age, year of enrollment and major, to more complex qualitative observations such as situational and personality attributes (Rettinger & Kramer, 2009), ethical and religious orientations (Quah et al., 2012), ethnicity and acculturation (Martin et al., 2011), effectiveness of honor code in complex social contexts (McCabe et al., 2002), as well as academic standing, ease of access, lack of time, procrastination habits, convenience and workload (Chien, 2016; Comas-Forgas & Sureda-Negre, 2010). Other studies broke down the factors into broader reasons, such as Devlin and Gray's (2007) interview-based research that reported eight such categories: inadequate admission criteria, poor understanding of plagiarism, poor academic skills, teaching/learning issues, laziness/convenience, pride in plagiarizing, pressures and education costs.

Walker (2010) examined the impact of student characteristics on the percentage of plagiarism in assignments by factors such as gender, nationality, study mode, age and year of enrollment. Significant differences were found in the different types of plagiarism across different nationalities, study modes and ages, while gender proved not to be an indicator of plagiarism.

Findings of some other survey-based reports asserted that sex, year in school, GPA, academic major, fraternity–sorority membership and extracurricular participation were significantly related to cheating to varying degrees (Baird, 1980; Crown & Spiller, 1998). Ledesma (2011) concluded that an increase in GPA from 3.5 to 4.5 (on a 4.5 scale) lowered the likelihood of cheating by 14.1% or 17% (according to two different models). Other individual variables such as neutralizing attitudes, low grades and laziness were all associated with increased cheating (Young, 2013).

While self-reported survey or self-selected interview data were commonly used in the above reviewed literature, other recent studies capitalized on the growing use of text-matching software. A majority of these studies relied on data provided by Turnitin, a popular Internet-based text-matching service, to measure the extent of plagiarism as well as to see whether such software could be used as a cheating deterrent (Ison, 2012; Martin et al., 2011; Walker, 2010). In these studies, plagiarism was measured either in terms of the percentage of reports found plagiarizing (i.e., the number of reports that have a similarity index exceeding a certain threshold level) or the average similarity index of the reports as measured by Turnitin. For example, Barrett and Malcolm (2006) used a similarity index threshold of 15% and found that, by this definition, 41% of the 182 submissions containing plagiarism. The highest findings were by Martin et al. (2011) with 61% of 158 postsecondary business research papers found to be plagiarized, albeit using a very low conservative threshold of 3%. Martin et al. (2011) further found the mean similarity index of this sample to be 10.61%, while Batane (2010) reported an average 20.5% similarity index in a sample of 138 papers from students at the University of Botswana. Ison (2012) calculated the mean similarity index of 100 dissertations randomly selected from the ProQuest database to be 15.1%. These results, however, should be interpreted and compared with caution since the reported similarity indices depend on the optional settings of the software, which were often not specified and could vary among each study.

The role of Turnitin and similar services as a deterrent to plagiarism has drawn mixed conclusions. For example, Martin (2005) found that the use of such tools reduced plagiarism, but Walker's (2010) and Warn's (2006) findings did not support this conclusion. The inconsistency of these results indicates that, instead of relying only on text-matching tools, a combination of the use of such tools with appropriate policies and educational programs might be needed to effectively address the problem.

Other researchers took a different approach, developing theoretical models to describe and explain cheating behaviors. Specifically, Collins et al. (2007) analyzed three simple models that explained a student's decision to plagiarize, and indicated that the amount of plagiarism decreases as the penalty and possibility of getting caught increase. More recently, Quandt (2012) proposed two cost–benefit models exploring the motivations for plagiarism and bribery in academic institutions. Quandt's plagiarism model stipulates that the aggregate demand for plagiarism will positively correlate with the potential benefits the students expect to gain from plagiarizing (increasing the grades and saving

time) and negatively correlate with the endowment of the students as well as the probability of being caught and level of penalty.

