Automated essay scoring in applied games: reducing the teacher bandwidth problem in online training

Westera, W., Dascali, M., Kurvers, H., Ruseti, S. & Trausan-Matu, S. (2018). Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training. Computers & Education, 123, 212-224. Doi: 10.1016/j.compedu.2018.05.010.

Abstract

This paper presents a methodology for applying automated essay scoring in educational settings. The methodology was tested and validated on a dataset of 173 reports (in Dutch language) that students have created in an applied game on environmental policy. Natural Language Processing technologies from the *ReaderBench* framework were used to generate an extensive set of textual complexity indices for each of the reports. Afterwards, different machine learning algorithms were used to predict the scores. By combining binary classification (pass or fail) and a probabilistic model for precision, a trade-off can be made between validity of automated score prediction (precision) and the reduction of teacher workload required for manual assessment. It was found from the sample that substantial workload reduction can be achieved, while preserving high precision: allowing for a precision of 95% or higher would already reduce the teacher's workload to 74%; lowering precision to 80% produces a workload reduction of 50%.

Keywords: intelligent tutoring systems; architectures for educational technology system; interactive learning environments; simulations; distance education and telelearning

1. Introduction

This paper presents the development and use of an automated essay scoring methodology in an applied game. Over the last decade, applied games, viz. games for learning and teaching, have received increased interest from researchers and practitioners. Applied games are among the most challenging, most dynamic and most interactive e-learning environments, as they offer learners rich and highly interactive content, large degrees of control, freedom of movement and responsibility for the actions undertaken.

As applied games are readily positioned as the attractive alternative of traditional teaching practice, their creators tend to minimise the resemblances with school and classes, and avoid traditional approaches such as direct instruction, reflection, and, importantly, writing exercises (Schank, Berman, & MacPherson, 1999; Saveski, Westera, Yuan, Hollins, Fernández Manjón, Moreno Ger, et al., 2015; Westera, 2015; Westera, 2017). It is widely acknowledged that pedagogical interventions during the game, such as e.g. writing assignments, may affect the learners´ "flow", their motivation and "click-and-go" fun, and thereby they may undermine the rationale of using games at all (Shute, 2011). Because of these potentially disturbing effects, writing assignments are scarce in games. They are also scarce, because they are not scalable. While online connectivity enables and suggests personalised, one-to-one communication, online training in practice is constrained by the limited "teacher bandwidth" (Wiley, 2002), which refers to the maximum number of online students that a teacher is capable to support and guide. The many one-to-one communications that individual learners in online training expect from their teachers ("Why don't they just

answer my email?") readily leads to teachers´ work overload. Without having the opportunity of including efficient plenary classroom sessions, it is a big challenge to provide all learners with personalised feedback. This would also apply for writing assignments.
Still, there are multiple reasons for including writing assignments in applied games. First, various authors argue that game-based learning is highly experience-based and thereby tends to promote implicit, tacit knowledge rather than deep and explicit understanding (Verpoorten, Castaigne, Westera, & Specht, 2014). Writing assignments, such as open answer tests or student essays, accommodate deeper knowledge processing since they require explicit consideration of learned concepts, principles and their relationships, reflection about the significance and appraisal of the experiences, and the creative synthesis of argumentation. Second, writing assignments are an excellent diagnostic tool for detailed assessment of learning progress. That is why schools and universities often require students´ reports or theses as proofs of mastery. Third, in many cases – and also in the environmental policy consultancy game that is presented here - writing skills are included in the learning goals: a trainee consultant should not only learn to analyse a problem, but also to record the analysis and formulate potential solutions.
In this study, we present the results of an automated essay scoring approach in the VIBOA environmental policy game, a bachelor level game that was jointly developed by the Open University of the Netherlands, Utrecht University, and Radboud University Nijmegen. In the VIBOA game, students in the environmental sciences academic programme adopt the role of an environmental consultant in a (fictitious) consultancy agency. As a consultant, students are confronted with a set of authentic, environmental problem cases that are loaded with conflicting views, conflicting interests and conflicting demands. Thereby, the VIBOA game offers them the opportunity to (learn to) act as a professional at an academic level. Students are required to record their findings in various (intermediate) reports (in Dutch), which then are manually scored by teachers. Students with a low score should redo part of the game or review other learning resources, otherwise students proceed with the next level.
In this paper, we investigate how automated essay scoring methods can help to reduce the teachers´ work load. In a preliminary study the automated essay scoring method for the reports in the VIBOA game demonstrated a validity level of 75.1% in correctly predicting a pass or a failure (Dascalu, Westera, Ruseti, Trausan-Matu, & Kurvers, 2017). Although this rate is substantial, it is considered insufficient for practical application. However, the reported precision of 75.1% essentially reflects the average precision level across the entire sample. By taking into account the variability of predictions the current study aims to provide a methodology for identifying a subset of reports that can be automatically classified with greater validity. It combines probability theory and binary classifiers to describe how automated score predictions can be evaluated in order to decide upon the subset of essays that is suited for automated assessment. Consequently, our methodology allows to decide and estimate what the reduction of teacher work load (for manual assessment) will be. The methodology was validated with the very same data set of the VIBOA game.
First, we present an overview of the state of the art on assessment in games, text-based assessment approaches, and the use of Natural Language Processing (NLP) technologies in education. Then we explain the context and methods used in our study. Finally, we present and discuss our findings.

2. Background

In this section, we present the state of the art in assessing learning progress in games, compare the implications of multiple choice tests and open answer assignments, and discuss the use of Natural Language Processing technologies in education, particularly in games, respectively.

2.1 Assessing learning progress in games

Any school or teacher would endorse that regular monitoring and assessment of a student's learning progress is essential. The many traces that students leave in an applied game seem to offer many opportunities for the detailed monitoring and assessment of their progression toward the learning objectives. The traces could provide data about navigation, strategic choices about how to address the challenges in the game, or the selection of tools and resources used. Also, spent time on each task and the number of task trials are relevant indicators for assessing progress (Westera, Nadolski, & Hummel, 2014). In practice, however, the extraction of learning progress from data traces is far from straightforward. First, many suggested learning analytics approaches (Buckingham-Shum, & Ferguson, 2012) tend to focus on population data and the identification of anomalous behaviours as compared to population averages. Indeed, learner data analysis may provide useful insights at the macro- and meso-levels of classes, educational curricula and courses, or for playtesting and tweaking the game challenges, but do not necessarily provide detailed diagnosis about the learning needs of individuals.

Second, many (applied) games use score systems, be it that these tend to conform to gameplay standards rather than educational standards, and to focus on events rather than the underlying skills, knowledge or competence frameworks. Consequently, game score systems seldom comply with the strict requirements of validity, and fairness of educational assessments. Game score systems are often designed and used as a motivational add-on, in fact a reward mechanism, rather than a valid diagnostic tool.

