**Doctoral thesis submitted to**
**the Faculty of Behavioural and Cultural Studies**
**Heidelberg University**
**in partial fulfillment of the requirements of the degree of**
**Doctor of Philosophy (Dr. phil.)**
**in Psychology**

Title of the publication-based thesis

# *Improving Learning and Teaching at Universities: The Potential of Applying Automatic Essay Scoring with Latent Semantic Analysis*

presented by
Dipl.-Psych. Eva Seifried

year of submission
2016

Dean: Prof. Dr. Birgit Spinath
Advisor: Prof. Dr. Birgit Spinath

## Acknowledgments

First and foremost, I would sincerely like to thank my advisor Prof. Dr. Birgit Spinath for her constant support, her continuous interest in my work, her feedback, and for giving me the freedom to develop and pursue my research interests and ideas. I am also grateful to Prof. Dr. Jörg Zumbach for agreeing to act as a referee for this thesis.

I would sincerely like to thank Prof. Dr. Wolfgang Lenhard for our pleasant and successful collaborations.

Moreover, I would like to thank all my colleagues for all of our interesting and fruitful discussions and for creating a pleasant work atmosphere. In particular, and in no specific order, I would like to thank Heike Dietrich, Christine Eckert, and Patrick Schaller. I also wish to thank Dr. Herbert Baier and Fabian Grünig for their technical support with the learning platforms, our group's teaching assistants, and Dr. Jane Zagorski for her proofreading.

Finally, I gratefully acknowledge the financial support I received from the Innovation Fund FRONTIER at Heidelberg University (project number D.801000/10.25).

Special thanks are due to my family for supporting me in everything I do.

## Summary

The aim of this dissertation was to find ways to improve learning and teaching at universities by analyzing whether the application of new technologies would facilitate the implementation of an effective teaching-learning format in which students write essays and are given feedback. More specifically, Latent Semantic Analysis (LSA) as a semantic technology that can be used for automatic essay scoring (AES) was applied for several purposes to facilitate essay writing in large university courses as part of an overarching strategy to improve learning and teaching at universities, that is, evidence-based teaching (EBT).

In this dissertation, I will summarize and discuss findings regarding good learning and teaching (i.e., EBT) as well as why and how essay writing should be used in university courses, and regarding AES and LSA. Further, I will provide my own empirical findings on different ways to apply LSA in university courses: First, when students write essays at home, cheating must be expected, detected, and avoided. Thus, in Paper I, we analyzed whether LSA could be used to detect cheating in a large university course. Second, due to capacity constraints, instructors might need to focus their time and energy on students who are in need of special guidance. Thus, in Paper II, we investigated whether LSA could be used to identify poorly performing students. Third, before applying LSA for essay scoring, the effects of LSA-based evaluations should be explored. Thus, in Paper III, we analyzed the effects of LSA-based scores on students' acceptance of automatic assessments and on learning-related characteristics. I will discuss these findings critically and conclude that LSA should not be used alone but is useful for *assisting* university instructors in different ways.

## List of Papers Included in the Publication-Based Dissertation

**I. Paper**

Seifried, Eva, Lenhard, Wolfgang & Spinath, Birgit (2015). Plagiarism detection: A comparison of teaching assistants and a software tool in identifying cheating in a psychology course. *Psychology Learning and Teaching, 14*, 236-249, SAGE Publications. doi: 10.1177/1475725715617114

**II. Paper**

Seifried, Eva, Lenhard, Wolfgang & Spinath, Birgit (2016). Filtering essays by means of a software tool: Identifying poor essays. *Journal of Educational Computing Research 0735633116652407, SAGE Publications, first published on June 6, 2016 as doi:10.1177/0735633116652407*.

**III. Paper**

Seifried, Eva, Lenhard, Wolfgang & Spinath, Birgit (accepted pending revisions). Automatic essay assessment: Effects on students' acceptance and on learning-related characteristics. *Psihologija*.

This is a publication-based dissertation. The above-mentioned publications are the three core publications.
Moreover, the following publications are closely related to the dissertation's topic (none of them were part of my diploma thesis):

Eckert, C., Seifried, E., & Spinath, B. (2015). Heterogenität in der Hochschullehre aus psychologischer Sicht: Die Rolle der studentischen Eingangsvoraussetzungen für adaptives Lehren. In K. Rheinländer (Hrsg.), *Ungleichheitssensible Hochschule* (S. 257-264). Heidelberg: Springer.

Seifried, E., Eckert, C., & Spinath, B. (2014a). Eingangs- und Verlaufsdiagnostik von Lernvoraussetzungen und Lernergebnissen in der Hochschullehre. In M. Krämer, U. Weger & M. Zupanic (Hrsg.), *Psychologiedidaktik und Evaluation X* (S. 267-274). Aachen: Shaker.

Spinath, B. & Seifried, E. (2012). Forschendes Lehren: Kontinuierliche Verbesserung einer Vorlesung. In M. Krämer, S. Dutke & J. Barenberg (Hrsg.), *Psychologiedidaktik und Evaluation IX* (S. 171-180). Aachen: Shaker.

Spinath, B., Seifried, E., & Eckert, C. (2014). Forschendes Lehren: Ein Ansatz zur kontinuierlichen Verbesserung von Hochschullehre. *Journal Hochschuldidaktik, 25(1-2)*, 14-16.

Spinath, B., Seifried, E. & Eckert, C. (in press). Forschendes Lehren: Ein Ansatz zur kontinuierlichen Verbesserung von Hochschullehre. In M. Heiner, B. Baumert, S. Dany, T. Haertel, M. Quellmelz & C. Terkowsky (Hrsg.), *Was ist gute Lehre – und was kann die Hochschuldidaktik dazu beitragen?*

## Contents

# 1. Introduction

The scenery in higher education is changing: Because students' cognitive and motivational prerequisites are becoming more divergent, university instructors are challenged to enhance their students' learning and to ensure their motivation. Further, due to the Bologna process, it is necessary to assess students more often. Many instructors use easy-to-score methods (e.g., multiple-choice items), and they use such methods only once, that is, on an exam at the end of the semester. However, a major goal of universities is to enable students to analyze and evaluate complex contents and to foster sustainable learning. It is questionable whether these aims can be ensured by the traditional approach mentioned above.

An alternative to the common teaching format might be to let students write essays. With students writing essays continuously throughout the semester, instructors can both help students to apply effective learning techniques and assess their performance in the form of a formative evaluation. This alternative and its implications are in line with several ideas about good teaching. However, reading and assessing essays is time-consuming and cost-intensive so that many instructors refrain from applying essay writing. Automated essay scoring (AES) might be a useful alternative or might at least provide assistance when teachers are responsible for large classes. There are various techniques that can be applied to score essays (semi-)automatically, but most have been developed and tested in the English language. Several techniques have been evaluated positively and are gaining more popularity, for example, Latent Semantic Analysis (LSA), with which the content of an essay can be assessed. In Germany, however, there have been only some first trials in which LSA was applied for

AES. Because these first studies have provided encouraging results, it is worthwhile to test whether and in which ways LSA might be used to assist university instructors.

In the following, I will begin with a chapter that includes some general remarks about learning and teaching at universities today. I will explain the meaning and relevance of evidence-based teaching (EBT) and derive why and how essays should be used in university courses. The problems involved in using essays in large classes will lead to the chapter on AES with its pros and cons and LSA as a possible AES approach. In the following chapter, I will present three empirical studies that tested the application of LSA-based scores at universities: LSA was used to detect cheaters (Paper I: Seifried, Lenhard, & Spinath, 2015) and to identify poorly performing students (Paper II: Seifried, Lenhard, & Spinath, 2016). Finally, the effects of LSA-based evaluations on students' acceptance of automatic assessments and on learning-related characteristics were investigated (Paper III: Seifried, Lenhard, & Spinath, accepted pending revisions). There are also some other publications on our first attempts to use LSA to score complex German student-authored texts (Seifried, 2010; Seifried, Lenhard, Baier, & Spinath, 2012) and our concept of practicing EBT (i.e., *Forschendes Lehren*; Eckert, Seifried, & Spinath, 2015; Seifried, Eckert, & Spinath, 2014a; Spinath & Seifried, 2012; Spinath, Seifried, & Eckert, 2014, in press). These are not included in this dissertation but contributed to the idea of how to improve teaching and learning at universities. I will end my dissertation with a final discussion about the application of LSA to detect plagiarism and to score essays, which will result in a general conclusion on how LSA can be used to assist university instructors to improve their teaching and their students' learning.

## 2. Learning and Teaching at Universities

In this century, there have been some major changes in the higher education sector (see e.g., Altbach, Reisberg, & Rumbley, 2009; OECD, 2015).

> It is clear that over the last 10 years, real momentum for change in university approaches to teaching and learning has emerged in at least some parts of the world. The challenges of producing those changes across systems, institutions, and disciplines, however, are significant. The traditional research-based university will still exist, but privatization, massification, and commodification greatly increase the need for prioritizing teaching, learning, and assessment, and for effecting changes that are is [sic] anchored in credible scholarship and proven strategies. (Altbach, et al., 2009, p. 120)

Due to the massification of higher education and globalization (e.g., student mobility and movements of internationalization such as the Bologna Process in Europe), more students and more diverse students are coming to universities. Thus, it is important for university instructors to consider their students' heterogeneity and adapt their teaching to their students' prerequisites (Eckert et al., 2015). According to Biggs and Tang (2011), instructors can cope with academic diversity by improving teaching and learning. These authors argue that instructors can reduce the gap between the "academic Susans" and the "nonacademic Roberts" by helping the Roberts to learn more like the Susans. To this end, teachers should use active teaching methods that force students to use deep approaches to learning. Further, in Europe, the Bologna Process has led to outcome-based learning and teaching, and thus, there is now a greater need for teachers to assess their students' learning more often. This change is beneficial in that it might

lead to a permanent reflection of one's "impact" as a teacher (see Hattie, 2015) but might also be a heavy burden when classes are large.

In this chapter, I will introduce methods of evidence-based teaching (EBT) as a means for practicing good teaching and for coping with the problems mentioned above, followed by ideas on why and how to use essay writing in university courses.

## 2.1 Evidence-Based Teaching (EBT)

Several authors have called for EBT (e.g., Benassi, Overson, & Hakala, 2014; Cranney, 2013; Dunn, Saville, Baker, & Marek, 2013; Schwartz & Gurung, 2012). They claim that teaching and learning can be improved if instructors apply certain principles that have been shown to be effective empirically. There have been some synopses of theoretically based and empirically investigated principles of learning and teaching. For example, Graesser, Halpern, and Hakel (2008) listed "25 principles of learning", Pashler et al. (2007) made seven recommendations to improve student learning, and Dunn et al. (2013) examined five areas of evidence (for details, see Table 1).

Table 1

*Summary of Theoretically Based and Empirically Investigated Recommendations for Learning and Teaching*

| Graesser et al. (2008): 25 principles of learning | | Pashler et al. (2007): Seven recommendations | Dunn et al. (2013): Five areas of evidence |
|---|---|---|---|
| 1. Contiguity effects | 14. Desirable difficulties | 1. Space learning over time. | 1. The testing effect |
| 2. Perceptual-motor grounding | 15. Manageable cognitive load | 2. Interleave worked example solutions with problem-solving exercises. | 2. Spaced learning |
| 3. Dual code and multimedia effects | 16. Segmentation principle | 3. Combine graphics with verbal descriptions. | 3. Metacognition: Thought about thinking |
| 4. Testing effect | 17. Explanation effect | 4. Connect and integrate abstract and concrete representations of concepts. | 4. Writing to learn |
| 5. Spacing effect | 18. Deep questions | 5. Use quizzing to promote learning. | 5. Interteaching |
| 6. Exam expectations | 19. Cognitive disequilibrium | 6. Help students allocate study time efficiently. | |
| 7. Generation effect | 20. Cognitive flexibility | 7. Ask deep explanatory questions. | |
| 8. Organization effect | 21. Goldilocks principle | | |
| 9. Coherence effect | 22. Imperfect metacognition | | |
| 10. Stories and example cases | 23. Discovery learning | | |
| 11. Multiple examples | 24. Self-regulated learning | | |
| 12. Feedback effects | 25. Anchored learning | | |
| 13. Negative suggestion effects | | | |

In addition to these specific principles that instructors might apply, they should also follow a general approach to improve their teaching and learning, that is, they should monitor their actions and effects. From his prominent synopsis of over 800 meta-analyses on the effects of 128 influences on student achievement, Hattie (2009) argued for monitoring one's actions in teaching with the help of empirical data: There is no single principle that will work in all situations at every time. Rather, teachers should critically reflect on their actions in light of evidence. "[T]hose teachers who are students of their own effects are the teachers who are the most influential in raising students' achievement" (Hattie, 2009, p. 24). Thus, "Know thy impact" is his mantra and recommendation for teachers (Hattie, 2012, p. 169). In this sense, teachers should

analyze whether their students are making progress and should adapt their teaching to the needs of their students. Further, teachers should help students learn to recognize on their own when they are not making progress and should help students learn to figure out how to improve their performance (see also Hattie, 2011, for "three claims for higher education"). These recommendations are also in line with ideas by Biggs and Tang (2011) who stated that teachers should focus on what the students do and that (transformative) reflective practice is important for effective teaching (i.e., [repeatedly] reflect, plan, apply, evaluate; also called *action research*). Teachers should collect students' feedback to see how teachers might improve their teaching (see also the "scholarly practitioner"; Cranney, 2013). Further, monitoring one's own progress is important for students as well; Biggs and Tang (2011) refer to this as "metacognitive control" and "reflective learning" (p. 60).

To systematically improve one's teaching by applying effective techniques and using one's competencies as an educational psychologist was also the idea behind a concept called *Forschendes Lehren*, which was developed in Heidelberg (Spinath & Seifried, 2012; Spinath et al., 2014, in press). Forschendes Lehren can be understood as an approach that is applied to continuously improve teaching in higher education by systematically investigating one's own didactic actions in an iterative cycle. This idea is similar to the concepts of the educator-scientist model (Bernstein et al. 2010) or the scholarship of teaching and learning (e.g., Boyer 1990; Huber, 2011; Huber, Pilniok, Sethe, Szczyrba, & Vogel 2014; Hutchings, Huber, & Ciccone 2011). The cycle includes seven steps: becoming acquainted with the principles of good teaching and learning (Phase 1), comparing whether these are embedded in one's own teaching and testing the effects of one's teaching (Phase 2), and then initiating empirical studies: Phase 3 includes deducing questions and hypotheses about how to improve the current

teaching-learning arrangement on the basis of theory and empirical work, and these are tested with an adequate research design in Phase 4. When a method has proven successful in practice, it should be implemented (Phase 5). Phase 6 includes the contribution to theory development, the deduction of further questions and hypotheses that may arise from the former study. Ultimately, Phase 7 includes the iterative process by which Phases 4 to 6 are performed again and again. With a process like this, the goal is not only to improve one's own teaching but also to produce generalizable knowledge about good teaching and learning that can then be shared within the scientific community through publications (i.e., the "scientist practitioner" as "the 'gold standard' for psychology research and practice"; Cranney, 2013; p. 1). There are several aims of Forschendes Lehren for learners, instructors, and teaching quality (e.g., increases in learning, motivation, and satisfaction as well as enforced EBT; see Spinath et al., 2014).

Because educational psychologists are experts in the field of teaching and learning, they are predestined to both practice EBT and encourage students to use techniques that have been shown to be effective empirically. Helping students use effective techniques is particularly important because students do not seem to know what is good for them: Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) reviewed 10 learning techniques and estimated their utility by evaluating their generalizability. They found that some of the learning techniques that students assume to be effective and thus use very frequently were not very effective in reality (e.g., highlighting/underlining, rereading). Thus, instructors might force and thereby help students to apply some useful learning techniques by modifying the teaching arrangement.

To summarize thus far, it can be said that there are several learning and teaching techniques that have proven useful. Further, educational psychologists should be the

first to practice what they preach and to continuously improve their teaching. In the next paragraph, I will outline why and how essay writing might be used to achieve this aim.

## 2.2 Essay Writing in University Courses

Although there are many positive findings for testing students with multiple-choice-like quizzes (for a differentiated synopsis of results, see e.g., Nguyen & McDaniel, 2015), there are some advantages of including writing in teaching. For example, McGovern and Hogshead (1990) reviewed writing activities and listed four objectives for including them in psychology courses, namely, assessing students, promoting learning, developing student writing skills, and facilitating analytic and creative thinking and problem solving (p. 6). They referred to the writing-across-the-curriculum movement to list some further examples of why teachers should ask their students to write (e.g., to foster involvement, provide the instructor with information on how well students are doing).

Further, with their book about teaching for quality learning at university, Biggs and Tang (2011) argued for constructive alignment; that is, teachers should state their intended learning outcomes, ensure that students perform learning activities that are helpful for achieving the intended learning outcomes (because learners construct meaning from what they do to learn), and assess students' performance against the intended learning outcomes. The authors stated that multiple-choice questions should be avoided or should be used only for quizzes because such questions encourage students to apply a surface approach to learning. However, if teachers had higher order goals for their students, teachers should not allow students to get away with surface approach strategies but should encourage learning activities that imply a deep approach. Although multiple-choice questions can theoretically be used to assess higher order outcomes and

to trigger retrieval processes, they seldom do so because it is very hard to generate appropriate items (see e.g., Little, Bjork, Bjork, & Angello, 2012). Multiple-choice examinations might also send the wrong signals to students: Students presume that multiple-choice question examinations versus assignment essays require rather low- versus high-level cognitive processes, respectively, and so students apply them accordingly (e.g., Scouller, 1998). Hence, when university instructors want to align their teaching activities, their students' learning activities, and the assessment tasks with the (higher level) intended learning outcomes – as suggested by Biggs and Tang (2011) – instructors should use different tasks. Tasks that aim to capture functioning knowledge are needed. There is room for declarative knowledge and its assessment because students need to have knowledge about something to be able to apply such knowledge. However, in general, instructors should focus on application. Essays seem to be a good way to ask for actions that are at least relational according to the SOLO model (Biggs & Collis, 1982), for example, compare and contrast, analyze, or apply. In conclusion, Biggs and Tang (2011) stated:

> MCQ [multiple-choice question] items are best avoided. Too readily they address lower order ILOs [intended learning outcomes]. Essays have a better potential for assessing higher level understanding of declarative knowledge such as explain, argue, analyze, and compare and contrast. (p. 226–227)

In sum, there are good reasons to use essays both as learning activities and assessment tasks. Thus, there have been many ideas about how to include writing in university courses (for some examples, see Paper II). Considering the recommendations for learning and teaching mentioned above, I have deduced the following recommendation for the best way to apply essay writing:

> During the semester, students should write essays that answer challenging questions, include key concepts, and call for applications of the material that was taught, and students should receive timely feedback on their ideas.

A teaching format like this would meet several "characteristics of good learning contexts" (Biggs & Tang, 2011, p. 60) and include several recommendations mentioned above: First, timing matters: Essays can be used to distribute students' learning. When students are asked to write an essay every week, they have to space out their learning in small portions and can be exposed to key concepts several times (see e.g., the following recommendations mentioned above: space learning over time, Pashler et al., 2007; spacing effect, Graesser et al., 2008; spaced learning, Dunn et al., 2013). Second, active engagement matters: When writing an essay, students have to produce something actively. This is advantageous because production has been shown to be advantageous over recognition, rereading material, or other passive strategies (see e.g., the following recommendations mentioned above: generation effect, organization effects, Graesser et al., 2008). Third, the type of question matters: Essays can be used to foster a deep approach to learning. Questions that ask for elaboration (e.g., explanations, analyses, comparisons, applications, or other higher level understanding) can be addressed by essays (see Biggs & Tang, 2011). If instructors set corresponding tasks that go beyond the pure reproduction of factual knowledge, this falls in line with some recommendations made by Pashler et al. (2007; e.g., ask deep explanatory questions) and Graesser et al. (2008; e.g., use deep questions, explanation effects, anchored learning). The questions should be challenging enough to make use of the effect of desirable difficulties but still "just right" to fulfill the Goldilocks Principle (see corresponding recommendations by Graesser et al., 2008). Most likely, the questions

might also induce cognitive disequilibrium if there is a problem stated within the essay question (see recommendation by Graesser et al., 2008). Beyond these recommendations that have been shown to enhance *learning*, another aspect that should be considered when creating essay questions is students' *motivation*. Assigning a certain value to a task and expecting to be able to successfully complete it should have positive effects on students' motivation (see expectancy-value theory; Wigfield & Eccles, 2000). Thus, using meaningful application-oriented and challenging but attainable tasks should enhance students' motivation. Instructors should make sure that students enjoy the tasks, that students are challenged by them, identify with them, and see their use for students' professional lives (e.g., by taking examples from the students' present or future lives). Fourth, the reaction to the essay matters: If students receive feedback on their essays, this should also foster their learning. If students are not asked to repeat their knowledge but rather to relate new knowledge to their prior knowledge and question some former beliefs in light of evidence (e.g., by comparing their opinions with empirical studies), the essays can be used to reveal students' misconceptions, and feedback might help to correct them and transform the students' knowledge. To achieve such a restructuring of knowledge (see also Pearsell, Skipper, & Mintzes, 1997) or to make them apply their knowledge to practical situations, students have to be active and must be given formative feedback (see also Hattie, 2011). Students might see the essays as a way to test themselves and to realize what they have (not yet) understood (for the positive effects of feedback in general, see e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; see also, e.g., the following recommendations mentioned above: feedback effects and the possibility of avoiding negative suggestion effects by offering immediate feedback, Graesser et al., 2008; help students allocate study time efficiently, Pashler et al., 2007). Thus, in

combination with continuous learning, writing essays and receiving feedback might help students to monitor and reflect on their progress. And fifth, writing itself matters: Writing to learn has been listed as a tool for EBT as such (Dunn et al., 2013).

In sum, essay writing can be applied in such a way that it is in line with many principles that have been shown to be advantageous for enhancing learning (and motivation).[1] Further, asking students to work continuously throughout the semester and continuously gathering data on students' understanding and other learning-related characteristics (e.g., students' motivation) involves the opportunity to monitor the students' progress and one's own impact as a teacher (i.e., an important part of Forschendes Lehren: to monitor one's teaching and critically reflect on it on the basis of empirical evidence; see also Hattie, 2009).

Because time in class is scarce, instructors might give students the choice of whether to attend the lecture sessions but ask them to write essays at home. Granting such liberties might be seen as a sign of high trust and might encourage high-value outcomes (Biggs & Tang, 2011). However, there also are some problems with this procedure. First, asking students to write essays at home might provide leeway for cheaters. There are some ideas about how to minimize this risk. For example, asking questions that refer to personal experiences should help students become or remain motivated and should also weaken the risk of plagiarism (see e.g., recommendations by Warn, 2006). Further, instructors should pose essay questions that require a deeper understanding, for example, questions that ask students to analyze complex relations, to build a personal opinion on the basis of empirical studies, to abstract ideas and find concrete examples of principles, or to connect different facets. Such questions are useful

[1] Our own research shows that answering open-ended questions throughout the semester is a good way to prepare for an exam with different item formats (i.e., forced-choice items and open-ended questions; see Eckert, Seifried, & Spinath, 2014; Seifried, Eckert, & Spinath, 2014b) and that it fosters sustainable learning (Spinath, 2011; see also Blümel, 2013; Lange, 2014).

because, to answer questions like these, students cannot simply write down what was presented in the lecture, and they cannot find the answers in a textbook or on the Internet. Rather, they have to think about the material and its implications. However, although a concept like the one described is certainly beneficial for enhancing students' learning and motivation and at the same time for reducing plagiarism, there still might be some persons who try to cheat (especially by copying from another student), and these persons must be detected. Second, besides the problem of plagiarism, which needs to be solved, there is also a capacity problem: Although letting students write essays and giving them feedback is a desirable teaching-learning format, it is difficult to implement in large courses. When the number of students increases, this might lead instructors to refrain from using corresponding teaching practices or to leave students with minimal feedback only. Having only a little teacher-learner interaction might be especially detrimental for struggling students because they cannot self-regulate their learning and need special guidance in order to improve. Thus, it might be helpful to identify them in order to provide them with more and individual feedback.

The goal of this dissertation was to analyze whether a software tool that can (semi)automatically evaluate essays might be helpful for solving these problems and might hence offer a way to improve teaching and learning at universities (for other ways to use technology to support learning and teaching in higher education, see e.g., Fisher, Exley, & Ciobanu, 2014).

# 3. Automatic Essay Scoring (AES)

AES has been defined as "the ability of computer technology to evaluate and score written prose" (Shermis & Burstein, 2003, p. XIII). Through the use of learning platforms, students can hand in their texts electronically, and AES might be helpful for detecting cheaters and identifying poorly performing students who are in need of special guidance. Thus, AES might solve both problems mentioned above. One way to fulfill these tasks is to use a special approach from the field of automatic language processing, that is, LSA. In this chapter, the pros and cons of AES in general are outlined, followed by an illustration of LSA as a specific AES method.

