Informatics in Education, 2011, Vol. 10, No. 1, 47–64

Six Key Topics for Automated Assessment Utilisation and Acceptance

Torsten REINERS[1,2], Carl DREHER[2], Heinz DREHER[2]
[1]*Institute of Information Systems, University of Hamburg*
Hamburg, Germany
[2]*Curtin Business School, Curtin University*
Perth, Western Australia
e-mail: reiners@econ.uni-hamburg.de, c.dreher@curtin.edu.au, h.dreher@curtin.edu.au

Received: January 2011

Abstract. Automated assessment technologies have been used in education for decades (e.g., computerised multiple choice tests). In contrast, Automated Essay Grading (AEG) technologies: have existed for decades; are 'good in theory' (e.g., as accurate as humans, temporally and financially efficient, and can enhance formative feedback), and yet; are ostensibly used comparatively infrequently in Australian universities. To empirically examine these experiential observations we conducted a national survey to explore the use of automated assessment in Australian universities and examine why adoption of AEG is limited. Quantitative and qualitative data were collected in an online survey from a sample of 265 staff and students from 5 Australian universities. The type of assessment used by the greatest proportion of respondents was essays/reports (82.6%), however very few respondents had used AEG (3.8%). Recommendations are made regarding methods to promote technology utilisation, including the use of innovative dissemination channels such as 3D Virtual Worlds.
Keywords: automated assessment, automated essay grading, technology acceptance, benefits realisation, mixed methods research.

1. Introduction

One of the core research focuses in Information Systems is the full-automation of organizations by providing information and communication systems to support the gathering, processing, storing, distribution, and use of information (O'Brien and Marakas, 2008). While stand-alone systems dominated the infrastructure of most organizations until a few years ago (performing merely technical support to handle structured documents), we experienced a formidable shift during the Web 2.0 era towards social networks, cloud computing, web-based services, and distributed storage. Here, the paradigm of *everyone is a producer* enhanced collaboration and communication in a *flat world*, but, again, with the user as the main (and generally only) intelligent component in the system. Nowadays, we are experiencing the next shift, this time towards Web 3.0 (also described as Semantic Web), where software agents become intelligent, *aware* of unstructured content, and fully responsible participants in (business) processes (Murugesan, 2009).

Web 3.0 represents multiple subjects of importance (e.g., ontologies, reasoning, semantic analysis, and conceptualization). The present authors' research is fed by the interest in grasping the meaning of documents allowing software (agents) to understand, process, and compare documents without the need of external interferences. The range of applying such technology is broad, highly interdisciplinary and includes, for example, machine translation (improvement and verification of translated documents), plagiarism checking (exposing rephrased documents or copied concepts rather than word-by-word copies), intelligent information gathering based on vague specifications (intelligent and autonomous search bots), and automated grading of assessment (in educational institutions or advanced training).

Based on a sound research methodology and first *proof of concept* implementations in our research group (Dreher *et al.*, 2011; Dreher, 2007; Reiter *et al.*, 2010; Williams, 2006; Williams and Dreher, 2005), we discuss a highly important field of application (enhancing educational systems and advanced training at lower cost and higher quality) by demonstrating the advances and prospects based on a national survey in Australia. This study was motivated by an ostensible discrepancy we have observed between the sophisticated automated assessment technologies available and a lack of utilisation, acceptance, and subsequent benefit, in particular regarding Automated Essay Grading (AEG). There is an apparent discrepancy between theory and practice: AEG is good in terms of pedagogical and management theory (the technology can work as accurately as human markers, it can save time and money, and can enhance formative feedback), but it is not being put into common practice.

Assessment is crucial for all participants in the educational system, albeit from different perspectives. *Students* conduct assessments to gain credits, and perhaps less frequently, to receive qualitative feedback from lecturers in a formative assessment process. When *educators* need to measure students' outcomes of learning a process, summative assessment is used (Black and Wiliam, 1998), which also provides the educational *administrator/manager* with operational and performance data. Aspects like frequency, type, and format of assessment depend on the kind of learning being appraised, the individual preferences of educators and, especially, the applied pedagogical model. However the application of the pedagogical model is often restricted by pragmatic realities, including: an increased workload for educators when performing high quality formative assessment; economic pressure (for administrators/managers) in a competitive market and; dissatisfaction (for students) with poor quantity-quality ratios where assessments are evaluated on simplistic levels. The true perfection in mastering all factors relevant to successful educational practice for all roles is finding the balance point representing the pareto optimum of pedagogical assessment with regard to students' learning outcomes and universities' resources (e.g., quality control regarding both formative and summative assessment, educators' skills and effort, time, and costs).

To outline the remainder of this paper, the following section discusses both extant and emergent automated assessment technologies, and subsequent sections present this study's rationale and method. This national survey of staff and students at Australian universities explored the current state of assessment practices, including: respondents'

use and perceptions of various human and automated assessment approaches, and participants' desires for automated assessment technologies. The paper is concluded with an outlook on future research, including: extending future experiments and taking a sneak peek into novel methods to demonstrate, apply, and promote automated assessment technologies. In addition, we discuss the importance of our findings with respect to the current status quo at Australian universities and how the results can be used to enhance development and integration of modern technology in learning and education.

2. Advanced assessment in education and information science

Assessment and its automation can be used wisely or detrimentally (see Black and Wiliam, 1998, for a review). If used wisely the automation of assessment offers a number of benefits over manual assessment in the provision of formative assessment, including self-assessment and immediate feedback. Discussed below are the extant and emerging automated assessment technologies and their roles in education.

