A Cautionary Note on Checking

Software Engineering Papers for Plagiarism

Cem Kaner and Rebecca L. Fiedler

ABSTRACT

Several tools are marketed to the educational community for plagiarism detection and prevention. This article briefly contrasts the performance of two leading tools, TurnItIn and MyDropBox, in detecting submissions that were obviously plagiarized from articles published in IEEE journals. Both tools performed poorly because they do not compare submitted writings to publications in the IEEE database. Moreover, these tools do not cover the ACM database or several others important for scholarly work in software engineering. Reports from these tools suggesting that a submission has "passed" can encourage false confidence in the integrity of a submitted writing. Additionally, students can submit drafts to determine the extent to which these tools detect plagiarism in their work. Because the tool samples the engineering professional literature narrowly, the student who chooses to plagiarize can use this tool to determine what plagiarism will be invisible to the faculty member. An appearance of successful plagiarism prevention may in fact reflect better training of students to avoid plagiarism detection.

Index Terms – copyright, academic honesty, plagiarism, plagiarism detection, intellectual property, editorial manuscript review, TurnItIn, MyDropbox

I. INTRODUCTION

In recent years, student plagiarism scandals have rocked higher education and brought unwanted attention to some universities. In the United States, *The Chronicle of Higher Education* [1] and other media outlets reported on plagiarism in Ohio University's Mechanical Engineering program. In Australia, RMIT's 'mytutor' case [2] exposed student cheating in the Computer Science department. Plagiarism problems aren't limited to student authors, however. Both the ACM and IEEE Codes of Ethics appear to oppose plagiarism [3], yet both professional societies report an increase in plagiarism incidents. Professional society leaders, including IEEE's Mintzer [4] and ACM Publications Board co-chairs Boisvert and Irwin [5], used columns in their journals to discuss the rise of plagiarism they see in submissions to their journals, and called on colleagues to adhere to the professional practice of proper citation of previous work. The ACM recently published the first *ACM Policy and Procedures on Plagiarism* [6] to clarify and codify their position on the matter.

How can academics, in their roles as authors, instructors, editors, and reviewers, enforce anti-plagiarism stances? Manual searches are labor intensive and time consuming. The nature of the plagiarism detection task is appropriate for an automated solution and a number of tools have been developed in response to this need. Some authors suggest these tools can appropriately be used by students, asserting these students can develop proper citation methods by allowing the tools to find their errors [7]; can use the tools to warn them when they are in danger of being charged for infractions [7]; and can receive automated feedback on their citation practices [8].

Manuscript received January 1, 2007; revised May 27, 2007. This work is partially based on research supported by NSF Grant IIS-0629454: "Learning Units on Law and Ethics in Software Engineering." Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

C. Kaner is with the Department of Computer Sciences, Florida Institute of Technology, Melbourne, Florida, USA (email: kaner@kaner.com).

R. L. Fiedler is with the Department of Education, St. Mary-of-the-Woods College, St. Mary-of-the-Woods, Indiana, USA (email: rfiedler@smwc.edu).

Gotterbarn, Miller, and Impagliazzo [3] suggest that using plagiarism detection tools would benefit the scholarly publication process by helping authors examine their own work and by allowing reviewers and editors to detect and deter plagiarism prior to publication.

How well do plagiarism detection tools work for detecting plagiarized work? This paper reports on a simple exercise designed to answer that question about two popular commercial tools. But first, the authors describe how the tools are generally used in academic settings.

At the Florida Institute of Technology, as a matter of policy, the Department of Computer Science checks all dissertations and theses for plagiarism using TurnItIn.com. The Department Chair and most faculty will usually accept a passing report from *TurnItIn* as definitive unless there are other obvious suggestors of plagiarism, such as distinct shifts in formatting or writing style or blatant inconsistencies between the student's apparent knowledge as reflected in the paper versus oral discussion. A similar process is typically followed, as a matter of faculty discretion rather than policy, for undergraduate essays, such as those submitted for the Computer Law, Ethics & Society course. Informal discussions with faculty at other universities suggest that this approach is widespread. Certainly, this is consistent with vendor guidance and with success stories reported on one popular vendor's website, which includes such assertions as "*TurnItIn's* plagiarism prevention is often so successful that institutions using our system on a large scale see measurable **rates of plagiarism drop to almost zero**." [emphasis as in the original] (See Fig. 1).