Third, as in all learning-by-doing situations, applied games' scoring systems are likely to confuse performance with learning. They often enforce the achievement of performance goals, e.g. the swift completion of tasks, avoiding errors, and the use of proven methods for reducing risks. This stimulates learners to demonstrate high ability and speed, and to avoid poor performance, while learning would be better served with spending sufficient time for in-depth understanding, having sufficient opportunities for reflection, revision, and self-evaluation, and being allowed to make mistakes since errors and failure are productive sources of learning (Mathan, & Koedinger, 2005).

Fourth, Shute and Ventura (2013) have convincingly demonstrated the possibility to extract player's progress indicators from user traces. Their stealth assessment model offers an unobtrusive alternative for intermediate tests or questionnaires, which are often perceived as unwanted interruptions of game play. But the approach is laborious, if not impracticable. It combines Evidence-Centred Assessment (Mislevy, Steinberg, & Almond, 2003) and Bayesian score models, which require a prior analysis of the activities in the game, a detailed mapping of activities to a competence model and a Bayesian net that is trained by a set of test data. Few game developers will be prepared to understand, create and implement such complex and laborious models.

Fifth, only few game examples are available that use student open-ended answers. The VIBOA game on environmental policy, which is presented in this paper, requires students to compose and post written reports as part of the game scenario. Another notable example would be the series of simulation games used in Dutch vocational education for the training of IT-system design, where students after consultations of their customers and end-users have to report on identified system requirements, the system´s functional design, the technical design and a test plan, respectively (Nadolski, & Hummel, 2016). In these cases, however, the written reports are manually assessed and scored by a teacher, thus encumbering the teacher bandwidth.

Altogether, the assessment of learning progress in games, in particular the use of open-ended answers, seems a suppositious child, which would deserve more attention. The option of

recording players' open answers, either in text or audio, would open up a new dimension to assessing learning progress in games.

2.2 Multiple choice tests versus open ended questions

In education, multiple choice testing - students are asked to choose the correct answer to a question from 3 or more options – is an established practice for assessing learner's learning outcomes. In many applied games, quiz-like mechanisms are used to implement multiple choice test. Likewise, discrete decision points and branching scenarios that are used in many games are technically equivalent with a set of multiple choice items. Such player decisions can be easily tracked and used for performance scoring. For about a century, researchers have studied and compared multiple choice testing (MC) with open answer tests, the latter requiring candidates to formulate their own answers which might be a short answer, an essay, or even a diagram (Ruch, & Stoddard, 1925; Heim, & Watts, 1967). Multiple choice tests allow for efficient and unambiguous scoring. There is a large body of literature about multiple choice testing, e.g. focusing on test validity, test reliability, interpretation of test results, and adaptation algorithms for enhanced efficiency and accuracy (Cronbach & Gleser, 1959). It is widely recognised that multiple choice items are difficult to write well, that they cannot measure all types of skills (e.g. "conducting a scientific study"), and that reading abilities of the learners may influence the observed outcomes. Furthermore, guessing or random marking may lead to unjust high scores (Funk, & Dickson, 2011). Generally, open answers are found to be more informative, because compared to conventional MC tests guessing is minimized and the correct solution cannot be derived by successive elimination (Gibbs, 1995). In a study about vocabulary learning, Heim and Watts found significant differences between multiple-choice tests and open-ended scores, the former being substantially higher than the latter. Also Funk and Dickson (2011) found higher scores in multiple choice tests, which was said to be plausible because students are far more likely to recognise the correct response than they are to work it out for themselves. In a study on procedural knowledge of fraction addition arithmetic, Birenbaum and Tasuoka (1987) considerable differences were found in favour of the open-ended test format. Ozuru, Briner, Kurby and McNamara (2013), using a text comprehension task, concluded that multiple choice tests and open-ended testing measure different aspects of comprehension: performance on multiple choice tests was correlated with prior knowledge, while performance of open ended tests was correlated with the amount of active processing (e.g., as measured by the quality of self–explanations). Overall, open answer tests provide more detailed insights in students´ learning achievements. However, scoring open answers or essays is time consuming. Teachers in schools and universities are incited to spend a great deal of their time to the individual grading of student reports. Therefore, more and more open answer exams tend to be replaced with multiple-choice questions, which can be easily processed by computers. Recent advances in Natural Language Processing open up new opportunities of automated scoring of students´ open answers.

2.3 Natural Language Processing in education

Natural Language Processing (NLP) involves the use of semantic, statistical, linguistic, rule-based and machine learning approaches to support the interactions between computers and human language(s). NLP covers the analysis and interpretation of human texts, text and speech synthesis, translations, as well as text-to-speech and speech-to-text conversions. One notable example of NLP in education has been the intelligent tutoring system, which engages into personalised dialogues with students for supporting and guiding them in their learning. Intelligent tutoring systems have received wide interest since the 1970s. At that time, the emergence of expert systems and microcomputers created high expectations about the potential role of intelligent tutoring systems in stretching the teacher bandwidth. A

breakthrough failed to occur, however, because of the underestimated complexity of creating intelligent dialogue systems.
After the turn of the century, influenced by advances in computer processing hardware and artificial intelligence, some successful natural language tutoring systems emerged, such as AutoTutor (Nye, Graesser, & Hu, 2014), which has demonstrated to produce learning gains across multiple domains (e.g., computer literacy, physics, critical thinking). Still, intelligent tutoring systems have never become common. To date, internet search engines are probably the most widely used NLP applications in education. Also, anti-plagiarism software is abundantly used for checking student projects. General purpose syntactic and grammar analysis is used by writing pals (Roscoe, & McNamara, 2013), which assist students at spelling and grammar issues when they have to write reports. Automated essay scoring is less common.
Although some commercial services for the automated assessment and marking of open texts are available such as E-rater (Burstein, Tetreault, & Madnani, 2013), the field is still in the making. The E-rater service is mainly used for quality control on human scores: upon large deviations, an additional human rater is assigned (Born, 2015). In addition, the system was used for scoring different genres, including scientific writing (e.g. Liu et al., 2016). The inherent complexity and multi-layeredness of human expression, and the contextual, cultural and language dependencies of extracting relevant indicators from texts requires a multifaceted approach based on syntactical and grammatical error detection, discourse coherency and sentiment analysis, to be captured by dozens of quality metrics that should be calibrated and checked for validity (Dascalu, 2014; Crossley, Dascalu, Trausan-Matu, Allen, & McNamara, 2016). Advances in such holistic approaches have made machine scoring of essays a realistic option (Zhang, 2013). As will be shown in this paper, automated essay scoring can be configured effectively in such a way that it would reduce the need for manual assessment.
Developments in Natural Language Processing technologies cannot be viewed independently from worldwide advances in artificial intelligence. IBM (Watson), Microsoft (Project Oxford), Alphabet (Deepmind), Baidu (Minwa), and Apple (BNNS) all revert to deep learning with neural networks to be able to provide intelligent services to the global consumer markets of smartphones and tablets. Natural language interfaces with enhanced knowledge and reasoning capabilities are expected to advance the interaction modes for conversational agents and digital assistants (Alexa, Siri, Google Assistant), automated vehicle systems and social robots. For games, such conversational agents are of particular interest for the communication with virtual characters. To date, however, dialogues with game characters are mostly implemented by using pre-recorded human texts or speech in fixated dialogue branches. Ultimately, the research could breathe new life into intelligent tutors.

