Automatic essay assessment: Effects on students’ acceptance and on learning-related characteristics

Eva Seifried¹, Wolfgang Lenhard², and Birgit Spinath¹

¹Heidelberg University, Heidelberg, Germany
²University of Würzburg, Würzburg, Germany

When capacity constraints hinder university instructors’ ability to give feedback, software tools might provide a remedy. We analyzed students’ acceptance of automatic assessments and the development of learning-related characteristics such as motivation, achievement aspirations, and subjective learning. We randomly assigned university students to four groups that differed with regard to the real and assumed source of assessment of students’ texts (i.e., teaching assistant or software tool). Data from \( N = 300 \) 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 they were being assessed by the software tool. Nevertheless, students saw some general merits to assessment by computers, and the development of learning-related characteristics was not affected by the real or assumed source of assessment. Thus, combining feedback from software tools and human graders seems to be a feasible way to expand feedback capacities in higher education.

Keywords: automatic essay scoring, acceptance, motivation, achievement aspirations, subjective learning, higher education

Essay assignments are widely used at universities: ‘Writing-to-learn’ has been shown to be effective in improving learning (e.g., Nevid, Pastva, & McClelland, 2012) and has been identified as an evidence-based teaching technique (Dunn, Saville, Baker, & Marek, 2013). Although even ungraded writing assignments can foster learning (e.g., Drabick, Weisberg, Paul, & Bubier, 2007; Nevid et al., 2012), receiving feedback seems desirable to help students monitor their learning (see Hattie, 2009). However, in large classes (i.e., a hundred or more participants), it is not possible for a sole instructor to

Corresponding author: 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 and Fabian Grünig for his work on βASSIST (i.e., the follow-up system of ASSIST).
read all assignments. Progress in the sector of automatic essay scoring (AES) has made it possible for students to receive feedback on their performance even in large courses. Since the 1960s, there have been attempts to score essays automatically by computers (e.g., Page, 1966), and recent technologies can be used for both summative and formative purposes, for high-stakes and low-stakes assessments (Shermis & Hamner, 2013). Despite evidence of the validity of AES (see Shermis & Burstein, 2003), there is little research on the acceptance of AES by students, especially at the 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 further analysis is needed on 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. If AES were accepted and had no negative effects on learning-related characteristics, AES might make a desirable teaching-learning format possible even in large lectures: having students write essays and giving them feedback.

Examining the acceptance of AES

There seem to be some concerns regarding AES not only within the scientific community (e.g., Ericsson & Haswell, 2006), but also among those who are being assessed. 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 already signed this petition. According to Gierl, Latifi, Lai, Boulais, and De Champlain (2014), AES “has been described as ‘robo-scoring’, ‘robo-reading’, ‘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, see also Stevenson & Phakiti, 2014), but their extent and impact need to be further analyzed.

Although not focusing on the acceptance of AES, some studies have reported interesting yet mixed results on this front (Lai, 2010; Lenhard, Baier, Hoffmann, & Schneider, 2007; Lipnevich & Smith, 2009a, 2009b). Lai (2010) found that English as a foreign language learners preferred to receive feedback from peers over feedback from a computer tool. Lenhard et al. (2007) found that students perceived computer-generated feedback as helpful but as not really reflecting the quality of their texts. Lipnevich and Smith (2009a) found the perceived source of feedback (i.e., a computer or the instructor) had little impact, but students who assumed they had received feedback from 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 being more 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 been fairer and more unbiased than the professor. Some students also felt relieved that it was not the professor who 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 they had received their feedback from the instructor, although their 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.

To our knowledge, there is no study that has directly dealt with the acceptance of AES and thereby segregated the real and assumed source of an assessment. Thus, it is not clear whether it is being assessed or believing that one is being assessed by a software tool that explains the results mentioned above. Further, it is not clear whether university students accept computers in teaching in general and whether there is any effect on learning-related characteristics when they only assume they are being assessed compared to actually being assessed by a software tool. 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 in their motivation, their aspirations and their subjective learning (i.e., the personal perception of how much one has learnt). According to expectancy-value theory (Wigfield & Eccles, 2000), motivation can be divided into two constructs: ability beliefs (i.e., the belief that one can do well in something) and three components of subjective values (i.e., seeing the usefulness and importance of something, and being interested in it). If students do not accept AES, this might negatively affect these constructs and hence their motivation with regard to a course and its content. To investigate the development of these learning-related characteristics, it is necessary to collect data before and after students are or assume that they are assessed by either a human grader or a software tool.