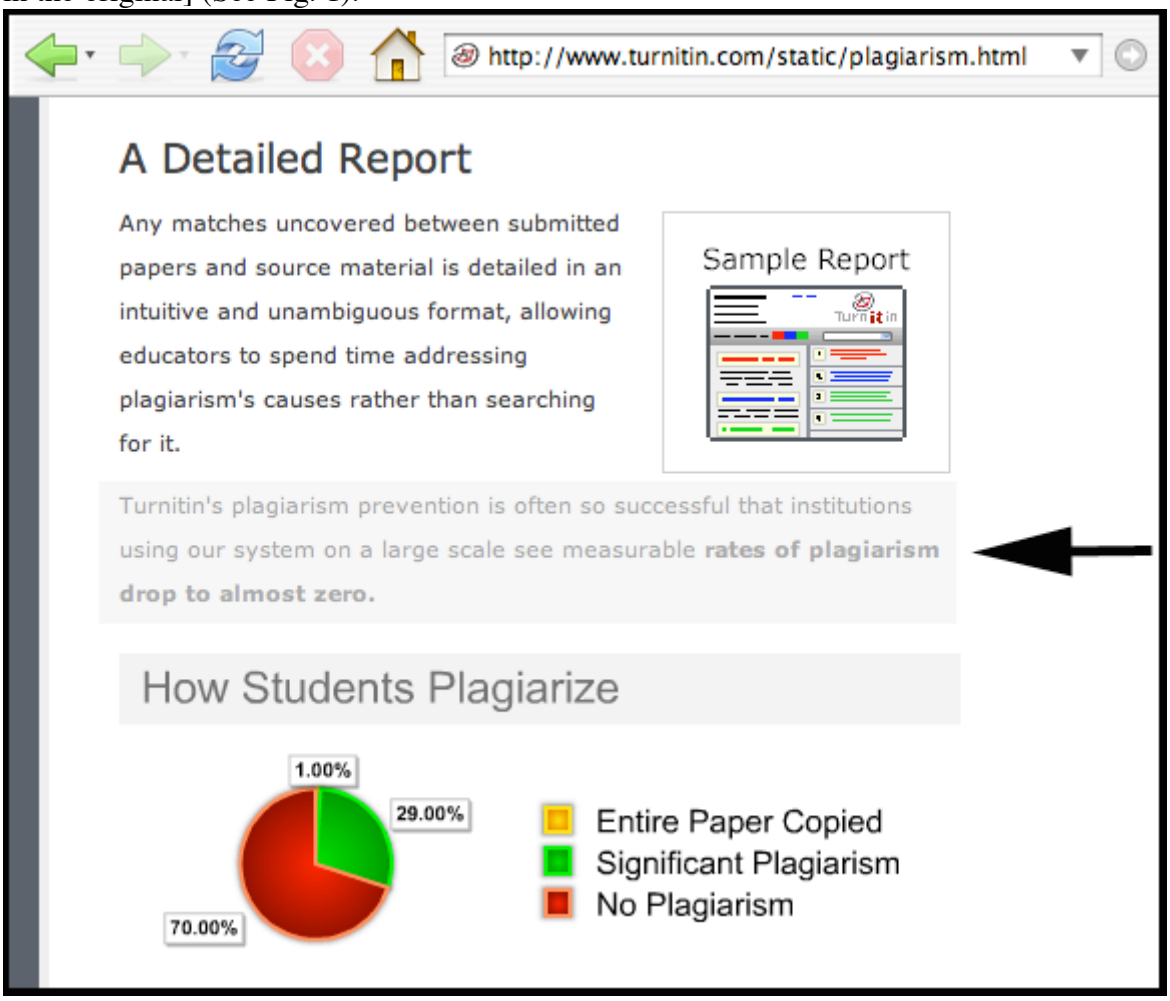

Fig. 1: Screenshot of *TurnItIn's* website

TurnItIn provides a valuable and convenient service. However, *TurnItIn* works by comparing writings to articles in its proprietary database and in some commercial or academic databases. If a given article is published in a journal not in the *TurnItIn* database, has not been posted on the web, and has not been submitted to *TurnItIn* for a plagiarism check in a way that allows *TurnItIn* to archive a copy of it, a plagiarized section of that article will not be detected by *TurnItIn*.