3. Context

3.1 Developing advanced open source game technologies

The study presented in this paper is part of the RAGE project (rageproject.eu), which is the principal research and innovation project in Applied Games funded under the Horizon 2020 Programme of the European Commission. RAGE makes available a reusable set of applied game software components that can be used across the wide variety of game engines, game platforms and programming languages that game developers have in use (Van der Vegt, Westera, Nyamsuren, Georgiev, & Martínez Ortiz, 2016). Technology components cover affective modeling, learning analytics, emotion detection, game adaptation, gamification and natural language processing, among other things. Natural language technology components are developed within the *ReaderBench* framework (Dascalu, 2014; Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014), which introduces a generalised, multi-lingual, automated

model applicable to both essay-like or story-like texts as well as conversations in multi-user games, Computer Supported Collaborative Learning (CSCL), or Communities of Practice. All technology components created by RAGE are publicly available without barriers (rageproject.eu).

3.2 The VIBOA game

The current study is linked to the VIBOA environmental policy game, which is part of the scientific bachelor degree programmes of the Open University of the Netherlands, Utrecht University and (until recently) the Radboud University. The VIBOA game is composed of a series of inquiry-based games that allow students apply the basic knowledge of theories, methods, and models learned in the early courses of the bachelor programme in practical, realistic conditions. The learning objectives include the mastery of methodologies of analysis, evaluation, and design of environmental policies. The study load is around 50 hours. All games are situated in a (fictitious) consultancy agency, called VIBOA. In the games students adopt the role of a consultant and they receive various assignments to investigate topical problem cases. They have to apply scientific methodologies and theories in a context that is imbued with conflicting views, conflicting interests, conflicting demands and conflicting information. They have to make a thorough analysis of the problems and devise solutions for these by collecting, assessing and combining relevant information from reports, scientific papers, interviews, texts of law, formal documents and other sources. All these resources are publicly available authentic materials. The interviews, which are composed of a large set of pre-recorded video answers from real stakeholders, such as policy makers, researchers, politicians, duped citizens and some more, are interactive sessions that are controlled by the students who can decide what issues to raise and how to proceed. Videos also allow players to attend meetings with experts and stakeholders (cf. Figure 1).

Figure 1. Screenshot of attending a meeting with experts and stakeholders in the VIBOA game

Fictitious colleagues in the consultancy agency are impersonated by non-playing characters that can be likewise consulted in video interviews; they have a supportive role by covertly guiding and hinting students to relevant aspects that they might take into consideration. The game scenario is based on a structure of triggered events such as incoming notifications or (pseudo) email messages that provide new information, announce new events, provide hints

or prompt for certain actions. All these scenario elements contribute to enhanced realism and sense of urgency.

In the VIBOA game five separate games are linked together in a single run that requires about 50 hours of study load. The problem cases include 1. Wadden Sea, 2. Wind energy, 3. Lake Naarden, 4. Micro pollution, 5. River management, all of which cover topical problem areas. Our study on essay scoring focuses on the Wadden Sea, which is the largest natural reserve in the Netherlands. The Wadden Sea is an intertidal zone in the southern part of the North Sea. Confined by the continental coast and a series of small islands it constitutes a body of water with sandbanks, tidal flats and wetlands. The discovery of natural gas under the Wadden Sea has led to persistent political debates whether or not extraction of gas should be permitted, given the importance of this ecosystem. Different governments took different decisions in the trade-off between the exploitation of gas resources and the preservation of nature, by partly granting permissions, while at the same time establishing a fund for sustainable development and the protection of nature. Also, commercial fishery of shellfish in the Wadden Sea was subject to varying regulations. Students in this game component have to analyse the genesis of the Wadden Sea policy and try to answer and explain questions such as: What exactly is the policy? What models and theories should be used to analyse and explain the Wadden Sea policy? What are the principles? What are relevant factors and stakeholders? Who is responsible for the policy? What will be the effects of the policy? And what are the expectations about future policy developments? All findings should be laid down in a report that is then uploaded to a server for manual assessment by a teacher. Here, the teacher bandwidth problem surfaces: teacher workload explodes and feedback to students who proceed with the next game is delayed. This is where essay scoring could be of help.

4. Method

4.1 Target group and Sample

The corpus in this study was composed of a set of 173 anonymised Wadden Sea essays from students of Utrecht University that had to play the game as part of the academic bachelor programme in environmental sciences. Although the VIBOA game is also in use by the Open University of the Netherlands, we restricted the sample of participants to Utrecht University, which offered the largest user group, for increased homogeneity with respect to participants´ age (20- 25 years) and the overall educational programme. All collected essays were already scored by human tutors on the basis of a formalised assessment framework. The score ranges extend from 1 (utterly weak) to 10 (excellent), while using 0.5 points increments. For passing the assessment student should achieve 6 points or more. The size of the reports was typically 2000 words. As no strict template was used in VIBOA, all reports had to be manually corrected in terms of appropriate usage of heading styles, before being ready for automated assessment. In practice, such manual operation can be easily avoided, however, by using a report template with fixed headings or sections.

4.2 Essay scoring pipeline

In a previous study (Dascalu, et al., 2017), the *ReaderBench* framework (Dascalu, 2014; Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014) was used to evaluate the quality of the technical reports from the previous dataset. A standard NLP pipeline for the Dutch language was used to process each essay, followed by the computation of a series of textual complexity indices. The pipeline includes content word detection, based on the E-Lex lexicon (formerly named TST-lexicon) (CGN Consortium, 2017), stop words elimination and lemmatization using lists and part-of-speech tagging. Latent Semantic Analysis (LSA)

(Landauer, & Dumais, 1997) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) semantic models were trained for Dutch using the Corpus of Contemporary Dutch (Hedendaags Nederlands; 1.35 billion words; http://corpushedendaagsnederlands.inl.nl), which is the most favourable alternative in terms of dimension, breadth of topics, as well as novelty of comprised documents. After performing a thorough text pre-processing, the Corpus was reduced to around 500 million content words from approximately 11.5 million paragraphs. Complementary to the semantic models, the Open Dutch WordNet was integrated in the system, which was used to compute various semantic distances (Wu-Palmer, Leacock-Chodorow and path length) (Budanitsky, & Hirst, 2006) and to identify lexical chains (Galley, & McKeown, 2003). Having included these procedures in the *ReaderBench* framework allows for extracting a wide range of complexity indices from reports in Dutch. These indices are categorized according to their textual analysis scope and focus on text cohesion and discourse connectivity. Most of these indices are computed for each discourse analysis level (sentence, paragraph, document).