Methodology

Data

In this study, the dependent variable used was Turnitin's similarity index. This index, which depends on optional settings at report submission, reflects the percentage of a submitted report that is copied from sources recognized by the service's online databases, including Internet sources, publications and other paper submissions. The use of Turnitin's similarity index as a proxy measure for student plagiarism is consistent with recent research trends – for example, see Batane (2010), Ison (2012), and Martin et al. (2011) – and allows for comparing findings. However, it should be noted that not all texts flagged by Turnitin are plagiarized materials, and the interpretation of the similarity index should be subject to close examination by individual instructors (Maurer, Kappe, & Zaka, 2006). In this study, the optional settings included matches of more than 10 continuous words including quoted materials, which might result in an overstated similarity index. On the other hand, since this was the first time that Turnitin had ever been used at a Vietnamese university with student papers in the Vietnamese language, the repositories used by the software to check for plagiarism were mostly Internet sources and therefore excluded some offline sources such as printed materials and students' papers shared amongst themselves. As a result, the similarity indices reported for papers in Vietnamese are potentially lower compared to indices reported for English submissions.

Several characteristics of the submitted reports were extracted from the Turnitin reports for the purpose of our analyses. These characteristics consisted of the length of the report as measured by its word count (*wc100*) and the type of report (*typeR*). In this study, three types of reports were considered: graduation theses (*typeR* = 1), internship reports (*typeR* = 2) and term papers (*typeR* = 3). Theses and internship reports were individual assignments, while term papers were group submissions. To measure the impact of the new academic integrity policy implemented from the beginning of the academic year of 2013–2014 with the introduction of Turnitin, a variable (*afterAI*) was used to distinguish papers submitted before (*afterAI* = 0) and after (*afterAI* = 1) the start of this academic year. In addition to these report characteristics, a number of variables were included to characterize the student authors in the analysis. These variables – retrieved from the Student Information System available at the university – categorized students by sex (*sex*), program (*program*) and cumulative grade-point average (*grade*). Students from six programs were studied: Human Resources Management, Tourism Management, Business Administration, Finance and Banking, Accounting and Marketing Management (*program* = 1 through 6, respectively). For students' grades, we used their cumulative grade-point averages (CGPA; on a 4.0 scale) up to the semester their associated reports were submitted into the Turnitin system, thus reflecting their academic performance and capabilities at the time of submission. Some additional variables were also extracted from the Turnitin reports for further exploratory analysis: the source most often used by the students in their reports, and their percentage usage. The appendix provides an explanation of all variables used and their summary statistics.

Of the 681 reports from the sample, the 286 graduation theses and 330 internship reports were individual papers, and the other 65 term papers were group papers. This explains why, in the sample data, the report characteristics (report type, word count and submission time) were available to all 681 observations, but the student characteristics (gender, program and CGPA) were available only to 616 observations. With the sample male-to-female ratio at roughly 1:2 (compared to the 2:3 ratio of the full university population), gender distribution in the sample was relatively representative and remained so across the types of reports, and to a lesser extent, across the programs. The distribution of the sample by gender, program and type of report is summarized in Table 1.

Although the university under study did not have a dedicated research ethics committee to review projects involving human participants, the use of the students' reports for the purpose of this study was treated with great consideration. We discussed the use of past submissions for the research with the deans of the faculties concerned for approval, with assurance that the results would be anonymized and not be reported to any individual or functional unit, to avoid any direct effect on the students involved. The research proposal was then submitted to the Vice President for Research and Academic Affairs for final approval to ensure academic and ethical rigor.

Hypotheses

With the purpose of empirically assessing the determinants of similarity index reported by Turnitin, our hypotheses were formulated based on the plagiarism models of Collins et al. (2007) and Quandt (2012) mentioned above. First, grade-point average was used to measure the endowment of the students, leading to the following hypothesis:

H1: The grade of the students (*grade*) negatively correlates with their similarity index.

Since the time saved by plagiarism is arguably higher for longer and more time-consuming assignments, the time-saving motivation for plagiarism implied the following hypothesis:

H2: Similarity index positively correlates with the length of the reports submitted (*wc100*).

On the other hand, grade improvement motivation implied more plagiarism for more important assignments (due to higher gain to be expected). Since graduation theses are generally more important to students than internship reports which in turn are more important than group term papers, we postulated the third hypothesis as follows:

Table 1. Distribution of the sample by gender, program and type of report.