## 3.1 Pros and Cons

In general, the positive aspects of AES are manifold: (Semi-)Automatically generated scores might be more accurate, producible with lower costs and in a shorter amount of time, and be used to address several research questions involving, for example, the observation of yearly trends or group differences (Page & Peterson, 1995). The arguments regarding costs and efficiency seem to be obvious: Computerized feedback might be applied more frequently and more quickly because grading a large number of essays with a computer can be done in (milli-)seconds, whereas human graders need at least several minutes to assess only one essay. Further, replacing a human grader by a computer might save a substantial amount of money. In general, having students submit essays electronically might also save material costs such as paper. In today's universities, students usually have personal or institutional access to computers and the Internet. Thus, electronic text is available as an input for online platforms (for some

further arguments about the influences of technology on AES, see Shermis, Burstein, & Bursky, 2013).

Another more general argument in favor of AES – which is well-founded in the literature mentioned earlier (e.g., Biggs & Tang, 2011) – is the following: If AES can be used to offer immediate performance feedback, this might encourage instructors to use essays as both a learning tool and an assessment tool and therefore ensure the alignment of tasks and a shifting away from multiple-choice examinations toward methods that concentrate on deeper understanding (for this argument, see also Landauer, Laham, & Foltz, 2003a; Williamson, 2013). In this regard, computerized feedback might be (perceived as) less biased and more objective and be taken less personally than feedback given by an instructor (see e.g., Hattie, 2009). This might be especially important for students who are performing poorly (see also Lipnevich & Smith, 2009a, 2009b).

Probably one of the most critical aspects of the use of AES is its reliability and validity. Several authors have addressed the question of human graders' and/or AES's reliability and validity, both theoretically and/or empirically (e.g., Attali, 2013; Bridgeman, 2013; Chung & Baker, 2003; Cizek & Page, 2003; Keith, 2003; Williamson, 2013). When judging AES's capacities, it should be noted that intra- and interrater reliability among human graders – which is most important when assessing essays – is far from perfect. Biggs and Tang (2011) called the reliability of assessment "the downside of the essay" (p. 231). Human graders might not use the same criteria or might disagree about their relative importance. And even if human graders use the same criteria or receive training, their agreement is not perfect but is influenced by certain biases (e.g., halo effects, fatigue; see e.g., Engelhard, 1994; Lumley & McNamara,

1995, for summaries and discussions, see also e.g., Attali, 2013; Williamson, 2013).[2] However, AES is often assessed by analyzing whether the automatic scores agree with human graders' scores as highly as human graders agree with each other. Analyses have shown that several AES systems have passed this reliability check (e.g., Dikli, 2006; Keith, 2003; Shermis & Hamner, 2013). Thus, software-based scores are as reliable as human graders' scores and might even have some advantages over instructors' scores (see above; e.g., savings in costs and efficiency or being perceived as less biased).

However, there are also more general concerns and suspicions about AES. Page and Peterson (1995; see also Page, 2003) listed three early objections against computer grading, that is, humanist objections (i.e., the belief that only human graders are capable of understanding and judging texts), defensive objections (i.e., the fear that students might trick a computer system), and construct objections (i.e., the fear that computers will not take into account the variables that are really important). These objections are still present, both in the academic literature among writing professionals (e.g., Ericsson & Haswell, 2006) and in the broader public (see e.g., the petition against AES under http://www.humanreaders.org/petition/index.php): It is claimed that computers cannot understand a text and thus that computers cannot properly evaluate aspects that are based on such an understanding. In the same vein, Weigle (2013) summarized her concerns about using AES for summative assessments in the classroom (e.g., that computers cannot "read" essays and instead focus on the wrong skills; see also Attali, 2013, and Elliot & Klobucar, 2013, for the position statements from the Conference on College Composition and Communication in 2004 and 2009; for suspicions about computers being able to provide feedback on writing, see also Stevenson & Phakiti, 2014). However, concerns such as these might especially be dominant when the scores

[2] See also Koch (2014) for findings on teaching assistants' objectivity, reliability, validity, and on students' perceptions of assessments of quality.

are based on linguistic features or other so-called proxes (i.e., approximations or correlates of variables that truly are of interest; see Page, 1966). When content is the basis of evaluation, computer grading might find more acceptance.

These general pros and cons should be considered when selecting a specific approach for AES because different approaches might have specific pros and cons as well. There are several scoring engines that use different ways to produce their evaluations, for example, Project Essay Grade (PEG; see e.g., Page, 2003), E-rater® (see e.g., Burstein, 2003; Burstein, Tetreault, & Madnani, 2013), IntelliMetric™ (see e.g., Elliot, 2003; Schultz, 2013), or the Intelligent Essay Assessor (IEA; see e.g., Foltz, Streeter, Lochbaum, & Landauer, 2013; Landauer et al., 2003a, 2003b). Most systems use some kind of regression approach to derive a score for an ungraded essay by establishing a scoring model on the basis of prescored essays (for an overview, see Koskey & Shermis, 2013; Shermis & Daniels, 2003). Further, most of the tools focus on grammar, style, or mechanics. However, there is an exception: The IEA primarily focuses on content – and it does so by using LSA (for details on LSA, see next paragraph). Landauer et al. (2003a) showed that – although there seem to be high correlations between different writing components – content seems to be the most important aspect, and other aspects do not contribute much unique variance to the prediction of human graders' scores beyond the LSA component. Thus, this approach to AES seems to be the most useful one, especially when the intended use of the AES is to determine whether students have covered the relevant contents of a course well enough in their writings. Referring to essential features (i.e., primarily a text's content) might also be important for averting some general concerns about AES, for example, that AES might be outwitted or that it might encourage a focus on formal aspects of writing (for these and other concerns, see Shermis et al., 2013). Landauer et al. (2003a) claim that

content-based scores "will have greater face validity, be harder to counterfeit, more amenable to use in diagnosis and advice, and be more likely to encourage valuable study and thinking activities" (p. 87). Thus, instead of using superficial proxes or by applying criteria that are commonly used for essay scoring, essays might best be evaluated directly on their content because this might also make it more difficult to apply teaching-to-the-test strategies or cheating to get a good score. Another advantage of LSA is its range of comparison possibilities: When using LSA, there are several approaches that can be applied to derive scores for essays of unknown quality. That is, beyond using prescored essays (i.e., like the other scoring engines do), LSA can also base its scores on a comparison with either ideal or expert model essays, on knowledge source materials, or even on an internal comparison among the unscored essays (Landauer et al., 2003a). Thus, LSA seems to be the best approach for evaluating texts when the goal is to see whether students have covered the most relevant aspects of the topic in their writings. Thus, the next chapter will deal with how LSA works.

## 3.2 Latent Semantic Analysis (LSA)

LSA is a special approach from the field of automatic language processing. It is a computational method that aims to represent the meanings of words on the basis of their occurrence in large text corpora within n-dimensional vector spaces by using linear algebra methods. By applying mathematical similarity computations, LSA can be used to evaluate texts on their content (see Landauer, Foltz, & Laham, 1998; for details, see also Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Martin & Berry, 2007). Thus, LSA's underlying mechanism is related to Wittgenstein's idea that "the meaning of a word is its use" (Wittgenstein, 1953) and the idea that meaning is constructed from experience with language (for some thoughts

about LSA's underlying psychological theory, see e.g., Landauer, 1998, 2007; Landauer & Dumais, 1997).[3] All words in a passage contribute to the meaning of the passage, but the same meaning can be conveyed by different words.[4] However, observing co-occurrence is just one step toward extracting the meaning of words or their relations out of a collection of texts.

To derive LSA-based scores, a so-called semantic space has to be created. Quesada (2007) provided advice on how people can create their own LSA space (i.e., software-related issues such as parsing text, computing singular value decomposition, operating with vectors, and software-independent issues such as selection of the corpus, weighting, and dimension optimization). In general, the creation of a semantic space requires a large body of electronically stored literature (i.e., book chapters, articles) from the targeted knowledge domain that will make up the text corpus. The text corpus has to be split into smaller units (i.e., text fragments, also called documents). There is no strict advice on the length of the text fragments, but first and foremost, it seems important to split the original texts (i.e., the book chapters, etc.) into units of meaning (i.e., every document should include a separate unit of meaning). In continuous texts, paragraphs appear to represent this quite well, and thus, documents should reasonably include about 50 to 500 letters. Then, a raw frequency matrix that is made up of *terms* (rows) and *documents* (columns) as well as the frequency of each term in each

---

[3] This idea is in line with the fact that some words have changed their meaning, for example, "awful," which once meant "full of awe" (i.e., "inspiring wonder") but has a negative connotation today.

[4] One might think of a system of equations that help to define the single characters by analyzing relationships. For example, "wiggle and tiggle are yoggle" and "wiggle and toggle are yoggle" is similar to A+B=C and A+D=C, and it can easily be deduced that tiggle and toggle as well as B and D are the same. However, to derive a solution for all parts of the equation or all words, one would need more equations or more passages, that is, experience with language (for an example of how to infer further relations from a small amount of information via induction, see also Landauer et al., 1998).

document (cells) is computed; Table 2 illustrates a very simple and small term-by-document matrix.

Table 2

*7 x 3 Term-by-Document Matrix Based on "Filtering Essays by Means of a Software Tool: Identifying Poor Essays"*

| Terms | Documents | | | |
|---|---|---|---|---|
| | Paragraph 1 | Paragraph 2 | Paragraph 3 | Paragraph 4 |
| Essay(s) | 12 | 1 | 1 | 0 |
| Feedback | 6 | 0 | 0 | 5 |
| Learn(ing) | 5 | 9 | 0 | 3 |
| Psychology | 0 | 1 | 1 | 0 |
| Student(s) | 10 | 6 | 1 | 3 |
| University | 3 | 0 | 0 | 1 |
| Write/writing | 3 | 10 | 6 | 2 |

In reality, this term-by-document matrix is very large and includes unnecessary information. In order to eliminate this noise, to optimize data consumption, and to extract the underlying (i.e., latent) relations among the words (i.e., their meaning), additional steps are needed. First, words that do not carry specific information have to be excluded (e.g., prepositions or articles). Moreover, words that occur only once or twice in the corpus do not carry reliable information and might be simple misspellings. These should be excluded as well. Second, a weighting function (i.e., a local and global log-entropy weighting) is applied to the remaining word frequencies to emphasize the words that are specific to a context and to deemphasize words that are used frequently but are not specific to a certain context. The third and last step includes the computation of a singular value decomposition (e.g., by means of the iterative algorithm by Lanczos, 1950) and the reduction of the dimensionality. The raw frequency matrix is decomposed into orthogonal components and then reduced to a smaller number of independent dimensions. Each word is given a coordinate in the n-dimensional vector space on the

basis of its semantic content, with the vector's direction indicating the topic, and the vector's length indicating the amount of information. The vector by itself does not mean anything; rather, meaning is relational: Words that occur frequently in similar contexts are placed close to each other, whereas semantically unrelated words are represented as vectors that are at 90° angles to each other. The optimal dimensionality is an empirical issue, but dimensionalities of about 300 have proven successful in different languages (e.g., Dennis, 2007; Lenhard, Baier, Hoffmann, & Schneider, 2007; Quesada, 2007). Thus, 300 independent dimensions seem to be sufficient for capturing essential semantic content in natural language. By reducing the dimensionality, the latent relations between words are revealed. This last step is the one that distinguishes LSA from other tools: LSA does not derive its scores from simple contiguity frequencies, co-occurrence counts, or correlations in usage but uses a reduced rank vector space model and thus rather depends on a mathematical analysis of deeper relations that are hidden but make up the (gist of the) meaning of a text (i.e., where LSA's name comes from: *latent semantics*). For a detailed mathematical but also very illustrative description of the entire process, see also Martin and Berry (2007).

LSA has been shown to mimic human behavior in a variety of tasks, for example, in passing multiple-choice tests (Landauer & Dumais, 1997; Landauer et al., 1998). Further, there is a wide area of application for LSA (for an overview, see Parts III and IV of the LSA handbook by Landauer, McNamara, Dennis, & Kintsch, 2007). Originally, it had been invented as a technique for automatic indexing and retrieval to improve the detection of relevant documents (Deerwester et al., 1990). However, one of the most prominent fields of application is essay assessment (e.g., Foltz, Laham, & Landauer, 1999; Landauer, Laham, & Foltz, 2000; Miller, 2003). Meanwhile, several

tools and commercial products use LSA with a combination of other methods (for an overview, see e.g., Foltz et al., 2013).

In general, to evaluate the content of essays with LSA, the essays are represented as vectors in the semantic space by adding up the single vectors that represent the words in a text. Then, a new essay can be assessed by comparing it with previously scored essays and taking some of them (those that are most semantically similar to the target essay) into account to derive the new assessment. This approach is called *nearest neighbors*; there is also another approach that is called *gold standard*. With the gold standard approach, the scoring is based on a single text. Usually, the cosine between texts is used to derive a score for a new text. The cosine is a standard measure in the application of LSA and can be interpreted like a correlation: "0" indicates complete semantic independence and "1" maximum similarity. As an indicator of the validity of the evaluations, one usually refers to the correlation between a score given by the system and a score given by a human grader, which is (at least) as high as the correlation between the scores given by two human graders.

Although LSA cannot analyze syntax, grammar, logic, or some other facets of writing, there is a high agreement between LSA-based scores and human graders' scores, whereas other aspects do not contribute much unique variance to the prediction of human graders' scores beyond LSA[5]: Landauer et al. (2003a) reported diverse analyses that showed evidence of the validity of the IEA's and especially LSA's scores, for example, their agreement with both single and resolved rater scores for both standardized tests and classroom studies and their agreement with external criteria. Thus, although not everything is captured by LSA, it seems to be sufficient enough to

[5] A very good example of the fact that syntax is not essential for understanding is Yoda from Star Wars. You might also think of foreign language students whose grammar and syntax might not be correct but whose texts can be understood nevertheless (though with more effort; thus, the primary purpose of syntax is probably to ease understanding, but it is not necessary for our understanding).

work quite well (see also Landauer, Laham, Rehder, & Schreiner, 1997; for the relative importance of word choice instead of word order, especially in larger texts, see also Landauer, 2002, 2007).

Because previous research has focused on school children and rather narrow questions (e.g., summaries, see e.g., Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007; Wade-Stein & Kintsch, 2004), it seemed important to also test LSA's ability to evaluate more complex texts that involve a large amount of analytical reasoning and evaluation and that are authored by university students. Further, LSA-based systems depend on their basic text corpus and hence are language-dependent. LSA has predominantly been used in the English language. In Germany, there have been some first trials in which the application of LSA has been tested, and the results of these trials have been positive (Lenhard, Baier, Hoffmann, & Schneider, 2007; Lenhard, Baier, Hoffmann, Schneider, & Lenhard, 2007). Thus, it seemed important to expand the literature by testing whether and in what ways LSA might be used to assist university instructors. Thus, in a paper based on my diploma thesis, we analyzed the reliability and validity of LSA-based evaluations in a university course. Specifically, we tested whether LSA would be capable of scoring complex texts (i.e., texts that include much more than a summary of a factual text) written in German and authored by university students (Seifried et al., 2012). Results showed that, in line with previous studies on more factual texts, correlations between human graders' scores and LSA-based scores equalled interrater correlations between human graders and reached an acceptable level of agreement. This was independent of the method used by LSA, that is, using a gold standard, (i.e., comparing a new essay with one single example of a standard solution) or nearest neighbors (i.e., comparing a new essay with a sample of previously scored essays). Thus, LSA-based and human

evaluations agreed with each other to a satisfactory degree in terms of a measure of consistency, that is, Pearson product-moment correlation coefficients. Further, LSA-based evaluations of students' essays predicted students' results on a final exam (i.e., an external criterion; predictive validity).

## 4. Using LSA in University Teaching

After LSA's general ability to score complex German texts was shown, I analyzed LSA's potential to improve teaching and learning (especially in large courses) at universities by conducting the empirical studies that are included in this dissertation. Before conducting the studies, some preparatory work had to be done, that is, creating an appropriate semantic space and finding a way to convert the similarity values produced by LSA into scores that are comparable to those provided by human graders. To this end, a semantic space that had been used before was enriched, and a very simple and partly norm-referenced approach to derive scores was developed (details are described in the empirical papers). Further, students were asked to use a learning platform called *ASSIST* in which LSA was integrated and where they could submit their essays electronically. These steps were performed in close collaboration with Dr. Wolfgang Lenhard and Dr. Herbert Baier who also provided the platform *ASSIST* at the University of Würzburg. *ASSIST* has now been replaced by *βASSIST*, with great effort by Fabian Grünig, and is now located at Heidelberg University.

In the following, I will include my initial thoughts about the three empirical studies that examined different aspects of how LSA might be used to assist university instructors to solve the problems mentioned above (the papers are included at the end of this document). First, LSA might be used to detect plagiarism (Paper 1). Second, LSA might help instructors to focus their limited time and capacities on students who need special assistance (Paper 2). Further, before applying LSA for essay scoring, one should analyze the effects of LSA-based evaluations, that is, whether they are accepted by students and whether they influence students' development of learning-related characteristics (Paper 3).

### 4.1 Detecting Cheaters (Paper I: Seifried, Lenhard, & Spinath, 2015)

A great deal of research has addressed a broad range of different aspects of academic dishonesty or cheating at universities. Already 20 years ago, Franklyn-Stokes and Newstead (1995) analyzed *who* does *what* and *why* in undergraduate cheating, as well as the incidence and causes of student cheating (Newstead, Franklyn-Stokes, & Armstead, 1996). Further research analyzed the *prevalence* of and *increases* in cheating (e.g., McCabe, 2005; Vandehey, Diekhoff, & LaBeff, 2007; Whitley, 1998). Some authors referred to a new type of plagiarism in the form of *online plagiarism or cyber-cheating* arising from the opportunities offered by the Internet (e.g., Austin & Brown, 1999; Selwyn, 2008). Some recent research has further contributed to knowledge about the *characteristics of cheaters* (e.g., Giluk & Postlethwaite, 2015; Hensley, Kirkpatrick, & Burgoon, 2013; Williams, Nathanson, & Paulhus, 2010) and *perceptions* of (different kinds of) plagiarism/cheating (staff perceptions: e.g., Bennett, Behrendt, & Boothby, 2011; Flint, Clegg, & Macdonald, 2006; student perceptions: e.g., Ashworth, Bannister, & Thorne, 1997; Sutton & Taylor, 2011; for both perspectives, see e.g., Barrett & Cox, 2005; Wilkinson, 2009). Additional studies have investigated *why* students might engage in plagiarism (e.g., Bennett, 2005; Park, 2003), *how to deter* students from committing plagiarism (for a comparison of the effectiveness of some plagiarism reduction strategies, see Owens & White, 2013), and *how to detect* those who do, for example, by using software tools (for a review of tools and some tests of or comparisons between systems, see e.g., Kakkonen & Mozgovoy, 2010; Lancaster & Culwin, 2005; Maurer, Kappe, & Zaka, 2006; McKeever, 2006; Purdy, 2005; Weber-Wulff, Möller, Touras, & Zincke, 2013). Many studies have focused on *unintentional plagiarism*, which results from a lack of academic skills in citing and paraphrasing, and how to minimize this kind of plagiarism (e.g., Belter & Du Pre, 2009; Elander, Pittam,

Lusher, Fox, & Payne, 2010; Estow, Lawrence, & Adams, 2011; Landau, Druen, & Arcuri, 2002; Schuetze, 2004; Walden & Peacock, 2006). Some have also used detection software (mostly Turnitin) for educational purposes (e.g., Graham-Mathesona & Starr, 2013; for the perceptions of students and staff, see e.g., Buckley & Cowap, 2013; Dahl, 2007; Sutherland-Smith & Carr, 2005).

However, although there are many ideas about how to avoid unintentional plagiarism, there might also be *intentional* plagiarism. Giving students the freedom to write essays during the semester at home is desirable because this might encourage deeper learning outcomes (Biggs & Tang, 2011). However, this approach is feasible only if we can ensure that students will not cheat (i.e., commit plagiarism). There is much advice on how to design tasks to minimize the risk of (both unintentional and intentional) plagiarism, for example, educating students on what constitutes plagiarism/cheating, regularly changing the tasks and asking open-ended questions that require students to apply their knowledge by asking them to analyze, evaluate, or synthesize and by using open-ended questions for which many solutions are possible (see e.g., Carroll & Appleton, 2001). These recommendations are perfectly in line with the recommendations for how to best apply essay writing mentioned above and make plagiarism from other sources rather useless. However, of course, it is still possible that students will copy another student's text.[6] Owens and White (2013) found that person-to-person plagiarism was usually an act of friendship in which one student willingly gave his or her assignment to a friend who seemed to be in trouble (e.g., ill, having trouble with the language or the task; see also Ashworth et al., 1997). Although this

[6] Whether or not copying another student's text is plagiarism is not easy to say because, in general, the literature shows that "plagiarism" is not easy to define (see e.g., the categorizations by Badge & Scott, 2009; Culwin & Naylor, 1995, cited in Culwin & Lancaster, 2001; Park, 2003, 2004; Walker, 1998). Also, the borderline between collaboration and collusion seems hazy, and collusion appears to be seen as more acceptable than plagiarism (Barrett & Cox, 2005). However, in this dissertation, copying from another student is understood as a form of intentional intracorporal plagiarism.

reason might be easy to comprehend, this form of intentional intracorporal plagiarism as a form of cheating must be expected, detected, and avoided. It is important for both the instructors (e.g., to ensure that the right students are passing their courses and to reduce plagiarism) and the students (e.g., for aspects of fairness; for further reasons that address multiple aspects and perspectives, see Park, 2004) that plagiarism is not ignored. Telling and showing students that an efficient method is being applied to detect plagiarism might also contribute to its reduction (e.g., Braumoeller & Gaines, 2001). Although there has been a broad range of research on plagiarism, we found no study that directly compared the capacities of human graders and a software tool to detect plagiarism. Thus, in our first study, we tested the potential of teaching assistants and LSA to detect cheating in a psychology course.

### 4.2 Identifying Poorly Performing Students (Paper II: Seifried, Lenhard, & Spinath, 2016)

Feedback in general and formative feedback in particular are important for improving learning (see e.g., Bangert-Drowns et al., 1991; Hattie, 2009; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). In the same vein, Gibbs and Simpson (2004) made recommendations about how and why to provide feedback (e.g., frequently, in a timely and detailed way, to correct errors, to help students develop understanding, and to encourage students to continue studying). Thus, improving students' understanding and performance is at the heart of formative feedback. Of course, this is especially important for students who cannot yet grasp the material and therefore need to improve more than others. From the instructor's perspective, students should also not pass the course if they have not achieved a basic understanding. Further, students who feel

unable to cope with the tasks might be prone to cheating (e.g., Whitley, 1998). Thus, it seems to be important to identify poorly performing students.

Identifying those who are in need of special guidance is not a straightforward process. When confronted with a very large group of students or their essays, the instructor might have no other choice than to rely on chance to select some students for individual feedback. However, if essays could be quickly scored by LSA, it might be possible to identify the poorest performing students more reliably. Another possibility would be to select essays on the basis of their length because word count has been shown to offer a good way to predict scores assigned by human graders (see e.g., Page & Peterson, 1995; Burstein, Chodorow, & Leacock, 2004). This seems to make sense in that students would not want to write nonsense if they expect that a human grader will read their essay (for this argument, see Landauer et al., 2003a). The relationship between essay length and human graders' scores might be logarithmic (Shermis, Burstein, & Leacock, 2005), indicating that essay length might be especially important in the bottom sector. Because word count is easily computed, choosing essays by their length might be a good method by which to identify those students who might receive the lowest scores from human graders. Because LSA has sometimes been criticized as a "bag-of-words" technique (see e.g., Landauer, 2007), it seems important to test the potentials of both LSA and text length to identify poor texts to show that LSA goes beyond merely adding up all the words in a text. Landauer et al. (2003a) emphasized that although LSA-based measures such as vector length are often highly correlated with essay length, this is not always the case (e.g., when simply repeating some words). They also reported that vector length has sometimes explained variance independent of essay length in predicting human graders' scores. Further, as Attali (2013) stated: "Because essay length is highly predictive of human scores, correlations of machine

scores with essay length can also serve as evidence of divergence – that machine scores are not overly dependent on this easily computable feature" (p. 193). Thus, in our second study, we tested the potential of both LSA and essay length to identify poorly performing students.

### 4.3 The Effects of LSA-Based Evaluations (Paper III: Seifried, Lenhard, & Spinath, accepted pending revisions)

Research on AES has focused on psychometric issues, and whilst there are several concerns about AES (see above), not many studies have directly analyzed students' acceptance of AES. Rather, there are some studies that have provided insights into this topic as a byproduct (e.g., Lenhard, Baier, Hoffmann, & Schneider, 2007; Lipnevich & Smith, 2009a, 2009b). The results were mixed, with computerized feedback being perceived as less accurate and helpful on the one hand but also probably more fair and less biased than an instructor's feedback on the other hand.