Technological advances in automated assessment carry the potential to improve the benefits for all stakeholders in the assessment process. Students can receive immediate and objective feedback, educators can focus on teaching and giving formative feedback, and administration/management can be afforded lower costs – e.g., more accurate planning by cost per marking and less personnel for grading – and increased esteem in society (Dreher *et al.*, 2011). Automated assessment systems, heretofore, operate on a recall of memorized *knowledge* without checking understanding of the taxonomy of educational objectives (Bloom, 1956). However emerging technologies intend to support interpretation of short answer and essay type questions by automating grading and annotation of assignments with formative qualitative feedback. Such approaches would support *interpretation* and *problem-solving levels* (Krathwohl, 2002).

Computerised assessment of fixed-choice response formats (e.g., Multiple-Choice or M-C) has been standard practice at universities for many years (Haladyna *et al.*, 2002). More recently, plagiarism assessment (text-string checking like Turnitin) has gained popularity (Rees and Emerson, 2009). The automation of these approaches presents certain benefits to students (e.g., self-assessment to monitor learning) and staff (e.g., less or no manual marking). However M-C tests have been criticised for assessing lower-order forms of learning, and plagiarism assessment does not assess learning or application of concepts/knowledge. In contrast, essays assess higher-order learning (Nicol and Macfarlane-Dick, 2006). However they are labour intensive for markers, which reduces the rate and/or amount of formative feedback provided to students, and makes them impractical in large courses. The scoring of essays using computers offers advantages, such as enhanced formative feedback (Williams and Dreher, 2005). Automated essay scoring/grading was first developed in 1960s by Page (2003). Subsequently a variety of approaches have been developed for AEG, including E-rater Scoring Engine, Intelligent Essay Assessor, Intellimetric, and text categorisation (Dikli, 2006; Shermis and Burstein, 2003).

AEG software uses various techniques to compare students' essays with a model solution. Here, we briefly introduce MarkIT, which uses normalized word vectors to derive a conceptual footprint of essays. Normalization in this context refers to the process where words (and their frequency) from the essay are mapped to their corresponding root word in a thesaurus. The created footprint can be compared to other sources, such as a model solution for grading or other documents for plagiarism checking on a semantic level; see Williams (2006) and Williams and Dreher (2005). Note that attributes like spelling, grammar, or style are also considered for the result.

3. Rationale

"No systems, no impact" (Nievergelt, 1994, p. 299). Building a sound theory and methodology within the ivory tower of universities might enhance the research credibility, but can also broaden the gap between theory and practice if systems development lags behind theory or if systems are not accepted by the stakeholder. With automated assessment, and especially AEG, we have experienced at Curtin University (and suspect the same holds throughout Australian universities) a certain scepticism about the technology. This may be due to a prevalent view that regards human markers as being superior to computers at the tasks of understanding content and making comparisons between student essays and a model solution. When the academic community does not adopt state-of-the-art assessment technology, it forgoes the subsequent benefits, including: improved learning outcomes for students, job satisfaction for staff, and quality assurance and financial benefits for universities.

The pragmatic reality that universities are run as businesses leads to certain factors which challenge educators, including that large classes are common, that workloads are increasing, and the importance of quality assurance (i.e., quality management) of education and assessment. For automated assessment technologies to be utilized, a change is required in the academic culture surrounding assessment practices. Indeed, automated assessment has the power to beneficially change the socio-technological process of assessment in educational organizations. However, currently such change is ostensibly resisted.

Specific aims of this research were to: (1) survey the human and automated assessment practices in Australian universities; specifically, the educational roles of users (e.g., students, educators, management, IT-support, HR-administration), assessment types used, and mode of marking – human vs. computer); (2) determine preferences-for-use of assessment types; (3) explore: the pros and cons of automated assessment and AEG; the desired elements of automated assessment technologies by staff who have used them, and; the barriers-to-use of automated assessment technologies by staff who have not used them.

In examining the acceptance/adoption of technology via our survey, we did not use extant measures or specific constructs (e.g., those associated with the Technology Acceptance Model, TAM; Venkatesh *et al.*, 2007) because we decided to use an approach that prioritised the inductive principle (operationalized through both qualitative and quantitative questions). We wished to prioritise respondents' subjective impressions and did not

want to constrain our measurement to *a priori* constructs. Hence the survey used many open-ended questions. Where closed-ended questions were used (e.g., those suggesting options to choose from), the last option was labelled *other* and a text-box was provided for open-ended responses.

4. Method

Described below are the measure, procedure, participants, and analysis that comprise this study. This paper continues the discussion of survey results, of which other parts are already reported in Dreher *et al.* (2011). Both, the method and participants' demographics are reported (though worded uniquely) in both papers. Note that the overlap is limited because both publications focus on different subjects.

4.1. *Measure, Procedure, Participants*

We used an anonymous web-based survey to collect the data for this study. The survey tool (EFS Survey) allowed the application of content filters such that each respondent was presented with only relevant questions based on previous responses or their educational experiences (e.g., their educational role and prior use of automated assessment). We developed the content of the survey based on our academic knowledge of and practical experience with educational assessment.

We contacted Australian universities ($N = 40$) via email and obtained organisational consent from five universities (yielding a 12.5% response rate). The participating universities are located in three states (Victoria, Queensland, and Western Australia) being diversely situated across the continent. While consent was given to contact students at only three universities, all five universities gave permission to contact staff members.

The methods for contacting participants were chosen in collaboration with each institution and differed by university and educational role (i.e., staff vs. student). Students (with a minimum age of 14 years) were contacted by a student website ($n = 1$ university), an email distribution list ($n = 1$), and an unspecified method ($n = 1$). Staff were contacted by notices on email distribution lists ($n = 4$ universities) and an online newsletter ($n = 1$). Individuals' consent was indicated by responding to the survey. The study was approved by the Curtin University Human Research Ethics Committee. A sample of 265 (57.5%) out of a pool of 461 individuals who began the online survey, completed the survey. The sample ($N = 265$) comprised 60.0% ($n = 159$) females. Demographic variables are presented in Fig. 1 (by frequency and percentage with modal categories marked with bold lines). Further data was collected, but is not shown here, including Country of Birth (27.9% non-Australian), Highest Level of Education (49.9% have a Masters or PhD, 26.4% do not have an academic degree), Country of Education (16.2% not in Australia); see also Dreher *et al.* (2011).