The *TurnItIn* website does not make clear which professional databases are included and which are excluded from their indexing. However, personal experience of the authors checking the writings of Florida Tech students and manuscripts submitted for publication by the Association for Software Testing caused the authors to wonder whether or not the service systematically checks ACM, IEEE or Springer databases. If not, these services might miss most plagiarism from the professional-level publication in software engineering.

II. RESEARCH QUESTION

Will plagiarism detection services (*TurnItIn* and *MyDropbox*) correctly identify obviously plagiarized submissions to their service?

III. METHOD

To explore this question, the authors selected thirteen papers from IEEE journals without consideration of the probability of their detection by the plagiarism detection services under investigation. The papers were selected because they looked interesting to read and relevant to other projects the authors were working on. Selections included [9-13] from a search of IEEE *Xplore* for articles on "whistle blowers"; [14-17] from a similar search for "plagiarism"; and [18-21] from the latest year's *IEEE Transactions on Education*.

Each selected paper was downloaded to one of the authors' computers. The PDF file of each complete downloaded paper was then submitted to two leading plagiarism detection services (*TurnItIn* and *MyDropbox*) in the same way a professor would submit student work to check for plagiarism. Thus, each experimental submission was 100% plagiarized from an IEEE paper and the authors of this paper expected the plagiarism detection services to flag each experimental submission as 100% plagiarized. Please note that the present authors are not asserting that the original papers were plagiarized. Instead, the experimental submissions in which the present authors played the role of a student submitting a published paper were 100% plagiarized.

IV. RESULTS

TurnItIn color-codes the results of their reports on a scale that runs low to high (blue/green/yellow/orange/red). For 10 of the 13 experimental submissions that were 100% plagiarized from IEEE papers, *TurnItIn* reported a similarity code of "blue" or "green" indicating there was little similarity in the experimental submission to the works the detection services searched even though the experimental submissions were 100% plagiarized.

Additionally, each plagiarism detection service reports a percentage of similarity between a submitted paper and other sources indexed by the service. Fig. 2 presents the raw data from the experimental submissions reported here. (Copies of the original reports are available as PDF files on request.) Notice that the similarity percentages reported by *TurnItIn* are often higher than those reported by *MyDropbox*, but *TurnItIn's* results are slightly inflated because they include

matches to quotations, matches to bibliographic entries, and many matches of short snippets of text to articles that contain no other matches to the submission.