In general, there are two theoretical perspectives on how computers might be seen, either as social actors with people responding to computers in the same way as they do to humans (see e.g., Reeves & Nass, 1996) or as neutral cognitive tools that are perceived as being free from bias and thus more trustworthy (e.g., Earley, 1988). Although they did not focus on this question but rather on the effects of differential feedback on students' examination performance, Lipnevich and Smith (2009a) provided some insights into which perspective might be true. They crossed groups that received no detailed feedback, feedback perceived as coming from the instructor, and feedback perceived as coming from a computer with receiving a grade or not and with receiving praise or not. Their results tended to support the view of computers as social actors because no difference was found in the sources of the feedback in that the feedback

given by the computer and the feedback given by the instructor caused similar effects. However, the interactions that were found between the source of feedback, grade, and praise showed that this perspective was only partially supported. If the second perspective was true, one would have expected that the feedback from the computer would be trusted more because it should be seen as more neutral and objective/unbiased. By contrast, in this study, students rated the feedback from the instructor as more accurate and helpful. However, because the detailed feedback in truth was a weighted score of the AES system E-rater and a human grader, it is not sure whether it was really the source of feedback that mattered (for this argument in general, see also Stevenson & Phakiti, 2014). Thus, in our third study, we applied a 2 x 2 experimental design with the real and assumed sources of feedback fully crossed to analyze the effects of LSA-based evaluations (i.e., acceptance of automatic assessments and development of learning-related characteristics such as motivation).

# 5. Final Discussion

Three empirical studies were conducted to investigate LSA's potential to improve learning and teaching at universities in different ways (i.e., detecting cheaters, identifying poorly performing students, giving (semi-)automatic assessment feedback). In the following, I will discuss the results in a broader context. I will start with some further ideas on the application of LSA for plagiarism detection and turn to LSA's potential to score essays afterwards. In the end, I will consider both aspects for some final and general conclusions.

## 5.1 Applying LSA to Detect Plagiarism

In our study, we had teaching assistants and LSA both detect unauthorized collaboration among students in vivo and in vitro (i.e., during the semester and within a specific sample). Thus, our study sheds light on the capacity of both human graders and LSA to detect plagiarism. We found that one particular responsible teaching assistant either could not or did not detect duplicates during the semester. That is, in two cases, it was impossible for her to detect the duplicates because she did not read all relevant texts but she also did not detect duplicates when she objectively had the chance to detect them. Further, most of the 14 teaching assistants did not notice that they had read the same essay twice in a specific sample of essays. However, with LSA, it was quite easy to identify the duplicates. Further, adding to the literature on intrarater agreement in addition to interrater agreement, not all teaching assistants scored the duplicates as equally good, whereas LSA was perfectly reliable. Thus, Paper 1 showed that LSA is helpful for detecting cheating in a university course and gave another hint about LSA's potential to score essays with a very simple scoring approach (i.e., by using the

semantic similarity to a model solution and transferring the cosine measure into a score by applying a normal rank transformation).

However, two specific conditions of our approach should be emphasized as limitations before deriving general conclusions. First, LSA was used to compare texts that were submitted in a particular course (and can be used to compare texts that are submitted in different courses) but not to check against the Internet and so forth. Thus, if students copied from files outside of our corpus, they might not have been detected. However, following the recommendations on how to implement essay writing deduced above, the Internet should not have been of much use. Further, if there was a potential database and students used it, there might soon be more than one copy in our text corpus and thus, these instances would be detected nevertheless. Thus, I conclude that this limitation is of minor importance. Second, LSA was not used alone but was combined with another algorithm, namely, the Smith-Waterman algorithm (Irving, 2004). This combination made the examination of suspicious cases much easier. If LSA is applied by itself, the texts that are most semantically similar will be identified, but it will not be easy to see the parts that share the same content. Thus, a combination of the two methods seems to be beneficial to compensate for each method's shortcomings. Again, this limitation seems to be of minor importance, but it should be remembered when judging LSA's potential to make plagiarism detection easier. There might be several reasons for the fact that the duplicates were not ranked in the first positions of the plagiarism check (see the paper for details). Most likely, another check based on verbatim overlap and in smaller units should be performed in addition.

There are many other plagiarism detection systems that might also check texts against the Internet (for a review of tools and some tests of or comparisons between systems, see the references cited above in Chapter 4.1). However, most of them

primarily detect verbatim plagiarism and this might be what students know or expect. Further, when using tasks such as those mentioned above, students cannot make use of the Internet. Thus, the most common form of plagiarism will be copying from another student and will most likely involve changing some words or paraphrasing contents (i.e., intentional intracorporal plagiarism). Structural and semantic changes such as these are said to be the most common forms of plagiarism (Britt, Wiemer-Hastings, Larson, & Perfetti, 2004), and it has been shown that LSA is able to detect these sophisticated duplicates (see e.g., Britt et al., 2004; Cosma & Joy, 2012). Another positive aspect of our general approach and the system that we are using is the fact that we need to detect plagiarism only within and across student cohorts, and thereby, we can avoid legal and ethical problems (see e.g., Purdy, 2005) by using an external server. Because our system rests on semantic similarity in first place, presumably, it will also not be easy to outwit it. Rather, it might be easier to compose one's own essay. However, the performance of an LSA-based system is not predictable but relies on several parameters (for some technical considerations when using LSA, see e.g., Rehder et al., 1998; for an analysis of parameters that drive the effectiveness of AES with LSA, see e.g., Wild, Stahl, Stermsek, & Neumann, 2005); thus, comparisons with other tools or generalizations are not easy to make (see also Cosma & Joy, 2012; Mozgovoy, Kakkonen, & Cosma, 2010).

Given the superiority of LSA in our study and the fact that human graders might not be capable of identifying plagiarism – due to both capacity constraints (e.g., considering quadratic growth in the number of single comparisons in large classes) and organizational issues (e.g., because teaching assistants often teach only parts of a course) – it can be concluded that LSA can definitely be helpful for detecting plagiarism in large university courses. However, it should be noted that human confirmation and

inspection remains necessary because LSA will give back only a rank order of the texts on the basis of their semantic similarity, and human inspection will then be necessary to make decisions about its severity. The results can also be used to educate students about plagiarism, for example, to show them what constitutes plagiarism and why students should not share their essays with friends. In recent semesters, we have also found that the number of acts of plagiarism has declined and that students who were accused of plagiarism confessed their misconduct. These findings add to the validity and importance of applying a plagiarism detection method that is based on LSA.

### 5.2 Applying LSA to Score Essays

In two studies, we analyzed LSA's potential to score essays (semi-)automatically. Paper 2 showed that LSA-based evaluations might provide a way to identify poorly performing students in large courses. With the help of LSA-based evaluations, in different samples and in all except one analysis, a larger number of poorly performing students were identified than by random sampling (i.e., which might be done instead). By contrast, text length as an indicator of a text's quality was not more helpful than random sampling would have been, and the number of essays correctly identified by LSA equaled or exceeded the number that were identified by text length. Further, regarding the detection ratio for and the credibility of the selection methods, LSA seemed to be superior to text length as well. Thus, LSA proved to be more than the mere addition of all the words in a text (i.e., more than a mere "bag-of-words" technique; see also Landauer, 2007). Considerations of costs (i.e., the number of additional essays to be examined) and benefits (i.e., the number of additional hits) led to the conclusion that combining the methods would not be more efficient than using LSA alone, although it should be noted that we were not able to test this explicitly. However,

some texts remained undetected by both methods and still remained undetected even when the two methods were combined. Further, our definition of "poor" (i.e., 25% or 12.5% of the texts) might seem arbitrary and can be attacked on the basis of its reliance on a social comparison. However, although the absolute level would not be considered with this approach, selecting the worst performing students for feedback seems worthwhile.

Paper 3 included an experiment that was conducted to investigate students' acceptance of LSA-based scores, their opinions about the use of computers in teaching in general, and their development of learning-related characteristics (i.e., motivation, achievement aspirations, and subjective learning). The real and assumed sources of assessment (i.e., scores between 0 and 10 points) were fully crossed. It was found that students' acceptance of their text's score was lower when they assumed they had been assessed by the software tool – although the real source did not matter at all. In general, acceptance was at a medium level and higher when students received a higher score. Although students preferred human graders over computers for most situations in teaching in general, this preference was weakened for two situations in which students assumed their text had been assessed by the software tool. Students also saw some general merits in computer assessments (i.e., speed and objectivity) but opted for human graders when it came to reliability and validity. Further, the (real or assumed) source of assessment did not influence the development of learning-related variables; there was only a general decline for most variables (probably due to the high starting level and the self-evaluations becoming more realistic), which was again partly influenced by the level of the score that was received.

Combining the results of these two studies, the conclusion that might be reached is that LSA-based evaluations can be helpful for instructors but that our scores are not

perfect and that the students have some acceptance problems (at least in their heads and when they are directly affected by AES). Thus, the best use of LSA might be as a tool that can work in the background to help university instructors quickly identify students who are in need of individual feedback or in order to provide a second opinion (which might be perceived as more objective by the students). Because there is no negative effect on important learning-related characteristics, combining scores with comments might be a good way to combine the capacities of instructors and computer tools in an efficient manner (with the amount of feedback that is deemed necessary depending on students' level of achievement).[7]

The conclusion that LSA-based scores should be used only in the background can also be reached when looking at the quality of the LSA-based scores in general. In both studies, the scores were not perfectly reliable or valid and occasionally led to false conclusions (e.g., not identifying all poor texts as such or producing scores that differed from the teaching assistants' scores to an unacceptable extent). When using LSA, hundreds of texts can be scored in milliseconds, and students might receive more frequent and very quick or even immediate feedback on the quality of their work. However, to give feedback to students or for high-stakes assessment, the scores must be reliable and valid. In this regard, there are some aspects that need to be thought of when using LSA in general: It is essential to "teach" LSA all relevant words so that it can "understand" their meaning because LSA's "knowledge" and thus, the scoring results, always depend on the semantic space that is used (i.e., the magnitude and representativeness of the literature). Further, misspellings in students' essays should be reduced because LSA might again not "know" a word when it is not spelled properly, and this might result in incorrect scores.

[7] See also Engelhardt (2011) for differential effects of different feedback types that depend on students' prior performance.

Further, the reliability and validity of the LSA-based scores might be influenced by aspects that are germane to the specific approach that we used in our studies: First, we used the comparison with one text only to derive the LSA-based evaluations. The quality of this comparison text is very important: If its quality is low, the LSA-based scores might not be valid. There are other techniques that might be used to avoid the strong dependence on the quality of a single text, for example, using previously scored essays as a basis for assessment. However, our previous study showed that this nearest neighbors approach was not superior to the gold standard approach (Seifried et al., 2012). Further, the nearest neighbors approach requires previously scored essays and thus, this method needs much more preliminary work than comparing new essays with a single model solution. This also holds true for even more complex analysis strategies and machine learning algorithms such as the *Neural Networks* or *Support Vector Machines*, where hundreds of prescored essays are necessary as a training base. For these reasons, we used the comparison with one text only as the assessment method for all analyses that were reported in this dissertation. However, if LSA is to be applied in university contexts with this approach, the suitability of the comparison text should be well established. Second, correlation analyses show that there is a high level of agreement between LSA-based scores and human graders' scores; it is as high as the agreement between several human graders. However, there is no absolute level when considering correlations. To translate the similarity scores that LSA produces into the raw point scoring system used by human graders, we applied a very simple approach: The essays were ranked according to their similarity to the model solution, and then, the rank was transferred by applying a normal rank transformation by computing the accordant z-score by means of the inverse normal cumulative distribution. Then, we manually assessed two essays (i.e., the essays at the 10th and 90th percentiles to avoid

giving outliers a heavy weight) and adjusted the scores of the remaining essays via linear regression. This procedure comes with several pitfalls. First, because the assessment of an essay depends on the other essays that are in the to-be-scored sample, the assessment is not 100% reliable. Second, if the two anchor assessments are not valid, the overall scoring will not be either. Third, deriving the scores for all but the anchor texts in the way described above relies on the assumption that the quality of the texts is normally distributed. If this assumption is not true, again, the scoring will not be valid. It might seem rational to assume a normal distribution, and the empirical studies mentioned above seem to indicate that this approach can be valid. However, re-analyses with our samples indicated that data do not always follow a normal distribution (especially data from samples of psychology students seem to be rather skewed to the right; i.e., many texts are of good quality). However, if LSA is used only in the background to help instructors detect plagiarism and identify poorly performing students, knowing the underlying distribution and having exact scores are not necessary, but the similarity scores and the ranking on the basis of these scores might be sufficient.

In sum, it seems important to include some further checks before using LSA-based scores for high-stakes assessments or before reporting the scores to students in the form of a formative assessment; for example, at least a subset of essays should be evaluated by human graders to confirm the automatic evaluations and to ensure that the distribution of scores is similar for both human graders and LSA (see also Williamson, 2013). The potential consequences of inaccurate scores should be thought of before applying AES: If the scores are used only in the background, inaccurate scores seem to be less important than when the scores are used as feedback for students. Nevertheless, other backup methods are necessary to ensure that unusual essays are detected (e.g., essays that are off topic, essays that consist of the repetition of only a few sentences,

plagiarism) and that the scores are reliable even if an essay is highly unusual, for example, because it is very creative (see e.g., functioning of the IEA; Landauer et al., 2003a, 2003b).

We have not yet asked students to try to outwit our system. Rather, I have only done so myself in a preliminary analysis in which two special texts were given the lowest scores by LSA (i.e., song lyrics with odd content in the context of a psychology course and another text with technical terms that were relevant for the topic in question but used in the wrong context; Seifried, 2010). Of course, dealing with special texts might be an interesting issue for future research because, for example, an essay might be good but not comparable to the comparison text (or even previously scored essays). Thus, some "safeguards" should be implemented to select specific texts for human inspection to avoid giving incorrect feedback (again, see e.g., functioning of the IEA; Landauer et al., 2003a, 2003b).

Further, Attali (2013) makes an important point about the use of AES in general when he states:

> In making decisions about whether and how to incorporate AES in the evaluation of essays, an assessment program has to take into account the range of evidence for the validity of machine scores. It has to weigh the possible benefits in cost savings and reliability against the possible risks of shifting the measured construct, changes in subgroup differences, and susceptibility to large errors in scoring. (p. 193).

Attali suggested two models of implementation (i.e., a contributory model or a check score/confirmatory score model) and pointed out the necessity of the evaluation of the intended and unintended consequences of AES use. An example of the latter would be when students change their writing strategies to receive better scores from the AES

system. However, Attali referred to a study by Powers (2011), which indicated that students think the best strategy would be to improve spelling, grammar, form, or structure, use more transition words and diverse vocabulary – but not to change their essays' length, use of complex sentences, long words, or to focus less on content or logic. This is promising news for the application of AES, but of course, the (un)intended consequences should regularly be inspected when applying AES.

There are some further general aspects concerning AES, and LSA in particular, that also affected the studies mentioned above to some extent. The first concerns AES's possible range of application. When it comes to analyzing the quality of arguments, the capacity of AES is limited (see e.g., Attali, 2013). Thus, it might be an interesting topic for future research to use sentiment analysis for essays that include opinions (see Burstein, Beigman-Klebanov, Madnani, & Faulkner, 2013; for extensions to LSA that might also be interesting, see Part V of the LSA handbook by Landauer et al., 2007). Second, although the cosine is the standard measure in the literature on LSA and was therefore used in the empirical studies in this dissertation, researchers might also wish to further analyze whether other measures produce better scores (e.g., the Euclidian distance between texts). Third, there are several metrics that might be used to measure agreement between AES or LSA's scores and human graders' scores beyond correlations, for example, kappa, weighted kappa, or exact and adjacent agreement. Whereas some authors speak against correlations (e.g., Cizek & Page, 2003), others list arguments against the alternative metrics (e.g., the dependence of the coarseness of the scoring system when using exact or adjacent agreement; Keith, 2003; see also, e.g., Shermis & Daniels, 2003) and argue in support of correlation coefficients (e.g., that machine scores do not have to be rounded but can be kept continuous; Attali, 2013). Correlation coefficients are the common measure in the literature and were therefore

used in this dissertation as well. The fourth aspect is the reliance on human graders' scores as the gold standard in general. For the three studies mentioned above, we used human graders' scores to assess LSA's validity. However, when using humans as a gold standard, we might transfer human biases to AES (see Chung & Baker, 2003). Thus, although agreement with human graders has been the most important evidence of the reliability and validity of AES – and LSA passes this test very well – human agreement is not sufficient, and agreement with external criteria should be considered as well (as is done for the IEA; Foltz et al., 2013). Hence, as was done in the preliminary study (Seifried et al., 2012), the validity of the LSA-based scores might have been tested against other external criteria in the studies as well (see also Keith, 2003). However, a strength of the studies was that LSA-based scores could be based on or compared with the average of the scores given by several (i.e., three to 14) trained human graders because this procedure should increase the ability to measure an essay's "true score" and thus allows LSA to be tested against a valid measure.

## 5.3 General Conclusion

I have argued that university instructors should practice EBT and should encourage students to use learning techniques that have been shown to be effective. On the basis of these considerations, I have discussed why essay writing is important in university courses and how it should be applied. With students writing essays continuously throughout the semester, instructors can both help students to apply effective learning techniques and assess their performance in the form of a formative evaluation. However, because letting students write essays might mean an unmanageable amount of effort for university instructors, I have reflected on the pros and cons of AES in general

and introduced LSA as one AES approach. Three empirical studies were reported and discussed to reflect on LSA's potential to assist university instructors in several ways.

Recently, there has been a shift away from AES (i.e., Automatic Essay *Scoring*) to AEE (i.e., Automatic Essay *Evaluation*; see Whithaus, 2013). AES has been included in richer technologies that are able to provide feedback beyond a mere score and interact with learners (for an overview, see e.g., Shermis et al., 2013; for an evaluation, see Shermis & Hamner, 2013). Using AES for educational or formative purposes (e.g., allowing more than one attempt and improvement based on immediate feedback, probably including hints about further readings; for an example of how LSA can be used to match readers and texts, see Wolfe et al., 1998) rather than for judging or summative purposes (e.g., deciding whether or not a student should pass a course) might contribute to its acceptance – and this acceptance might be necessary for online and distance education, which are becoming more popular (e.g., Massive Open Online Courses; MOOCs).

It is most likely not necessary to explain to students how the system works to increase its acceptance. Revealing the scoring mechanism might raise concerns about the fact that the system might not use the same process as human graders do and might facilitate attempts to outwit the system. Although it might be an interesting topic for future research to identify the attributes of a text that result in it receiving a higher score by LSA than actually deserved and to invite students to incorporate these aspects into their writing in order to consider such attributes for an advanced scoring mechanism, we might also leave students in the dark about the scoring technique. We have accepted some other things in our daily lives even though we do not really know *how* they work as long as we are convinced *that* they do. In their book about AEE, Shermis and Burstein (2013) used the metaphor of a microwave: In the beginning, people might have

been skeptical about this black box and probably did not prefer it over an oven. Microwaves might also not be able to do everything that an oven can (and not in the same way), but today, microwaves are highly accepted. Another analogy might be to compare LSA with a prosthesis: It cannot do anything the same way as the original (i.e., a human grader or a real part of a body, respectively) but it resembles the original quite well. Thus, first and foremost, we should probably convince students and the broader public that AES systems can do the things we want them to do – in whatever way – to raise their acceptance.

In the end, we need to know the intended use of the results to judge their validity:

> High reliability or agreement between automated and human scoring is a necessary, but insufficient condition for validity. Evidence needs to be gathered to demonstrate that the scores produced by automated systems faithfully reflect the intended use of those scores. For example, automated essay scoring for the purpose of improving instruction should yield information that is usable by teachers about students who need improvement. (Chung & Baker, 2003, p. 29)

We intended to improve teaching and learning at universities by making it possible to achieve desirable teaching-learning formats, that is, asking students to write essays and giving them feedback. In particular, we intended to use LSA to detect plagiarism and identify poorly performing students. Reflecting on the studies mentioned above, it can be concluded that LSA-based scores are useful for these purposes: Whereas teaching assistants were not able to identify cheating reliably both during the semester and within a specific sample, the duplicates could easily be found with the help of a semantic comparison based on LSA. Further, those students who would have received the lowest scores out of a larger group of students could better be identified by using LSA-based

scores than by relying on essay length (i.e., a feature that might also be faked easily). Moreover, there were no negative effects on the development of students' learning-related characteristics when applying LSA.

For my studies, I collaborated with different groups (i.e., psychologists, computer scientists, and students) to cover different perspectives. For example, βASSIST was created by a student to adapt the former system to the specific needs of our teaching-learning format and our students. Thus, we developed a learning platform that helps us collect and compare students' essays and that makes it easy for our teaching assistants to give feedback to the students. I have used various statistical methods in my studies (including correlations, chi-square tests, and a specially built test statistic) and included the studies in a real university course, which adds to the ecological validity and generalizability of the findings. On the basis of my studies and these considerations, I conclude that the application of AES can be helpful for three groups that are active in higher education, namely, the students, the instructors, and the teaching assistants. However, as stated above, there are also several caveats about our special technique, and further studies and safeguards are necessary to ensure the reliability and validity of the LSA-based scores for the intended purposes.

Thus, I agree with those authors who state that AES might best be used as a complement to rather than as a replacement for human graders (e.g., Attali, 2013): AES might be used to detect plagiarism, to detect poorly performing students, to flag unusual essays, and to obtain a coarse overview of the overall quality of a set of essays. Beyond these possibilities, AES might also be used to monitor human raters (see Bridgeman, 2013).

On the basis of the three studies mentioned above, it can be concluded that it is too early to use our LSA-based scores for high-stakes assessments. The empirical

studies included a basic application of LSA, which was quite successful. However, although LSA-based evaluations might be helpful for identifying the poorest essays out of a larger group or to detect plagiarism, the scores are not precise enough to be used for direct feedback. LSA-based evaluations can be useful for assisting but not for replacing humans. In this regard, however, they can help make possible teaching-learning arrangements that are highly desirable. Referring to the recommendation on how to best apply essay writing that I have deduced above, I want to close by saying:

> *By using AES in the background*, during the semester, students *can* write essays that answer challenging questions, include key concepts, and call for applications of the material that was taught, and students *can* receive timely feedback on their ideas.

## References

Altbach, P.G., Reisberg, L., & Rumbley, L.E. (2009). *Trends in global higher education: Tracking an academic revolution*. Report for the UNESCO 2009 World Conference on Higher Education.

Ashworth, P., Bannister, P., & Thorne, P. (1997). Guilty in whose eyes? University students' perceptions of cheating and plagiarism in academic work and assessment. *Studies in Higher Education, 22*, 187–203.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 181–198). New York, NY: Routledge.

Austin M., & Brown L. (1999). Internet plagiarism: Developing strategies to curb student academic dishonesty. *The Internet and Higher Education, 2*, 21–33.

Badge, J., & Scott, J. (2009). Dealing with plagiarism in the digital age, [online] Available at: http://evidencenet.pbworks.com/Dealing-with-plagiarism-in-the-digital-age

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.

Barrett, R., & Cox, A. L. (2005). "At least they're learning something": The hazy line between collaboration and collusion. *Assessment & Evaluation in Higher Education, 30*, 107–122.

Belter, R. W., & Du Pre, A. (2009). A strategy to reduce plagiarism in an undergraduate course. *Teaching of Psychology, 36*, 257–261.

Benassi, V. A., Overson, C. E., & Hakala, C. M. (Eds.) (2014). *Applying science of learning in education: Infusing psychological science into the curriculum.* Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/asle2014/index.php

Bennett, R. (2005). Factors associated with student plagiarism in a post-1992 university. *Assessment & Evaluation in Higher Education, 30*, 137–162.

Bennett, K. K., Behrendt, L. S., & Boothby, J. L. (2011). Instructor perceptions of plagiarism: Are we finding common ground? *Teaching of Psychology, 38*, 29–35.

Bernstein, D. J., Addison, W., Altman, C., Hollister, D., Komarraju, M., Prieto, L., Rocheleau, C.A., & Shore, C. (2010). Toward a scientist-educator model of teaching psychology. In Halpern, D. F. (Ed.). *Undergraduate education in psychology. A blueprint for the future of the discipline* (pp. 29–45). Washington, DC: American Psychological Association.

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: the SOLO taxonomy.* New York: Academic Press.

Biggs, J. B., & Tang, C. S. (2011). *Teaching for quality learning at university: What the student does*. Maidenhead: Open University Press.

Blümel, J. (2013). *Besteht ein Zusammenhang zwischen dem Verfassen schriftlicher Stellungnahmen als Prüfungsformat und nachhaltigem Lernformat?* (Unpublished bachelor thesis). Heidelberg University, Heidelberg.

Boyer, E. L. (1990). *Scholarship reconsidered. Priorities of the professoriate*. Princeton, NJ: The Carnegie Foundation for the Advancement of Teaching.

Braumoeller, B. F., & Gaines, B. J. (2001). Actions do speak louder than words: Deterring plagiarism with the use of plagiarism-detection software. *PS Online, 34*, 835–839.

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 221–232). New York, NY: Routledge.

Britt, M. A., Wiemer-Hastings, P., Larson, A. A., & Perfetti, C. A. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education, 14*, 359–74.

Buckley, E., & Cowap, L. (2013). An evaluation of the use of Turnitin for electronic submission and marking and as a formative feedback tool from an educator's perspective. *British Journal of Educational Technology, 44*, 562–570.

Burstein, J. (2003). The E-rater® scoring engine: Automated essay scoring with natural language processing. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, N.J.; London: Erlbaum Associates.