- *Surface, lexicon, and syntax analyses*
  These indices are based on the first attempts of measuring textual complexity proposed by Page (1966) and later on refined by Wresch (1993), while adding other frequent surface measures used in essay scoring systems (e.g. average length of words/sentences/paragraphs, number of content words, number of commas, word entropy, or number of pronouns for each type).
- *Semantic analysis and discourse structure*
  Text cohesion is evaluated at local and global levels, by using the CNA cohesion graph (Dascalu, McNamara, Trausan-Matu, & Allen, 2017; Dascalu, Trausan-Matu, McNamara, & Dessus, 2015). This graph is constructed using the semantic similarities computed with the LSA and LDA models, as well as WordNet-based distances. In addition, several characteristics of the detected lexical chains (e.g. average and maximum span, average number of lexical chains, percentage of words contained in a chain) were also introduced, together with discourse connector types based on cue phrases extracted from the Referentiebestand Nederlands (RBN, 2018).
- *Word complexity*
  These indices aim to capture the complexity of the text by estimating the complexity of individual words. They vary in terms of depth of the analysis, from the mere counting of characters and measuring differences from each inflected form to its corresponding lemma, to WordNet-based metrics (e.g., distance to the hypernym tree root or number of senses assigned to each word).

4.3 Descriptive probabilities

In the VIBOA case essay scoring aims to reduce the teacher workload by identifying the essays for which the automated scores predict a pass or a fail with sufficient certainty. Only the remaining essays, viz. the essays that cannot be qualified as a pass or a fail with sufficient certainty, would then be left for manual assessment by the teacher. Practically, this means that automated scores at the extreme ends of the scale (low scores and high scores) are accepted, while automated scores in the middle segment would require manual assessment. The principal question would be to identify the boundaries of this middle segment for the given essay assignment: essays outside this interval are assessed automatically with sufficient validity, while essays within the interval range would require manual assessment. Below, a brief technical explanation of the applied procedure is presented.

The problem of deciding about a pass or a fail is essentially a binary classification problem. Let O be the observed score (the score prediction) and let R be the real score (the reference score). Let T be the critical score threshold between passing and failing. As a prediction O can either be true or false, four situations should be distinguished (cf. Table 1).

Table 1. Four distinct situations in binary classification (fail/ pass).

	Prediction is true (O and R on the same side of T)	Prediction is false (O and R on opposite sides of T)
Prediction O is a pass (positive)	True positive	False positive
Prediction O is a fail (negative)	True negative	False negative

Each observation O is subject to uncertainty: let ΔO be the mean error of the prediction. The value of O is the best empirical estimate of the real score R. If the probability density of the predicted score can be described by a normal distribution f(O, ΔO), centered around the observed value O then the probability of a true fail, that is, the probability of correctly predicting a fail (R<T), is given by the shaded area under the curve in figure 2.

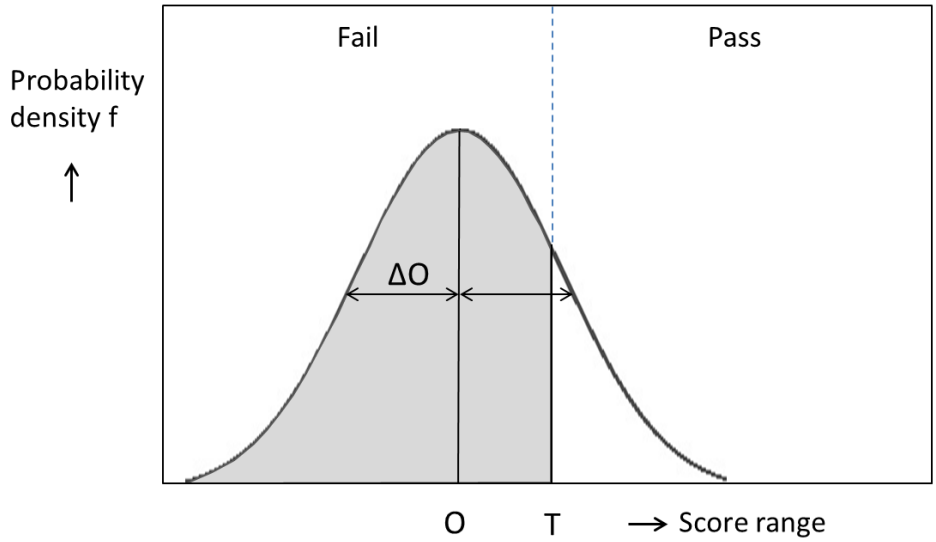

Figure 2. The probability that the assigned fail score O is correct is given by the shaded part of the area under the curve.

This true negative probability $P_{true\_negative}$ that the fail judgement (O<T) is correct (R<T), can be expressed in terms of the cumulative distribution function F (cf. figure 2):

$$P_{true\_negative}(O) = F(T, O, \Delta O) \tag{1}$$

with

$$F(T, O, \Delta O) = \frac{1}{\sqrt{2\pi \cdot \Delta O^2}} \cdot \int_{-\infty}^{T} e^{-\frac{(S-O)^2}{2 \cdot \Delta O^2}} \cdot dS \tag{2}$$

Consequently, the probability that the fail judgement (O<T) is false (R>T) can be expressed as

$$P_{false\_negative}(O) = 1 - F(T, O, \Delta O) \tag{3}$$

Similar expressions can be used for true positive and false positive observations. Figure 3 presents the graphical representation of the true positive and true negative probability of

observations as determined by these expressions for the full range of essay scores (assuming that ΔO is a constant over the full range of scores).

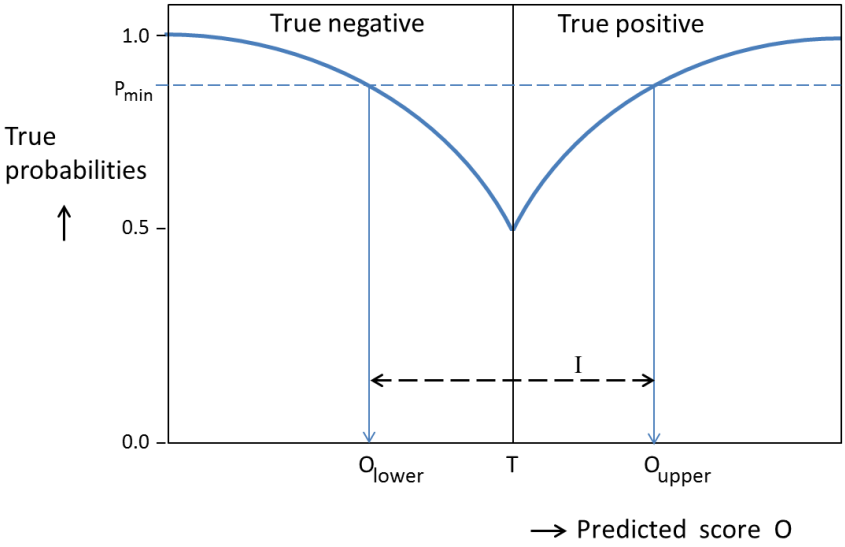

Figure 3. Exemplary probability of correct fail or pass judgements as a function of score.