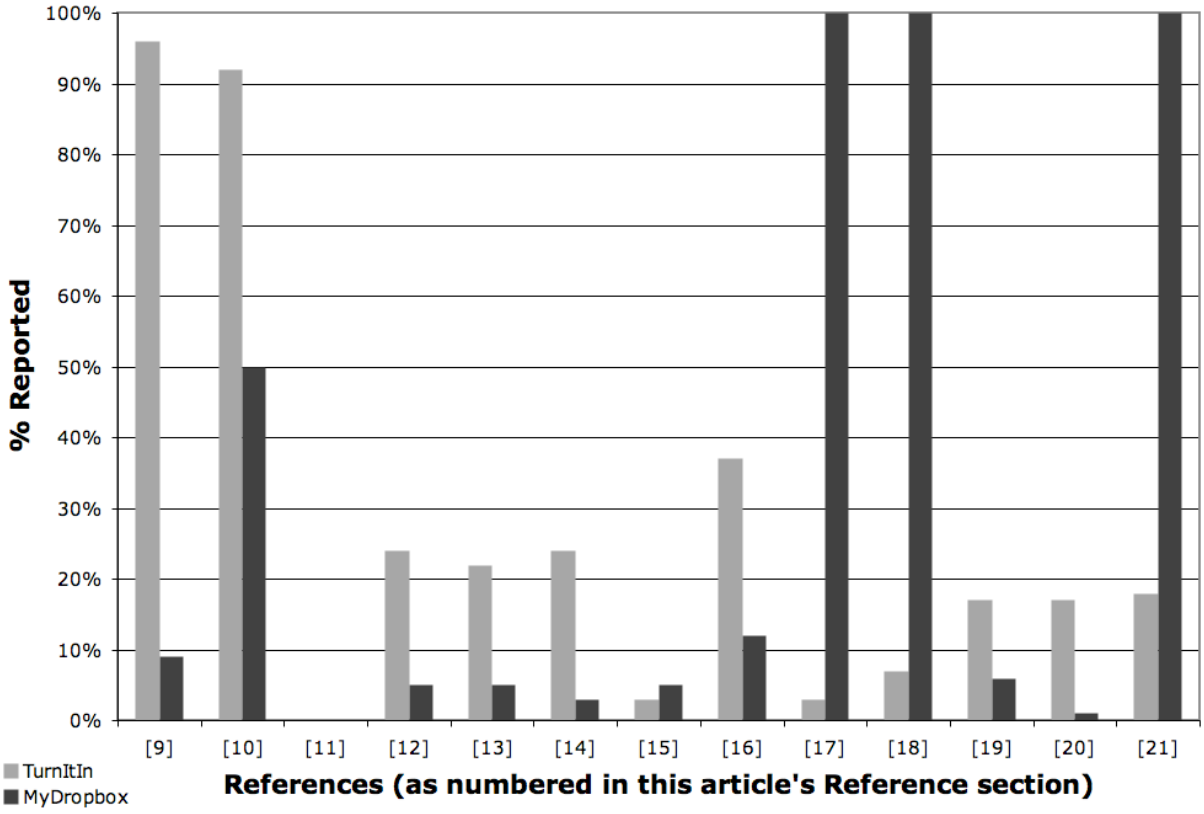

Fig. 2: Percentage of similarity reported in plagiarism detection services.

V. ANALYSIS

The authors evaluated the reports from each service, ignoring matches to the abstract, copyright notice, journal page numbers and running heads/footers. Based on their subjective assessment of the plagiarism detection services' reports, the authors categorized the results into three tiers in which each service reported an experimental submission was (a) obviously plagiarized, (b) possibly plagiarized and worth further careful study, or (c) apparently not plagiarized. Table 1 presents the results of that subjective assessment.

Reference	**Authors**	**TurnItIn**	**MyDropbox**
9	House, Watt & Williams (2004)	OP	ANP
10	Kumagai (2004)	OP	PP
11	Park (1996)	ANP	*
12	Tan, Smith, Keil & Montealegre (2003)	PP	ANP
13	Adams (2005)	ANP	ANP
14	Allen (2003)	ANP	ANP
15	Jansen, Van Lijf & Toussaint (2001)	ANP	ANP
16	Pertsemlidis & Garner (2004)	OP	ANP
17	Frincke	ANP	OP
18	Day & Foley	ANP	OP
19	Grigoriadou, Kanidis, Gogoulou (2006)	ANP	ANP
20	Massey, Ramesh, Khatri (2006)	ANP	OP
21	Nickerson (2006)	ANP	OP

Legend
OP Obviously plagiarized
PP Possibly plagiarized
ANP Apparently not plagiarized
* MyDropbox did not scan this paper
References as numbered in References section of this paper

Table 1: Three-tiered subjective assessment

TurnItIn exposed 3 papers and *MyDropBox* exposed 4 papers as obviously plagiarized. In each case, the plagiarism checker found what was actually an exact match, the paper having been posted to the public web by the author or the journal. Interestingly, the two services didn't find matches to the same paper. Both found some of the articles that were available on the open web and missed others.

Combining papers that are obviously plagiarized and papers that appear to warrant follow-up, *TurnItIn* missed 9 of 13 papers and *MyDropBox* missed 8 of 13. A strategy of submitting all papers to both engines yields slightly better success—combined, they miss only 5 of 13.

Finally, the authors skimmed each paper, looking for one or more memorable phrases to conduct a manual, full-text search in the IEEE *Xplore* database. Within 90 minutes, using no more than three matches per paper, the authors had found a memorable phrase that matched the original source, thereby exposing the plagiarism in each case. (See Table 2). Conducting a manual search is a time-consuming, subjective process that takes much longer to expose plagiarism when the student samples from many papers rather than merely copying one, but such searches do find sources that *TurnItIn* and *MyDropBox* miss.