Type of report			HR	Tourism	BA	Finance	Accounting	Marketing	Total
						Program/major			
Graduation thesis	Sex	M	15	14	21	16	11	12	89
		F	30	48	25	31	28	35	197
	Total		45	62	46	47	39	47	286
Internship report	Sex	M	5	5	8	54	27	7	106
		F	15	17	17	78	77	20	224
	Total		20	22	25	132	104	27	330
Total	Sex	M	20	19	29	70	38	19	195
		F	45	65	42	109	105	55	421
	Total		70	84	71	179	143	74	616

H3: Across the types of reports, similarity index is higher for graduation theses (*typeR = 1*) and lower for group term papers (*typeR = 3*).

The final hypothesis for this research concerns the impact of the launch of the Academic Integrity Program at the university beginning in the 2013–2014 academic year, which includes the introduction of the Academic Integrity Policy, the establishment of the Academic Integrity Committee and the use of Turnitin in major student assignments. For this purpose, our study differentiated between two kinds of student reports: those submitted before and those submitted after this launch of the Academic Integrity Program, leading to the fourth hypothesis:

H4: Similarity index declines after the introduction of the Academic Integrity Program (*afterAI = 1*).

Regression models

Hypotheses H1–H3 were tested using regression analysis with the similarity index reported by Turnitin (*similarity*) as the dependent variable. Since the independent variables could be classified into groups of student and report characteristics, the general regression model was:

$$P = \beta_0 + \beta_1 * S + \beta_2 * R + \varepsilon,$$

where P was the amount of plagiarism, represented by the similarity index (*similarity*) as provided by Turnitin; S was the student characteristics, including the cumulative GPA of the student (*grade*), their gender (*sex*) and their declared major (*program*); R was the report characteristics, including the length of the paper as represented by its word count in 100s of words (*wc100*) and the type of report (*typeR*).

In order to control for all possible variables and to account for their impact on the similarity index, the regression was first run with only S, and then with only R, and finally with both sets of independent variables.

Similarly, hypothesis H4 was tested using a regression model with one additional variable T (*afterAI*) reflecting the introduction of Turnitin and the new academic integrity policies at the university:

$$P = \beta_0 + \beta_1 * S + \beta_2 * R + \beta_3 * T + \varepsilon.$$

The impact of T on the similarity index was controlled first with just S, and then with just R, and finally with both S and R.

Findings

Overall observations

From this sample data, the issue of plagiarism appeared to be very serious at the university under study. Indeed, of the 681 reports, the similarity index had a mean of 29.06% with a standard deviation of 13.02% and a median of 29%. These numbers were much higher when compared with the means ranging between 10.61% and 20.5% as reported by Batane (2010), Ison (2012), and Martin et al. (2011). At the same time, 73% of our

observations had a similarity index of over 20%, which, by the guidelines of the Academic Integrity Policy of the university under study, indicated a high likelihood of plagiarism. With a threshold of 15%, this percentage of plagiarized reports jumped to 84%, compared to 41% reported by Barrett and Malcolm (2006). In addition, 11.7% of the reports in our sample had over 10% of their content copied from a single source on the Internet. Although there were numerous types of Internet sources, Internet report mills such as luanvan.net.vn, slideshare.net and tailieu.vn, which allow users to read and download various dissertations and graduation reports at no or very low cost, represented the sources most often used for roughly one fifth of all the reports.

The comparative findings above should be interpreted with care since the submission options on Turnitin (e.g., the inclusion of quoted matches, the length of similar texts to be flagged) might have been set differently in different studies. Also, it should be noted that the sample for this study consisted of mainly business students who in previous researches have been found to be more tolerant toward dishonesty than non-business students (Crown & Spiller, 1998; Klein, Levenburg, McKendall, & Mothersell, 2007; Roig & Ballew, 1994).

Table 2 presents the results of the regression analysis when first only student characteristics were included as predictors (column 1), then only report characteristics (column 2) and finally with both types of independent variables (column 3).

Regression with student characteristics (*grades, sex* and *programs*) in column (1) confirmed hypothesis H1: a grade-point increase correlated to a 6.1% decrease in similarity index with $p < .05$. These results supported earlier findings by Ledesma (2011) and Young (2013). Interestingly enough, female students were found to have slightly higher

Table 2. Regression results with student (S) and report (R) characteristics.