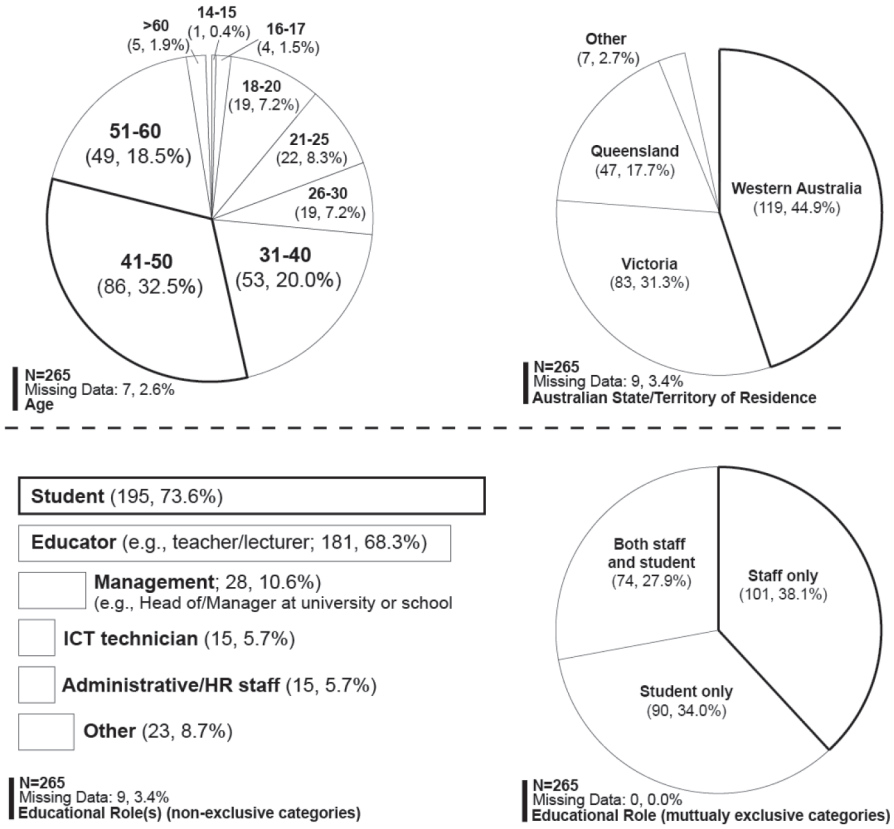


Fig. 1. Respondents' demographic variables by frequency and percentage.

4.2. *Analysis*

A mixed-method approach was employed in the design of this study (Charmaz, 2000; Teddlie and Tashakkori, 2009). Consequently the survey questions comprised both fixed-response and open-response formats yielding quantitative and qualitative data respectively. The qualitative data were analysed using a constructivist grounded theory method, of which some results are reported in Dreher *et al.* (2011). Two main processes were used in the present qualitative analysis: coding and categorising themes. First, coding was used to create potential themes from the open-ended responses. Then codes were assigned to specific themes line-by-line. Themes were adapted throughout the analysis (using such processes as typification, revision, and contradistinction) based on the response as well as respondents' demographics, educational role, and prior use of technology. In the second process, an analytic framework to explain the data was developed by categorising themes: firstly, using focused/selective coding and, secondly, by specifying categories of themes. The quantitative data were summarised using descriptive statistics (presenting frequency charts and highlighting modes). A selection of these results is reported here.

5. Results

The results are summarised here according to the following six topics: (1) use of assessment in general and (2) automated assessment; (3) its usefulness ratings (4) preference-for-use by type of automated assessment; (5) barriers to use, and; (6) desired elements of automated assessment. The topics in this paper depict one unique part of the complete survey with respect to use and benefits of automated assessment; see also Dreher *et al.* (2011).

5.1. *Survey Topic 1: Use of Assessment in General*

The survey began with a series of questions about assessment practices in general (i.e., we did not distinguish between automated or human assessment). Respondents were asked to specify the types of assessment that they had used (for staff) or experienced (for students), and subsequently were only asked questions about these types of assessment, and about the frequency with which they had used/experienced each type of assessment (on a 4-point ordinal scale labelled *rarely*, *sometimes*, *frequently*, and *most of the time*). Figure 2 shows: the number (and percentage) of respondents in the total sample who had used each type of assessment, and; the modal frequency-of-use for each type of assessment.

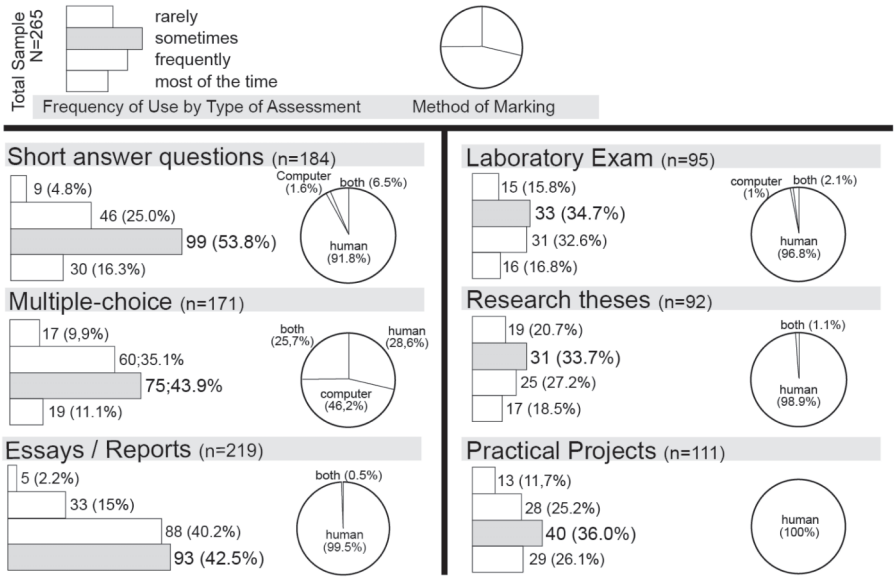


Fig. 2. Number and percentage of respondents by type of assessment used, frequency of use, and method of marking.