The present study

This study sought to explore whether software-generated assessments are accepted by students, and whether being assessed by a software tool has any effect on students’ learning-related characteristics. We wanted to know whether students accepted the application of a software tool to assess their texts in a university course 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. Thus, in general, we were interested in the effects of AES on students’ acceptance and on learning-related characteristics.

Previous studies revealed some acceptance problems for AES (e.g., Lenhard et al., 2007; Lipnevich & Smith, 2009a, 2009b). However, these studies could not segregate whether it was the real or the assumed source of assessment that resulted in lower acceptance. We suspect that it is only the assumed and not the real source of assessment that leads to lower acceptance. As to using computers in teaching in general, some previous studies have shown that students seem to prefer humans over computers (see e.g., Lai, 2010); thus, we expected this finding, too. Furthermore, despite the evidence that AES has some acceptance problems, mixed results have been found when students believed they were assessed by either source (see e.g., Lipnevich & Smith, 2009b); therefore, and due to the assumption that the software-based assessments are not worse than a human grader’s assessments (see validity of AES; e.g., Shermis & Burstein, 2003), we did not expect any negative effects on students’ learning-related characteristics. Thus, 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. The (real or assumed) source of the assessment will not negatively influence the development of learning-related characteristics, that is, neither the real nor the assumed source of the assessment are expected to have a negative effect on student outcomes such as 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 regarding the implementation of computers in teaching in general. Further, we applied a 2x2 design with the real and assumed sources of assessment fully crossed (see Table 1), thus following the recommendation of Stevenson and Phakiti (2014) that the kinds of feedback should be comparable so that analyses about whether it is really the source of feedback that matters can be conducted.
Table 1
Combination of Real and Assumed Source of Feedback Within the Four Groups Based on The Experimental Condition

<table>
<thead>
<tr>
<th>Assumed Source of Assessment</th>
<th>Real Source of Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software Tool</strong></td>
<td><strong>Teaching Assistant</strong></td>
</tr>
<tr>
<td><strong>Score:</strong></td>
<td><strong>Score:</strong></td>
</tr>
<tr>
<td>LSA-based software tool</td>
<td>Teaching assistant</td>
</tr>
<tr>
<td>Text:</td>
<td>Text:</td>
</tr>
<tr>
<td>“Your assignment was assessed with the help of the software tool.”</td>
<td>“Your assignment was assessed with the help of the software tool.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Teaching Assistant</strong></th>
<th><strong>Teaching Assistant</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Score:</strong></td>
<td><strong>Score:</strong></td>
</tr>
<tr>
<td>LSA-based software tool</td>
<td>Teaching assistant</td>
</tr>
<tr>
<td>Text:</td>
<td>Text:</td>
</tr>
<tr>
<td>“Your assignment was assessed by one of the teaching assistants.”</td>
<td>“Your assignment was assessed by one of the teaching assistants.”</td>
</tr>
</tbody>
</table>

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; for details see below) were available for $N = 300$ students, with 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 completed the course in a previous semester received training on providing feedback on the assignments. All teaching assistants were female and were studying psychology as a major (six were in the fourth semester/second year, and six were in the sixth semester/third year of their bachelor’s studies, two were in the first year of their master’s studies); their ages ranged from 21 to 26 years, with a mean age of 22.36 years ($SD = 1.69$), and their time as a teaching assistant in our department ranged from 1 to 6 semesters, with a mean of 2.43 semesters ($SD = 1.50$). Every assignment was assessed with the help of a specimen model solution and a scoring scheme by one of the teaching assistants. Texts were assessed on a 10-point scale with a 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; $N = 367$ of the $N = 405$ students submitted an assignment for the session relevant for the current study. ASSIST uses Latent Semantic Analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007) to perform specific tasks, for example, detecting plagiarism or scoring 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, see also Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Martin & Berry, 2007). Several authors have
successfully used LSA for automatic essay assessment (e.g., Landauer, Laham, & Foltz, 2003; Seifried, Lenhard, Baier, & Spinath, 2012).