Variables	(1) S only	(2) R only	(3) Both RS
grade	−0.0610**		−0.0941***
	(0.0253)		(0.0258)
sex	0.0235**		0.0262**
	(0.0106)		(0.0105)
2.program	−0.0471**		−0.0383*
	(0.0191)		(0.0197)
3.program	−0.0134		−0.00956
	(0.0178)		(0.0179)
4.program	0.0770***		0.0929***
	(0.0166)		(0.0177)
5.program	0.0822***		0.0970***
	(0.0157)		(0.0162)
6.program	−0.0355**		−0.0275
	(0.0174)		(0.0178)
wc100		0.000503***	0.000239**
		(9.22e−05)	(0.000106)
2.typeR		0.0672***	−0.0150
		(0.0135)	(0.0161)
3.typeR		−0.0478***	
		(0.0171)	
Constant	0.443***	0.204***	0.510***
	(0.0756)	(0.0173)	(0.0749)
Observations	615	681	615
R-squared	0.181	0.113	0.210

Note: Standard errors in parentheses.
***$p < .01$, **$p < .05$, *$p < .1$.

similarity indices compared to their male counterparts. This impact of gender on similarity index, although significant in this study, is inconsistent with findings of Pritchett (2010) or most of earlier studies reviewed by Crown and Spiller (1998). For the students majoring in Business Administration (*program* = 3), the level of plagiarism was not significantly different from that of the students in the baseline HR major (*program* = 1). However, Tourism and Marketing majors (*program* = 2, 6, respectively) had significantly lower similarity (4.7% and 3.5%, respectively). Meanwhile, the average level of plagiarism of Finance and Accounting majors (*program* = 4, 5, respectively) were 7.7% and 8.2% higher than the baseline group at a significance level of $p < .01$.

Regression analysis: regression outputs, interpretations, explanations and implications

Testing hypothesis H2 was based on report characteristics (column 2). Regressing on the length of the papers while controlling for the type of report, the coefficient for word length was statistically significant but negligible, implying an average of 0.5% increase in plagiarism for every additional 1000 words in length. This result supports the findings of Devlin and Gray (2007) that showed evidence of a positive relationship between report length and academic plagiarism.

For hypothesis H3, the coefficients on the *typeR* dummy variables (column 2) showed that controlling for just the report length, compared to the baseline Graduation Theses (*typeR* = 1), Internship Reports (*typeR* = 2) were plagiarized 6.7% more while term papers (*typeR* = 3) were plagiarized 4.8% less, all significant at $p < .01$, apparently disagreeing with hypothesis H3.

Our most comprehensive model combined the two sets of regression models on student and report characteristics, accounting for all measured controls (column 3). The effect of CGPA on the extent of plagiarism was magnified, implying a 9.4% decrease in similarity index for every grade-point increase, all else being equal with $p < .01$. This negative coefficient confirmed hypothesis H1 and supported Ledesma's (2011) conclusion that GPA is negatively correlated to academic dishonesty. Similarly, the positive effect of report length on similarity index, while relatively small at 0.24% increase per 1000 words, remained statistically significant with $p < .05$, supporting hypothesis H2. Coefficients for program indicated very low or no significant differences between programs 1, 2, 3 and 6 (HR, Tourism, BA and Marketing, respectively), while programs 4 and 5 (Finance and Accounting) showed 9.3% and 9.7% higher rates of plagiarism than the baseline HR major at $p < .01$. The results suggested different levels of plagiarism among different majors, corroborating more general comparative findings by Baird (1980) and Klein et al. (2007). Meanwhile, in contrast to the previous model in column 2, the coefficient for Internship Reports reversed its sign to negative as predicted by hypothesis H3. However, the result was not statistically significant and thus H3 cannot be supported.

In order to test hypothesis H4, the similarity indices of reports submitted before the academic year 2013–2014 (*afterAI* = 0) were compared with those of the reports submitted during the new academic year (*afterAI* = 1) after the launch of the university's Academic Integrity Program. Again, linear regression was used, first without any control variables, then controlling for either student characteristics or report characteristics, and finally with all control variables. Table 3 (column 4) indicates that, controlling for all independent

Table 3. Regression results with impacts of Turnitin introduction (afterAI).