Burstein, J., Beigman-Klebanov, B., Madnani, N., & Faulkner, A. (2013). Automated sentiment analysis for essay evaluation. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 281–297). New York, NY: Routledge.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine, 25(3)*, 27–36.

Burstein, J., Tetreault, J, & Madnani, N. (2013). The E-rater® automated essay scoring system. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 55–67). New York, NY: Routledge.

Carroll, J., & Appleton, J. (2001). Plagiarism: A good practice guide. Oxford Brookes University and Joint Information Systems Committee (JISC).

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23–40). Mahwah, N.J.; London: Erlbaum Associates.

Cizek, G. J., & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 125–145). Mahwah, N.J.; London: Erlbaum Associates.

Cosma, G., & Joy, M. (2012). An approach to source-code plagiarism detection and investigation using Latent Semantic Analysis. *IEEE Transactions on Computers, 61*, 379–394.

Cranney, J. (2013). Toward psychological literacy: A snapshot of evidence-based learning and teaching. *Australian Journal of Psychology, 65,* 1–4.

Culwin F., & Lancaster T (2001). Plagiarism, prevention, deterrence and detection. Higher Education Academy.

Dahl, S. (2007). Turnitin®: The student perspective on using plagiarism detection software. *Active Learning in Higher Education, 8*, 173–191.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Dennis, S. (2007). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57–70). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*. Retrieved 08.12.2009 from http://www.jtla.org.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58.

Dunn, D. S., Saville, B. K., Baker, S. C., & Marek, P. (2013). Evidence-based teaching: Tools and techniques that promote learning in the psychology classroom. *Australian Journal of Psychology, 65,* 5–13.

Earley, P. C. (1988). Computer-generated performance feedback in the magazine-subscription industry. *Organizational Behavior and Human Decision Processes, 41*, 50–64.

Eckert, C., Seifried, E. & Spinath, B. (2014). Ist das Verfassen semesterbegleitender Texte für Studierende eine gute Vorbereitung für Klausuren mit offenen und geschlossenen Aufgaben? In E. Seifried, C. Eckert, B. Spinath & K.-P. Wild (Chairs), Verbesserung von Hochschullehre: Beiträge der pädagogisch-psychologischen Forschung. Symposium auf dem 49. Kongress der Deutschen Gesellschaft für Psychologie, Bochum, September 2014.

Eckert, C., Seifried, E. & Spinath, B. (2015). Heterogenität in der Hochschullehre aus psychologischer Sicht: Die Rolle der studentischen Eingangsvoraussetzungen für adaptives Lehren. In K. Rheinländer (Hrsg.), *Ungleichheitssensible Hochschullehre. Positionen, Voraussetzungen, Perspektiven* (S. 257–264). Heidelberg: Springer.

Elander, J., Pittam, G., Lusher, J., Fox, P., & Payne, N. (2010). Evaluation of an intervention to help students avoid unintentional plagiarism by improving their authorial identity. *Assessment and Evaluation in Higher Education, 35*, 157–171.

Elliot, S. (2003). Intellimetric$^{TM}$: From here to validity. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, N.J.; London: Erlbaum Associates.

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 16–35). New York, NY: Routledge.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93–112.

Engelhardt, L. (2011). *Entwicklung von Lernkurven unter verschiedenen Rückmeldungsarten* (Unpublished diploma thesis). Heidelberg University, Heidelberg.

Ericsson, P. F., & Haswell, R. H. (Eds.) (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Estow, S., Lawrence, E. K., & Adams, K. A. (2011). Practice makes perfect: Improving students' skills in understanding and avoiding plagiarism with a themed methods course. *Teaching of Psychology, 38*, 255–258.

Fisher, A., Exley, K., & Ciobanu, D. (2014). Using technology to support learning and teaching. New York, NY: Routledge.

Flint, A., Clegg, S., & Macdonald, R. (2006). Exploring staff perceptions of student plagiarism. *Journal of Further and Higher Education, 30*, 145–156.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*. Retrieved from http://imej.wfu.edu/articles/1999/2/04/

Foltz, P., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68–88). New York, NY: Routledge.

Franklyn-Stokes, A., & Newstead, S. E. (1995). Undergraduate cheating: Who does what and why? *Studies in Higher Education, 20*, 159–172.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street®: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33*, 53–80.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1*, 3–31.

Giluk, T. L., & Postlethwaite, B. E. (2015). Big Five personality and academic dishonesty: A meta-analytic review. *Personality and Individual Differences, 72*, 59–67.

Graesser, A. C., Halpern, D. F., & Hakel, M. (2008). *25 principles of learning*. Washington, DC: Task Force on Lifelong Learning at Work and at Home.

Graham-Mathesona, L., & Starr, S. (2013). Is it cheating or learning the craft of writing? Using Turnitin to help students avoid plagiarism. *Research in Learning Technology, 21*(1).

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hattie, J. (2011). Which strategies best enhance teaching and learning in Higher Education? In D. Mashek & E.Y. Hammer (Eds.), *Empirical research in teaching and learning: Contributions from Social Psychology* (pp.130–142). London: Blackwell Publishing.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning.* London: Routledge.

Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology, 1,* 79–91.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hensley, L. C., Kirkpatrick, K. M., & Burgoon, J. M. (2013). Relation of gender, course enrollment, and grades to distinct forms of academic dishonesty. *Teaching in Higher Education, 18*, 895–907.

Huber, L. (2011). Forschen über (eigenes) Lehren und studentisches Lernen – Scholarship of Teaching and Learning (SoTL). Ein Thema auch hierzulande? In Das Hochschulwesen, 59. Jg., S. 118–124.

Huber, L., Pilniok, A., Sethe, R., Szczyrba, B., & Vogel, M. (Eds.). (2014). *Forschendes Lehren im eigenen Fach. Scholarship of Teaching and Learning in Beispielen*. Bielefeld: W. Bertelsmann Verlag.

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered. Institutional integration and impact*. San Francisco, CA: Jossey-Bass.

Irving, R. (2004). Plagiarism and collusion detection using the Smith-Waterman Algorithm. DCS Technical Report. Department of Computing Science, University of Glasgow. Retrieved from http://www.dcs.gla.ac.uk/publications/PAPERS/7444/TR-2004-164.pdf

Kakkonen, T., & Mozgovoy, M. (2010). Hermetic and web plagiarism detection systems for student essays - An evaluation of the state-of-the-art. *Journal of Educational Computing Research*, 42(2):135–139.

Keith, T. Z. (2003). Validity and automated essay scoring systems. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–167). Mahwah, N.J.; London: Erlbaum Associates.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street®: Computer-guided summary writing. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 263–277). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Koch, F. D. (2014). *Zur Güte tutorieller Textbewertungen* (Unpublished masters thesis). Heidelberg University, Heidelberg.

Koskey, K. L. K., & Shermis, M. D. (2013). Scaling and norming for automated essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 199–220). New York, NY: Routledge.

Lancaster, T., & Culwin, F. (2005). Classifications of plagiarism detection engines. *Innovation in Teaching and Learning in Information and Computer Sciences* 4(2). Available at http://journals.heacademy.ac.uk/doi/pdf/10.11120/ital.2005.04020006

Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards, 45,* 255–282.

Landau, J. D., Druen, P. B., & Arcuri, J. A. (2002). Methods for helping students avoid plagiarism. *Teaching of Psychology, 29*, 112–115.

Landauer, T. K. (1998). Learning and representing verbal meaning: The Latent Semantic Analysis theory. *Current Directions in Psychological Science, 7*, 161–164.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In B. H. Ross (Ed.), The psychology of learning and motivation (pp. 43–84). New York: Academic Press.

Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–34). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems, 15*, 27–31.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003a). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, N.J.; London: Erlbaum Associates.

Landauer, T. K., Laham, D. & Foltz, P. W. (2003b). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10,* 295–308.

Landauer, T. K., Laham, D., Rehder, B. & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In Shafto, M. G. & Langley, P. (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (412-417). Mahwah, NJ US: Lawrence Erlbaum Associates.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Lange, S. (2014). *Nachhaltigkeit von Lernerfolg – Ein Vergleich von zwei Lehrmethoden* (Unpublished masters thesis). Heidelberg University, Heidelberg.

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse

[Automatic scoring of constructed-response items with latent semantic analysis]. *Diagnostica, 53*, 155–165.

Lenhard, W., Baier, H., Hoffmann, J., Schneider, W., & Lenhard, A. (2007). Training of summarisation skills via the use of content-based feedback. In F. Wild, M. Kalz, J. Van Bruggen, & R. Koper (Eds.), *Proceedings of the First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning* (pp. 26–27). Heerlen, the Netherlands: Open University of the Netherlands.

Lipnevich, A. A., & Smith, J. K. (2009a). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 15*, 319–333.

Lipnevich, A. A., & Smith, J. K. (2009b). "I really need feedback to learn:" students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability, 21*, 347–367.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*, 1337–1344.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54–71.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science, 12*, 1050–1083.

McCabe, D. L. (2005). Cheating among college and university students: A North American perspective. *Journal of Educational Integrity, 1*. Retrieved 5/22/2014 from http://www.ojs.unisa.edu.au/index.php/IJEI/article/view/14.

McGovern, T. V., & Hogshead, D. L. (1990). Learning about writing, thinking about teaching. *Teaching of Psychology, 17*, 5–10.

McKeever, L. (2006). Online plagiarism detection services - saviour or scourge? *Assessment & Evaluation in Higher Education, 31,* 155–165.

Miller, T. (2003). Essay assessment with Latent Semantic Analysis. *Journal of Educational Computing Research, 29,* 495–512.

Mozgovoy, M., Kakkonen, T., & Cosma, G. (2010). Automatic student plagiarism detection: Future perspectives. *Journal of Educational Computing Research, 43*, 507–527.

Newstead, S. E., Franklyn-Stokes, A., & Armstead, P. (1996). Individual differences in student cheating. *Journal of Educational Psychology, 88*, 229–241.

Nguyen, K., & McDaniel, M. A. (2015). Using quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology, 42*, 87–92.

OECD (2015). *Education at a Glance 2015: OECD Indicators.* OECD Publishing.

Owens, C., & White, F. A. (2013). A 5-year systematic strategy to reduce plagiarism among first-year psychology university students. *Australian Journal of Psychology, 65*, 14–21.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, N.J.; London: Erlbaum Associates.

Page, E. B, & Petersen, N. S. (1995). The computer moves into essay grading: updating the ancient test. *Phi Delta Kappan, 48*, 238–243.

Park, C. (2003). In other (people's) words: Plagiarism by university students – literature and lessons. *Assessment & Evaluation in Higher Education, 28*, 471–488.

Park, C. (2004). Rebels without a clause: Towards an institutional framework for dealing with plagiarism by students. *Journal of Further and Higher Education, 28*, 291–306.

Pashler, H., Bain, P. T., Bottge, B., Koedinger, K., McDaniel, M., & Metcalf, J. (2007). *Organizing instruction and study to improve student learning*. Washington, DC: National Center for Education Research, Institute of Education Science, U.S., Department of Education.

Pearsell, N. R., Skipper, J. E., & Mintzes, J. J. (1997). Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. *Science Education, 81*, 193–215.

Purdy, J. P. (2005). Calling off the hounds: Technology and the visibility of plagiarism. *Pedagogy, 5*, 275–296.

Quesada, J. (2007). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 71–85). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25,* 337–354.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers. television, and new media like real people and places.* New York, NY Cambridge University Press.

Schuetze, P. (2004). Evaluation of a brief homework assignment designed to reduce citation problems. *Teaching of Psychology, 31*, 257–259.

Schultz, M. T. (2013). The IntelliMetric™ automated essay scoring engine – A review and an application to Chinese essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 89–98). New York, NY: Routledge.

Schwartz, B. M., & Gurung, R .A. R. (Eds.) (2012). *Evidence-based teaching for higher education*. Washington, DC: American Psychological Association.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*, 453–472.

Seifried, E. (2010). *Evaluation der Latenten Semantischen Analyse (LSA) bezüglich der Bewertung komplexer Texte* (Unpublished diploma thesis). Heidelberg University, Heidelberg.

Seifried, E., Eckert, C., & Spinath, B. (2014a). Eingangs- und Verlaufsdiagnostik von Lernvoraussetzungen und Lernergebnissen in der Hochschullehre. In M. Krämer, U. Weger & M. Zupanic (Hrsg.), *Psychologiedidaktik und Evaluation X* (S. 267-274). Aachen: Shaker.

Seifried, E., Eckert, C., & Spinath, B. (2014b). Is answering open questions throughout the semester a good means to prepare for an exam with forced-choice items and open questions? Paper presented at the 6th International Conference on Psychology Education (ICOPE6), Flagstaff, Arizona, USA, August 2014.

Seifried, E., Lenhard, W., Baier, H., & Spinath, B. (2012). On the reliability and validity of human and LSA-based evaluations of complex student-authored texts. *Journal of Educational Computing Research, 47*, 67–92.

Selwyn, N. (2008). 'Not necessarily a bad thing …': a study of online plagiarism amongst undergraduate students. *Assessment & Evaluation in Higher Education, 33*, 465–479.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, N.J.; London: Erlbaum Associates.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. New York, NY: Routledge.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 1–15). New York, NY: Routledge.

Shermis, M. D., Burstein, J., & Leacock, C. (2005). Applications of computers in assessment and analysis of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 403–416). New York, NY: Guilford Publications.

Shermis, M. D., & Daniels, K. E. (2003). Norming and scaling for automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 169–180). Mahwah, N.J.; London: Erlbaum Associates.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 313–346). New York, NY: Routledge.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189.

Spinath, B. (2011). Initiating sustainable learning in future teachers by means of asking complex questions. Presentation in an invited symposium at the Psychology

Learning and Teaching Conference (PLAT2011), 12. European Congress of Psychology (ECP2011), Istanbul, July 2011.

Spinath, B., & Seifried, E. (2012). Forschendes Lehren: Kontinuierliche Verbesserung einer Vorlesung. In M. Krämer, S. Dutke, & J. Barenberg (Hrsg.), *Psychologiedidaktik und Evaluation IX* (pp. 171–180). Aachen: Shaker.

Spinath, B., Seifried, E., & Eckert, C. (2014). Forschendes Lehren: Ein Ansatz zur kontinuierlichen Verbesserung von Hochschullehre. *Journal Hochschuldidaktik, 25(1-2)*, 14-16.

Spinath, B., Seifried, E., & Eckert, C. (in press). Forschendes Lehren: Ein Ansatz zur kontinuierlichen Verbesserung von Hochschullehre. In M. Heiner, B. Baumert, S. Dany, T. Haertel, M. Quellmelz, & C. Terkowsky (Hrsg.), *Was ist gute Lehre – und was kann die Hochschuldidaktik dazu beitragen?*

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.

Sutherland-Smith, W., & Carr, R. (2005). Turnitin.com: Teachers perspectives of anti-plagiarism software in raising issues of educational integrity. *Journal of University Teaching & Learning Practice, 2*(3).

Sutton, A., & Taylor, D. (2011). Confusion about collusion: Working together and academic integrity. *Assessment & Evaluation in Higher Education, 36*, 831–841.

Vandehey, M. A., Diekhoff, G. M., & LaBeff, Emily E. (2007). College cheating: A twenty-year follow-up and the addition of an honor code. *Journal of College Student Development, 48*, 468–480.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333–362.

Walden, K., & Peacock, A. (2006). The i-Map: A process-centered response to plagiarism. *Assessment & Evaluation in Higher Education, 31*, 201–214.

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

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

Weber-Wulff, D., Möller, C., Touras, J., & Zincke, E. (2013). Plagiarism detection software test 2013. Retrieved from http://plagiat.htw-berlin.de/software-en/test2013/report-2013/

Weigle, S. C. (2013). English as second language writing and automated essay evaluation. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 36–54). New York, NY: Routledge.

Whithaus, C. (2013). Foreword. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. vii-ix). New York, NY: Routledge.

Whitley, B. E. Jr., (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*, 235–274.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81.

Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). *Parameters driving effectiveness of automated essay scoring with LSA*. Proceedings of the 9th International Computer Assisted Assessment Conference, 485–494.

Wilkinson, J. (2009). Staff and student perceptions of plagiarism and cheating. *International Journal of Teaching and Learning in Higher Education, 20*, 98–105.

Williams, K. M., Nathanson, C., & Paulhus, D. L. (2010). Identifying and profiling scholastic cheaters: Their personality, cognitive ability, and motivation. *Journal of Experimental Psychology: Applied, 16*, 293–307.

Williamson, D. M. (2013). Probable cause: Developing warrants for automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 153–180). New York, NY: Routledge.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford, England: Macmillan.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes, 25,* 309–336.

## List of Tables

# List of Abbreviations

| Abbreviation | Long Version |
|---|---|
| AES | Automatic Essay Assessment |
| EBT | Evidence-Based Teaching |
| IEA | Intelligent Essay Assessor |
| LSA | Latent Semantic Analysis |

**Declaration in accordance to § 8 (1) c) and d) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies**

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

**FAKULTÄT FÜR VERHALTENS-
UND EMPIRISCHE
KULTURWISSENSCHAFTEN**

**Promotionsausschuss der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg**
**Doctoral Committee of the Faculty of Behavioural and Cultural Studies, of Heidelberg University**

**Erklärung gemäß § 8 (1) c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften**
**Declaration in accordance to § 8 (1) c) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies**

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe.
I declare that I have made the submitted dissertation independently, using only the specified tools and have correctly marked all quotations.

**Erklärung gemäß § 8 (1) d) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften**
**Declaration in accordance to § 8 (1) d) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies**

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe.
I declare that I did not use the submitted dissertation in this or any other form as an examination paper until now and that I did not submit it in another faculty.

Vorname Nachname
First name Family name _____

Datum, Unterschrift
Date, Signature _____

## Publications of the Publication-Based Dissertation

**Paper I**

This is the accepted manuscript (Version 2) of the article

Seifried, Eva, Lenhard, Wolfgang & Spinath, Birgit (2015). Plagiarism detection: A comparison of teaching assistants and a software tool in identifying cheating in a psychology course. *Psychology Learning and Teaching, 14*, 236-249, SAGE Publications. doi: 10.1177/1475725715617114

# Plagiarism detection: A comparison of teaching assistants and a software tool in identifying cheating in a psychology course

Eva Seifried

Heidelberg University

Wolfgang Lenhard

University of Wuerzburg

Birgit Spinath

Heidelberg University

## Author Note

This research was supported by the Innovation Fund FRONTIER at Heidelberg University (project number D.801000/10.25). We want to thank Dr. Herbert Baier for his work on ASSIST (i.e., a former version of βASSIST at the University of Wuerzburg) and Fabian Grünig for his work on βASSIST, Jane Zagorski for proofreading the manuscript and Steve Newstead, former editor of PLAT, for handling this paper.

Correspondence concerning this article should be addressed to Eva Seifried, Department of Psychology, Heidelberg University, Hauptstrasse 47-51, D-69117, Heidelberg, Germany, Phone: ++49 (0) 6221 / 547728, Fax: ++49 (0) 6221 / 547326, E-Mail: Eva.Seifried@psychologie.uni-heidelberg.de

**Notes on Contributors**

**Eva Seifried** received her diploma in psychology at Heidelberg University in 2010. She now is a doctoral student and research associate at the Department of Psychology at the same university. Her research interests are learning and teaching in higher education, especially psychology and teacher education.

**Wolfgang Lenhard**, PhD, studied special education and psychology. Since 2005, he has worked at the Institute for Psychology in Wuerzburg. His research interests are diagnoses and interventions in the field of reading comprehension, computer-based assessment, intelligent tutorial systems, the fostering of mathematical abilities, cognitive training, and the diagnosis of attention-deficit hyperactivity disorder.

**Birgit Spinath** is Professor of Educational Psychology in the Department of Psychology at Heidelberg University. Her research interests concern learning and teaching in schools and higher education, the motivational prerequisites of learning and achievement, individual differences as determinants of learning and achievement, teacher education, and the psychology of learning and teaching.

Abstract

Essays that are assigned as homework in large classes are prone to cheating via unauthorized collaboration. In this study, we compared the ability of a software tool based on Latent Semantic Analysis (LSA) and student teaching assistants to detect plagiarism in a large group of students. To do so, we took two approaches: The first approach was in vivo; that is, we observed whether LSA and the teaching assistants could detect plagiarism during the term. The second approach was in vitro; that is, we had 14 teaching assistants and LSA evaluate, after the term had ended, a sample of $N = 60$ essays of which two essays were identical Results showed that the responsible teaching assistant did not detect the duplicates during the term (in vivo) and that the majority of the teaching assistants did not notice that they had read two identical essays (in vitro). Some of them even scored the duplicates in markedly different ways. However, the duplicates were easily identified and evaluated as equally good by LSA. We conclude that using LSA can improve assessment at universities in terms of detecting plagiarism.

*Keywords:* Latent Semantic Analysis, LSA, plagiarism detection, cheating, collusion

Plagiarism detection: A comparison of teaching assistants and a software tool in identifying cheating in a psychology course

The advantages of assigning written essays in teaching have been highlighted by several authors (e.g., Foltz, Gilliam, & Kendall, 2000; McGovern & Hogshead, 1990; Miller, 2003; Wade, 1995). In addition, there are some problematic aspects of traditional exams (e.g., multiple-choice tests), for example, that they do not resemble future tasks and that they promote shallow learning (e.g., Ritter, 2000) or that "they are insufficient for teaching, learning and measuring the full range of study and knowledge-application skills that competent adults need" (Landauer & Psotka, 2000, p. 73). Consequently, Isaksson (2008) justified the teaching practice of having students write some kind of essay instead of using a final exam because the final-exam format fosters surface learning.

Although there are benefits to having students write essays, there is also a risk: If students write the essays at home, they might try to cheat and copy from other sources; that is, they might commit plagiarism. According to Park (2003), plagiarism can be seen as a form of cheating or academic misconduct or dishonesty. It is obvious that any behavior falling in this category is not acceptable; hence, it is best avoided and should be detected if it occurs.

A great deal of research has addressed a broad range of different aspects of academic dishonesty or cheating, and especially important for us, it has been shown that it occurs—and that some forms seem to increase—at university level (e.g., McCabe, 2005; Newstead, Franklyn-Stokes, & Armstead, 1996; Vandehey, Diekhoff, & LaBeff, 2007; Whitley, 1998). There is a large literature on *unintentional plagiarism* and how to educate students to avoid it (e.g., Belter & Du Pre, 2009; Elander, Pittam, Lusher, Fox, & Payne, 2010; Landau, Druen, & Arcuri, 2002).

However, another—probably more severe—problem is *intentional* plagiarism. Strategies that teachers can use to avoid (intentional) plagiarism include regularly modifying the tasks, asking for analysis, evaluation, or synthesis, and using open-ended questions for

which many solutions are possible (e.g., guidelines published by Nottingham Trent University: http://www.ntu.ac.uk/adq/document_uploads/teaching/137785.pdf, referring to e.g., Carroll & Appleton, 2001). However, none of these strategies can ensure that no student will copy another student's text. Irrespective of its reasons—there might be understandable reasons to give away one's essay, for example, to help a friend (e.g., Ashworth, Bannister, & Thorne, 1997; Franklyn-Stokes & Newstead, 1995; Newstead et al., 1996; Owens & White, 2013)—, there is a deliberate intent to cheat in these cases and hence this form of cheating must be expected, detected, and avoided.

With this paper, we compare the abilities of student teaching assistants and a computerized system in detecting plagiarism in a large university course. The current paper is not aimed at addressing common forms of intentional plagiarism by which students copy from textbooks, the Internet, and so forth. Rather, we are suggesting a solution for large courses in which students answer questions that include giving one's personal opinion, analyzing and evaluating research findings, and connecting these with personal experiences or the like. These answers cannot directly be found in a textbook or the course material; consequently, students cannot copy from such external sources. Thus, although this concept includes factors that make plagiarism less likely (e.g., Austin & Brown, 1999; Culwin & Lancaster, 2001a; Warn, 2006), students might work together and fail to provide independent work.

Because it is impossible for a university instructor to read all essays, it is a well-established policy to employ teaching assistants who read the essays and give feedback to the students. However, if two students work together or one student copies another student's text and these two students are not supervised by the same teaching assistant, plagiarism cannot be detected. Even if the same teaching assistant is responsible for the work of both students, he or she might not notice plagiarism. Moreover, there is quadratic growth in the number of single comparisons that can be made between students' essays. To directly compare all of the essays written for a course, n*(n-1)/2 single comparisons would be necessary. In a course

with 500 students, this would result in 124,750 comparisons—a task that cannot be carried out manually. This is where a software tool might be helpful if it could detect plagiarism more reliably.

**Plagiarism Detection**

There are several plagiarism-detection methods and services that have relative advantages and disadvantages (for a review of these tools and some tests of or comparisons between systems, see e.g., Kakkonen & Mozgovoy, 2010; Lancaster & Culwin, 2005; Maurer, Kappe, & Zaka, 2006; McKeever, 2006; Purdy, 2005; Weber-Wulff, Möller, Touras, & Zincke, 2013). The functioning of most systems is similar: Their algorithm is based on the assumption that two writers will usually not use the same words and thus, the system identifies overlapping word strings. Most systems can check for duplicates across the submitted texts, and some of the products claim that they can also detect slight linguistic modifications. However, the majority of the systems primarily detect verbatim plagiarism.