On the basis of our positive results on the evaluation of LSA-based scores (Seifried et al., 2012) and approaches for identifying cheaters (Seifried, Lenhard, & Spinath, 2015) or poorly performing students with the help of LSA-based scores (Seifried, Lenhard, & Spinath, 2016), 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 prior studies; it was originally based on 41 psychology textbooks and extended with material on the specific topics of our lecture. Prior to the construction of the space, the original text corpus was split into smaller units and analyzed according to word frequency and the occurrence of individual words. The resulting frequency or term-by-document matrix is made up of terms (rows) and documents (columns) as well as the frequency of each term in each document (cells). Next, unnecessary words (i.e., words that do not carry specific information or appear only very rarely) were excluded. Subsequent steps included the application of weighting functions to (de)emphasize (un)important words, a singular value decomposition as well as a dimension reduction (to 300 dimensions) to reveal the (gist of the) meaning of the texts (for references about the procedure in general and further details about the contents of the specific semantic space, see Seifried et al., 2012 and Seifried et al., 2016). To evaluate their content, the essays were represented as vectors in the resulting 300-dimensional semantic space and their proximity to a comparison text (i.e., the model solution used by the human graders) was assessed by means of the cosine between them. The essays were ranked according to their proximity to this “gold standard”. 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 10th and 90th percentile were evaluated by human graders to adjust the rubric for the scores of the remaining essays via linear regression. Thus, for LSA, it took just a few seconds to score the essays (for details on the procedure, see also Seifried et al., 2016). 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 for their opinion about this assessment in another survey; N = 341 of the students who had submitted an assignment also took both surveys. What students did not know was that the experimental design included randomly assigning students to four groups that differed regarding the real and the assumed source of assessment (fully crossed experimental design). This means that only half of the students who thought their feedback came from a software tool actually received software feedback, with the other half receiving feedback from a teaching assistant (the same was true for students who thought they received feedback from a teaching assistant; N = 315 of the students who had submitted an assignment and taken both surveys passed a corresponding manipulation check; see below). To make these conditions especially credible, students who had been told that their feedback was generated by a software tool received feedback within one day of the submission deadline, whereas those who thought that their 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, feedback could not be given as intended (i.e., for \( N = 15 \) of the students who had fulfilled all previous criteria): 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. Thus, in the end, we had full data sets from \( N = 300 \) students (for details on these students, see above).

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 (see above).

To analyze students’ acceptance of computer-generated scores (Hypothesis 1), they were asked to rate on a 5-point scale ranging from (1) absolutely not to (5) very the degree to which 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 in different contexts, regarding both weekly submitted assignments and examinations. The applications regarding weekly submitted assignments included the following: (a) pass / no pass decisions on weekly submitted assignments, (b) assessments of weekly submitted assignments, (c) feedback on weekly submitted assignments, and (d) providing a model solution for weekly submitted assignments. The situations regarding examinations included the following: (a) pass / no pass decisions on an examination, (b) assessing an ungraded examination, and (c) assessing a graded examination. Additionally, because students had already learnt about criteria of scientific measurements within the course, students were asked about the relative advantages and disadvantages of assessments by human graders or computers with regard to these aspects (i.e., objectivity, reliability, validity, and speed). Again, they were asked to indicate what or who would be better with regard to 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 to me”; competence beliefs: “I do well in educational psychology”). Students indicated agreement on a 5-point scale ranging from (1) completely disagree to (5) completely agree. Students also rated their achievement aspirations, both for the weekly assignments by indicating the number of points that they wanted to achieve in 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. In addition, they were asked to rate their subjective learning (i.e., answer the question “How would you assess your current knowledge in educational psychology?”) on a 5-point scale with (1) = low and (5) = high.
Results

Descriptive statistics

Table 2 shows descriptive statistics for the assessments by the teaching assistants and LSA as well as descriptive statistics for the learning-related variables. The scores were highly correlated ($r = .62$, $p < .001$).

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

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 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 they had been assessed by a teaching assistant rather 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 score on 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 a few students had no preference in most occasions (see Table 3 for details).