Note: n = number of respondents who had used each type of assessment, and (%) = percentage of the respondents who had used each type of automated assessment.

These data indicate that *essays or reports* are the type of assessment that were used by the largest proportion of the sample, and that they were used *most of the time*.

Respondents were asked what methods were used to mark each type of assessment that they had experienced previously (i.e., human, computer, or both; see Fig. 2). Figure 2 also shows that *human grading* is the modal method of marking for each type of assessment with the exception of M-C questions, which are most often marked by computer (46.2%). However, a relatively large number of respondents had experienced M-C questions that were marked by human markers only (28.6%) or both computer and human markers (25.7%).

5.2. *Survey Topic 2: Use of Automated Assessment*

Respondents were asked what types of automated assessment they had used before. Of the 265 respondents, 60 (22.6%) indicated that they had not used automated assessment before. The types of automated assessment that were most commonly used before were M-C questions (scored by computers) ($n = 186;\ 70.2\%$) and plagiarism checking ($n = 125;\ 47.2\%$). Less frequently used types of automated assessment include: marking computer programming/code ($n = 16;\ 6.0\%$); "other" types of automated assessment not listed in the survey ($n = 16;\ 6.0\%$); AEG ($n = 10;\ 3.8\%$), and marking mathematical proofs ($n = 7;\ 2.6\%$). Note that many respondents indicated that they had used multiple types of automated assessment before, therefore the number of respondents reported above sums to > 265 and the percentage of the total sample ($N = 265$) sums to $> 100\%$.

The role(s) in which respondents used automated assessment were reported (as distinct from general educational roles discussed in the method section). The modal role for use of automated assessment was that of a student. The specific frequencies and percentages (of the total sample, $N = 265$) are as follows: student role ($n = 111;\ 41.9\%$); educator ($n = 98;\ 37.0\%$); marker ($n = 35;\ 13.2\%$); information and communication technology technician ($n = 3;\ 1.1\%$); administrative assistant ($n = 4;\ 1.5\%$); managerial role ($n = 6;\ 2.3\%$); other ($n = 6;\ 2.3\%$); have not used automated assessment before ($n = 60;\ 22.6\%$). Note that many respondents indicated using multiple types of assessment before, therefore the frequency of respondents' sums to > 265 and the percentages sum to > 100%. In summary, these results illustrate that a high proportion of respondents were engaged in student roles (41.9%), educator roles (37.0%), marker roles (13.2%), or had not used automated assessment before (22.6%). Due to the fact that many respondents had multiple roles for using automated assessment, it is useful to examine these roles further. Viewing the data from this perspective, one notices that a high proportion of the sample indicated they acted in staff roles only ($n = 94;\ 35.5\%$) or student roles only ($n = 91;\ 34.3\%$) while using automated assessment. A relatively small proportion had used automated assessment in both student and staff roles ($n = 20;\ 7.5\%$).

The educational contexts in which automated assessment was used are reported below according to the frequency of respondents and percentage of the total sample ($N = 265$) endorsing each educational context, these being: have not used automated assessment before ($n = 60;\ 22.6\%$); classroom teaching (i.e., normal face-to-face methods, as in

internal education in classrooms, lectures or laboratories; $n = 89;\ 33.6\%$); fully online learning (i.e., external or distance education) ($n = 52;\ 19.6\%$); computer-assisted classroom teaching (i.e., the main teaching method is face-to-face, but computers are used in-class; $n = 57;\ 21.5\%$); blended learning (i.e., both *take-home* online lessons and *in-class* face-to-face methods; $n = 84;\ 31.7\%$); other ($n = 21;\ 7.9\%$). Because respondents indicated using automated assessment in multiple contexts, the total frequency is > 265 and the total percent is > 100%.

The purposes for using automated assessment were investigated by asking the question: *For what educational purpose(s) was the automate assessment used?* Three options were given, of which respondents could choose one: summative, formative, and both summative and formative. The frequency and percentage of respondents are as follows ($N = 265$): have not used automated assessment before ($n = 60;\ 22.6\%$); summative (counted towards the mark for the subject; $n = 71;\ 26.9\%$); formative (primarily used to assist learning and give feedback on progress; $n = 20;\ 7.5\%$); both summative and formative ($n = 112;\ 42.3\%$), and missing data ($n = 2;\ 0.7\%$).

5.3. *Survey Topic 3: Usefulness Ratings of Automated Assessment*

Here ratings are presented firstly for the general usefulness of automated vs. human assessment, and subsequently for educational usefulness of each type of automated assessment respondents had experienced. Respondents were asked to quantitatively rate the utility of automated assessment in comparison with human assessment. On a four-point Likert-type scale, respondents were asked to rate how useful they found automated assessment in comparison with normal (human) assessment (from *counterproductive* to *very useful*). Table 1 presents the proportion of respondents endorsing each of the Likert-type rating points. A greater proportion of respondents rated automated assessment as either *somewhat useful* or *very useful* (56.6%) than did those who rated it as *counterproductive* or *neither useful nor counterproductive* (19.3%).