The graph of the correct classification as a function of the predicted score is a non-monotonous, non-differentiable curve that is symmetrical around the threshold score T. The probability drops to a minimum of 0.5 when the predicted score O is equal to the critical threshold T. This corresponds with the baseline probability of a random draw. In practical cases, one would require a higher probability for the classifier. In figure 3, such required probability level is illustrated with the horizontal line at P_{min}, suggesting the minimum probability that would be considered acceptable. The intersections with the curve define to associated scores on the horizontal scale: O_{lower} and O_{upper}. These scores define an interval $I=[O_{lower}, O_{upper}]$ on the score axis given by the horizontal, dashed line in figure 3. The significance of interval I is that, given the required minimum probability level P_{min}, the predicted scores *O* in this interval are rejected and manual assessment by the teacher is required. Reversely, scores outside the interval, can be accepted directly and removed from the manual assessment pool, thus reducing teacher workload. Figure 3 also allows to explain the trade-off between validity and efficiency: if a higher probability P_{min} would be desired, the interval increases, leaving more essays to be manually corrected.
The challenge is to find what the lower bound and upper bound of interval *I* should be to have a maximum number of reports automatically scored while only accepting prediction quality that is above the preferred minimum probability level P_{min}.

4.4 Interval optimisation: identifying successful pass or fail predictions

Our main purpose is to reduce the workload of teachers by automatically scoring essays that are either too good, or too bad. The challenge is to find what the lower bound and upper bound of interval I (dashed line in figure 3) should be to have a maximum number of reports (outside this interval) automatically scored, while only accepting prediction quality that is above a preferred minimum probability level P_{min}. Practically, it means that given a critical pass/fail threshold T, score predictions should be evaluated with interval *I* as an independent variable, or to be more precise: to use the lower bound (O_{lower}) and upper bound (O_{upper}) of interval *I* as independent variables.
In order to evaluate our method on such intervals the binary classification schema is slightly adjusted, while distinguishing between observations O that are either within the interval I or outside the interval (cf. Table 2).

Table 2. States of binary classification either inside interval I or outside interval I (¬I).

	Prediction is true (O and R on the same side of T)	Prediction is false (O and R on opposite sides of T)
Prediction O is outside I	True positive	False positive
Prediction O is inside I	False negative	True negative

As we have teacher-assigned reference scores R available for the whole sample the probability of correct classification, given interval I, is easily calculated by:

$$Probability\ (I) = Precision(I) = \frac{\#true\_positives(I)}{\#true\_positives(I)+\#false\_positives(I)} \quad (4)$$

Note that in information retrieval the probability metric is usually referred to as "Precision": the fraction of correct items in the sample. Precision is often used in conjunction with the "Recall" metric, which indicates the fraction of correct items selected:

$$Recall(I) = \frac{\#true\_positives(I)}{\#true\_positives(I) + false\_negatives(I)} \quad (5)$$

While the Precision metric takes into account the number of false hits, the Recall metric includes the number of missed hits. As a false hit (which implies that an essay is unjustly subjected to an automated pass) is more detrimental than a missed hit (which implies manual and thereby correct assessment), Precision would be the best metric for evaluation of our methodology. Still for balancing off prediction and workload reduction, a global optimization metric is introduced, which is the classical harmonic mean of precision, recall and coverage (the number of essays that were automatically scored, disregarding whether correctly or not).

$$Global = \begin{cases} \frac{3}{\frac{1}{precision}+\frac{1}{recall}+\frac{1}{coverage}}, & precision, recall, coverage \neq 0 \\ 0, & otherwise \end{cases} \quad (6)$$

The Global metric thus accounts for both precision and recall, while also compensating for bias of sample size.

5. Results

5.1 The dataset

The validation of the method was done on the VIBOA dataset, which contains 173 essays scored by the teacher. The mean score of the whole dataset was 6.91 with a standard deviation of 1.19. About 20% of the input dataset was randomly chosen as test data and the rest was used for training and validation. The test set contained scores of 34 student reports. The mean score of the sample was 6.94 with a standard deviation of 1.18. A Shapiro-Wilk test on the sample could not reveal deviations from normality (W = 0.959, p = 0.221). Kurtosis (excess = -0.037) was likewise acceptable.

5.2 Model Selection

First, the most appropriate regression algorithm was chosen based on the cross-validation results and score coverage. For each essay, over 200 complexity indices were computed using the *ReaderBench* framework, as described in Dascalu, et al. (2017). The most relevant ones were selected based on different statistical measures (i.e., linguistic coverage - at least 20% of values are different than 0 for all documents, normality and multi-collinearity checks), thus limiting the number of features to 15 as presented in Table 4. The remaining textual complexity indices are related to three major categories: a) *surface* – frequently used indices related to different counts at multiple levels (document, paragraph, sentence), and word entropy which denotes a more diversified vocabulary; b) *word lists* – counts using predefined lists of words, including specific pronouns and discourse connectors; c) *cohesion* – perceived at both global level (e.g., lexical chains spanning throughout the entire document), as well as local cohesion (e.g., intra-paragraph cohesion determined as sentence-paragraph semantic distances). The only value that is negatively correlated with the student scores is the average sentence-paragraph cohesion which denotes that more elaborated ideas are introduced per paragraph, thus reducing the similarity between each individual sentence and the corresponding paragraph.