Variables	(1) only *afterAI*	(2) *afterAI* & S	(3) *afterAI* & R	(4) *afterAI* & RS
afterAI	−0.0174	−0.0276**	−0.0206*	−0.0300**
	(0.0135)	(0.0120)	(0.0123)	(0.0121)
grade		−0.0867***		−0.0968***
		(0.0329)		(0.0335)
sex		0.0164		0.0179
		(0.0131)		(0.0131)
2.*program*		−0.0603***		−0.0535**
		(0.0205)		(0.0210)
3.*program*		−0.0203		−0.0195
		(0.0213)		(0.0213)
4.*program*		0.123***		0.114***
		(0.0221)		(0.0228)
5.*program*		0.144***		0.140***
		(0.0219)		(0.0221)
6.*program*		−0.0306		−0.0268
		(0.0209)		(0.0210)
wc100			0.000657***	0.000150
			(7.51e−05)	(9.72e−05)
Constant	0.281***	0.544***	0.181***	0.549***
	(0.00921)	(0.0980)	(0.0142)	(0.0979)
Observations	351	285	351	285
R-squared	0.005	0.394	0.184	0.400

Note: Standard errors in parentheses.
***$p < .01$, **$p < .05$, *$p < .1$.

variables (both student and report characteristics), the similarity index declined by 3.0% after this program's introduction, supporting H4 with $p < .05$. Comparing with earlier conflicting results by Martin (2005), Walker (2010), and Warn (2006), our findings confirmed the notion that a combination of text-matching tools with appropriate policies and educational programs could be effective to address the problem of plagiarism in higher education.

One final note is that the value of the R-squared statistic in this study, ranging from 21% to 40% for the regressions with full control variables, is consistent with similar studies on the topic – for example, McCabe et al. (2002) with R-squared values between 19% and 22% and Quah et al. (2012) with R-squared values between 15.8% and 16.6%.

Conclusion

A caveat is in order before the findings and implications are summarized as conclusion: the similarity index reported by Turnitin does not measure directly the level of plagiarism. In fact, this index reflects only the percentage of texts in a submitted report that are found to be matched with those found in the service's databases. However, while not all texts flagged by Turnitin are necessarily plagiarized materials, the index is still the best available proxy measure of actual plagiarism of student reports, and has been widely used in recent empirical research on the problem (Barrett & Malcolm, 2006; Batane, 2010; Ison, 2012; Martin et al., 2011).

The findings of this study were multifaceted and multi-layered. For one, with the use of actual data on text-matching reported by Turnitin, we presented empirical tests of the different hypotheses based on Collins et al. and Quandt's plagiarism models. We have

found a significant negative correlation between students' overall academic performance and their copy-and-paste inclinations, meaning that students with higher cumulative GPAs tend to plagiarize less. In terms of programs, the students could be grouped into two separate sets of relatively homogeneous plagiarism behaviors, with the students from the Finance and Accounting majors having higher levels of similarity index than other programs.

On the relationship between the similarity index and report length, our analysis confirmed the positive correlation hypothesized by Quandt (2012) and Collins et al. (2007). However, the influence of perceived importance of an assignment on the extent of plagiarism was not supported.

Furthermore, the results provide data-driven support for the impact of the use of Turnitin (and with it, the introduction of academic integrity policies and education) to counter plagiarism. When all the independent variables were controlled, the analysis confirmed the deterrent effect of the combination of such policies and tools. Despite the fact that it was the first semester the university under study adopted academic integrity policies and put into actual use a text-matching software, these results may encourage other universities to consider adopting similar measures.

Acknowledgements

Disclosure statement

References

Baird, J. S., Jr. (1980). Current trends in college cheating. *Psychology in the Schools*, *17*(4), 515–522. doi:10.1002/1520-6807(198010)17:4<515::AID-PITS2310170417>3.0.CO;2-3

Balingit, J. A. (2008). *Internet plagiarism as flash point and folklore: Do high school students plagiarize more from Internet sources than from print-based sources?* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (Doc. ID 250822007).