To examine the perceived utility of each type of automated assessment, respondents were asked to rate how educationally useful they found each type of automated assess-

Table 1

Usefulness of automated vs. human assessment

Usefulness	n	%
Counterproductive	10	3.8
Neither useful nor counterproductive	41	15.5
Somewhat useful	83	31.3
Very useful	67	25.3
Missing data	4	1.5
Have not used automated assessment before	60	22.6
Total	265	100

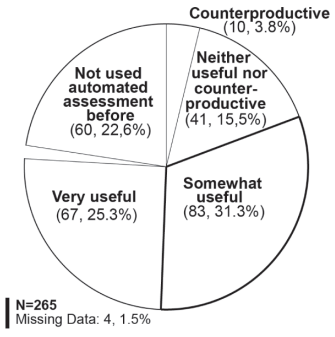
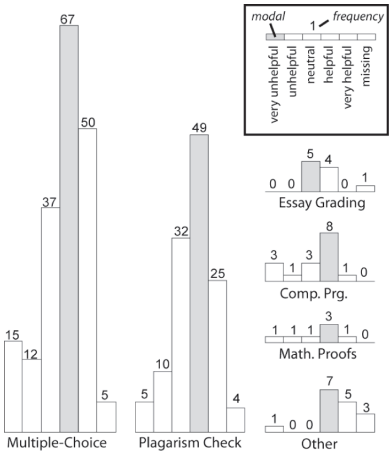


Table 2

Educational usefulness by type of automated assessment

Type of Automated Assessment	Frequency of Ratings: Very unhelpful	Unhelpful	Neutral	Helpful	Very helpful	Missing
Multiple-Choice True/False	15	12	37	67	50	5
Plagiarism checking	5	10	32	49	25	4
Essay grading	0	0	5	4	0	1
Marking computer programming / code	3	1	3	8	1	0
Marking mathematical proofs	1	1	1	3	1	0
Other	1	0	0	7	5	3



Note: Respondents were presented with an item for each type of automated assessment that they had experienced; *Missing* equals the number of respondents who indicated using a given type of automated assessment, but did not rate it; the modal frequency is highlighted in grey.

ment they had experienced on a 5-point Likert-type scale ranging from *very unhelpful* to *very helpful*; see Table 2 for results. For each type of automated assessment the modal rating for educational usefulness was *helpful*, with the exception of AEG, which had a modal rating of *neutral*. Caution must be used when interpreting the data for those assessment types that were rated by only a few respondents. These being AEG ($n = 9$), computer/programming code ($n = 13$), mathematical proofs ($n = 7$), and *other* ($n = 13$). In contrast, a larger number of respondents rated both M-C questions ($n = 179$) and plagiarism checking ($n = 121$), which means they are less likely to be biased by sample artefacts. Caution is best used when generalising these results beyond the sample as the sample is not representative of Australian universities.

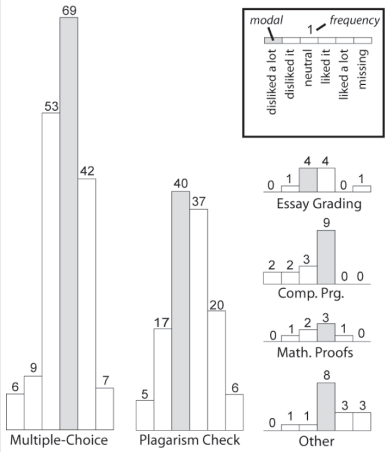
5.4. *Survey Topic 4: Preference-for-Use by Type of Automated Assessment*

Respondents were asked to rate their preference for using each type of automated assessment that they had experienced. For each type of automated assessment they indicated having used, they were presented with a 5-point Likert-type scale (ranging from *disliked a lot* to *liked a lot*). Table 3 presents the frequency of endorsement for each rating by type of automated assessment (with modes highlighted in grey). For each type of automated assessment, there is a clear skew away from negative ratings and towards neutral and positive ratings. Again, we must interpret with caution those types of assessment that have few respondents rating them. However, it is clear that the majority is either neutral toward, or have a preference for, M-C questions and plagiarism checking.

Table 3

Preference for using each type of automated assessment experienced

Type of Automated Assessment	Frequency of Ratings: Disliked a lot	Disliked it	Neutral	Liked it	Liked a lot	Missing
Multiple-Choice True/False	6	9	53	69	42	7
Plagiarism checking	5	17	40	37	20	6
Essay grading	0	1	4	4	0	1
Marking computer programming / code	2	2	3	9	0	0
Marking mathematical proofs	0	1	2	3	1	0
Other	0	1	1	8	3	3



Note: Respondents were presented with an item for each type of automated assessment that they had experienced; Missing equals the number of respondents who indicated using a given type of automated assessment, but did not rate it here; the modal frequency is highlighted in grey.

5.5. *Survey Topic 5: Barriers and Pathways to Use of Automated Assessment*

Participants who indicated that they had not used automated assessment before were asked no further questions about automated assessment. This is with the exception of staff members, who were asked a set of questions designed to determine reasons for their lack of use and possible methods that may be effective in promoting its use. Respondents who were staff ($n = 38$ out of $N = 265$) that had not used automated assessment before were asked the following three questions.

They were asked, *What are the main reasons that discourage you from using online/automated methods of collecting and marking basic assessments (e.g., multiple choice)?* Based on their free-response answers, we extracted 5 themes (where n = number of staff giving a particular response theme, and % = percentage of this subsample, $n = 38$): unawareness of available tools to perform automated assessment ($n = 7$; 18.4%); a belief that automated assessment is only available for basic assessment types like M-C ($n = 11$; 28.9%); a belief that automated assessment is not suitable for testing higher-order knowledge and skills as this requires human judgement ($n = 18$; 47.4%); high error rates and concerns about legitimacy ($n = 5$; 13.2%); lacking support and funding by the university ($n = 1$; 2.6%); being unexperienced ($n = 4$; 10.4%). In coding these responses, multiple themes occurred for some individual's answers, which resulted in the total frequency of themes being greater than the number of participants.