*Table 3. Correlations between textual complexity indices reported by ReaderBench and student score (significance levels: * <0.05, **<0.01).*

***Category* – Textual complexity index**	***r***	***p***
Surface – Logarithm (# words)	.461**	<.001
Cohesion – Average lexical chains identified using WordNet	.340**	<.001
Cohesion – Average sentence-paragraph cohesion using Wu-Palmer semantic distance (WordNet)	-.278**	<.001
Word lists – Average condition connectors per paragraph	.261**	.001
Word lists – Average circumstance connectors per paragraph	.253**	.001
Surface – Word entropy	.252**	.001
Word lists – Average indefinite pronouns per sentence	.252**	.001
Cohesion – Coverage of lexical chains	.250**	.001
Word lists – Average concession connectors per paragraph	.233**	.003
Word lists – Average third person pronouns per sentence	.201**	.008
Surface – Average sentence length (# characters)	.198**	.009
Surface – Average unique words per sentence	.188*	.014
Word lists – Average circumstance connectors per sentence	.185*	.015
Surface – # Paragraphs	.162*	.034
Word lists – Average condition connectors per sentence	.153*	.046

These features were used with different classifiers in a 5-fold cross-validation, and the best model was selected for the next step. For this purpose, the Weka machine learning library (Frank, Hall, & Witten, 2016) was used. In this step, three algorithms were tested, each with different configurations: linear regression with varying attribute selection methods (M5 and greedy), a multilayer perceptron (MLP) with one hidden layer, and Support Vector Regression (SVR; Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) with different kernels (i.e., polynomial and radial basis function - RBF -, and Pearson VII kernel function - PUK). The algorithms were selected to create a strong baseline as they are the most frequently used methods for continuous variable predictions. The results of the cross-validation are presented

in Table 4. In the case of the neural network, only the results for a hidden layer of size 1 and 2 are depicted because bigger networks over-fitted the small dataset.

Table 4. Accuracies of score prediction for different algorithms.

Model	**Mean absolute error (MAE)**	**Root mean squared error (RMSE)**	**Absolute error standard deviation (AESD)**
Linear regression – M5	0.86	1.12	0.72
Linear regression – Greedy	0.84	1.10	0.71
MLP – 1	0.83	1.03	0.61
MLP – 2	0.83	1.09	0.71
SVR – polynomial	0.81	1.05	0.68
SVR – RBF kernel	0.86	1.08	0.65
SVR – PUK	0.86	1.07	0.64

As can be observed from Table 3, the accuracies for all approaches are in similar ranges. Based on the previous results, the multilayer perceptron with only one hidden neuron was the best performing model considering the root mean squared error, but the overall differences are not significant due to the small corpus size. At closer inspection, the model learned only to predict passing scores, since the dataset is unbalanced towards higher scores. Because of this limitation, a MLP network with 2 hidden neurons within the hidden layer was more suitable for interval selection because it has a higher variance and score coverage, although it was not the best performing model with regards to MAE.

5.3 Finding the optimal interval

The previously selected MLP network with 2 hidden neurons was tested with different intervals on the test partition by varying the lower limit and upper limit of interval I in order to select the optimal solution comprising of both high accuracy and a remarkable effort reduction from the teacher.

The critical score threshold T for deciding about pass or fail was set to 6. For a wide range of intervals, the values of Preision, Recall and Global, respectively, were calculated with equations (4)-(6). Lower bounds and upper bounds of the interval I were independently changed with increments of 0.1. Figure 4 shows the results of Precision as a function of lower bounds (O_{lower}) and upper bounds (O_{upper}) of the interval.

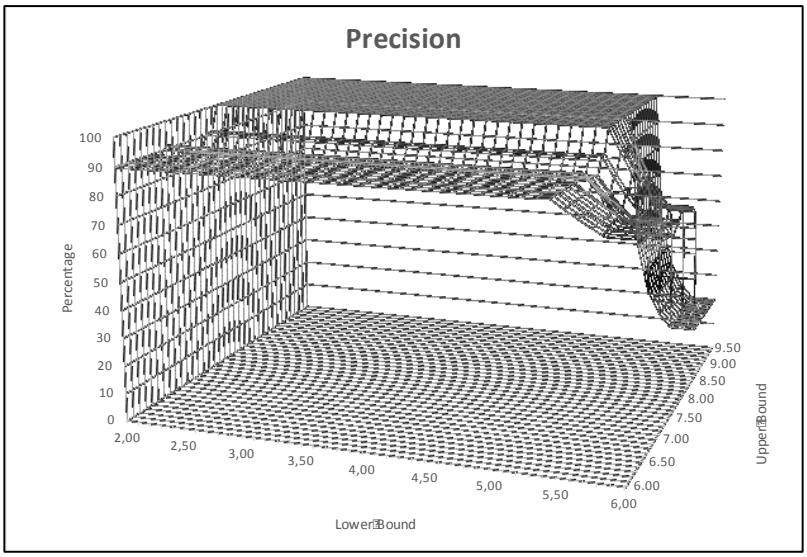

Figure 4. Precision of score prediction as a function score interval bounds.

As a result of the smaller sample and the bias in the scores towards higher scores, some intervals are not represented by predictions, where the precision reaches 100% (outside the 5.4-8.5 interval). Likewise, the Recall metric in this interval is zero.
The results of the Global optimisation metric are presented in figure 5.

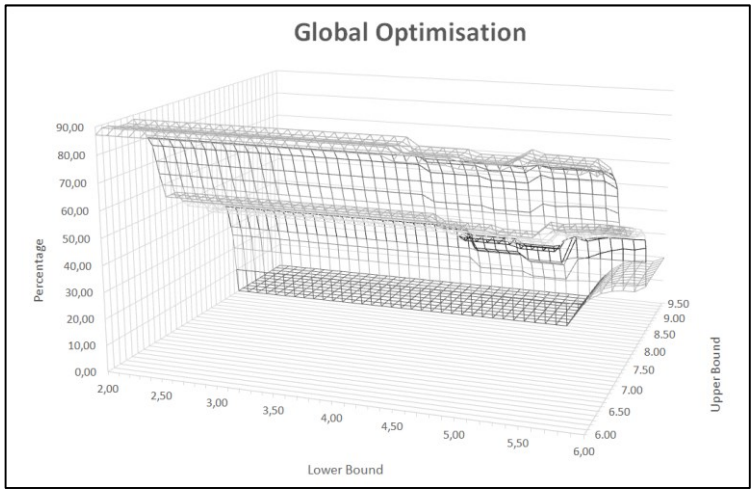

Figure 5 - Global prediction optimisation metric as a function of score interval bounds.

For the Global metric (Figure), the best results are obtained for the interval [4.4,6.1] (88.9%). Lowering the lower bound does not change this value because there were no essays scores below 4.4 in the test set. The upper bound is almost the same as the pass threshold (6), showing that the model can easily predict passing scores: this is inherently influenced by the fact that the majority of observable examples had scores above the critical threshold T=6.

5.4 Teachers´ workload reduction

Once the interval for manual scoring is set, the teacher´s reduction of workload can be calculated by taking into account the score distribution in the sample. The Global-based interval [4.4,6.1] is found to include 5 reports out of 34, which means that the workload would be reduced to 5/34=15%. However, as will be shown, this goes at a price. As explained in section 4, the value of Precision would be the most relevant metric for estimating validity. Precision of individual observations O can be directly derived with equations (1)-(3) and the resulting probabilities as illustrated in figure 3. For this purpose, the error of prediction ΔO

was set equal to the root mean squared error of the selected MLP model (cf. Table 3): ΔO=1.19. Figure 6 presents the Precision of the reports as a function of report score. As can be readily concluded, the precision of observed scores that are outside the interval [4.4, 6.1], which are the ones to be scored automatically, show overall variation between 0.6 and about 1.0.