Barrett, R., & Malcolm, J. (2006). Embedding plagiarism education in the assessment process. *International Journal for Educational Integrity*, *2*(1), 38–45.

Batane, T. (2010). Turning to Turnitin to fight plagiarism among university students. *Journal of Educational Technology & Society*, *13*(2), 1–12.

Chien, S.-C. (2016). Taiwanese college students' perceptions of plagiarism: Cultural and educational considerations. *Ethics & Behavior*. doi:10.1080/10508422.2015.1136219

Collins, A., Judge, G., & Rickman, N. (2007). On the economics of plagiarism. *European Journal of Law and Economics*, *24*(2), 93–107. doi:10.1007/s10657-007-9028-4

Comas-Forgas, R., & Sureda-Negre, J. (2010). Academic plagiarism: Explanatory factors from students' perspective. *Journal of Academic Ethics*, *8*(3), 217–232. doi:10.1007/s10805-010-9121-0

Crown, D. F., & Spiller, M. S. (1998). Learning from the literature on collegiate cheating: A review of empirical research. *Journal of Business Ethics*, *17*, 683–700.

Devlin, M., & Gray, K. (2007). In their own words: A qualitative study of the reasons Australian university students plagiarize. *Higher Education Research & Development*, *26*(2), 181–198. doi:10.1080/07294360701310805

Higbee, J. L., Schultz, J. L., & Sanford, T. (2011). Student perspectives on behaviors that constitute cheating. *Contemporary Issues in Education Research*, 4(10), 1–8. doi:10.19030/cier.v4i10.5968

Ison, D. C. (2012). Plagiarism among dissertations: Prevalence at online institutions. *Journal of Academic Ethics*, 10(3), 227–236. doi:10.1007/s10805-012-9165-4

Josien, L., & Broderick, B. (2013). Cheating in higher education: The case of multi-methods cheaters. *Academy of Educational Leadership Journal*, 17(3), 93–105.

Kayaoğlu, M. N., Erbay, Ş., Flitner, C., & Saltaş, D. (2016). Examining students' perceptions of plagiarism: A cross-cultural study at tertiary level. *Journal of Further and Higher Education*, 40(5), 682–705. doi:10.1080/0309877X.2015.1014320

Klein, H. A., Levenburg, N. M., McKendall, M., & Mothersell, W. (2007). Cheating during the college years: How do business school students compare? *Journal of Business Ethics*, 72(2), 197–206. doi:10.1007/s10551-006-9165-7

Ledesma, R. G. (2011). Academic dishonesty among undergraduate students in a Korean university. *Research in World Economy*, 2(2), 25–35. doi:10.5430/rwe.v2n2p25

Martin, D. F. (2005). Plagiarism and technology: A tool for coping with plagiarism. *Journal of Education for Business*, 80(3), 149–152. doi:10.3200/JOEB.80.3.149-152

Martin, D. E., Rao, A., & Sloan, L. R. (2011). Ethnicity, acculturation, and plagiarism: A criterion study of unethical academic conduct. *Human Organization*, 70(1), 88–96. doi:10.17730/humo.70.1.nl775v2u633678k6

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science*, 12(8), 1050–1084. doi:10.3217/jucs-012-08-1050

McCabe, D. L., & Treviño, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education*, 64(5), 522–538. doi:10.2307/2959991

McCabe, D. L., & Treviño, L. K. (1997). Individual and contextual influences on academic dishonesty: A multicampus investigation. *Research in Higher Education*, 38(3), 379–396. doi:10.1023/A:1024954224675

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics & Behavior*, 11(3), 219–232. doi:10.1207/S15327019EB1103

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2002). Honor codes and other contextual influences on academic integrity: A replication and extension to modified honor code settings. *Research in Higher Education*, 43(3), 357–378. doi:10.1023/A:1014893102151

Ngo, T. L. (n.d.). *Nguồn Gốc Văn Hóa Của Đạo Văn* [The cultural roots of plagiarism]. Retrieved from http://www.viet-studies.info/NgoTuLap_NguonVanHoaCuaDaoVan.htm