Regarding automated essay grading, these 38 staff members were asked, *What is mainly stopping you from using online/automated methods of collecting and marking*

essays? In total, $n = 33$ participants gave free-text answers which we categorized as follows: being unaware of automated essay grading software in general ($n = 6;\ 18.2\%$); here only $n = 2$ respondents overlapped with the same theme in the previous question); no support or funding by their institution ($n = 4;\ 12.1\%$); essay grading should be done by humans as computers are not capable of this task ($n = 13;\ 39.4\%$); being *cyberphobic* ($n = 7;\ 21.1\%$); the time required to set up the system ($n = 1;\ 3\%$), and; *other* reasons ($n = 4;\ 12.1\%$).

We also asked these 38 staff who had not yet used automated assessment, *Might any of the following be useful in assisting educators to use and benefit from automated essay grading?* The frequency (and percentage of this sub-sample) of staff selecting the following fixed-response options were: *running a free trial of the automated essay grading in parallel to my normal marking* ($n = 26;\ 68.4\%5$); seeing results of a survey supporting the reliability/validity of automated essay grading ($n = 23;\ 60.5\%$); being aware of the benefits of automated essay grading ($n = 20;\ 52.6\%$), and; *other* options suggested by respondents ($n = 5;\ 13.2\%$), which included training and support, and seeing subject-specific examples.

The small subsample means that generalisations to the population of Australian university staff are not well founded. However these data are useful for identifying tendencies with which to build further strategies regarding the dissemination of assessment technologies. The results suggest that these respondents do not have an up-to-date understanding of the technologies and methodologies involved. This may be due to the natural human affinity towards familiar technology and insufficient awareness of research outcomes.

5.6. *Survey Topic 6: Desired Elements of Automated Assessment*

Finally, we had a closer look at the $n = 93$ staff members who had already used automated assessment (35.1% out of $N = 265$ total participants). We asked them about the features that they would look for when choosing or using an automated assessment or marking tool. Their open-ended responses were qualitatively analysed, which resulted in 8 themes: ease of use ($n = 30;\ 32.3\%$); efficiency (i.e., shorter marking time or higher quality in the same time; $n = 28;\ 30.1\%$); accuracy and reliability without manual verification of each assessment ($n = 22;\ 23.7\%$); enhanced feedback for the students and reports for the staff and administration ($n = 17;\ 18.3\%$); advanced pedagogical opportunities such as assessing higher order thinking skills ($n = 10;\ 10.8\%$); higher flexibility and individualization while setting up assessments ($n = 8;\ 8.6\%$); commitment from the institution to apply automated assessment ($n = 3;\ 3.2\%$), and; choosing not to use a particular system due to not seeing real benefits therein ($n = 8;\ 8.6\%$). Other responses ($n = 5;\ 5.4\%$) indicated that respondents looked for integration with existing systems, and administrative features to help organize and archive assessments.

Additionally, staff members already using automated essay grading ($n = 8;\ 3.1\%$) out of $n = 265$ total participants) were asked what they found useful in using this technology (using a fixed-choice format question). The modal answer was *freeing time/energy for*

other educational tasks ($n = 6;\ 75\%$), followed by *marking the assessment in a shorter time* ($n = 5;\ 62.5\%$), increasing the accuracy of assessment ($n = 4;\ 50\%$); reducing the cost ($n = 4;\ 50\%$), and; *improving the feedback to students* ($n = 3;\ 37.5\%$). Note that because many respondents indicated multiple benefits, the total frequency is > 8 and the total percent is $> 100\%$.

6. Discussion

6.1. *Discrepant Use of Human and Computerised Assessment*

The results indicated that large proportions of this sample had used certain types of automated assessment before (i.e., 70.2% had used M-C or true/false questions scored by computers, and 47.5% had used plagiarism checking software). In contrast, all other types of automated assessment had been used by much smaller proportions of the sample (e.g., 6.0% for marking computer programming/code, and 3.8% for AEG). Furthermore, the survey explored the current use of assessment in general (conflating human and computerised assessment): essays and short answer questions were the most commonly reported types of assessments used in this sample. Therefore we can see that these most commonly reported types of assessments do not seem to be supported by marking with computers (AEG had been used by only 3.8% of the sample).

This discrepancy is further informed by examination of survey questions that asked respondents about their reasons for not having used automated assessment. Despite the existence of sophisticated proofs of concept (Dikli, 2006) and the demonstration of integration into the curriculum (Dreher *et al.*, 2008), it appears that many stakeholders are surrounded by walls of worries and doubt about automated assessment, particularly regarding less commonly used approaches such as AEG. The results suggest that the reasons might not be due to the technology itself, but may be due to limitations in access to, understanding of, and doubts about automated assessment. Understandably, no one is comfortable with unproven technology, including innovations in their early stages, and the survey supports the need for improved technology understanding, acceptance, and dissemination. We anticipate that such improvements will be instrumental in substantially altering the use of state-of-the art automated assessment technologies such as AEG. As highlighted in the rationale section, we cannot achieve an impact without the system being applied in pedagogical praxis, but sophisticated systems do exist for AEG. Furthermore stakeholders cannot utilise and benefit from a system without learning about the technology, and integrating it into their environment.

The results of the survey highlighted an apparent ambivalence among stakeholders regarding automated assessment. On one hand, the majority of respondents reported seeing advantages in automated assessment over human marking; 56.6% of respondents considered automated assessment as *somewhat useful* or *very useful*, compared to just 19.3% who rated it as *counterproductive* or *neither useful nor counterproductive* compared to human marking. On the other hand, this sample's use of automated assessment was limited mainly to M-C questions and plagiarism checking. Note that M-C is considered generally to assess the lowest level of Blooms' taxonomy (i.e., recall), however this depends

on the kind of questions that are written. In contrast, other forms of assessment (e.g., essays/reports) require much more input from students and can more easily be used to assess higher learning outcomes on Blooms taxonomy (e.g., analysis and synthesis). However in this sample, the majority of essays were marked by humans (with only 3.8% of the sample reporting having used AEG).