For plagiarism detection, Culwin and Lancaster (2001a, 2001b) suggested a four-stage process: collection, detection, confirmation, and investigation. With regard to the system used in this work, the first two steps can be facilitated by the use of software tools. We used the self-developed learning platform ASSIST at the University of Wuerzburg which was a former version of βASSIST which is now available at the University of Heidelberg (http://assist.psi.uni-heidelberg.de/). By (β)ASSIST students can hand in essays electronically so that the essays are collected immediately. Further, because this learning platform uses Latent Semantic Analysis (LSA; e.g., Landauer, McNamara, Dennis, & Kintsch, 2007), it is possible to detect both verbatim and semantic plagiarism (see next paragraph)[1]. However, as there is a risk of false positives, human judgment will always be necessary to decide whether a text really is a duplicate (confirmation) or whether the author has cited another person's

[1] Readers who want to apply LSA will find some helpful information at http://www.psychometrica.de/context_lsa_en.html

ideas correctly and whether there is enough evidence to accuse the author of cheating (investigation).

Thus, although software tools might be helpful within the plagiarism-detection process, one should also be aware of the terms and conditions of their use. Whenever software tools are used to detect plagiarism, legal and ethical aspects such as intellectual property and copyright are important (e.g., Butakov & Barber, 2012; Foster, 2002; Mozgovoy, Kakkonen, & Cosma, 2010; Purdy, 2005). However, these might not be as significant when the students' texts are stored within the university only.

**Using an LSA-Based System to Detect Plagiarism**

LSA is a statistical technique that can be used to generate automatic evaluations of texts on the basis of their semantic similarity. To do so, texts are represented as vectors within a semantic space (for details of the modus operandi, see Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998; Martin & Berry, 2007). LSA has been shown to be powerful in essay assessment in the English language (e.g., Foltz, Laham, & Landauer, 1999; Landauer, Laham, & Foltz, 2000; Landauer, Laham, & Foltz, 2003a, 2003b; Landauer, Laham, Rehder, & Schreiner, 1997) as well as in the German language (e.g., Lenhard, Baier, Hoffmann, & Schneider, 2007; Seifried, Lenhard, Baier, & Spinath, 2012).

Further, LSA has also been used to detect plagiarism: Cosma and Joy (2012) used LSA to detect source-code plagiarism (with their tool called PlaGate), and Britt, Wiemer-Hastings, Larson, and Perfetti (2004) integrated LSA into their Sourcer's Apprentice Intelligent Feedback system (SAIF), which is aimed at enhancing students' sourcing and integration skills. Both author groups stated that LSA can detect cases in which sentences have been reordered (structural changes) and cases in which synonyms have been substituted/renamed (semantic changes); these changes are said to be the most common types of plagiarism (Britt et al., 2004). However, the performance of an LSA-based system is not

predictable in general but relies on several parameters (e.g., the corpus), and this also renders comparisons with other tools that depend on string-matching algorithms almost impossible (Cosma & Joy, 2012; Mozgovoy et al., 2010). Thus, it seemed worthwhile to us to test the ability of an LSA-based system to detect plagiarism in our course and compare its performance with our student teaching assistants' performance.

With LSA, it is possible to execute a *semantic* comparison of each essay in a given set with every other essay in that set. Thus, it is possible to detect verbatim as well as semantic plagiarism; that is, LSA will detect plagiarism even if students try to cheat by substituting synonyms for some words or by paraphrasing the content. If there is a pool of n essays, LSA can compute all possible n*(n-1)/2 single comparisons within (milli)seconds.

By means of our learning platform, (β)ASSIST, texts can be compared within or across cohorts. Thus, it should be helpful for detecting what we want to detect, that is, intentional intracorporal plagiarism. After the number of comparisons is defined, (β)ASSIST will return a rank order of (all pairs of) texts that is based on the semantic similarity of the texts; this similarity is also expressed in a similarity score. Further, the authoring students' names, e-mail addresses, and a link to their submissions are provided. With this link, it is possible to see details about the submissions (e.g., the exact date and time of the submissions). If a text is identified as a duplicate, the text can be marked as plagiarism.

After having identified semantically similar texts, an in-depth analysis of the text surface is implemented using the Smith-Waterman algorithm (Irving, 2004), an approach commonly used in genetics to identify similar genes. This allows teachers to retrieve text passages, even when words are omitted or replaced or their sequence is altered. By combining the LSA-based ranking with the surface analysis, the restrictions of the two approaches are compensated for and false positives are avoided. Identical or similar text passages are highlighted in the same color, which facilitates the visual inspection of plagiarized text.

Further, there is a percentage score that represents the proportion of colored text (i.e., possible plagiarism) in each text.

However, human judgment is indispensable because (β)ASSIST does not classify a text as conspicuous or inconspicuous but only gives back a list of (all pairs of) texts—ordered according to their semantic similarity. Some highly ranked text pairs might not include real plagiarism but might rather be similar due to the fact that some students copied the question into their text and thus their texts shared a large number of words. Thus, the decision where to draw the line between plagiarism and random or irrelevant similarity is up to human judgement. Some studies have shown that automatic plagiarism detection is very helpful as the systems "found" (i.e., indicated as conspicuous) texts that had already been identified as plagiarism as well as further undetected cases (e.g., Badge, Cann, & Scott, 2007). Moreover, studies have shown that automatic plagiarism-detection systems contribute to the reduction of plagiarism when students are told that a plagiarism-detection technique will be applied to their papers (e.g., Braumoeller & Gaines, 2001).

However, although it might be a widespread belief that computer programs are better than humans at "detecting" plagiarism, direct evidence of this superiority is scarce. We could not find any study that directly compared the ability of a software system versus human graders in detecting plagiarism or that at least systematically analyzed either the ability of software or human graders to detect plagiarism in a real learning setting. Park (2004) stated that marker vigilance has been the traditional way to detect plagiarism and that some markers might be more vigilant than others, but this idea has yet to be tested. Landauer et al. (2003a) reported an instance involving a professor who did not notice the similarity of two essays that he had read only some minutes apart. Further, Shermis, Raymat, and Barrera (2003) stated only that it is very difficult for human scorers to detect students' plagiarism. Finally, in their paper on the Ferret copy detector, Lyon, Barret, and Malcolm (2006) alluded to the potential problem that there are many graders when cohorts are large, and thus, that plagiarism

detection is very difficult (unless the potential instances of plagiarism are graded by the same person). In their paper, they also included a paragraph on the "Comparison of Plagiarism Detection By Man and Machine", but they did so on a very theoretical basis: They looked at differences in language/memory processing (i.e., humans remember the semantics, machines store exact word strings). Hence, we wanted to add to the literature by testing an idea that is almost part of the folklore of higher education: that is, that software tools are superior to human graders in detecting plagiarism.

## Method

### Research Questions

In the present paper, we investigated the potential of LSA and student teaching assistants to detect plagiarism. We analyzed whether teaching assistants and LSA were sensitive to duplicates. The research question was whether LSA and human teaching assistants could detect partially and completely identical essays, and even if they could not detect plagiarism, whether they at least rated completely identical essays as equally good. Specifically, we investigated the following:

1. Is our software tool superior to human teaching assistants in detecting plagiarism, that is, can LSA identify plagiarism more reliably than human teaching assistants both in vivo and in vitro, that is, in (partially identical) essays during the term and in (completely identical) essays in a specially constructed sample of essays?
2. Do human teaching assistants and LSA evaluate duplicates as equally good (i.e., are the evaluations of human teaching assistants and LSA reliable)?

### Participants

The setting for this research was a university psychology course for preservice teachers. To pass the lecture "Introduction to Educational Psychology," preservice teachers answered two or three complex questions about the lecture material every second week. Fourteen teaching assistants who had attended the lecture in a previous term and who had

received a training for teaching assistants provided feedback to the students. Every teaching assistant had to supervise about 20 to 30 students (i.e., 10 to 15 students every week) and the same students throughout the term.

**Text Material and Procedure**

**Analyses in vivo**. We smuggled a fake person in the lecture whose name, Lina-Tessa Gropp, was an anagram of the German word for a person who commits plagiarism (i.e., "Plagiatsperson"). She seemed to be a regular student to the teaching assistants and was supervised by one of them throughout the term (i.e., teaching assistant 10). Under Lina-Tessa Gropp's name, we submitted some fake essays. To create the fake essays, we used passages from essays that had already been submitted by fellow students and deleted or added or substituted some words (i.e., minor changes). We ensured that the fake essays were of average quality (i.e., inconspicuous in this regard). Thus, their content was meaningful but not original because the fake essays were partially identical to some of the real essays which were submitted by ordinary students. Lina-Tessa Gropp became cockier during the term: For Session 1 and 2, the fake essays included slightly modified passages from students whose texts were not read by the teaching assistant who read the fake person's essays. Thus, it was almost impossible that she would be detected. However, the third fake essay—handed in for Session 3—included slightly modified passages from two other students who were supervised by the same teaching assistant who also was responsible for the fake person (i.e., teaching assistant 10). Thus, at least this case of plagiarism could be detected by the responsible teaching assistant. These data were used for the in vivo analysis to obtain a first impression of the efficiency of teaching assistants and the software tool for detecting plagiarism in a natural setting.

**Analyses in vitro**. To investigate the plagiarism-detection abilities of the teaching assistants and the software tool more systematically, after the term had finished, we had the 14 teaching assistants and LSA evaluate a sample of $N = 60$ essays on the topic of one lecture,

retrospectively. In this sample, there were two completely identical essays (apart from a heading in one text). These data were used for the in vitro analysis. We observed whether the teaching assistants and the software tool could detect plagiarism and whether these two completely identical essays were evaluated as equally good.

**Human evaluations.** The $N = 60$ essays that were evaluated by the teaching assistants after the term were anonymized and presented in a random order. The teaching assistants were told to score the essays independently of each other but with the help of a specimen model solution and a scoring scheme. The essays could be assigned a minimum of 0 and a maximum of 10 points. The scheme listed the requirements for allocating a certain number of points and included the maximum number of points for each of the two tasks (with a graduation of 0.5 points).

**LSA-based evaluations.** A full description of how LSA works is beyond the scope of this article (see e.g., Landauer & Dumais, 1997; Martin & Berry, 2007). For the present investigation, we used a semantic space that was previously used in another study (Seifried et al., 2012).

The students' essays were represented as vectors in this semantic space. LSA based its scores on a comparison with an ideal answer (i.e., a "gold standard"). The ideal answer was the specimen model solution that was available to the teaching assistants as well. The cosine between each text and the ideal answer was computed, and texts were ranked according to their proximity to this ideal answer. In order to project the rank of each essay to the raw point score used by the human graders, a normal rank transformation was applied by computing the respective z-score by means of the inverse normal cumulative distribution. One teaching assistant's scores for two texts (i.e., those texts at the 10th and 90th percentiles) were used to adjust the scores for the remaining essays via linear regression.

To apply the LSA check for plagiarism, three teaching assistants were told to check all texts for plagiarism by using LSA after the term. The teaching assistants were instructed to

compare all texts within ASSIST and list potential cases of plagiarism. They were not told that there were some partially or completely identical texts that they were supposed to identify. Thus, the check was conducted under real conditions.

## Results

### Detection of Plagiarism

In vivo (i.e., during the term), we had one fake person submit partially identical fake essays. None of them raised the suspicion of the responsible teaching assistant who supervised the fake person throughout the term. This had to be expected for Sessions 1 and 2 because the fake essays handed in for these Sessions were made up of passages by students who were supervised by other teaching assistants and hence the duplicates were read by different teaching assistants. However, even when the fake essay included passages from students who were supervised by the same teaching assistant and hence this teaching assistant was able to detect plagiarism as she read both the original texts and the duplicate (Session 3), she did not detect it.

In vitro, all 14 teaching assistants could detect the duplicates within the sample of $N = 60$ essays that they had to evaluate after the term. When they sent back their evaluations, four of them had noticed that two texts were identical (Teaching Assistants 4, 9, 10, and 14). Another teaching assistant wrote a comment that two texts were almost identical but nevertheless gave them scores that differed by 0.5 points (Teaching Assistant 6). The evaluations of the duplicates of three other teaching assistants also differed (Teaching Assistants 3, 5, and 7; for the exact evaluations, see Table 1 below). Thus, it is clear that they did not realize that they had read the same text twice. The remaining six teaching assistants scored the essays as equally good but did not mention anything about noticing plagiarism. Therefore, it is possible that they did not detect plagiarism but were reliable in their evaluations.

When LSA was applied, the suspicious texts were easily identified by the plagiarism check at the end of the term. This was true for both the completely identical texts from the sample (in vitro analysis) and the complex duplicates that were composed of passages from several other students' essays (including those partially identical essays that were used for the in vivo analysis). The completely identical texts that had been scored by all teaching assistants were at the top of the rank list for the respective session in ASSIST; the modified copies of other students' texts were ranked at the top of the list for the respective session when it was copied from one other student only (Session 1) and at Ranks 81 and 29 (Session 2) or Ranks 83 and 5 (Session 3) when the texts were composed of modified copies from two different fellow students. These lower rankings were probably due to the facts that (a) the texts were ranked with respect to their semantic similarity rather than with respect to verbatim overlaps, (b) the essays included passages from more than one text (i.e., from one text for each subquestion), and (c) the plagiarism check was conducted on the texts as a whole. However, skimming the texts led to a clear suspicion of plagiarism because of the colored text passages within the subtasks, and thus, these clusters of similar responses were identified as illicit teamwork. Further, because of the check at the end of the term, some more cases of plagiarism that had been undetected during the term were identified. By contrast, there were no cases of plagiarism that were identified by the teaching assistants but not identified by LSA.

**Evaluation of Identical Essays**

The single teaching assistants' evaluations, the teaching assistants' averaged evaluations, the LSA-based evaluations, as well as the differences between the evaluations of the two completely identical essays in the in vitro sample of $N = 60$ essays are shown in Table 1.[2]

[2] Because the partially identical essays that were submitted during the term included copies, abbreviations, rearrangements, and slightly modified passages from one or two fellow students, it is not possible to say that

Although the majority of the teaching assistants (i.e., 10 teaching assistants) evaluated the duplicates as equally good, four teaching assistants did not score the essays the same: The identical essays were scored with a difference of 0.5 points by two teaching assistants (Teaching Assistants 5 and 6) and with a difference of 1.5 by Teaching Assistant 7 and 2.0 points by Teaching Assistant 3. When LSA was applied, the identical texts were scored exactly the same.

## Discussion

The present study was conducted to test whether human teaching assistants and LSA could detect plagiarism. If LSA was superior to human teaching assistants, there would be a benefit of using a software tool within the teaching format of having students write essays because cheaters might be detected more reliably. Further, applying a software tool to detect plagiarism might also help to reduce plagiarism in the future because studies have shown that "in deterrence, actions speak louder than words" (i.e., students are not deterred from plagiarism by verbal or written warnings, but they are deterred when they know that teachers will check for it; Braumoeller & Gaines, 2001, p. 836).

The analyses computed to address Research Question 1 revealed that LSA was superior to human teaching assistants in detecting plagiarism; that is, LSA identified plagiarism more reliably than human teaching assistants. The in vivo data showed that the teaching assistant who had read partially identical texts during the term was not skeptical about this plagiarism. Further, although two texts in a sample of essays were completely identical, most of the teaching assistants did not notice the plagiarism (in vitro). This finding is in line with previous findings that showed that human graders are not good at detecting plagiarism (e.g., Landauer et al., 2003a; Shermis et al., 2003). On the other hand, all duplicates were easily detected by LSA.

---

these texts should have been scored the same as the original texts. Thus, these texts served as an indicator of only the potential of the teaching assistants and LSA to detect plagiarism in vivo (see paragraph above).

The analyses computed to address Research Question 2 on the in vitro data showed that some of the human teaching assistants reached only a poor reliability and assigned considerably different grades to the identical essays. The same essay quality was not evaluated as equally good by four of the 14 teaching assistants. It is known that different markers can assign different marks to the same essay and that criteria and marking schemes might be helpful to improve inter-rater agreement (e.g., Newstead & Dennis, 1994). However, our study indicates that there might also be problems of intra-rater agreement despite the use of scoring schemes. It might be an interesting topic for future research to identify the features of a text or a marker that make for different evaluations of the same essay. However, the majority of the teaching assistants assigned the same score to both texts and thus, their evaluations were reliable. When LSA was applied, the two identical texts were evaluated absolutely identically good or bad. This is a fact that might be an indicator of LSA's potential to evaluate even our complex essays (semi-)automatically and thus for another possible field of application for the software tool. We have addressed this in another paper (Seifried, Lenhard, & Spinath, submitted). While ranking the texts according to their semantic similarity might not be the best way to identify verbatim plagiarism, it is definitely useful to identify attempts to conceal plagiarism and to score essays based on their content.

**Practical Implications**

In our courses, we have one or two pairs of people who work together too much (i.e., who copy one another's texts) every semester. The results of the present study imply that the common practice of employing teaching assistants to accompany a lecture might be improved by the use of a software tool. It is impossible for teaching assistants to detect plagiarism if they do not read all texts that make up the collusions. However, duplicates that are authored by several students who are not supervised by the same teaching assistant and therefore cannot be detected by teaching assistants will easily be identified with the help of our software tool. The same is true for sophisticated duplicates that are made up of the ideas of

several students or that include synonyms or paraphrasing to disguise plagiarism because LSA is sensitive to *semantic* similarity. However, Mozgovoy et al. (2010) and Cosma and Joy (2012) pointed to the gaps in research concerning LSA's use in plagiarism detection and to the fact that there are several parameters that influence the power of LSA. Thus, the results cannot easily be generalized.

However, because LSA was reliable in its evaluation of two completely identical texts, whereas this was not true for some of the teaching assistants, an implication for educational contexts might comprise the use of our software tool for scoring essays. We have already conducted studies that show that the correlation between LSA-based scores and human scores does not differ significantly from the interrater correlations of human graders (Seifried et al., 2012) and that LSA can be used to identify poor essays (Seifried et al., submitted). Another aspect might be to use LSA as a "second opinion" to achieve objective scores as was suggested, for example, by Landauer et al. (2003b).

The present study shows that LSA can be useful for detecting plagiarism and possibly as a reliable second marker. Further, texts are easily collected within the system, texts can be compared within or across cohorts of students, and feedback can be assigned directly to the texts by teaching assistants. By using complex questions that ask students to give their personal opinion or analyze or criticize aspects, we can be quite sure that students will not find the answers to the questions in a textbook or on the Internet. Thus, it is sufficient to compare the texts only within our own database and therefore, we do not have to deal with legal concerns that arise when using an external system (Purdy, 2005). However, students should be informed that their texts will be collected and stored for reasons of plagiarism detection. In our experience, students who are accused of plagiarism usually deny their misconduct at first, but then it is interesting to see that the accused persons come to defend themselves together in couples even though we have told them only that there has been "considerable overlap with another student's text." Often, one of the students then states that

he/she is the one who is guilty and that the other person only wanted to help him/her. This (i.e., helping a friend) is a common reason for committing plagiarism (e.g., Ashworth et al., 1997). So, however or because of this, it is important to have students sign a pledge stating that they will not give away their own texts to another student (to optimize honor codes in different ways, see Gurung, Wilhelm, & Filz, 2012).

If students are aware of the application and efficiency of the software tool, they will most likely not dare to hand in duplicates of other students' ideas. This might lead to a reduction in plagiarism (e.g., Braumoeller & Gaines, 2001). In fact, in the last few semesters, we noticed a decrease in the plagiarism rate as the only clearly plagiarized texts were self-plagiarisms (i.e., students were asked to improve their text and obviously used their former text as the basis of their new text). We also think that it is not easy to outwit a system that bases its evaluations and comparisons on semantic similarity: Attempts to conceal plagiarism by the use of synonyms or the like should not influence the performance of LSA. Thus, although there is a fear that students might adapt to the software in order to avoid detection (Warn, 2006), this seems unlikely in our case because cheating the detection of plagiarism would require as much work as writing the essay on one's own. Plagiarism would lose its function as a labor saver in this way (also see Owens & White, 2013). Thus, our results show that a desirable teaching format (i.e., having students compose essays, giving feedback to them, as well as assessing students' achievements) can clearly be improved by the use of software tools.

References

Ashworth, P., Bannister, P., & Thorne, P. (1997). Guilty in whose eyes? University students' perceptions of cheating and plagiarism in academic work and assessment. *Studies in Higher Education, 22*, 187-203.

Austin M., & Brown L. (1999). Internet plagiarism: Developing strategies to curb student academic dishonesty. *The Internet and Higher Education, 2*, 21-33.

Badge, J. L., Cann, A. J., & Scott, J. (2007). To cheat or not to cheat? A trial of the JISC Plagiarism Detection Service with biological sciences students. *Assessment & Evaluation in Higher Education, 32,* 433-439.

Belter, R. W., & Du Pre, A. (2009). A strategy to reduce plagiarism in an undergraduate course. *Teaching of Psychology, 36*, 257-261.

Britt, M. A., Wiemer-Hastings, P., & Perfetti, C. A. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education, 14*, 359-374.

Braumoeller, B. F., & Gaines, B. J. (2001). Actions do speak louder than words: Deterring plagiarism with the use of plagiarism-detection software. *PS Online, 34*, 835-839.

Butakov, S., & Barber, C. (2012). Protecting student intellectual property in plagiarism detection process. *British Journal of Educational Technology, 43*, E101-E103.

Carroll, J., & Appleton, J. (2001). Plagiarism: A good practice guide. Oxford Brookes University and Joint Information Systems Committee (JISC).

Cosma, G., & Joy, M. (2012). An approach to source-code plagiarism detection and investigation using Latent Semantic Analysis. *IEEE Transactions on Computers, 61*, 379-394.

Culwin F., & Lancaster T (2001a). Plagiarism, prevention, deterrence and detection. Higher Education Academy.

Culwin, F., & Lancaster, T. (2001b) Plagiarism issues for higher education, *Vine, 31*, 36–41.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391-407.

Elander, J., Pittam, G., Lusher, J., Fox, P., & Payne, N. (2010). Evaluation of an intervention to help students avoid unintentional plagiarism by improving their authorial identity. *Assessment and Evaluation in Higher Education, 35*, 157–171.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*, 111-129.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*. Retrieved from http://imej.wfu.edu/articles/1999/2/04/

Foster, A. L. (2002). Plagiarism-detection tool creates legal quandary: When professors send students' papers to a database, are copyrights violated? *Chronicle of Higher Education*, 17 May, A37.

Franklyn-Stokes, A., & Newstead, S. E. (1995). Undergraduate cheating: Who does what and why? *Studies in Higher Education, 20*, 159-172.

Gurung, R. A. R., Wilhelm, T. M., & Filz, T. (2012). Optimizing honor codes for online exam administration. *Ethics & Behavior, 22*, 158-162.

Irving, R. (2004). Plagiarism and collusion detection using the Smith-Waterman Algorithm (available: http://www.dcs.gla.ac.uk/publications/PAPERS/7444/TR-2004-164.pdf). DCS Technical Report. Dept of Computing Science, University of Glasgow.

Isaksson, S. (2008). Assess as you go: The effect of continuous assessment on student learning during a short course in archaeology. *Assessment & Evaluation in Higher Education, 33,* 1-7.

Kakkonen, T., & Mozgovoy, M. (2010). Hermetic and web plagiarism detection systems for student essays - An evaluation of the state-of-the-art. *Journal of Educational Computing Research*, 42(2):135-139.

Lancaster, T., & Culwin, F. (2005). Classifications of plagiarism detection engines. *Innovation in Teaching and Learning in Information and Computer Sciences* 4(2). Available at http://journals.heacademy.ac.uk/doi/pdf/10.11120/ital.2005.04020006

Landau, J. D., Druen, P. B., & Arcuri, J. A. (2002). Methods for helping students avoid plagiarism. *Teaching of Psychology, 29*, 112-115.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems, 15*, 27-31.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003a). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis, & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective* (pp. 87-112). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003b). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*, 295-308.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In Shafto, M. G. & Langley, P. (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (412-417). Mahwah, NJ US: Lawrence Erlbaum Associates.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Landauer, T. K., & Psotka, J. (2000). Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments, 8,* 73-86.

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse [Automatic scoring of constructed-response items with latent semantic analysis]. *Diagnostica, 53*, 155-165.

Lyon, C., Barrett, R., & Malcolm, J. (2006). Plagiarism Is Easy, But Also Easy To Detect. *Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification*, 57-65.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35-55). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science, 12*, 1050-1083.

McCabe, D. L. (2005). Cheating among college and university students: A North American perspective. *Journal of Educational Integrity, 1*. Retrieved 5/22/2014 from http://www.ojs.unisa.edu.au/index.php/IJEI/article/view/14.

McGovern, T. V., & Hogshead, D. L. (1990). Learning about writing, thinking about teaching. Teaching of Psychology, 17, 5-10.

McKeever, L. (2006). Online plagiarism detection services - saviour or scourge? *Assessment & Evaluation in Higher Education, 31,* 155-165.

Miller, T. (2003). Essay assessment with Latent Semantic Analysis. *Journal of Educational Computing Research, 29,* 495-512.