O'Dwyer, M., Risquez, A., & Ledwith, A. (2010). Entrepreneurship education and plagiarism: Tell me lies, tell me sweet little lies. *Journal of Small Business and Enterprise Development*, 17(4), 641–651. doi:10.1108/14626001011088778

Owunwanne, D., Rustagi, N., & Dada, R. (2010). Students' perceptions of cheating and plagiarism in higher institutions. *Journal of College Teaching & Learning*, 7(11), 59–68. doi:10.19030/tlc.v7i11.253

Phan, L. H. (2006). Plagiarism and overseas students: Stereotypes again? [Peer commentary on "Plagiarism in ESOL students: Is cultural conditioning truly the major culprit?" by D. Liu, and "Plagiarism and the culture of multilingual students in higher education abroad" by C. Sowden]. *English Language Teaching Journal*, 60(1), 76–78. doi:10.1093/elt/cci085

Pritchett, S. (2010). *Perceptions about plagiarism between faculty and undergraduate students* (Doctoral dissertation). Alliant International University, Faculty of the Hufstedler School of Education. Retrieved from ProQuest Dissertations & Theses Global (Doc. ID 205430996)

Quah, C. H., Stewart, N., & Lee, J. W. C. (2012). Attitudes of business students' toward plagiarism. *Journal of Academic Ethics*, 10(3), 185–199. doi:10.1007/s10805-012-9157-4

Quandt, R. E. (2012). Some models of academic corruption. *European Journal of Law and Economics*, 34(1), 63–75. doi:10.1007/s10657-010-9162-2

Rettinger, D. A., & Kramer, Y. (2009). Situational and personal causes of student cheating. *Research in Higher Education*, 50(3), 293–313. doi:10.1007/s11162-008-9116-5

Roig, M., & Ballew, C. (1994). Attitudes toward cheating of self and others by college students and professors. *Psychological Record*, 44(1), 375–382.

Scanlon, P. M., & Neumann, D. R. (2002). Internet plagiarism among college students. *Journal of College Student Development*, 43(3), 374–385.

Stone, T. H., Jawahar, I. M., & Kisamore, J. L. (2009). Using the theory of planned behavior and cheating justifications to predict academic misconduct. *Career Development International*, 14 (3), 221–241. doi:10.1108/13620430910966415

AQ3 Straw, D. (2002). The plagiarism of generation why not. *Community College Week*, 14(24), 4–6.

Tran, T. T. (2012). The perceptions and attitudes of international students towards plagiarism. *The ACPET Journal for Private Higher Education*, 1(2), 13–21.

Walker, J. (1998). Student plagiarism in universities: What are we doing about it? *Higher Education Research & Development*, 17(1), 89–106. doi:10.1080/0729436980170105

Walker, J. (2010). Measuring plagiarism: Researching what students do, not what they say they do. *Studies in Higher Education*, 35(1), 41–59. doi:10.1080/03075070902912994

Warn, J. (2006). Plagiarism software: No magic bullet! *Higher Education Research & Development*, 25(2), 195–208. doi:10.1080/07294360600610438

Yeo, S. (2007). First-year university science and engineering students' understanding of plagiarism. *Higher Education Research & Development*, 26(2), 199–216. doi:10.1080/07294360701310813

Young, D. (2013). Perspectives on cheating at a Thai university. *Language Testing in Asia*, 3(6), 1–15. doi:10.1186/2229-0443-3-6

Appendix

Table A1. Summary statistics.

Variable	Definition	Obs.	Mean	Std. Dev.	Min.	Max.
similarity	Similarity index, in percentage, via Turnitin	681	29.056	13.023	0	68
typeR	Type of report: Graduation Thesis (1), Internship Report (2), Term Paper (3)	681				
wc100	Word count, in 100 words	681	119.906	75.682	15.27	470.54
sex	Dummy: male (0), female (1)	616				
grade	Cumulative GPA	616	3.076	0.211	2.54	3.59
program	Program/Major enrolled: HR (1), Tourism (2), BA (3), Finance (4), Accounting (5), Marketing (6)	616				
afterAI	Dummy: before (0), after (1) implementation of Academic Integrity policies at start of academic year 2013–2014	681				