Reference*	Paper	Memorable phrase
9	House, Watt & Williams (2004)	revealing corporate dishonesty with sacrificing one's life
10	Kumagai (2004)	anonymous remailers let people send
11	Park (1996)	self-regulation envisages a degree of responsibility
12	Tan, Smith, Keil & Montealegre (2003)	incentive to shirk because their interests diverge
13	Adams (2005)	FBI seized Scarfo's computer
14	Allen (2003)	affluent nations with well-established information infrastructures
15	Jansen, Van Lijf & Toussaint (2001)	acknowledgements are obligatory for moral reasons
16	Pertsemlidis & Garner (2004)	refinement of retrieved hits through iteration
17	Frinke & Bishop (2004)	running a honeynet or sniffing network data
18	Day & Foley	web lecture intervention
19	Grigoriadou, Kanidis, Gogoulou (2006)	means of a manual simulation of the cache
20	Massey, Ramesh, Khatri (2006)	campus wireless network and mobile devices MAD
21	Nickerson (2006)	underlying technical and social mechanisms of integration
* Reference as numbered in References section of this paper		

Table 2: Memorable phrases used to conduct a manual search for IEEE papers

VI. RECOMMENDATIONS

One cannot draw statistical conclusions from the work reported here because the sample was intentionally small and is not necessarily representative. The intent of this research is to bring attention to what should be an obvious problem, not to quantify it. However, based on these results and other informal experience, the authors suggest the following:

1. Plagiarism-checking services are convenient, if not necessarily powerful for work submitted in the engineering field. When each copied paragraph comes from a different source, services that compare articles against a large database can make the detection of plagiarism much

easier than is the case when a manual search is used. However, such services miss sources that are freely available on the web. Surprisingly, even though both services search the public web, this demonstration shows that they don't search it in the same way or find the same matches. Additionally, the services' lists of sources indicate that their access to professional databases – the primary collections of research publications – only partially overlap.

Recommendation: Submit writings to multiple services rather than to only one.

2. Because their databases miss a large portion of the professional engineering literature, a report of little similarity by current plagiarism detection services is untrustworthy. A manual search of the professional literature, for example searching for matches to a few memorable phrases, can expose papers that the plagiarism services do not report.

   Recommendation: Especially if there are any other suggestions of copying, follow up with a full-text search of IEEE Xplore and the other appropriate professional databases such as the ACM Guide to the Computing Literature or SpringerLink.

3. When so much of the professional literature is missed by the plagiarism-checking service, allowing students to submit drafts for checking creates a training ground for plagiarists. They can readily switch copying from articles that are detected by the service to articles the service does not find. Rather than concluding, in an engineering course, that "rates of plagiarism drop to almost zero" when the plagiarism-detection service is used, colleagues should consider the possibility that rates of plagiarism *detection* drop to zero as students learn what the service will and will not detect.

   Recommendation: Do not make it easy for students to use detection tools to check their drafts for plagiarism.

4. For these tools to be genuinely useful, instead of falsely reassuring, they must have access to professional research literature.

   Recommendation: The professional societies must work out a licensing structure that gives plagiarism detection services access to the professional literature so that teachers, editors, and manuscript reviewers, can time-efficiently determine whether a submitted work has been plagiarized.

REFERENCES

[1] P. Wasley, "The Plagiarism Hunter," in *The Chronicle of Higher Education*, vol. 52 (49), p. 7. Washington, DC, 2006.

[2] J. Zobel, ""Uni cheats racket": A case study in plagiarism investigation," presented at Sixth Conference on Australasian Computing Education, Dunedin, New Zealand, 2004.

[3] D. Gotterbarn, K. Miller, and J. Impagliazzo, "Plagiarism and scholarly publications: An ethical analysis," in *36th ASEE/IEEE Frontiers in Education Conference*. San Diego, CA USA, 2006.

[4] F. Mintzer, "Join the fight against plagiarism [President's Message]," *IEEE Signal Processing Magazine*, vol. 22, p. 4, 2005.

[5] R. F. Boisvert and M. J. Irwin, "ACM plagiarism policy: Plagiarism on the rise," *Communications of the ACM*, vol. 49, pp. 23-24, 2006.

[6] A.S.S.E.N. staff, "Frontmatter (ACM policy and procedures on plagiarism)," *ACM SIGSOFT Software Engineering Notes*, vol. 31, pp. 2-3, 2006.