Mozgovoy, M., Kakkonen, T., & Cosma, G. (2010). Automatic student plagiarism detection: Future perspecitves. *Journal of Educational Computing Research, 43*, 507-527.

Newstead, S.E. & Dennis, I. (1994). Examiners examined: The reliability of exam marking in psychology. *The Psychologist, 7*, 216-219.

Newstead, S. E., Franklyn-Stokes, A., & Armstead, P. (1996). Individual differences in student cheating. *Journal of Educational Psychology, 88*, 229-241.

Nottingham Trent University, Centre for Academic Development and Quality. (2013). CADQ Guide: Plagiarism and other academic misconduct. Retrieved from http://www.ntu.ac.uk/adq/document_uploads/teaching/137785.pdf

Owens, C., & White, F. A. (2013). A 5-year systematic strategy to reduce plagiarism among first-year psychology university students. *Australian Journal of Psychology, 65*, 14-21.

Park, C. (2003). In other (people's) words: Plagiarism by university students – literature and lessons. *Assessment & Evaluation in Higher Education, 28*, 471–488.

Park, C. (2004). Rebels without a clause: Towards an institutional framework for dealing with plagiarism by students. *Journal of Further and Higher Education, 28*, 291-306.

Purdy, J. P. (2005). Calling off the hounds: Technology and the visibility of plagiarism. *Pedagogy, 5*, 275–296.

Ritter, L. (2000). The quest for an effective form of assessment: the evolution and evaluation of a controlled assessment procedure (CAP). *Assessment & Evaluation in Higher Education, 25,* 307-320.

Seifried, E., Lenhard, W., Baier, H. & Spinath, B. (2012). On the reliability and validity of human and LSA-based evaluations of complex student-authored texts. *Journal of Educational Computing Research, 47*, 67-92.

Seifried, E., Lenhard, W., & Spinath, B. (submitted). Manuscript submitted for publication.

Shermis, M. D., Raymat, M. V., & Barrera, F. (2003). *Assessing writing through the curriculum with Automated Essay Scoring* (ERIC document reproduction service no ED 477 929).

Vandehey, M. A., Diekhoff, G. M., & LaBeff, Emily E. (2007). College cheating: A twenty-year follow-up and the addition of an honor code. *Journal of College Student Development, 48*, 468-480.

Wade, C. (1995). Using writing to develop and assess critical thinking. *Teaching of Psychology, 22*, 24-28.

Weber-Wulff, D., Möller, C., Touras, J., & Zincke, E. (2013). Plagiarism detection software test 2013. Retrieved from http://plagiat.htw-berlin.de/software-en/test2013/report-2013/

Whitley, B. E. Jr., (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*, 235–274.

Warn, J. (2006). Plagiarism software: No magic bullet! *Higher Education Research & Development, 25*, 195-208.

Table 1

*Single Teaching Assistants' Evaluations, Teaching Assistants' Average Evaluations, LSA-Based Evaluations, and Differences between the Evaluations of the Two Duplicates in the Sample of N = 60 Essays*

| Text | Teaching assistants | | | | | | | | | | | | | | | LSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | Average | |
| Original | 4.0 | 5.0 | 4.5 | 4.0 | 5.0 | 5.0 | 5.0 | 4.0 | 3.0 | 5.0 | 4.0 | 3.5 | 6.0 | 3.0 | 4.36 | 6.40 |
| Duplicate | 4.0 | 5.0 | 2.5 | 4.0 | 4.5 | 5.5 | 3.5 | 4.0 | 3.0 | 5.0 | 4.0 | 3.5 | 6.0 | 3.0 | 4.11 | 6.40 |
| Difference | 0.0 | 0.0 | 2.0 | 0.0 | 0.5 | 0.5 | 1.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.25 | 0.00 |

*Note.* Higher scores represent better evaluations of students' essays; essays could be evaluated with a minimum of 0 and a maximum of 10 points.

**Paper II**

This is the accepted manuscript (Version 2) of the article

Seifried, Eva, Lenhard, Wolfgang & Spinath, Birgit (2016). Filtering essays by means of a software tool: Identifying poor essays. *Journal of Educational Computing Research 0735633116652407, SAGE Publications, first published on June 6, 2016 as doi:10.1177/0735633116652407*.

Title: Filtering Essays by Means of a Software Tool: Identifying Poor Essays

Running head: IDENTIFYING POOR ESSAYS

Eva Seifried[1*], Wolfgang Lenhard[2], & Birgit Spinath[1]

[1] Department of Psychology, Heidelberg University, Germany

[2] Department of Psychology, University of Würzburg, Germany

[*]Corresponding author

Correspondence concerning this article should be addressed to Eva Seifried, Department of Psychology, Heidelberg University, Hauptstraße 47-51, 69117, Heidelberg, Germany, Phone: ++49 (0) 6221 / 547728, Fax: ++49 (0) 6221 / 547326, E-Mail: Eva.Seifried@psychologie.uni-heidelberg.de

**Acknowledgement**

This research was supported by the Innovation Fund FRONTIER at Heidelberg University (project number D.801000/10.25). We want to thank Dr. Herbert Baier for his work on ASSIST (i.e., a former version of βASSIST at the University of Würzburg), Fabian Grünig for his work on βASSIST and Jane Zagorski for proofreading the manuscript.

**Notes on Contributors/Short Bios**

**Eva Seifried** received her diploma in psychology at Heidelberg University in 2010. She now is a doctoral student and research associate at the Department of Psychology at the same university. Her research interests are learning and teaching in higher education, especially psychology and teacher education.

**Wolfgang Lenhard**, PhD, studied special education and psychology. Since 2005, he has worked at the Institute for Psychology in Würzburg. His research interests are diagnoses and interventions in the field of reading comprehension, computer-based assessment, intelligent tutorial systems, the fostering of mathematical abilities, cognitive training, and the diagnosis of attention-deficit hyperactivity disorder.

**Birgit Spinath**, PhD, is Professor of Educational Psychology in the Department of Psychology at Heidelberg University. Her research interests concern learning and teaching in schools and higher education, the motivational prerequisites of learning and achievement, individual differences as determinants of learning and achievement, teacher education, and the psychology of learning and teaching.

Abstract

Writing essays and receiving feedback can be useful for fostering students' learning and motivation. When faced with large class sizes, it is desirable to identify students who might particularly benefit from feedback. In this paper, we tested the potential of Latent Semantic Analysis (LSA) for identifying poor essays. Fourteen teaching assistants evaluated a sample of $N = 60$ German essays. Using the human graders' evaluations as the standard of comparison, more of the poor essays were correctly identified by LSA than by random sampling (i.e., selecting essays by chance). By contrast, selection by text length did not perform better than random sampling. When 3 different teaching assistants evaluated another sample of $N = 94$ essays, the results largely replicated those found in the first sample. We conclude that LSA can help university teachers to identify poorly performing students. Additional analyses were computed to investigate the potential of combining the methods in different ways.

*Keywords:* latent semantic analysis, LSA, automated essay scoring, filtering essays

# Filtering Essays by Means of a Software Tool: Identifying Poor Essays

In university teaching, it is a desirable goal to keep students engaged with the material and to foster a deep understanding as well as critical thinking. Thus, writing essays in university courses has several advantages. If students have to write essays across the entire semester, they are continuously engaged with the material, and in contrast to less effective massed learning, distributed practice is one of the most advantageous learning techniques (see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Further, learning progress should be especially high if students receive feedback on how to improve. Receiving feedback on one's performance facilitates the process of knowledge restructuring (e.g., Pearsall, Skipper, & Mintzes, 1997) and results in better consolidated knowledge structures. Nevertheless, reading all essays to ensure that every essay displays a minimum level of quality and providing a feedback to all students might be a hardly manageable effort. Thus, there is a need to reduce the amount of essays to be read and to select certain students for feedback. It is reasonable to provide feedback particularly to those students whose essays are of poor quality because a poor essay might indicate that the author has not understood the material and needs special advice. Further, it is important to supervise those students that do not show an adequate level of understanding because students should not pass the course if their work does not comply with basic standards. Because it is not clear a priori which essays are those of the poorest quality, it would be helpful to have a method that helps to filter texts quickly and without much workload for the instructor. This is especially true for Massive Open Online Courses (MOOCs) and other forms of distance learning, where individual feedback is hard to provide for all students. With the development of modern technology, students can now submit their essays via learning platforms. In this manner, the essays are available electronically and can be assessed with (semantic) technology. A special kind of automatic language processing, that is, Latent Semantic Analysis (LSA), can be applied to

perform various tasks, for example, to score essays automatically or to detect plagiarism in essays. Within this paper, we want to discuss a possibility on how to use LSA for handling large numbers of submissions so that writing can be applied even in large university courses.

## Writing Essays to Foster Learning and the Importance of Feedback

**Why and how to use writing in universities.** As early as 1977, Emig illustrated the similarities between writing and learning. Since then, several authors have highlighted the benefits of writing essays (e.g., Foltz, Gilliam, & Kendall, 2000; Miller, 2003; Nevid, Pastva, & McClelland, 2012). Wade (1995) stated, "writing is an essential ingredient in critical-thinking instruction" (p. 24) and listed further advantages. One example is to secure the active learning of every student, as it is known that active learning techniques are important for raising students' performance (Yoder & Hochevar, 2005). Gingerich and colleagues (2014) recently demonstrated that active-learning processes are the key for students to benefit from write-to-learn assignments. Some years earlier, McGovern and Hogshead (1990) referred to the assumptions of the writing-across-the-curriculum movement, that is, "that writing promotes learning and provides justification for writing in psychology" (p. 6). These authors reviewed literature that showed that writing fosters students' involvement and noted that such writing can also be used as an indicator of students' learning progress. In the same vein, other authors have reported positive effects of continuous assessment, which they stated, "is thought to promote deeper learning, greater motivation, and consequently [an] improved understanding of course material" (Carrillo-de-la-Peña & Pérez, 2012, p. 45). Thus, having students write continuously during a term has desirable effects that are due to more active and distributed learning (Cepeda, Pashler, Vu, Wixted, & Rohrer, 2006).

Consequently, there have been attempts to apply student writing in psychology courses (e.g., McGovern & Hogshead, 1990) despite the problem of the extra workload placed on instructors (e.g., Boice, 1990). Many studies have shown feasible ways to introduce

writing in class and outside of class; for example, journal writing (Connor-Greene, 2000; Hettich, 1990), portfolio assignments (Rickabaugh, 1993), summary writing (Radmacher & Latosi-Sawin, 1995), brief, focused, Internet-based writing assignments (Marek, Christopher, Koenig, & Reinhart, 2005), brief free writing (in class and ungraded; Drabick, Weisberg, Paul, & Bubier, 2007), "Five-minute" essays (Isaksson, 2008), microthemes (Stewart, Myers, & Culley, 2010), or creative designs such as PsychBusters (Blessing & Blessing, 2010).

**Why and whom to give feedback.** Although writing can foster learning, receiving feedback might improve this process even further. Meta-analyses have demonstrated that feedback can—if applied correctly—increase learning (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Gibbs and Simpson (2004) stated that "frequent assignments and detailed (written) feedback are central to student learning" (p. 8) but that resource constraints and increasing class sizes have reduced both the frequency of assignments and the desirable properties of feedback (i.e., quantity, quality, and timeliness). Because instructors' time is limited and university instructors in particular are faced with large classes, feedback cannot be given to every student every week. Thus, instructors might wish to focus their attention on particular students.

According to the functions listed by Gibbs and Simpson (2004), feedback might be particularly beneficial for students who are not performing well because feedback might serve to correct errors (e.g., Kulhavy, 1977), develop understanding through explanations, or encourage students to continue studying. Furthermore, it is important to supervise students who do not show an adequate level of understanding because students should not pass the course if their work does not comply with basic standards.

Although some methods can be applied to reduce instructors' workload or to make feedback in large classes feasible (e.g., Barber, Bagsby, Grawitch, & Buerck, 2011; Carkenord, 1998; McCabe, Doerflinger, & Fox, 2011), more detailed feedback for poorly

performing students seems to be desirable. Further, Isbell and Cote (2009) reported on the problems of connecting with and motivating students in large classes and asserted that negative effects might occur predominantly with struggling students.

## The Present Study

The literature summarized above suggests that writing in university courses is desirable, that feedback can enhance performance and that poorly performing students are natural candidates for feedback. However, when faced with a large number of essays of a priori unknown quality, identifying the poorest essays is not a straightforward process. Thus, instructors may be forced to select a certain number of essays by chance (i.e., pure random sampling). However, random sampling is not very promising and hence, alternative methods are needed. The current study was conducted to address the questions of whether automatic evaluations applied by a software tool can be used to identify poor essays and whether this method leads to better results than random sampling. We also assessed whether selecting essays by text length was better than random sampling.

### Methods to identify poor texts.

***LSA.*** The software tool we considered uses LSA (for an overview, see Landauer, McNamara, Dennis, & Kintsch, 2007)[1]. LSA is a special approach from the field of automatic language processing. It represents the meanings of words or texts on the basis of their occurrence in large text corpora within n-dimensional vector spaces. Using mathematical similarity computations, LSA can derive evaluations of texts (see Landauer, Foltz, & Laham, 1998; for details, see also Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990;

[1] We integrated LSA into a learning platform. Readers who want to apply the technology will find some helpful information at [deleted for anonymous review]. We did so in a platform called "ASSIST", which was provided by the University of [deleted for anonymous review], and in a new version of the system called βASSIST ([deleted for anonymous review]), which is now provided by the University of [deleted for anonymous review]. Students can directly submit their essays to βASSIST (by a given deadline), and this makes it possible to store texts, run plagiarism checks (based on semantic similarity; for an analysis of LSA's potential to detect plagiarism, see Authors, 2015), or give feedback.

Landauer & Dumais, 1997; Martin & Berry, 2007)[2]. There is a wide area of application for LSA, including cross-language information retrieval, intelligent tutoring systems, and semantic search engines (for some examples, see Landauer et al., 2007). Automatic essay assessment is a very prominent and interesting field of LSA application. It is possible to evaluate the quality of texts with respect to their content by measuring their semantic similarity to, for example, an especially good text (*gold standard*) or to other texts that have already been scored (*nearest neighbors*). LSA has been shown to be a powerful tool for essay assessment (e.g., Foltz, Laham, & Landauer, 1999). Previous studies have indicated the validity of LSA-based scores by showing that LSA's scores were correlated with those of single human graders to the same extent that single graders' scores were correlated with each other. For instance, this is the case for systems that can also be used to provide feedback: the IEA (e.g., Foltz, Streeter, Lochbaum, & Landauer, 2013; Landauer, Laham, & Foltz, 2000, 2003a, 2003b) and Summary Street (e.g., Wade-Stein & Kintsch, 2004). LSA has predominantly been applied in educational contexts in the English language and for the assessment of fact-based texts such as summaries (e.g., Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007; Wade-Stein & Kintsch, 2004). There is comparatively little experience with the application of LSA in other languages. This point is relevant because LSA-based systems rely strongly on their basic text corpus, which in turn depends on the underlying language. Thus, due to differences in language, it is possible that LSA is not equally suitable for application in every language. To date, there are only a few applications of LSA in German: Lenhard and colleagues created a German analogue of Summary Street

[2] There is also a free-access website for anyone who is interested in the application of LSA in the English language: http://lsa.colorado.edu/. This website is provided by the University of Colorado at Boulder. Although the website's content does not appear to have been updated since 2003, much helpful information about LSA in general and some applications are available for people who want to get a first impression of how LSA works. Information on how to use the LSA website is provided by Dennis (2007; the file can be downloaded from the website). Further, for instructors using the English language, the IEA, which is mentioned below, is available at Pearson Education/Pearson Knowledge Technologies. It is included in a variety of products (for more information, visit http://www.pearsonassessments.com/automatedlanguageassessment.html).

(conText; Lenhard et al., 2012; Lenhard, Baier, Endlich, Schneider, & Hoffmann, 2011; Lenhard, Baier, Hoffmann, Schneider, & Lenhard, 2007) and conducted further validation experiments (Lenhard, Baier, Hoffmann, & Schneider, 2007), all of which provided encouraging results concerning the suitability of LSA for the German language. Further, the potential for LSA to be used to score complex German texts was tested by comparing two scoring methods: nearest neighbors and gold standard (Authors, 2012). The results showed that the two scoring methods led to good and equally strong correlations with human graders' scores. Thus, although LSA is sometimes integrated in software tools that can provide automatic feedback (e.g., the IEA by Foltz et al., 2013; Landaueret al., 2000, 2003a, 2003b), there has been some evidence that automatic evaluations are not easily accepted by students (Lenhard, Baier, Hoffmann, & Schneider, 2007), and much effort would be necessary to ensure that the scores and feedback generated by such a system are valid and fair. Thus, the primary goal of the present article was restricted to another possible field of application of automatic evaluations, that is, to test the ability of LSA to identify poor texts.

***Text Length.*** Further, an alternative selection method was applied: It was analyzed whether choosing the shortest texts resulted in finding more poor essays than selecting essays by random sampling. This selection method was chosen for several reasons: First, the length of an essay can be easily computed by means of word count. Second, because students with a limited knowledge of a certain topic will typically not be able to produce a long text, the length of an essay might be an indicator for its quality (also see argumentation and data reported by Landauer et al., 2003a). Thus, choosing the shortest texts is an alternative selection method that can be implemented without much additional effort. Third, Landauer (2007) reports that LSA-based systems have sometimes been described as mere "bag-of-words" techniques. This is true insofar as LSA does not account for word order (i.e., syntactical information); however, words are treated differently by LSA than by keyword or

vector space models that can be classified as actual "bag-of-words" methods (see Landauer, 2007, p. 21f.). Thus, testing the potential of both LSA and text length to identify poor essays might underline that LSA goes beyond a mere addition of the words of a text.

**Research Questions.** The present study investigated the potential of LSA to identify the poorest essays among a larger number of complex German student-authored essays. The main research question was whether a selection of essays by means of LSA-based evaluations is able to identify more poor essays than random sampling. Further, we investigated whether a selection of essays by means of word count is able to achieve the same result. Specifically, the following was investigated:

1) Can LSA be used to identify poor essays? That is, does a selection of essays based on LSA-based evaluations result in the identification of more poor essays than selecting texts by chance (i.e., random sampling)?
2) Can text length be used to identify poor essays? That is, does a selection of essays based on text length result in the identification of more poor essays than selecting texts by chance (i.e., random sampling)?

## Method

### Overview

Two samples of essays were collected from a university psychology course. Essay quality was determined by averaging across teaching assistants' evaluations of the essays. The research questions included whether two methods, that is, filtering essays by LSA or by text length, were able to identify poor essays (i.e., the poorest 25% as identified by human evaluation). Using each of the two selection methods, either 25% or 12.5% of the essays were selected. The limit of 25% was set because, in a realistic setting in a large lecture, this limit should be high enough to include the essays written by the students who are struggling and because there might be sufficient capacity to read and evaluate about a quarter of the

submitted essays per week. The second limit of 12.5% was set because even if it were possible to provide feedback to 25% of the students, for reasons of fairness, it might be preferable not to concentrate exclusively on the worst-performing students but to also give feedback to every student at least once during a term. This can be achieved if only some of those students that have written a poor essay receive a feedback, while the remaining capacity of the instructor is shared among other and different students every week. The main interest of our study was the application of LSA, but the alternative selection method that was based on text length was applied because the length of an essay might be an easy-to-assess indicator of its quality.

**Participants**

The first sample of essays was collected from a psychology course for preservice teachers at a German university from which we investigated $N = 60$ essays. The second sample of essays was collected from a psychology course for psychology students at a German university from which we investigated $N = 94$ essays.[3]

**Text Material**

The essays consisted of answers to two (Sample 1) or three (Sample 2) complex questions concerning the topic of the lecture and required students to form their own opinions on the basis of the material, to analyze aspects, to abstract from the material or to find concrete examples of principles, or to connect different facets. This examination of the scientific material was intended to help students achieve a deeper understanding of the content by encouraging them to engage more deeply with the material and its implications.

**Measures**

**Human evaluations.** Fourteen (Sample 1) or three (Sample 2) teaching assistants evaluated each of the student-authored essays independently of each other. The essays were

[3] Reanalysis of data that are published elsewhere (Authors, 2012).

given to the teaching assistants in a randomized order, and the names of the students were deleted. The teaching assistants spent about 20 min grading each essay. The essays were evaluated with the help of a model solution and a scoring scheme (0 to 10 points). The average interrater correlations were $\bar{r}$ = .75 (Sample 1) and $\bar{r}$ = .76 (Sample 2). Thus, averaging the scores of all teaching assistants resulted in a reliable criterion for indicating text quality. For all analyses, 25% of all essays were classified as poor. Thus, the poor essays consisted of the 25% of the essays that had received the lowest human grader scores. Because percentage cutoffs might seem arbitrary and depend on both the quality of an essay and the distribution of students' skills, additional analyses that were based on a more objective criterion were also computed. Because achieving at least half of the maximum number of points (i.e., 5 points) is often used as a passing criterion, we chose this criterion for the additional analyses.

**LSA-based evaluations.** To determine the LSA-based scores, a semantic space had been created (for details of this procedure see e.g., Landauer, Foltz, & Laham, 1998; Quesada, 2007). The semantic space used for the present investigation has already been used in another study where the dimensionality had been set to 300 dimensions (Authors, 2012). This space was based on 41 psychology textbooks (covering the topics educational, social, abnormal, cognitive, organizational and developmental psychology) and was extended with material on the specific topics of the lecture. Prior to the construction of the space, the original text corpus was split into 55,973 passages that consisted of 256,407 different word forms.

To evaluate their content, the essays were represented as vectors in the semantic space and their proximity to a comparison text (i.e., a 'gold standard') was assessed by means of the cosine between them. The cosine is a measure that is often used in the application of LSA and may be interpreted as a correlation. The model solution used by the human graders was taken

as the comparison text. The essays were ranked according to their proximity to this 'gold standard'. In order to relate the rank of each essay to the raw point scoring system used by the human graders, we first applied a normal rank transformation by computing the accordant z-score by means of the inverse normal cumulative distribution. Then, the essays at the tenth and ninetieth percentile were identified as the essays that should be evaluated by human graders to adjust the scores of the remaining essays via linear regression. Thus, only two essays have to be scored manually as calibration anchors.

LSA computed the scores for all essays in a few milliseconds. The LSA-based scores were on the same scale as the human scores (0 to 10 points) but it should be noted that, in contrast to human evaluations, LSA-based evaluations—as determined here—still depend on the quality of the other texts even when an objective criterion (e.g., "less than 5 points") is applied. We selected the bottom 25% or 12.5% of the essays, respectively, or the essays with less than 5 points.

**Text length.** To determine the length of an essay, the number of words was calculated with a standard Excel command. Analogous to the selection by means of LSA, the shortest 25% or 12.5% of the essays were selected as potentially of poor quality.

**Statistical Analyses**

To assess the performance of the two selection methods (i.e., LSA and text length) in identifying the poor essays, the probability of finding an equal or higher number of poor essays by pure random sampling was calculated with the following test statistic:

$$p = \sum_{k=\hat{k}}^{\min(M,n)} P(k) \text{ with } P(k) = \frac{\binom{M}{k} * \binom{N-M}{n-k}}{\binom{N}{n}}$$

where $\hat{k}$ is the number of essays that were correctly identified by the selection method under consideration (i.e., the number of hits), $N$ is the total number of essays, $M$ is the number of essays classified as poor according to human standards (i.e., the human grader score), and $n$ is the total number of essays selected by each method.

The null hypothesis of this test states that the selection method can identify at most as many poor texts as random sampling. Under the null hypothesis, the exact distribution of the test statistic is given in terms of the hypergeometric distribution. Thus, if the probability $p$ does not exceed the common level of significance (i.e., a probability of 5% or .05), the selection method is significantly better than random sampling. Further, to include a measure of effect size, the observed-to-expected ($OE$) ratio was computed in addition to conducting null hypothesis significance testing.

## Results

### Descriptive Statistics

Table 1 presents the descriptive statistics for the essay evaluations that were based on the teaching assistants, LSA, and text length. Pearson correlations between the human grader scores, LSA-based evaluations, and word count are displayed as an indicator of agreement between evaluation methods.

Insert Table 1 about here

**Number of poor essays identified by each method in Sample 1.** For the first analysis in Sample 1 ($N = 60$, $M = 15$), the number of essays to be selected by LSA or word count, respectively, was set to $n = 15$ (25%). Under the null hypothesis, the expected value of $\hat{k}$ was $E(\hat{k}) = 3.75$ for these parameters. However, LSA correctly identified $\hat{k} = 7$ of the essays categorized as poor according to human standards. The probability of finding seven or more of the poor essays by pure random sampling was $p = .03$, $OE = 1.87$. By contrast, by using the selection method that was based on word count, $\hat{k} = 6$ of the poor essays were correctly identified. The probability of correctly identifying six or more of the poor essays by random sampling was $p = .12$. For the second analysis, the number of essays to be selected

was set to $n = 8$ (12.5%). For random sampling, $E(\hat{k}) = 2.00$ for these parameters. However, LSA correctly identified $\hat{k} = 5$, $p = .02$, $OE = 2.50$. With the help of word count, $\hat{k} = 3$ of the poor essays were correctly identified ($p = .32$). Thus, in both settings, the LSA-based selection method performed significantly better than random sampling, and the *OE* effect sizes were large, showing that LSA correctly identified 87% to 150% more of the poor essays than random sampling would have. By contrast, choosing the shortest texts did not result in the correct identification of a larger number of poor essays than choosing essays with random sampling.