Table 3

<table>
<thead>
<tr>
<th>Preference</th>
<th>Weekly submitted assignments</th>
<th>Examinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Decisions about passing</td>
<td>Assessment</td>
</tr>
<tr>
<td>Computer</td>
<td>99 (33.0%)</td>
<td>29 (9.7%)</td>
</tr>
<tr>
<td>Human Grader</td>
<td>160 (53.3%)</td>
<td>243 (81.0%)</td>
</tr>
<tr>
<td>No preference</td>
<td>41 (13.7%)</td>
<td>28 (9.3%)</td>
</tr>
<tr>
<td>Total</td>
<td>300 (100%)</td>
<td>300 (100%)</td>
</tr>
</tbody>
</table>

For further analyses, we decided to focus on students who 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 for the human grader).

Furthermore, 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 whether a student had passed them ($\chi^2 = 3.94, df = 1, p = .047$) and their assessment ($\chi^2 = 4.58, df = 1, p = .032$). For both situations, students generally preferred a human grader, but this tendency was weakened within the group who assumed that their text had been assessed by the software tool: Relatively more students preferred the computer and fewer preferred the human grader. Thus, interestingly, those who assumed that they had been assessed by the software tool had more favorable views regarding the computer.

In addition, 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 a few students had no preference (see Table 4 for details).
Thus, again, we decided to focus on the students who indicated a clear preference in further analyses. For all aspects, students had a significant preference: for the computer when it comes to the 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 the reliability ($\chi^2 = 8.27, df = 1, p = .004$) and validity of an assessment ($\chi^2 = 74.98, df = 1, p < .001$). Furthermore, 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 in the group who assumed that their text had been assessed by the software tool: Relatively more students voted for the computer and fewer for the human grader. Thus, interestingly, those who assumed they had been assessed by the software tool had again more favorable views with regard to the computer.

**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 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 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. In addition, 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 differences in 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 towards the assumed teaching assistants’ assessments. Thus, our first hypothesis was supported. With respect to general perceptions on 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 among those who thought they had been assessed by the software tool. Furthermore, 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 with regard to 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 for automatic assessments when students directly receive them but not necessarily when students are asked about the use of computers in general. However, a positive attitude towards AES might be essentially important in light of the up-coming expansion of Massive Open Online Courses, which will not be manageable without tools like LSA that can (semi-)automatically score texts, give feedback and select appropriate new topics based on the learning material.

Another main result of our study was that the assumed or real source of assessment had no negative effect on students’ development with regard to learning-related characteristics (i.e., students’ motivation, achievement aspirations, and perceived knowledge). Thus, our third hypothesis was supported. All variables but subjective learning showed a decline, but this is a rather typical general development that we face in every course. More importantly, the assumed or real source of assessment did not have a negative impact on this development; there was no difference in development among 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. In addition, the study was a small and basic experiment on the acceptance of scores only. Generally, students rated the scores as of medium acceptability. Looking at individual aspects, we found none were rated very positively. This might be because students were asked to evaluate nothing but a plain score. Hence, nothing can be concluded about more detailed feedback that might be produced by more advanced software tools. Furthermore, we provided students with feedback on a single assignment only and analyzed short-term developments. For ethical reasons, it would not have been possible to lie to students about their assessment source throughout the term and thus analyze long-term developments depending on the real or assumed source of the assessments. In addition, 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 could make some statements about both factors.

Practical Implications

We found that teaching assistants’ assessments were favored by students. Generally speaking, a plain score cannot replace detailed feedback. Certainly, we agree with Lipnevich and Smith (2009a) that descriptive feedback is better than evaluative feedback. Furthermore, 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. From the perspective of an instructor, it is useful to quickly know which students are performing poorly/worst in one’s course, so that particularly 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 detailed feedback because a poor essay might indicate that the author has not understood the material and needs more specific advice. There is some evidence that automatic assessments can be useful in this regard (Seifried et al., 2016).

Because there was 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 were no negative effects except those in students’ heads. 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 students who receive a satisfactorily high score, but for other students, detailed comments from 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 for computers for speed and objectivity of assessments) – and to identify the students who are in need of individual feedback. Then, detailed feedback (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


RECEIVED 27.04.2016.
REVISION RECEIVED 11.11.2016.
ACCEPTED 15.11.2016.