Free-text input vs menu selection: exploring the difference with a tutorial dialogue system.

Jenny McDonald[1], **Alistair Knott**[2] **and Richard Zeng**[1]
[1]Higher Education Development Centre, University of Otago
[2]Department of Computer Science, University of Otago
jenny.mcdonald@otago.ac.nz

Abstract

We describe the in-class evaluation of two versions of a tutorial dialogue system with 338 volunteers from a first-year undergraduate health-sciences class. One version uses supervised machine-learning techniques to classify student free-text responses; the other requires students to select their preferred response from a series of options (menu-based). Our results indicate that both the free-text and menu-based tutors produced significant gains on immediate post-test scores compared to a control group. In addition, there was no significant difference in performance between students in the free-text and menu-based conditions. We note specific analysis work still to do as part of this research and speculate briefly on the potential for using tutorial dialogue systems in real class settings.

1 Introduction

In large undergraduate classes (1500-1800 students), it is time-consuming, costly and seldom practical to provide students with feedback on their conceptions other than by computer-based marking of formative assessments. Typical examples of this include Learning Management System (LMS) based multiple-choice quizzes or similar. Most computer-assisted assessment involves students being able to recognise a correct response rather than recall and independently generate an answer. In the context of the first-year undergraduate health sciences course that we studied, currently all computer-assisted assessment takes this form. In 2008, the coordinator of a first year undergraduate health sciences class asked us about ways in which technologies might assist students to practice writing short-answer questions. As a result of this request, we wanted to investigate whether students answering questions with free-text or multiple-choice(menu-based) selections in a tutorial dialogue setting would result in performance gains on student test scores and whether there would be any difference in performance between students who generated free-text answers and those who selected their preferred answer from a number of options. In the next section we begin with a brief literature review from the fields of both Education and Cognitive Science. Next, we briefly describe the design and features of our tutorial dialogue system. The experimental design, and results are described in subsequent sections and we conclude with a discussion of our key findings.

2 Background

This study is situated at the boundaries between at least three established fields of inquiry: educational assessment research; psychological research, in particular the study of memory, recognition and recall; and finally intelligent tutoring systems (ITS) and natural language processing (NLP) research.

Since the 1920s the positive benefits on student performance of answering practice, or formative assessment, questions have been demonstrated in classroom studies (Frederiksen, 1984). Similar positive effects have been demonstrated in psychology laboratory studies since the 1970s. (McDaniel et al., 2007) Large meta-analytic educational studies looking at the impact of practice tests on student outcomes indicate that on average, the provision of practice assessments during a course of study does confer a clear advan-

Jenny Mcdonald, Alistair Knott and Richard Zeng. 2012. Free-text input vs menu selection: exploring the difference with a tutorial dialogue system. In *Proceedings of Australasian Language Technology Association Workshop*, pages 97−105.

tage, although the effect of increasing practice-test frequency is less clear. (Crooks, 1988). More recently, the role for computer-based assessment has been reviewed and Gipps (2005) writing in Studies in Higher Education has noted that,

> the provision of feedback to the learner, both motivational (evaluative) and substantive (descriptive), is crucially important to support learning. The developments in automated diagnostic feedback in short answer and multiple-choice tests are therefore potentially very valuable. If feedback from assessment could be automated, while maintaining quality in assessment, it could certainly be a powerful learning tool.

She goes on to say that use of computer-marking for anything other than MCQ-style questions, while showing some promise, is seldom used in practice in higher education institutions.

Recent research from the Cognitive Science and ITS domain, for example Chi (2009) and VanLehn (2011), suggests that tutor behaviour, human or machine, which encourages or promotes constructive or interactive behaviour by the student is likely to yield greater learning gains than passive or active behaviour. It also suggests that opportunities for extended interactive dialogue between teacher and student in a given domain are likely to produce the largest gains.

On the basis of this considerable body of research we felt that an opportunity to practice answering questions with formative feedback, in this case in a tutorial dialogue setting, should produce learning gains over and above those expected from working with existing study resources and formative tests. We were also interested to test whether there is a difference in performance between students who generate free-text responses and those who select an answer from a series of options in the course of a tutorial dialogue. There is some literature which specifically explores this, however the number of studies is limited and the results are inconclusive. Gay (1980) found that in retention tests students who practiced answering short-answer (SA) or free-text questions performed as well as or better than students who practiced MCQs but this effect was also dependent on the mode of retention testing. Specifically, retention test results where the test was conducted using SA were better for both MCQ-practice and SA-practice, whereas there was no difference between the two practice groups where the retention test mode was MCQ. In 1984, reviewing the education and psychology literature at the time, Frederiksen (1984) concluded that,

> testing increases retention of the material tested and that the effects are quite specific to what was tested. There is some evidence that short-answer or completion tests may be more conducive to long-term retention.