The reason for the limited use of automated assessment does not result from the participants' attitude towards automated assessment in general (which as discussed above was rated favourably compared to human assessment by the majority of the sample). Attitudes/beliefs that may limit adoption of particular technologies may be those which are more specific to them. For instance, we asked the 38 staff members who had not used automated assessment before what in particular prevented them from using AEG. The more commonly cited reasons they gave for not using online/automated methods of collecting and marking essays included: essay grading should be done by humans as computers are not capable of this task ($n = 13;\ 39.4\%$); being unaware of AEG software in general ($n = 6;\ 18.2\%$), and; being 'cyberphobic' ($n = 7;\ 21.1\%$).

6.2. *Technology Acceptance and Innovative Dissemination Channels*

The limited adoption of AEG might not be inherent to the technology, which has proved to be as accurate as human markers in specific applications (Williams, 2006). It may also be caused by missing or incomplete information about the current state-of-the-art (e.g., the belief that essay grading should be done by humans as computers are not capable of this task). Indeed this technology is contentious because it affronts the very qualification of educators by claiming to evaluate (and interpret) the written word. Thus, the dissemination of automated assessment technology should be accompanied by demonstrations, case studies, and hands-on experiences to learn about the benefits.

In summary, we can derive from the survey results and our professional experiences the following tasks to improve the acceptance of the automated assessment technology: (1) comparative experiments; (2) individual and domain specific demonstrations; (3) compelling benefits, and; (4) free (real-life) trials to demonstrate the existence and benefits of software.

Furthermore, while process documentation and statistics demonstrate the technical perspective, stakeholders are fond of practical demonstrations and, in particular, those applied to their courses. Regrettably, practical demonstrations are time consuming, require configuration, observation, and administration by human experts, and interfere with the course activities. Therefore, we argue that 3D Virtual Worlds (e.g., Second Life and Open Wonderland) are well suited for demonstrating how automated assessment and AEG can be conducted in a real-world-like scenario. They offer avatars to represent the different roles in the AEG process, handling of digital documents can be visualized, interfaces provide access to existing real-world systems, and recording of simulations allows for later review of the executed processes for training and evaluation. In addition, simulations of real world learning/vocational contexts increase opportunities to demonstrate specific scenarios that are difficult to achieve otherwise. In general, simulation reduces costs as

it can be executed in parallel to the real-world, requires less effort to be realized, and is more effective. Thus, 3D virtual worlds reduce the risks of large investments (cost and time) for demonstrations and having side-effects on the operational processes (Dreher *et al.*, 2009).

7. Conclusion, Knowledge Transfer, and Future Research

In this paper we have discussed a discipline (automated assessment) that is familiar to most lecturers and researchers in one form or another (e.g., computerised M-C tests), but remarkably few have utilised its advanced applications such as AEG. By conducting a national survey of Australian universities, this research examined the ostensible discrepancy between extant research/technology and limited utilisation of AEG. What we have observed in our professional practice was replicated in the survey data, which identified that state-of-the art automated assessment technologies were used by only a small proportion of this sample.

Regarding using and benefiting from AEG, this survey has indicated various barriers to use (e.g., lack of: awareness, support, funding and/or veridical knowledge) and pathways to use (e.g., free trials comparing humans and computers, demonstrating the accuracy, and being aware of the benefits). Our findings are congruent with the technology acceptance model (TAM), which focuses on perceived ease of use and perceived usefulness (Davis *et al.*, 1989). Experiments have shown that AEG can be as accurate as human markers in particular applications. AEG can also be faster, less expensive, and can enhance feedback (Dreher *et al.*, 2008). However AEG contradicts one of the main distinctions that we see between machines and humans – the view that computers cannot replace humans in tasks that require higher order intelligent reasoning. While this may be true for many endeavours, it is no longer true for grading essays. Therefore one direction for future research is to demonstrate that accurate AEG is achievable in commonplace academic settings. We propose various dissemination strategies to show that systems integrate smoothly into their processes and enhance their performance.

Dissemination strategies can branch in various directions. Firstly, we propose utilising emergent technologies (e.g., 3D Virtual Worlds) to create simulation environments for relevant stakeholders (e.g., educators and administrators) to learn and experience the assessment technology in real world scenarios without having high setup and execution costs. Secondly, we suggest promoting knowledge transfer into other disciplines in order to further validate advanced automated assessment technologies. Initial results in advanced plagiarism detection have been successful, and currently research is being conducted in the field of intelligent text processing for automated creation of semantic net databases and for improving verification of machine translation results.

Indeed, the tasks of processing and understanding unstructured documents are gaining vital importance in the emerging Web 3.0 era. In particular there is an increasing need for inter-cultural communication (via machine translation) in order to understand the endless stream of new documents (via text mining and autonomous intelligent *search*

bots). Therefore modern technology should support users in maximising their potential to work more efficiently at lower cost. Future research could adapt automated assessment to handle changing requirements of international educational systems. In addition to coping with multiple languages in distance education, we are confronted with manifold cultures that influence the interpretation of essays. Thus, the next phase of AEG research could extend conceptual analysis with domain models by mapping cultural influences and multiple languages.

References

Black, P., Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.

Bloom, B.S. (Ed.). (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. Book 1: Cognitive Domain. Longman, London.

Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Handbook of Qualitative Research (2nd ed.). Sage, London, 509–535.

Davis, F.D., Bagozzi, R.P., Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982–1003.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1), 1–35.

Dreher, C., Reiners, T., Dreher, H. (2011). Investigating factors affecting the uptake of automated assessment technology. Accepted for publication in *Journal of IT Education*, Informing Science Institute, Santa Rosa, California.