For an additional analysis with a more objective criterion, we examined the number of essays that were scored with fewer than half of the maximum points (i.e., 5 points) by the teaching assistants ($M = 20$) and by LSA ($n = 10$). With these parameters, the expected value of $\hat{k}$ was $E(\hat{k}) = 3.33$. Among the essays that had received less than half credit (5 points) by the teaching assistants, LSA correctly classified six ($\hat{k} = 6$, $p = .06$, $OE = 1.80$), that is, 80% more than random sampling.

**Number of poor essays identified by each method in Sample 2.** For the analyses in Sample 2, 25% of the $N = 94$ essays should have been classified as poor, as well. Because some essays were assigned the same score, $M$ was reduced to $M = 19$. With this parameter, no essays that had been assigned exactly the same scores by the human graders or by LSA or that were of exactly the same length had to be split into distinct categories.

When the LSA-based evaluation was set to select $n = 19$ essays, $\hat{k} = 10$ of these were also poor according to human standards. However, for these parameters, the expectation of $\hat{k}$ was $E(\hat{k}) = 3.84$, and the probability of correctly identifying 10 or more of the poor essays by random sampling was $p < .001$, $OE = 2.60$. The selection method based on word count performed equally as well as the LSA-based method ($\hat{k} = 10$; $p < .001$, $OE = 2.60$). Thus, both methods, that is, choosing the essays with the poorest LSA-based evaluation and the

shortest essays, were significantly better than choosing essays at random. For the second analysis, $n$ = 10 (12.5%) essays were selected. For these parameters, $E(\hat{k})$ = 2.02. However, LSA correctly identified $\hat{k}$ = 5 of the poor essays ($p$ = .03, $OE$ = 2.48). By contrast, the word count method correctly identified only $\hat{k}$ = 4 of the poor essays ($p$ = .11). These results can be summed up by stating that significantly more poor essays were correctly identified by LSA than would have been identified by random sampling (160% or 148%, respectively), but choosing the shortest texts did not result in the correct identification of a larger number of poor essays than selecting texts by random sampling for the second analysis.

Again, for an additional analysis, we examined the number of essays that were scored with fewer than 5 points by the teaching assistants ($M$ = 14) and by LSA ($n$ = 34). By selecting 34 essays randomly, correctly identifying $E(\hat{k})$ = 5.06 of the poor essays would have been expected. However, among the essays classified as poor according to the teaching assistants, 13 received fewer than 5 points by LSA as well ($\hat{k}$ = 13, $p < .001$, $OE$ = 2.57, 157% more than by random sampling).

**Number of poor essays identified by a hybrid method: Number of poor essays identified by at least one method (i.e., LSA-based evaluations or word count) or by both methods.** It is interesting that the essays correctly identified as poor by LSA and word count were not the same in either sample. Some poor essays were identified by both LSA and word count, whereas some essays were identified by one method only. Thus, in practice, one would probably want to use a hybrid method to increase the ability to identify poor essays. Alternatively, instructors might be interested in avoiding having to look at many "false positives"; that is, they might wish to be sure that the selected essays were actually the poor ones. Thus, we decided to analyze whether using a hybrid method would be better than using either method alone. In order to do so, we built two scores: Score 1 = $\frac{\hat{k}}{M}$ (i.e., the number of hits in relation to the number of poor texts) indicates the detection ratio for the selection

method. It answers the research question of how many poor texts were identified by the method. Score 2 = $\frac{\hat{k}}{n}$ (i.e., the number of hits in relation to the number of texts selected by a/both method/s) indicates the credibility of the selection method and specifies the number of hits from among the selected texts.

To consider both possible methods of combining an LSA-based selection and a selection that was based on word count, we looked at both LSA ∧ word count (i.e., a text is selected if it is poor according to both LSA *and* word count; or else it is not selected) and LSA ∨ word count (i.e., a text is selected if it is poor according to LSA *or* word count [or both]; if a text is not identified as poor by either method, it is not selected). The detailed results for each method alone and the two possible combinations of the selection methods in both samples are displayed in Table 2.

Insert Table 2 about here

There is no statistical test that could be applied to compare (the combinations of) the methods. However, when looking at the data, word count alone appears to yield the worst results in all but one analysis (i.e., in Sample 2 with $n$ = 19). Using LSA *and* word count does not seem to be superior to using LSA only either. However, combining LSA and word count with an "*or*"-condition might lead to the best results when an instructor's interest is to find as many poor essays as possible. However, considering the cost-benefit ratio, the number of additional essays that need to be examined should be compared with the number of additional hits. For example, for the first sample with $M$ = 15 and $n$ = 15 for each of the two methods, only four poor essays were identified by both LSA and word count, whereas three poor essays were identified by LSA only and two were identified by word count only; six essays were classified as poor by both LSA and word count but not by the human graders. Due to

the intersecting set of essays identified by both methods, a total of $n_{joint}$ = 20 essays would have been selected. Thus, compared with using LSA only, using LSA and word count with an "*or*"-condition resulted in five additional essays that had to be examined, and two additional poor essays were identified.

## Discussion

### Summary

The present study analyzed whether LSA-based evaluations could identify poor essays. Such a tool can facilitate the writing and submitting of essays in large classes because the instructor's attention can be focused on students who need special help. Our study shows that LSA-based evaluations can be used to identify essays of poor quality. More of the essays classified as poor according to human standards were correctly identified by LSA than would have been by random sampling. This was true for different samples and in all except one analysis. By contrast, in all but one setting, selecting texts by length did not perform better than random sampling would have. In addition, in both samples and in all analyses, the number of essays correctly identified as poor by the LSA-based evaluation equaled or exceeded the number of essays correctly identified as poor by means of text length. Thus, the results imply that LSA is superior to text length in identifying poor essays.

If an instructor is trying to decide whether to use either method or a combination of them, there are different aspects that should be considered. First, there is the detection ratio (operationalized as Score 1 in our analyses), which indicates how many poor essays will be found. Second, there is the credibility of a method (operationalized as Score 2 in our analyses), which indicates how many of the selected essays will truly be poor ones. Higher scores represent better results for both Score 1 and Score 2. However, there is a certain trade-off: Whereas Score 2 would be maximized by reducing the number of texts selected, Score 1 would be maximized by increasing the number of texts that are selected because the larger

the sample, the larger the number of hits that can be expected. Hence, third, when combining the methods, an instructor might consider the cost-benefit ratio. This ratio can be defined as the additional number of essays that would be selected in relation to the additional number of poor essays that would be identified by doing so. The results of our study indicate the superiority of LSA over word count—both when comparing a method's results with random sampling and when looking at Score 1 or Score 2, respectively. However, there is no statistical test to compare the scores between methods (e.g., Score 1 for LSA and Score 1 for word count) and thus to compare the results when the methods are combined. Looking at Score 1, it seems as though combining the methods by using an "*or*"-condition was superior to using LSA only. However, we think that although there was an increase in the number of poor essays identified in each case, this came at the expense of selecting more essays. Thus, we would say that it would not be more efficient to use both methods instead of LSA alone.

**Critical Reflection**

**Conclusion.** When an instructor requires his/her students to compose essays, such writing assignments can be used to directly address a student's misconceptions (e.g., proposed by Connor-Greene, 2000). Further, the writing assignments can be used in a formative way to help students understand the content (e.g., Nevid, Pastva, & McClelland, 2012). However, to put effective yet time-consuming teaching formats such as continuous essay writing into practice—especially in large classes—instructors need a way to reduce their effort and to focus on particularly important aspects of their teaching. Such tools enable instructors to give feedback specifically to those students who are likely to benefit the most from such (formative) feedback. Hence, there is a need to filter poor essays. Our study analyzed the potential of two selection methods that might be used to identify poor essays instead of relying on pure random sampling, that is, selecting essays by LSA or text length. LSA proved to be useful and probably more efficient than a combination of both methods.

Thus, we conclude that LSA can help university teachers to identify poorly performing students and hence make the application of writing possible even in large university courses.

**Limitations and issues for future research.** Our analyses show that LSA was superior to random sampling in identifying poor essays, whereas text length was not. This may be because LSA focuses on the content of a text irrespective of its length. This is true for human graders as well. However, there is a relation between the quality of a text and its length (which might be logarithmic; see Shermis, Burstein, & Leacock, 2005). Also, LSA does not directly take text length into account, but the length of an essay is included indirectly in LSA scores. To be given a high score by LSA, a text has to achieve a certain length. However, as mentioned above, the poor essays correctly identified by LSA and by word count were not always the same. Further, although more poor essays could be identified by LSA than would be found by random sampling, a certain number of poor essays remained undiscovered. It might be an interesting topic for future research to identify the features that allow poor-quality essays to achieve good LSA scores.

Another general issue concerns the definition of what a poor essay is. Dichotomizing continuous variables (e.g., the quality of an essay) has been criticized (e.g., MacCallum, Zhang, Preacher, & Rucker, 2002). All analyses presented above are affected by this criticism: Whenever a cutoff is drawn, students whose performance falls below the criterion might be more different than two persons whose performance just passes or misses the threshold so that one of them is classified as a poorly performing person whereas the other is not. Nevertheless, we think that the idea of identifying and helping poorly performing students—however defined—is worthwhile.

Despite this possibility, there are some caveats as well: First, the idea to help poorly performing students—or rather the poorest performing students in a course—is related to our definition of "poor" . It should be clear that with our very simple approach, the worst essays

in the sample were defined as “poor” no matter what their absolute level was. However, at worst, selecting essays in this way would mean that instructors would focus on essays that are not very poor. Second, the success of our method depends on the comparison used by LSA, that is, an essay’s LSA-based score depends on both the quality of the model solution and the quality of the other essays in the sample. If texts are compared with a specified comparison text, it is possible that students will write a text that is not similar to the model solution but that is nonetheless very good. Thus, it is important to carefully select the comparison text (i.e., the quality of the model solution is important for the success of the scoring) and to keep in mind that there might be some “missings” as well as “false positives” if one relies solely on LSA scores to determine the quality of a text. LSA might help in the identification of special essays, but it is not a perfect method for doing so. Further, the scoring method that we used can certainly be improved, for instance, if prescored essays are available, if additional information beyond the cosine is included in the computations, or if transforming the proximity scores into the raw point scores used by the human graders is based on more than two essays. Our very simple approach (see above) is based on a social criterion because the closer students’ essays are to the model solution, the higher the scores they will receive. However, the closeness depends on the other essays; that is, if many very good essays are handed in in the same course, a good essay will be graded worse than an equally good essay in a sample of rather bad essays. Further, because the essays are ranked according to how close they are to the model solution and assigned a score by a normal rank transformation, the scores will not be valid if the underlying distributional assumptions are not true. Thus, an essay’s LSA-based score depends on the quality of the other essays in the sample. The same is true for word count—we focused on the shortest essays because we could not set a clear boundary a priori to define a poor essay. A score based on a social comparison should not be used to assign a grade or be reported to the students, but if the intention is to identify the

students with the poorest performance, this approach seems justifiable. In contrast, when it comes to grading, an objective criterion will be indispensable for guaranteeing fair treatment. However, one would need a huge number of prescored essays or the like to do so. But if instructors want only to increase the probability of finding the worst essays in their course, the simple method based on a social comparison would be sufficient. Third, LSA is language-dependent. Results from research in English- and German-speaking areas are promising, but further research is needed to make a reliable statement about their generalizability. Fourth, LSA does not account for syntactical information that is not captured by inflections on the word level. Tasks that rely heavily on word order details will probably not be suitable for an LSA-based evaluation.

**Implications for educational practice.** For educational practice, using LSA to filter essays has a clear advantage over filtering essays by word count: Students can easily outwit a system that is based on word count. By contrast, to "trick" a system based on LSA, it should be necessary to write an essay with meaningful content. However, this is completely in line with educational objectives. Thus, both from an empirical and from a practical perspective, selecting essays by means of LSA-based evaluation is a valid method for identifying poor-quality texts. Because distance learning is and will continue to grow as an alternative to on-campus programs, developing and improving learning platforms or using software tools will be of great value in the future. We hope that—with advanced tools—the application of LSA or comparable technologies will become easier and will lead to even better results than can currently be achieved.

References

Authors (2015).

Authors (2012).

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.

Barber, L. K., Bagsby, P. G., Grawitch, M. J., & Buerck, J. P. (2011). Facilitating self-regulated learning with technology: Evidence for student motivation and exam improvement. *Teaching of Psychology, 38*, 303–308.

Blessing, S. B., & Blessing, J. S. (2010). PsychBusters: A means of fostering critical thinking in the introductory course. *Teaching of Psychology, 37*, 178–182.

Boice, R. (1990). Faculty resistance to writing-intensive courses. *Teaching of Psychology, 17*, 13–17.

Carkenord, D. M. (1998). Assessing the essay feedback technique of providing an example of a full-credit answer. *Teaching of Psychology, 25*, 190–191.

Carrillo-de-la-Peña, M. T., & Pérez, J. (2012). Continuous assessment improved academic achievement and satisfaction of psychology students in Spain. *Teaching of Psychology, 39*, 45–47.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354-380

Connor-Greene, P. A. (2000). Making connections: Evaluating the effectiveness of journal writing in enhancing students learning. *Teaching of Psychology, 27*, 44–46.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Dennis, S. (2007). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57–70). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Drabick, D. A. G., Weisberg, R., Paul, L., & Bubier, J. L. (2007). Keeping it short and sweet: Brief ungraded writing assignments facilitate learning. *Teaching of Psychology, 34*, 172–176.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58.

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication, 28*, 122-128.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*, 111–129.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*. Retrieved from http://imej.wfu.edu/articles/1999/2/04/

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and Applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88). New York: Routledge.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1*, 3–31.

Gingerich, K. J., Bugg, J. M., Doe, S. R., Rowland, C. A., Richards, T. L., Tompkins, S. A., & McDaniel, M. A. (2014). Active processing via write-to-learn assignments: Learning and retention benefits in introductory psychology. *Teaching of Psychology, 41*, 303-308.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hettich, P. (1990). Journal writing: Old fare or nouvelle cuisine? *Teaching of Psychology, 17*, 36–39.

Isaksson, S. (2008). Assess as you go: The effect of continuous assessment on student learning during a short course in archaeology. *Assessment & Evaluation in Higher Education, 33*, 1–7.

Isbell, L. M., & Cote, N. G. (2009). Connecting with struggling students to improve performance in large classes. *Teaching of Psychology, 36*, 185–188.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street®: Computer-guided summary writing. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 263–277). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232.

Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–34). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems, 15*, 27–31.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003a). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003b). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*, 295–308.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Lenhard, W., Baier, H., Endlich, D., Lenhard, A., Schneider, W., & Hoffmann, J. (2012). Computerunterstützte Leseverständnisförderung: Die Effekte automatisch generierter Rückmeldungen [Computer assisted reading comprehension instruction: Effects of automatically generated content feedback]. *Zeitschrift für Pädagogische Psychologie, 26*, 75–90.

Lenhard, W., Baier, H., Endlich, D., Schneider, W., & Hoffmann, J. (2011). Rethinking strategy instruction: Direct reading strategy instruction versus computer-based guided practice. *Journal of Research in Reading*. doi:10.1111/j.1467-9817.2011.01505.x

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse [Automatic scoring of constructed-response items with latent semantic analysis]. *Diagnostica, 53*, 155–165.

Lenhard, W., Baier, H., Hoffmann, J., Schneider, W., & Lenhard, A. (2007). Training of summarisation skills via the use of content-based feedback. In F. Wild, M. Kalz, J. Van Bruggen, & R. Koper (Eds.), *Proceedings of the First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning* (pp. 26–27). Heerlen, the Netherlands: Open University of the Netherlands.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19–40.

Marek, P., Christopher, A. N., Koenig, C. S., & Reinhart, D. F. (2005). Writing exercises for introductory psychology. *Teaching of Psychology, 32*, 244–246.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McCabe, J., Doerflinger, A., & Fox, R. (2011). Student and faculty perceptions of e-feedback. *Teaching of Psychology, 38*, 173–179.

McGovern, T. V., & Hogshead, D. L. (1990). Learning about writing, thinking about teaching. *Teaching of Psychology, 17*, 5–10.

Miller, T. (2003). Essay assessment with Latent Semantic Analysis. *Journal of Educational Computing Research, 29*, 495–512.

Nevid, J. S., Pastva, A., & McClelland, N. (2012). Writing-to-learn assignments in introductory psychology: Is there a learning benefit? *Teaching of Psychology, 39*, 272–275.

Pearsall, N. R., Skipper, J. E. J., & Mintzes, J. J. (1997). Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. Science Education, 81, 193–215.

Quesada, J. (2007). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 71–85). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Radmacher, S. A., & Latosi-Sawin, E. (1995). Summary writing: A tool to improve student comprehension and writing in psychology. *Teaching of Psychology, 22*, 113–115.

Rickabaugh, C. A. (1993). The psychology portfolio: Promoting writing and critical thinking about psychology. *Teaching of Psychology, 20*, 170–172.

Shermis, M. D., Burstein, J., & Leacock, C. (2005). Applications of computers in assessment and analysis of writing. In C.A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 403-416). Guilford Publications.

Stewart, T. L., Myers, A. C., & Culley, M. R. (2010). Enhanced learning and retention through "Writing to Learn" in the psychology classroom. *Teaching of Psychology, 37*, 46–49.

Wade, C. (1995). Using writing to develop and assess critical thinking. *Teaching of Psychology, 22*, 24–28.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333–362.

Yoder, J. D., & Hochevar, C. M. (2005). Encouraging active learning can improve students' performance on examinations. *Teaching of Psychology, 32*, 91–95.

Table 1

*Descriptive Statistics for the 25% or 12.5% Poorest or Shortest Essays and Correlations between the Average Human Grader Score, LSA-Based Evaluation, and Word Count*

| | *N* | *M* | *SD* | *Min.* | *Max.* | *Upper limit 25%* | *Upper limit 12.5%* | *r* | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | *2* | *3* |
| Sample 1 | | | | | | | | | |
| 1. Human grader score | 60 | 5.71 | 1.48 | 2.46 | 8.39 | 4.46 | | .61*** | .42*** |
| 2. LSA-based evaluations | 60 | 6.25 | 1.28 | 3.40 | 9.10 | 5.30 | 4.70 | | .76*** |
| 3. Word count | 60 | 586.78 | 136.43 | 286.00 | 837.00 | 492.00 | 411.00 | | |
| Sample 2 | | | | | | | | | |
| 1. Human grader score | 94 | 6.78 | 1.70 | 1.50 | 9.17 | 5.33 | | .58*** | .54*** |
| 2. LSA-based evaluations | 94 | 5.50 | 1.46 | 2.00 | 9.00 | 4.20 | 3.60 | | .71*** |
| 3. Word count | 94 | 865.26 | 325.12 | 336.00 | 2056.00 | 569.00 | 518.00 | | |

*Note.* Higher scores represent better evaluations of students' essays.

***$p \leq .001$.

Table 2

*Results for the Single Selection Methods and their Combinations across Samples*

| Sample 1: $N = 60$; $M = 15$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 15 essays selected by every single method | | | | 8 essays selected by every single method[a] | | | |
| | LSA | Word count | LSA ∧ word count | LSA ∨ word count | LSA | Word count | LSA ∧ word count | LSA ∨ word count |
| $n$ / $n_{joint}$ | 15 | 15 | 10 | 20 | 8 | 8 | 4 | 12 |
| $\hat{k}$ | 7 | 6 | 4 | 9 | 5 | 3 | 2 | 6 |
| Score 1 | 0.47 | 0.40 | 0.27 | 0.60 | 0.33 | 0.20 | 0.13 | 0.40 |
| Score 2 | 0.47 | 0.40 | 0.40 | 0.45 | 0.63 | 0.38 | 0.50 | 0.50 |

| Sample 2: $N = 94$; $M = 19$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 19 essays selected by every single method | | | | 10 essays selected by every single method[a] | | | |
| | LSA | Word count | LSA ∧ word count | LSA ∨ word count | LSA | Word count | LSA ∧ word count | LSA ∨ word count |
| $n$ / $n_{joint}$ | 19 | 19 | 11 | 27 | 10 | 10 | 4 | 16 |
| $\hat{k}$ | 10 | 10 | 6 | 14 | 5 | 4 | 2 | 7 |
| Score 1 | 0.67 | 0.67 | 0.40 | 0.93 | 0.33 | 0.27 | 0.13 | 0.47 |
| Score 2 | 0.53 | 0.53 | 0.55 | 0.52 | 0.50 | 0.40 | 0.50 | 0.44 |

*Notes.* $N$ = total number of essays; $M$ = number of essays classified as poor according to human standards (i.e., number of poor essays); $n$ / $n_{joint}$ = total number of essays selected by the selection method/s; Score 1 $= \frac{\hat{k}}{M}$; Score 2 $= \frac{\hat{k}}{n}$; $n$-$\hat{k}$ indicates the number of false positives; $M$-$\hat{k}$ indicates the number of missings.

[a] Note that the number of missings will be high and Score 1 will be low if only half of the poor essays are selected by the methods (i.e., the minimum number of missings is $M - n$ and the maximum for Score 1 is .50).

**Paper III**

This is the submitted version (Version 1) of the article

Seifried, Eva, Lenhard, Wolfgang & Spinath, Birgit (accepted pending revisions). Automatic essay assessment: Effects on students' acceptance and on learning-related characteristics. *Psihologija*.

# Automatic Essay Assessment: Effects on Students' Acceptance and on Learning-related Characteristics

Eva Seifried

Heidelberg University

Wolfgang Lenhard

University of Würzburg

Birgit Spinath

Heidelberg University

## Author Note

This research was supported by the Innovation Fund FRONTIER at Heidelberg University (project number D.801000/10.25). We want to thank Dr. Herbert Baier for his work on ASSIST (i.e., a former version of βASSIST at the University of Würzburg) and Fabian Grünig for his work on βASSIST.

Correspondence concerning this article should be addressed to Eva Seifried, Department of Psychology, Heidelberg University, Hauptstraße 47-51, D-69117, Heidelberg, Germany, Phone: ++49 (0) 6221 / 547728, Fax: ++49 (0) 6221 / 547326, E-Mail: Eva.Seifried@psychologie.uni-heidelberg.de

**Notes on Contributors**

**Eva Seifried** received her diploma in psychology at Heidelberg University in 2010. She now is a doctoral student and research associate at the Department of Psychology at the same university. Her research interests are learning and teaching in higher education, especially psychology and teacher education.

**Wolfgang Lenhard**, PhD, studied special education and psychology. Since 2005, he has worked at the Institute for Psychology in Würzburg. His research interests are diagnoses and interventions in the field of reading comprehension, computer-based assessment, intelligent tutorial systems, the fostering of mathematical abilities, cognitive training, and the diagnosis of attention-deficit hyperactivity disorder.

**Birgit Spinath** is Professor of Educational Psychology in the Department of Psychology at Heidelberg University. Her research interests concern learning and teaching in schools and higher education, the motivational prerequisites of learning and achievement, individual differences as determinants of learning and achievement, teacher education, and the psychology of learning and teaching.

Abstract

When capacity constraints restrain university instructors from giving feedback, software tools might provide a remedy. We analyzed students' acceptance of automatic assessments and development of learning-related characteristics such as motivation. We randomly assigned university students to four groups that differed regarding the real and assumed source of assessment of students' texts (i.e., teaching assistant or software tool). Data of $N = 300$ students were analyzed. Assessments were less accepted when presumably coming from the software tool. Students mostly preferred human graders over computers in teaching in general but this preference was weakened for some situations when students assumed to be assessed by the software tool. Nevertheless, students saw some general merits of assessments by computers and the development of learning-related characteristics was not affected by the real or assumed source of assessment. Thus, combining feedback by software tools and human graders seems feasible to enlarge feedback capacities in higher education.

*Keywords:* automatic essay scoring, acceptance, higher education, university students

Essay assignments are widely used at universities but when instructors are faced with large classes, assessing all essays can be an unmanageable effort. However, progress in the sector of automatic essay scoring (AES) makes it possible that students receive a feedback on their performance even in large courses. Despite the evidence on the validity of AES (e.g., Shermis & Burstein, 2003), there is little research on the acceptance of AES by students, especially at university level. In their review on the effects of computer-generated feedback on the quality of writing, Stevenson and Phakiti (2014) state that the relative effects of computer-generated feedback and teacher feedback are not clear yet and that it needs to be analyzed whether it is really the source of feedback that matters. The present study aims to close this gap by analyzing the acceptance of computer-based assessments with an experimental design.