In a related area in his comprehensive review of classroom evaluation practice, Crooks (1988) suggested,

> there is no strong evidence...to support widespread adoption of any one [question] item format or style of task. Instead, the basis for selecting item formats should be their suitability for testing the skills and content that are to be evaluated.

Support for this view is found in a meta-analysis of 67 empirical studies which investigated the contruct equivalence of MCQ and constructed-response (SA) questions (Rodriguez, 2003). Where the content or stem of the MCQ and short-answer questions were the same Rodriguez found a very high correlation between the different formats. In other words, where the questions relate to the same content they will measure the same trait in the student. However, even if the same traits are measured by performance on questions in different formats, this says nothing about whether using practice questions in different formats results in differential learning gains for the students on subsequent retention tests.

The closest studies to our current work examined the impact on student performance of constructing or generating free-text descriptions vs. selecting descriptions from a series of options in a Geometry Tutor (Aleven et al., 2004) and an Algebra Tutor (Corbett et al., 2006). The results from both these studies suggest that there is little difference between the two formats especially on immediate post-test but that the free-text option may yield some advantage for long-term retention and some benefit for performance in subsequent short-answer questions. These results are

consistent with the much earlier educational review conducted by Frederiksen (1984).

In real-class settings, there is considerable time and effort involved in developing and implementing tutors which can provide immediate feedback on student short-answer responses and in particular, natural language dialogue systems (for example, Murray (1999)). This means that it is crucially important to understand what the potential benefits of these systems could be for both students and teachers.

The tutor we describe in the next section is substantially different from the Geometry and Algebra Tutors. Unlike these systems, it is not a formal step-based tutor; that is, it is not asking students to explain specific steps in a problem-solving process and providing feedback at each step. Our Dialogue Tutor simply engages students in a conversation, much like an online chat-session, where the Tutor poses questions which are directly related to students' current area of study about the human cardiovascular system and the student either types in their response or selects a response from a series of options. Nevertheless, in common with other ITS, our tutor does provide immediate formative feedback to the student and offers a series of options for proceeding depending on the student response.

3 Natural Language Tutor Design

3.1 Tutorial dialogue design

The structure of the tutorial dialogue is determined entirely by the dialogue script. We wanted to use a finite-state model for the dialogue since this permits an organic authoring process and imposes no theoretical limit to how deep or broad the dialogue becomes. The script structure is based on Core and Allen's (1997) dialogue coding scheme and has been described previously (McDonald et al., 2011).

The current study utilises a single-initiative directed dialogue; however the opportunity for limited mixed-initiative is incorporated into the system design through classifying question contributions at any stage of the dialogue and searching for possible answers within the dialogue script.

Design of the tutorial dialogue began with the development of an initial script covering the curriculum on cardiovascular homeostasis. This was developed in close consultation with course teaching staff and was written by a medical graduate using lecture notes, laboratory manuals and self-directed learning material from the course itself. The initial script was refined through a series of pilot interactions with student and staff volunteers and released to the first year undergraduate class on a voluntary basis at the beginning of their module on the human cardiovascular system. The default position in this early script was to provide the correct answer and move on unless confidence was high that an appropriate match had been made, using minimum-edit distance between student response and model answers. A handful of dialogues were interrupted because of system-related problems but the majority that terminated before completion did so because the students simply ended their session. Feedback from course tutors and comments from the students supported our intuition that poor system 'understanding' of student dialogue contributions was a key reason for the fall-off in use. Nevertheless, student perceptions of this early tutorial were broadly positive and it served its purpose in capturing a reasonable quantity of student responses (between 127-242 responses to 50 tutorial questions) for the next stage of tutorial dialogue development.

The next step in dialogue development involved building classifiers for each dialogue contribution from the student corpus and revising the script depending on the nature of student responses. We followed the XML schema of the NPSChat corpus provided with the NLTK (Bird, 2006) in marking-up the corpus. The classes used are specific to each dialogue contribution although three generic classes are used throughout the dialogue where context-specific classification fails: *question*, *dont-know* and *dont-understand*. A flow diagram of the classification process is illustrated in Figure 1: There is a classifier for each dialogue contribution (DC-Classifier). A-D represent possible classes for student input. If classifier confidence falls below a certain threshold for assigning input to one of the possible classes then the unclassified input is passed on to a series of generic binary classifiers: Question, Dont-know and Dont-understand which identify whether the input string is likely to be a question (Q) or some variation on 'I don't know' (DK) or 'I don't understand the question' (DU). If the input remains unclassified after each of these generic classifiers has been tried, the dialogue moves to the next de-

fault node in the script (Default).