Dreher, C., Reiners, T., Dreher, N., Dreher, H. (2009). Virtual worlds as a context suited for information systems education: Discussion of pedagogical experience and curriculum design with reference to second life. *Journal of Information Systems Education* (JISE), 20(2), 211–224.

Dreher, H. (2007). Automatic conceptual analysis for plagiarism detection. *Journal of Issues in Informing Science and Information Technology*, 4, 601–614.

Dreher, H., Dreher, N., Reiners, T. (2008). Design and integration of an automated assessment laboratory: Experiences and guide. In: *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 2858–2863.

Haladyna, T.M., Downing, S.M., Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334.

Krathwohl, D.R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice*, 41(4), 212–218.

Murugesan, S. (2009). Handbook of research on Web 2.0, Web 3.0, and X.0: technologies, business, and social applications. *Information Science Reference*, Hershey, PA.

Nicol, D.J., Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.

Nievergelt, J. (1994). Complexity, algorithms, programs, systems: The shifting focus. *Journal of Symbolic Computation*, 17(4), 297–310.

O'Brien, J., Marakas, G. (2008). *Management Information Systems*. McGraw-Hill, New York.

Page, E.B. (2003). Project essay grade: PEG. In: Shermis, M.D., Burstein, J.C. (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Mahwah, NJ, 43–54.

Rees, M., Emerson, L. (2009). The impact that Turnitin®has had on text-based assessment practice. *International Journal for Educational Integrity*, 5(1), 20–29.

Reiter, E., Dreher, H., Guetl, C. (2010). Automatic concept retrieval with Rubrico. In: *Proceedings of Multikonferenz Wirtschaftsinformatik* (MKWI), 3–12.

Shermis, M.D., Burstein, J.C. (Eds.). (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Mahwah, NJ.

Teddlie, C., Tashakkori, A. (2009). *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Sage, Thousand Oaks, CA.

Venkatesh, V., Davis, F., Morris, M.G. (2007). Dead or alive? The development, trajectory and future of technology adoption research. *Journal of the Association for Information Systems*, 8(4), 267–286.
Williams, R. (2006). The power of normalised word vectors for automatically grading essays. *The Journal of Issues in Informing Science and Information Technology*, 3, 721–730.
Williams, R., Dreher, H. (2005). Formative assessment visual feedback in computer graded essays. *Issues in Informing Science and Information Technology*, 2, 23–32.

T. Reiners is a postdoctoral researcher at the University of Hamburg, Germany, and University Associate at the Curtin University in Perth, Western Australia. His research and teaching experiences are in the areas of operations research (meta-heuristics/simulations models for container terminals), fleet logistics, information systems and several topics in e-learning and software development. His PhD thesis "Simulation and OR with SmartFrame" demonstrated concepts for didactical models. Besides scientific publications, he conducts research in semantic networks to improve cross-border communication, e-learning and machine translation. Dr. Reiner's interests also include virtual worlds and their interconnectivity / exchange without barriers. This research includes the development of adaptive systems, automatic processing, analysis, and evaluation of documents, innovative platforms in combination with emerging technologies like mobile devices. Torsten Reiners is co-founder of the Second Life Island University of Hamburg and `students@work`, an initiative to promote education in Web 3D as well as the value of students' work.

C. Dreher holds a PhD in psychology, a masters in clinical psychology, and a graduate certificate in research commercialisation. Inter alia, Dr. Dreher enjoys conducting research regarding mindfulness-based interventions in health care and emerging technologies in information systems. He is a "digital native" who finds emerging socio-technological developments to be an excitingly innovative confluence between people and technology. Based on his own empirical observations, he has been known to claim that "smart phones are about just as much fun as you can have while being alone."

H. Dreher is a professor in informatics at the Curtin Business School, Curtin University, Perth, Western Australia. He has published in the educational technology and information systems domain through conferences, journals, invited talks and seminars; is currently the holder of Australian national competitive grant funding for a 4-year e-learning project and a 4-year project on automated essay grading technology development, trial usage and evaluation; has received numerous industry grants for investigating hypertext based systems in training and business scenarios; and is an experienced and accomplished teacher, receiving awards for his work in cross-cultural awareness and course design. In 2004 he was appointed adjunct professor for computer science at TU Graz, Austria, and continues to collaborate in teaching & learning and research projects with European partners. Dr. Dreher's research and development programme is supported by Curtin Business School Area of Research Focus funding – Semantic Analysis and Text Mining for Business and Education (www.eaglesemantics.com) in addition to other competitive funding obtained for individual projects.

Šeši svarbiausi klausimai apie automatinio vertinimo naudojimą ir priėmimą

Torsten REINERS, Carl DREHER, Heinz DREHER

Automatinio vertinimo technologijos švietime naudojamos jau kelis dešimtmečius (pvz., kompiuterizuoti kelių pasirinkimo variantų testai). Automatinės rašinių (esė) reitingavimo technologijos egzistuoja taip pat jau kelis dešimtmečius, tačiau Australijos universitetuose jos naudojamos palyginti retai. Šio straipsnio autoriai, norėdamai suprasti priežastis, kodėl automatinės rašinių reitingavimo sistemos retai naudojamos Australijos universitetuose, atliko nacionalinį tyrimą. Kiekybiniai ir kokybiniai duomenys internetinės apklausos būdu buvo surinkti iš 265 darbuotojų ir studentų penkiuose Australijos universitetuose. Didžiausia dalis respondentų vertinimui pateikia rašinius ir referatus (82,6%), tačiau automatinis reitingavimas buvo naudotas labai retai (3,8%). Straipsnyje pateiktos rekomendacijos taikyti metodams, kurie skatintų vertinimo technologijų naudojimą, įskaitant novatoriškus sklaidos kanalus, pavyzdžiui, trimačius virtualiuosius pasaulius.