'Writing-to-learn' has been shown to be effective to improve learning (e.g., Nevid, Pastva, & McClelland, 2012) and was identified as an evidence-based teaching technique in university teaching (Dunn, Saville, Baker, & Marek, 2013). Although even ungraded writing assignments can foster learning (Drabick, Weisberg, & Bubier, 2007; Nevid et al., 2012), receiving a feedback seems desirable to help students monitor their learning (e.g., Hattie, 2009). However, when faced with large classes (i.e., hundred or more participants), reading all assignments cannot be accomplished by the instructor only. Since the 1960s, there have been attempts to score essays automatically by computers (e.g., Page, 1966) and meanwhile, recent technologies are used both for summative and formative purposes, for high-stakes and low-stakes assessments (Shermis & Hamner, 2013). In general, research on AES has focused on psychometric issues (i.e., first of all its validation; e.g., Shermis & Burstein, 2003; authors, 2012) but little is known about the acceptance and the effects of AES.

Not only in the scientific community (e.g., Ericsson & Haswell, 2006) but also from those who are being assessed, there seem to be some concerns regarding AES. For example, there has been a petition, initially written by Haswell and Wilson in 2013, to stop using

computer scoring of student essays written during high-stakes tests (http://www.humanreaders.org/petition/index.php). The initiators list several reasons why machine scoring of essays is not defensible and refer to several research findings that substantiate their claim. More than 4,300 persons have yet signed this petition. According to Gierl, Latifi, Lai, Boulais and De Champlain (2014), AES "has been described as 'robo-scoring', 'roboreading', 'robo-grading' and 'auto-scoring'." (p. 959) These characterizations indicate that there are concerns regarding AES (for initial objections against AES see Page, 2003; Page & Peterson, 1995; for suspicions about the capability of computers to provide scores or feedback on writing also see Stevenson & Phakiti, 2014) but the extent and impact need to be further analyzed.

**Examining the Acceptance of AES**

Although not focusing on the acceptance of AES, some studies reported interesting, yet mixed results on this aspect nevertheless (Lai, 2010; Lenhard, Baier, Hoffmann, & Schneider, 2007; Lipnevich & Smith, 2009a, b). Lai (2010) found that English as a foreign language learners preferred to receive feedback from peers over a feedback from a computer tool. Lenhard and colleagues (2007) found that students perceived computer-generated feedbacks as helpful but not as really reflecting their texts' quality. Lipnevich and Smith (2009a) found the perceived source of feedback (i.e., a computer or the instructor) had little impact but students who assumed to have received a feedback by a computer rated their feedback as less accurate and helpful. In subsequent focus group discussions, Lipnevich and Smith (2009b) found that students who perceived that their feedback had come from a computer reported to be cautious or skeptical when hearing about the source of their feedback but then seeing its merits. Students indicated that the feedback was relevant for improving their essay and thought that the computer might have even be fairer and more unbiased than the professor. Some students also felt relieved that not the professor had read their essays. However, almost all students also reported that some of the comments did not apply to their

work and some decided to ignore the feedback. These expressions of doubt and rejection did not appear within the group that assumed to have received their feedback from the instructor although the comments were comparable. Some students within the perceived computer feedback group also perceived their grades as unfair (i.e., too low) because the computer might not be capable of scoring complex writing. Thus, there seem to be some concerns regarding AES but to our knowledge, they have not yet been analyzed systematically.

Experimental designs are needed to investigate whether (university) students accept AES and whether assessments by software tools influence the development of learning-related characteristics. If students do not accept AES, being assessed by a software tool might result in a decline of their motivation, aspirations and subjective learning.

**The Present Study**

This study aimed to explore the effects of software-generated assessments. We wanted to know whether students accepted the application of a software tool to assess their texts and whether there would be any further effects depending on the real or assumed source of assessment: We were also interested in students' perceptions on the use of computers in teaching in general, and whether an automatic assessment would negatively influence learning-related characteristics. Specifically, we had the following hypotheses:

1. The acceptance of a specific assessment will depend on the assumed source of assessment not the real source. The acceptance of the assessment will be lower when presumably coming from the software tool than when presumably coming from a teaching assistant. Scores coming from the software tools in truth will not be less accepted than scores coming from a teaching assistant in truth.
2. Students will prefer a person over a computer for different tasks in teaching in general.
3. There will be no negative effect on learning-related characteristics depending on the (real or assumed) source of the assessment, that is, neither the real nor the assumed

source of the assessment are expected to have a negative effect on students' outcomes such as their motivation, achievement aspirations, and subjective learning.

## Method

To enhance the ecological validity of the study, we included a small experiment within a lecture. We followed much of the study by Lipnevich and Smith (2009a) but extended the analyses to appraisals about the implementation of computers in teaching in general. Further, we applied a 2x2 design with the real and assumed sources of assessment fully crossed, thus following the recommendation by Stevenson and Phakiti (2014) that the kinds of feedback should be comparable so that it can be analyzed whether it is really the source of feedback that matters.

### Participants and Setting

The setting for this research was a university psychology course for preservice teachers (i.e., "Introduction to Educational Psychology"). Course requirements included answering complex questions about the lecture material every week and passing an examination at the end of the semester. Full data sets (i.e., submitted assignment, survey data, successful manipulation check, form of feedback as intended) were available from $N = 300$ students (age ranging from 18 to 41 years, with a mean age of 22.29 years ($SD = 3.39$); 189 (63.0%) participants were women, and 111 (37.0%) were men).

### Assessment

**Assessment by the teaching assistants**. Psychology students ($N = 14$) who had attended the course previously received training to provide feedback on the assignments. Every text was assessed with the help of a specimen model solution and a scoring scheme.

Texts were assessed on a 10-point scale with gradation of 0.5 points. It took about 6 hours to score the 33-34 texts per teaching assistant.

**Assessment by the software tool**. Students handed in their assignments electronically via a learning platform called ASSIST. ASSIST uses Latent Semantic Analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007) to perform special tasks, e.g., to detect plagiarism or to score texts. LSA is a special approach from the field of automatic language processing and aims to represent the meanings of words or texts in a so-called semantic space on the basis of the words' occurrence in large text corpora. Using mathematical similarity computations, LSA can derive evaluations of texts (see Landauer, Foltz, & Laham, 1998; for details also see Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Martin & Berry, 2007) . Several authors have used LSA for automatic essay assessment successfully (e.g., Landauer, Laham, & Foltz, 2003; authors, 2012).

On the basis of our positive results on the evaluation of LSA-based scores (authors, 2012) and the approaches for identifying cheaters (authors, 2015) or poorly performing students with the help of LSA-based scores (authors, submitted), we chose to test the acceptance of AES with LSA-based scores. For the present study, we used a semantic space which had already been used for other studies (for references about the procedure in general and the contents of the specific semantic space, see authors, 2012 and authors, submitted). For LSA, it took only some seconds to score the essays (for details on the procedure, see authors, submitted). Because the LSA-based assessments were continuous scores, they were adjusted upward or downward to the nearest gradation of 0.5 points to match the gradations by the teaching assistants.

## Procedure and Measures

At the beginning of the course, students took a survey that included questions about their motivation, their achievement aspirations, and their subjective learning (for details about

these measures, see below). We told students that they would receive either a software-generated assessment or an assessment by a teaching assistant for their first essay and that we would ask them about their opinion about this assessment. What students did not know was that the experimental design included that students were randomly assigned to four groups that differed regarding the real and the assumed source of assessment (fully crossed experimental design). That means, that only half of the students who thought their feedback came from a software tool actually received feedback from a software whereas the other half received feedback from a teaching assistant (the same was true for students who thought they received feedback from a teaching assistant. To make these conditions especially credible, students who had been told that their feedback was generated by a software tool received the feedback within one day after the submission deadline whereas those who thought that the feedback was generated by a teaching assistant received their feedback five days after the submission deadline. In all groups, feedback consisted of a score between 0 and 10 that indicated the degree to which the demands of the assignment were met. For some students, the feedback could not be given as intended: For ethical reasons, the teaching assistants were told to tell a student their own assessment if their score differed at least three points from the LSA-based score ($N = 14$) or if they thought that a text failed and the LSA-based assessment indicated that a text passed a minimum level of acceptance (or vice versa) so that the feedback was not completely unrealistic ($N = 1$). These data were excluded from further analysis.

Within one week after having received feedback, students were asked to complete a survey about their assessment, the implementation of computers in teaching in general and – like at the beginning of the term – their motivation, achievement aspirations, and subjective learning. We reminded students that they had received a score either from a teaching assistant or a software tool. Only data from students who indicated the source of their feedback correctly (manipulation check) were included in the following analyses.

To analyze students' acceptance of computer-generated scores (Hypothesis 1), they were asked to rate on a 5-point scale with (1) *absolutely not* and (5) *very* how much they perceived their assessment as (a) useful, (b) informative, (c) motivating, (d) clear and comprehensible, (e) helpful, (f) explicable and fair, (g) whether they thought that the score represented the quality of their text, and (h) how satisfied they were with their assessment. These data were integrated into an acceptance scale (Cronbach's $\alpha = .87$).

Further, students were asked to give their general opinion about automatic assessments and the implementation of computers in teaching in general (Hypothesis 2). Students were asked which source (i.e., a computer, a human grader or none) they would prefer for different opportunities, both regarding weekly submitted assignments and examinations. The applications regarding weekly submitted assignments included the following: (a) deciding about passing weekly submitted assignments, (b) assessing weekly submitted assignments, (c) giving feedback to weekly submitted assignments, and (d) providing a model solution for weekly submitted assignments. The situations regarding the examinations included the following: (a) deciding about passing an examination, (b) assessing an ungraded examination, and (c) assessing a graded examination. Additionally, because students had learnt about the criteria of scientific measurements within the course meanwhile, students were asked about the relative advantages and disadvantages of assessments by human graders or computers regarding these aspects (i.e., objectivity, reliability, validity, and speed). Again, they were asked to indicate what or who would be better regarding these aspects (i.e., a computer, a human grader or none).

Moreover, we wanted to monitor the development of learning-related characteristics (Hypothesis 3). Students' motivation was assessed according to expectancy-value theory (Wigfield & Eccles, 2000). Both students' values and their competence beliefs were assessed by three items each (e.g., value: "A sound knowledge of educational psychology is important for me"; competence beliefs: "I do well in educational psychology"). Students indicated

agreement on a 5-point scale with (1) *completely disagree* and (5) *completely agree*. Further, students rated their achievement aspirations, both for the weekly assignments by indicating the number of points that they wanted to achieve for their further texts (i.e., a score ranging between 0 and 10 points) and the examination at the end of the term by indicating whether they wanted (1) = *to be very good*, (2) = *to be good*, (3) = *to pass*. Further, they were asked to rate their subjective learning on a 5-point scale with (1) = *low* and (5) = *high*.

## Results

### Descriptive Statistics

Table 1 shows the descriptive statistics for the assessments by the teaching assistants and LSA as well as the descriptive statistics for the learning-related variables. The scores were highly correlated ($r = .62$, $p < .001$).

Insert Table 1 about here

### Acceptance of the Assessment

Across all experimental conditions, acceptance of the assessments was near the theoretical mean of the scale ($M = 2.77$, $SD = 0.77$). A 2 (real source of assessment) x 2 (assumed source of assessment) ANOVA revealed a significant effect main effect of the assumed source of assessment ($F(1,296) = 18.67$, $p < .001$, $\eta^2 = .06$). This effect indicated that students' acceptance was higher if they assumed to be assessed by a teaching assistant than by the software tool ($M = 2.98$, $SD = 0.77$ vs. $M = 2.60$, $SD = 0.73$). No significant effects were found for the main effect of the real source of assessment ($F(1,296) = 1.21$, $p = .272$) and the interaction between the real and the assumed source of assessment ($F(1,296) = 0.88$, $p = .349$).

The scores that the students had received ($M = 6.69$, $SD = 1.26$) correlated significantly with students' acceptance ($r = .54$, $p < .001$). Thus, to ensure that the different levels of acceptance of the assessments were not merely an effect of lower scores within one group, we ran a 2x2 analysis of covariance (ANCOVA) with real source of assessment, and assumed source of assessment as factors and the score for the essay as a covariate. The covariate was significant ($F(1,295) = 120.66$, $p < .001$, $\eta^2 = .29$), indicating that students' acceptance was associated with their level of achievement. However, the main effect of the assumed source of assessment remained significant ($F(1,295) = 18.15$, $p < .001$, $\eta^2 = .06$) after controlling for level of achievement while both the main effect of the real source of assessment ($F(1,295) = 0.62$, $p = .431$) and the interaction between the real and the assumed source of assessment remained insignificant ($F(1,295) = 0.09$, $p = .771$).

**Attitudes towards the Implementation of Computers in Teaching in General**

Regarding the implementation of computers in teaching in general, only few students had no preference for most occasions (see Table 2 for details).

Insert Table 2 about here

For further analyses we decided to focus on students who had indicated a clear preference. There was only one situation where students did not prefer a human grader over a computer, namely, providing a model solution for weekly submitted assignments ($\chi^2 = 1.69$, $df = 1$, $p = .193$; all other $p < .001$ for a significant preference of the human grader).

Further, we analyzed the differences between the experimental conditions. There was a significant difference in the distribution of preferences for two occasions depending – again – only on the assumed source of the assessment (for the real source of the assessment all $p > .05$). The two occasions related to weekly submitted assignments, namely, the decision about passing them ($\chi^2 = 3.94$, $df = 1$, $p = .047$) and their assessment ($\chi^2 = 4.58$, $df = 1$, $p = .032$).

For both situations, in total, students preferred a human grader but this tendency was weakened within the group who assumed their text to be assessed by the software tool: Relatively more students preferred the computer and less the human grader. Thus, interestingly, those who assumed to be assessed by the software tool were more favorable regarding the computer.

Further, we analyzed what or who (i.e., a computer or a human grader) students thought would accomplish different aspects better (i.e., a speedy, objective, reliable and valid assessment). Again, only few students had no preference (see Table 3 for details).

Insert Table 3 about here

Thus, again, we decided to focus on the students who had indicated a clear preference for further analyses. For all aspects, students had a significant preference: for the computer when it comes to speed of an assessment ($\chi^2 = 264.67$, $df = 1$, $p < .001$) and objectivity ($\chi^2 = 73.25$, $df = 1$, $p < .001$), and for the human grader when it comes to reliability ($\chi^2 = 8.27$, $df = 1$, $p = .004$) and validity of an assessment ($\chi^2 = 74.98$, $df = 1$, $p < .001$). Further, we analyzed the differences between the experimental conditions. There was a significant difference in the distribution of preferences for one occasion depending – again – only on the assumed source of the assessment (for the real source of the assessment all $p > .05$). This aspect was speed of an assessment ($\chi^2 = 4.42$, $df = 1$, $p = .036$). In total, students thought that a computer was faster than a human grader and this tendency was stressed within the group who assumed their text to be assessed by the software tool: Relatively more students voted for the computer and less for the human grader. Thus, interestingly, those who assumed to be assessed by the software tool were more favorable for the computer again.

**Development of Learning-Related Characteristics**

To analyze effects on learning-related characteristics, we performed a mixed ANOVA with assumed and real source of assessment as between-subject factors and learning-related characteristics as repeated-measures (i.e., motivation – separately for values and competence beliefs, achievement aspirations for further texts and for the examination, and subjective learning). The main effect of time was significant ($F(5,271) = 14.36$, $p < .001$, $\eta^2 = .21$): There was a decline for all variables but the subjective learning ($F(1,275) = 0.30$, $p = .586$; all other main effects of time $p < .001$). No other effects were significant.

To rule out the possibility that these results were due to students' level of achievement only, we additionally controlled for the scores that students had received. The covariate was significant ($F(5,270) = 3.98$, $p = .002$, $\eta^2 = .07$), indicating that the level of achievement actually had an impact on students' learning-related characteristics: Students receiving higher scores had higher competence beliefs and achievement aspirations for their texts. Moreover, the main effect of time remained significant after controlling for level of achievement ($F(5,270) = 6.57$, $p < .001$, $\eta^2 = .11$). Contrasts revealed that this was due to students' competence beliefs and achievement aspirations still declining over time when controlling for their achievement level. Further, the interaction between the covariate and time became significant as well ($F(5,270) = 4.79$, $p < .001$, $\eta^2 = .09$), indicating that receiving a low score was associated with a disproportional decline in students' competence beliefs whereas students' competence beliefs remained or increased when receiving a higher score. However, all other effects remained non-significant (all $p > .05$).

## Discussion

Our study provides insight into students' acceptance of automatic assessments, students' opinion about the use of computers in teaching in general and the development of students' learning-related characteristics depending on the source of an assessment. Our results indicate that the real source of feedback was not important at all but the assumed

source of feedback was important with regard to the acceptance of the assessments: Students were more positive for the assumed teaching assistants' assessments. Thus, our first hypothesis was supported. With respect to the general perception of the use of computers in teaching, students preferred a human grader over a computer in all but one situation. Thus, our second hypothesis was supported as well. Interestingly, for two situations, the tendency to prefer the human grader was weakened for those who thought they had been assessed by the software tool. Further, students thought that computers could perform both a speedy and an objective assessment better than human graders but opted for the human graders when it came to the reliability and validity of an assessment. Students who thought that they had been assessed by the software tool were even more convinced about the advantage of computers in speed than the students who thought that they had been assessed by a teaching assistant. We conclude that there is some kind of acceptance problem of automatic assessments if students directly receive them but not necessarily if students are asked about the use of computers in general. However, a positive attitude towards AES might be essentially important with the up-coming Massive Open Online Courses which will not be manageable without tools like LSA that can (semi-)automatically score texts, give feedback or select appropriate new topics based on the learning material.

Another main result of our study was that there was no negative effect on students' development concerning learning-related characteristics (i.e., students' motivation, their achievement aspirations, and their perceived knowledge) depending on the assumed or real source of assessment. Thus, our third hypothesis was supported. All variables but the subjective learning showed a decline but this is a rather typical general development that we face in every course. However, the assumed or real source of assessment did not have a negative impact on the development; there was no different development for the groups. The synopsis of the results shows that an LSA-based assessment is not worse per se but it is

perceived as worse. We conclude that there is some kind of acceptance problem but there are no negative effects on important learning-related characteristics.

**Limitations**

Although the integration of this study into a real university course adds to the ecological validity and generalizability of the results, it should be noted that the sample included preservice teachers from one university only. Further, the study was a small and basic experiment on the acceptance of scores only. Generally, students rated the scores as medium acceptable. Looking at the single aspects, we found none rated very positive. This might be obvious because students were asked to judge nothing but a plain score. Hence, nothing can be concluded about detailed feedbacks that might be producible by more advanced software tools. Further, we provided feedback to students for one single assignment only and analyzed the short-term developments. For ethical reasons, it would not have been possible to lie to students about their assessments' source throughout the term and thus, to analyze any long-term developments depending on the real or assumed source of the assessments. Further, the results of our study might underestimate the acceptance of assessment of software tools in general because we used one tool only (i.e., LSA-based scores) and students were not told any details about the tool itself or its modus operandi. However, in contrast to the study by Lipnevich and Smith (2009a), we fully crossed the assumed and real source of assessment and thus, we could make some statements about both factors.

**Practical Implications**

We found that teaching assistants' assessments were favored by students and generally, a plain score cannot replace a detailed feedback. Certainly, we agree with Lipnevich and Smith (2009a) in that descriptive feedback is better than evaluative feedback. Further, it has been shown that detailed feedback might be especially important for more complex tasks (e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991) – and our assignments

are quite complex. Looking through the eyes of an instructor, it is useful to quickly know which students perform poorly/the worst in one's course, so that especially at-risk-students can receive feedback. Selecting certain students for feedback might be necessary when courses are large and capacity is constrained. Students who perform poorly might especially benefit from a detailed feedback because a poor essay might indicate that the author has not understood the material and needs special advice. There are some hints that automatic assessments can be useful in this regard (authors, submitted).

Because there is no negative effect on important learning-related characteristics, the use of automatically generated assessments should be considered as a possibility to assist human graders. There have been no negative effects but those in the head of the students. Thus, combining scores with comments might be a good way to combine the capacities of instructors and computer tools in an efficient manner (also see Stevenson & Phakiti, 2014): A score alone might be sufficient for the students who receive a satisfactorily high score but for the other students, a detailed comment by the instructor or teaching assistants might be necessary to ensure learning and motivation. Thus, we think that it might be best to use software tools to assist human graders. Software tools can be used to score texts in the background – as a possible second objective opinion (see students' preference of computers for speed and objectivity of assessments) – and to identify the students who are in need of an individual feedback and then, detailed feedbacks (answering the three feedback questions "Where am I going?", "How am I going?" and "Where to next?" (Hattie, 2009; Hattie & Timperley, 2007) can be given by teaching assistants.

References

Authors (2015). Manuscript submitted for publication.

Authors (2015).

Authors (2012).

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391–407.

Drabick, D. A. G., Weisberg, R., Paul, L., & Bubier, J. L. (2007). Keeping it short and sweet: Brief ungraded writing assignments facilitate learning. *Teaching of Psychology, 34*, 172–176.

Dunn, D. S., Saville, B. K., Baker, S. C., & Marek, P. (2013). Evidence-based teaching: Tools and techniques that promote learning in the psychology classroom. *Australian Journal of Psychology, 65,* 5–13.

Ericsson, P. F., & Haswell, R. H. (Eds.) (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Gierl, M. J., Latifi, S., Lai, H., Boulais, A.-P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education, 48*, 950–962.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Lai, Y.-H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology, 41(3)*, 432–454.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis, & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective* (pp. 87–112). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse [Automatic scoring of constructed-response items with latent semantic analysis]. *Diagnostica, 53*, 155–165.

Lipnevich, A. A., & Smith, J. K. (2009a). Effects of Differential Feedback on Students' Examination Performance. *Journal of Experimental Psychology: Applied, 15*, 319–333.

Lipnevich, A. A., & Smith, J. K. (2009b). "I really need feedback to learn:" students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability, 21*, 347–367.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Nevid, J. S., Pastva, A., & McClelland, N. (2012). Writing-to-learn assignments in introductory psychology: Is there a learning benefit? *Teaching of Psychology, 39*, 272–275.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Hillsdale, NJ: Lawrence Erlbaum Associates.

Page, E. B, & Petersen, N. S. (1995). The computer moves into essay grading: updating the ancient test. *Phi Delta Kappan, 48*, 238–43.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346). New York, NY: Routledge.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81.

Table 1

*Descriptive Statistics for the Assessments by the Teaching Assistants and the Software Tool as well as for the Learning-related Variables*

| | ***N*** | ***Min.*** | ***Max.*** | ***M*** | ***SD*** |
|---|---|---|---|---|---|
| Teaching assistants | 300 | 3.00 | 10.00 | 6.62 | 1.38 |
| LSA | 300 | 4.00 | 9.50 | 6.78 | 0.97 |
| Values (t1) | 298 | 2.00 | 5.00 | 4.20 | 0.62 |
| Values (t2) | 299 | 1.00 | 5.00 | 3.97 | 0.72 |
| Competence Beliefs (t1) | 297 | 1.67 | 5.00 | 3.66 | 0.65 |
| Competence Beliefs (t2) | 299 | 1.33 | 5.00 | 3.45 | 0.57 |
| Achievement Aspirations Text (t1) | 291 | 1.00 | 10.00 | 7.29 | 1.23 |
| Achievement Aspirations Text (t2) | 291 | 1.00 | 10.00 | 6.95 | 1.22 |
| Achievement Aspirations Exam (t1) | 299 | 1.00 | 3.00 | 1.57 | 0.58 |
| Achievement Aspirations Exam (t2) | 299 | 1.00 | 3.00 | 1.72 | 0.61 |
| Subjective Learning (t1) | 299 | 1.00 | 4.00 | 2.31 | 0.71 |
| Subjective Learning (t2) | 298 | 1.00 | 5.00 | 2.34 | 0.75 |

Table 2

*Total Numbers and Percentages of the Preferences for a Computer or a Human Grader for Different Situations in Teaching in General*

| | Weekly submitted assignments | | | | Examinations | | |
|---|---|---|---|---|---|---|---|
| Preference | Decide about passing | Assess | Give feedback | Provide a model solution | Decide about passing | Assess when ungraded | Assess when graded |
| Computer | 99 (33,0%) | 29 (9,7%) | 13 (4,3%) | 116 (38,8%) | 50 (16,7%) | 67 (22,4%) | 17 (5,7%) |
| Human Grader | 160 (53,3%) | 243 (81,0%) | 272 (90,7%) | 97 (32,4%) | 226 (75,3%) | 177 (59,2%) | 268 (89,6%) |
| No preference | 41 (13,7%) | 28 (9,3%) | 15 (5,0%) | 86 (28,8%) | 24 (8,0%) | 55 (18,4%) | 14 (4,7%) |
| Total | 300 (100%) | 300 (100%) | 300 (100%) | 299 (100%) | 300 (100%) | 299 (100%) | 299 (100%) |

Table 3

*Total Numbers and Percentages of Who/What Would Accomplish Different Aspects Better*

| | Speed | Objectivity | Reliability | Validity |
|---|---|---|---|---|
| Preference | | | | |
| Computer | 285 (95,0%) | 199 (66,3%) | 100 (33,3%) | 58 (19,3%) |
| Human Grader | 7 (2,3%) | 61 (20,3%) | 145 (48,3%) | 196 (65,3%) |
| No preference | 8 (2,7%) | 40 (13,3%) | 55 (18,3%) | 46 (15,3%) |
| Total | 300 (100%) | 300 (100%) | 300 (100%) | 300 (100%) |