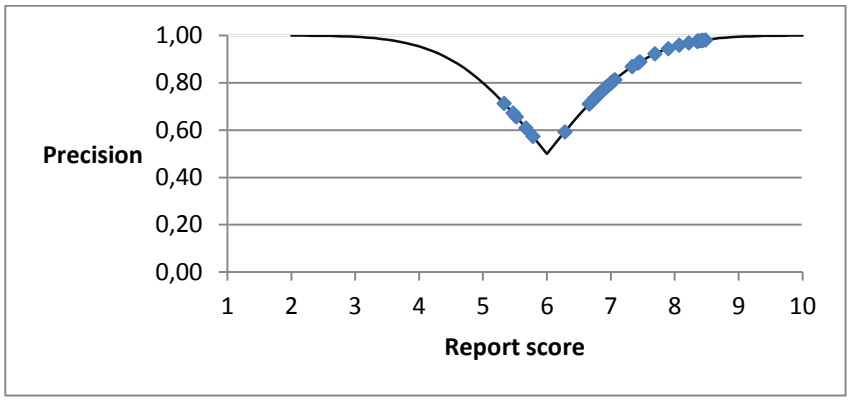

Figure 6. Calculated precision of reports in the test set as a function of report score.

The average precision of automatically rated reports in the sample, that is, outside the selected interval [4.4, 6.1], turns out to be 0.85 with a standard deviation of 0.11. It shows that although the Global optimisation method displays an acceptable overall precision of 0.85, quite some of the reports that are selected for automated assessment are subject to substantially, if not unacceptably, lower precision.

5.5 Meeting precision requirements

For preserving precision across the entire sample the selection process should be guided by a pre-defined minimum requirement of Precision. As can be observed directly from figure 6 above, checking for reports with higher Precision inevitably leads to larger intervals and consequently reduced numbers of automatically assessed reports. In figure 7, the relative number of manually assessed reports (viz. teacher workload) in the sample is presented as a function of Precision.

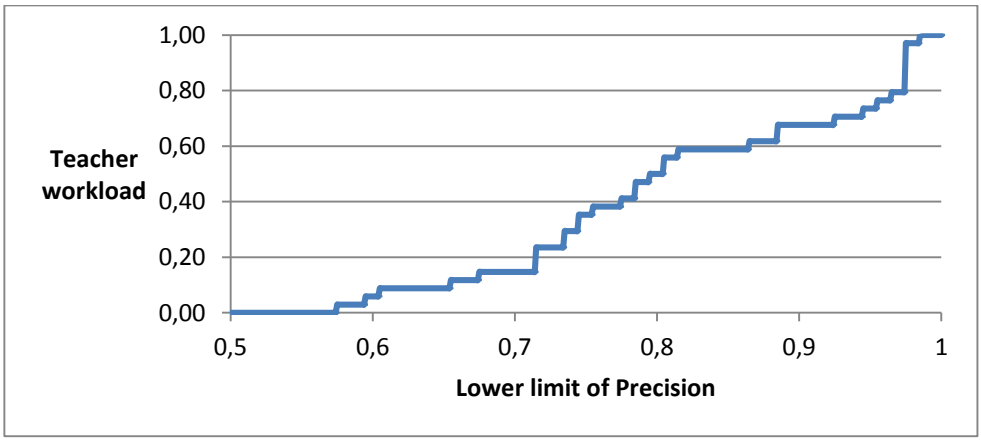

Figure 7. Workload reduction against pre-set lower limit of precision.

It appears that substantial workload reduction can be achieved, while preserving high Precision. For instance, a lower limit of Precision of 0.95 would already reduce the teacher's workload to 74%; a lower limit of 0.90 Precision would reduce workload to 68%; lowering Precision to 0.80 produces a workload reduction of 50%.

6. Discussion and conclusion

This paper presented an Automated Essay Scoring methodology that balances off validity and teacher work load reduction. Essentially, the methodology uses probability theory to estimate and control the validity of score prediction. By taking into account probabilities substantial work load reduction can be identified while validity is preserved. It opens up opportunities to extend fixed choice game patterns, such as quiz-based games or multiple choice-based scenarios with assignments that require textual expression. Moreover, in contrast to previous analyses performed using E-rater (Liu et al., 2016; Gerard et al., 2016), we rely on open-source alternatives that can be tailored for specific analyses in different languages. A wider application to online learning environments is feasible as well.
What validity level (Precision) should be considered acceptable is an open question. In our case we have postulated the teacher-assigned score as the 100% valid reference score. It should be noted, however, that in practice any assessment procedure aimed at assessing latent variables is subject to uncertainties. Even human teachers or other human experts are fallible and will never be able to attain full certainty. It is quite likely that validity of assessments will never exceed the level of 80% or 90%. This establishes the relevance of current study, as it favours approaches like the one presented.

However, the study is not without limitations. First, the dataset went with some restrictions. As was noted before, the dataset was biased toward positive scores and contained only few negative (fail) scores, which slightly restricted the investigations. Moreover, the optimisation metrics showed some intervals with slightly abnormal data due to the numerical instability of the multi-layered neural network trained on a rather small dataset. Also, the model choosing stage offered limited clues for model selection as only differences between different automated essay scoring approaches appeared minor (cf. Table 3). In addition, complex methods will probably perform better when trained on a larger collection of essays. Second, there are some statistical constraints. The assumption of normality of the data is a crucial one in our approach and testing for it cannot be omitted. Also, it is assumed that the error of prediction is constant across the full range of scores. Deviations from this assumption would lead to more complex calculations. Third, current study was not performed in real time. It used the recorded dataset of reports for carrying out post-practice calculations, which allowed to use teacher marks as a reference. Now that the model is in place it could run almost in real time. The NLP software service, which is being made available by RAGE can be installed and used without charge, typically returns score predictions of the reports within seconds. For automated scoring of other reports or open answers a new assessment model should be developed and validated first by re-applying the machine learning procedure to new sets of training data and test data.

Altogether, the successful application of automated essay scoring or even the assessment of automatically transcribed audio responses opens up new opportunities for teaching and learning that exploit the productive effects of self-expression and at the same time reduces teacher work load and operational costs.