[7] N. Ree-Lindstad, K. Roijen, and T. Vold, "Experience with a plagiarism control module," in *ITHET '06. 7th International Conference on Information Technology Based Higher Education and Training*. Ultimo, Australia: IEEE, 2006.
[8] S. Reisman, "Plagiarism or ignorance? You decide," *IT Professional*, vol. 7, pp. 7-8, 2005.
[9] R. House, A. Watt, and J. M. Williams, "Teaching Enron: The rhetoric and ethics of whistle-blowing," *IEEE Transactions on Professional Communication*, vol. 47, pp. 244-255, 2004.
[10] J. Kumagai, "The whistle-blower's dilemma," *IEEE Spectrum*, vol. 41, pp. 53-55, 2004.
[11] P. D. Park, "Whistleblowing: an employer's view. How vulnerable is your company?," *Engineering Management Journal* , vol. 6, pp. 183-186, 1996.
[12] B. C. Y. Tan, H. J. Smith, and M. Keil, "Reporting bad news about software projects: Impact of organizational climate and information asymmetry in an individualistic and collectivistic culture," *IEEE Transactions on Engineering Management*, vol. 50(1), pp. 64-77, 2003.
[13] C. W. Adams, "Legal requirements for the use of keystroke loggers," presented at First International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE '05), 2005.
[14] M. Allen, "Dematerialised data and human desire: the Internet and copy culture," presented at International Conference on Cyberworlds (W'03), 2003.
[15] F. Jansen, A. Van Lijf, and E. Toussaint, "A note on the evaluation of footnotes and other devices for background information in popular scientific texts," *IEEE Transactions on Professional Communication*, vol. 44, pp. 195-201, 2001.
[16] A. Pertsemlidis and H. R. Garner, "Engineering in genomics: text comparison based on dynamic programming," *IEEE Engineering in Medicine and Biology Magazine*, vol. 23, pp. 66-71, 2004.
[17] D. Frincke and M. Bishop, "Back to school [security education]," *IEEE Security & Privacy*, vol. 2, pp. 54-56, 2004.
[18] J. A. Day and J. D. Foley, "Evaluating a web lecture intervention in a human-computer interaction course," *IEEE Transactions on Education*, vol. 49, pp. 420-431, 2006.
[19] M. Grigoriaou, E. Kanidis, and A. Gogoulou, "A Web-based educational environment for teaching the computer cache memory," *IEEE Transactions on Education*, vol. 49, pp. 147-156, 2006.
[20] A. P. Massey, V. Ramesh, and V. Khatri, "Design, development, and assessment of mobile applications: The case for problem-based learning," *IEEE Transactions on Education*, vol. 49, pp. 183-192, 2006.
[21] J. V. Nickerson, "Teaching the integration of information systems technologies," *IEEE Transactions on Education*, vol. 49, pp. 271-277, 2006.

AUTHOR BIOS

Cem Kaner is a Senior Member of the IEEE. He received the B.A. degree in Arts and Sciences (primarily mathematics and philosophy) from Brock University, St. Catharines, Canada in 1974, the Ph.D. degree in experimental psychology from McMaster University, Hamilton, Canada, in 1984, and the J.D. degree from Golden Gate University, San Francisco, USA, in 1994.

He is Professor of Software Engineering and Director of the Center for Software Testing Education & Research at the Florida Institute of Technology, Melbourne, Florida. From 1983-2000, he worked in Silicon Valley as a programmer, tester, human factors analyst, project manager, test manager, technical publications manager, development director, independent consultant and attorney focused on the law of software quality at WordStar, Telenova, Power Up Software, Electronic Arts, kaner.com, and the Law Office of Cem Kaner. He is the editor of the *Journal of the Association for Software Testing*.

Dr. Kaner is a member of the American Bar Association, the American Law Institute, the American Psychological Association, the American Psychological Society, the American Society for Quality, the Association for Computing Machinery, the Association for Software Testing, the Human Factors and Ergonomics Society and the International Technology Law Association. He is licensed as an attorney in the State of California.

Rebecca L. Fiedler received an M.B.A. degree and Ph.D. degree in education from the University of Central Florida in 1995 and 2006, respectively.

She is an Assistant Professor of Education at St. Mary-of-the-Woods College in Indiana. She teaches in an online Master of Education degree and her research interests include online instruction, portfolio assessment, activity theory, and technology fluency.

Dr. Fiedler is a member of the American Educational Research Association, the International Society for Technology in Education, and the Association for Supervision and Curriculum Development.