References

Birenbaum, M., & Tatsuoka, K.K. (1987). Open-Ended Versus Multiple-Choice Response Formats - It Does Make a Difference for Diagnostic Purposes. *Applied Psychological Measurement*, 11(4), 385–395.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Born, R. (2015). 3 Key Facts about E-rater & Automated GRE Essay Scores. Online article retrieved from http://blog.powerscore.com/gre/e-rater-automated-essay-scoring-gre

Buckingham Shum, S., & Ferguson, R. (2012). Social Learning Analytics. *Educational Technology & Society*, 15(3), 3–26.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32*(1), 13–47.

Burstein, J., Tetreault, J., & Madnani, N. (2013) The E-rater Automated Essay Scoring System. In: Mark D. Shermis, M. D., & Burstein, J. (2013). *Handboek of Automated Essay Evaluation: Current Applications and New Directions* (pp. 55-77). New York: Routledge.

CGN Consortium. (2017). e-Lex, lexicale databank (lexical database). Retrieved April 2017, from Instituut voor Nederlandse Taal http://tst-centrale.org/en/tst-materialen/lexica/e-lex-detail

Cronbach, L.J., & Gleser, G.C. (1959). Interpretation of reliability and validity coefficients: Remarks on a paper by Lord. *Journal of Educational Psychology*, 50(5), 230 -237.

Crossley, S.A., Dascalu, M., Trausan-Matu, S., Allen, L., & McNamara, D.S. (2016). Document cohesion flow: Striving towards coherence. In: A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 764-769). Austin, TX: Cognitive Science Society.

Dascalu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* (Vol. 534). Cham, Switzerland: Springer.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 345–377). Cham, Switzerland: Springer.

Dascalu, M., McNamara, D.S., Trausan-Matu, S., & Allen, L.K. (2017). Cohesion Network Analysis of CSCL Participation. *Behavior Research Methods*, 1–16. doi: 10.3758/s13428-017-0888-4

Dascalu, M., Trausan-Matu, S., McNamara, D.S., & Dessus, P. (2015). ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395–423.

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017). ReaderBench Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch. In E. André, R. Baker, X, Hu, M.M.T. Rodrigo, & B. du Boulay (Eds.), *18th Int. Conf. on Artificial Intelligence in Education (AIED 2017)* (pp. 52–63). Wuhan, China: Springer.

Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J., & Vapnik, V. (1997). Support vector regression machines. *Advances in neural information processing systems*, 155–161.

Frank, E., Hall, M.A., & Witten, I.H. (2016). The WEKA Workbench *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (4th ed.). Online: Morgan Kaufman.

Funk, S.C., & Dickson, K.L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, 38(4), 273-277.

Galley, M., & McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. In G. Gottlob & T. Walsh (Eds.), *18th International Joint Conference on Artificial Intelligence (IJCAI'03)* (pp. 1486–1488). Acapulco, Mexico: Morgan Kaufmann Publishers, Inc.

Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., & Linn, M. C. (2016). Automated guidance for student inquiry. *Journal of Educational Psychology*, *108*(1), 60.

Gibbs, W.J. (1995). An Approach to Designing Computer-Based Evaluation of Student Constructed Responses: Effects on Achievement and Instructional Time. *Journal of Computing in Higher Education*, 6(2), 99-119.

Heim, A.W., & Watts, K.P. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology*, 37(3), 339–346.

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(2), 211–240.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215-233.

Mathan, S.A., & Koedinger, K.R. (2005). Fostering the intelligent novice: learning from errors with meta-cognitive tutoring. *Educational Psychology*, 40(4) 257–265.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective*, 10, 3-62.

Nadolski, R.J., & Hummel, H.G.K. (2016). Retrospective cognitive feedback for progress monitoring in serious games. *British Journal of Educational Technology*, 48(3), early online version, doi: 10.1111/bjet.12503

Nye, B.D., Graesser, A.C., & Hu, X. (2014). AutoTutor and family: a review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24, 427-469. doi:10.1007/s40593-014-0029-5

Ozuru, Y., Briner, S., Kurby, C.A., & McNamara, D.S. (2013). Comparing Comprehension Measured by Multiple-Choice and Open-Ended Questions. *Canadian Journal of Experimental Psychology*, 67(3), 215-227.

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

RBN (2018). Referentiebestand Nederlands. Retrieved from http://tst.inl.nl/producten/rbn/

Roscoe, R.D., & McNamara, D.S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4), 1010-1025. doi:10.1037/a0032340

Ruch, G.M., & Stoddard, G.D. (1925). Comparative Reliabilities of Five Types of Objective Examinations. *Journal of Educational Psychology*, 16(2), 89-103.

Saveski, G.L., Westera, W., Yuan, L., Hollins, P., Fernández Manjón, B., Moreno Ger, P., & Stefanov, K. (2015). What serious game studios want from ICT research: identifying developers' needs. In A. De Gloria and R. Veltkamp (Eds.), *Proceedings of the GALA 2015 Conference* (pp. 1-10), December 7-8, Rome, Italy, LNCS 9599. doi: 10.1007/978-3-319-40216-1_4

Schank, R.C., Berman, T.R., & Macpherson, K.A. (1999). Learning by doing. In C.M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory*, Vol. II (pp.161-181). Mahwah, NJ: Lawrence Erlbaum Associates.

Shute, V.J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, 55 (2), 503–524.

Shute, V.J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment.* Cambridge, MA: The MIT Press.

Van der Vegt, W., Westera, W., Nyamsuren, E., Georgiev, A. and Martínez Ortiz, I. (2016). RAGE Architecture for Reusable Serious Gaming Technology Components. International Journal of Computer Games Technology, doi:10.1155/2016/5680526. Retrieved from http://www.hindawi.com/journals/ijcgt/2016/5680526/ .

Verpoorten, D., Castaigne, J.-L., Westera, W., & Specht, M. (2014). A quest for meta-learning gains in a physics serious game. *Education and Information Technologies*, 19, 361-374. doi: 10.1007/s10639-012-9219-7

Westera, W. (2015). Games are motivating, aren´t they? Disputing the arguments for digital game-based learning. *International Journal of Serious Games*, 2(2). Retrieved from http://journal.seriousgamessociety.org/index.php

Westera, W. (2017). Why and how serious games can become far more effective. *Education, Technology and Society*. Accepted for publication.

Westera, W., Nadolski, N. & Hummel, H. (2014). Serious Gaming Analytics: What Students´ Log Files Tell Us about Gaming and Learning. *International Journal of Serious Games*, 1(2), 35-50. Retrieved from http://journal.seriousgamessociety.org/index.php

Wiley, D.A. (2002). The Instructional Use of Learning Objects, Bloomington, In: *Agency for Instructional Technology*. Retrieved from http://reusability.org/read/#1

Wresch, W. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition, 10*(2), 45–58.

Zhang, M. (2013). Contrasting Automated and Human Scoring of Essays. *Educational Testing Service R&D Connections*, 21, 1-11. Retrieved from http://origin-www.ets.org/Media/Research/pdf/RD_Connections_21.pdf