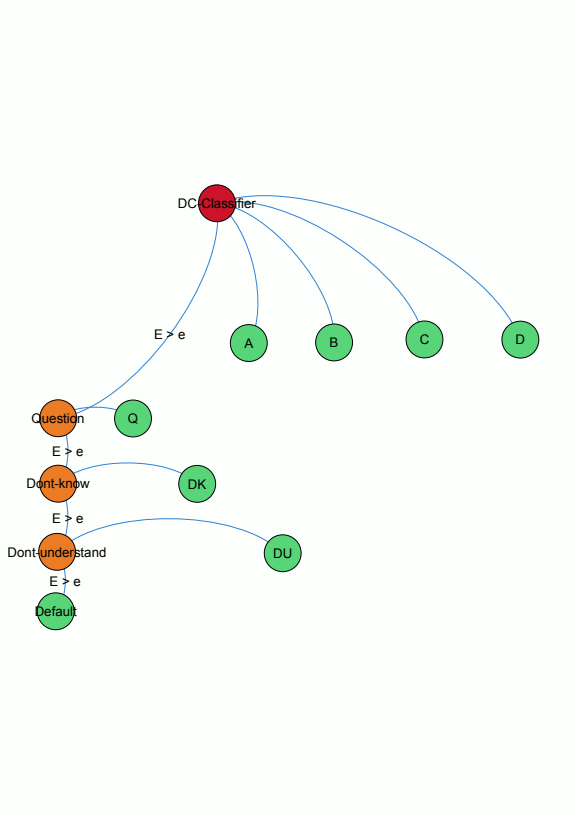

Figure 1: Classifier flow diagram.

For each dialogue context a training set was created from the corpus. Typically the first 100 student responses for each tutor question were classified by a human marker. This training set was divided into 5 folds and a Maximum Entropy classifier trained on 4/5 folds using simple *bag of words* as the featureset and then tested on the remaining fold. A 5-way cross-validation was carried out and accuracies for each of the 5 test sets calculated. The average accuracy across the 5 test sets and standard deviation was also recorded. This process was repeated using different combinations of featuresets (for example, bag of words, word length, first word, with/without stemming, with/without stopwords etc) until the highest accuracy and least variability in test set results was achieved. (The mean accuracy on test data across all 62 classifiers built for each dialogue context in the complete dialogue was 0.93. The minimum was 0.73, maximum was 1.00 and the first quartile was 0.89)

Tutor questions with multi-part answers, for example, *'Can you think of three main factors which affect cardiac contractility?'*, lent themselves to chaining together a series of binary classifiers, using the NLTK MultiBinary Classifier wrapper, rather than including all possible classes of response within a single classifier. This is the best approach given the relatively small amount of training data compared to the large number of possible classes of response. For example, in the question given above, three possible factors to list gives a total of eight possible classes of answer. For some combinations there are only very small, but nevertheless important training sets, and this leads to poor classfier performance overall. Training three binary classifiers which identify each of the factors sought as either present or not and then writing a function which returns a list of the factors found by all three binary classifiers for a given input effectively increases the amount of training data per factor. While this approach yielded some improvement, the *class imbalance problem* (Refer to, for example, Japkowicz(2000)) was still evident for some combinations.

The classifier is evaluated with previously unseen data and scored relative to a human marker. The entropy of the probability distribution (E) is calculated for each unseen response and this is used to determine appropriate thresholds for classification. For example, if E is close to zero the classifier confidence is generally very high. E $>$ 1 indicates low confidence and less difference between the class rankings. An appropriate entropy threshold (e) for each classifier is determined by the script designer. This is really a subjective judgement and is made based on the classifier performance as well as the dialogue script context and the likely impact of a false negative or false positive classification. (The mean accuracy on unseen test data across all 62 classifiers with manually set entropy thresholds was 0.95. The minimum was 0.70, maximum was 1.00 and the first quartile was 0.93) There is the potential to automate this process however this will require a method to assess the cost of false positive and false negative classification for each dialogue context.

Finally the classifier is serialised, along with its associated feaureset parameters and e value and saved for use in the dialogue system itself.

3.2 Dialogue system architecture

The dialogue system is written in Python and utilises several NLTK libraries, Peter Norvig's 'toy' spell checker, and the Asyncore and Asynchat libraries to manage multiple simultaneous client connections. The server can readily communicate with any web-application front end using XML-formatted messages and we have built a java-based web application through which multiple clients can connect to the tutorial server. Figure 2. provides an overview of our system architecture.

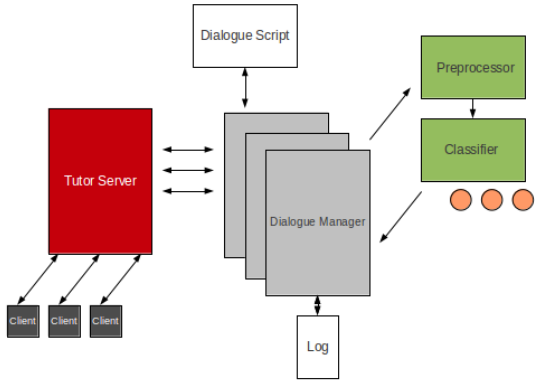

Figure 2: Architecture of Dialogue System.

Each client connection to the system creates an instance of the dialogue manager which sends tutor contributions to the client according to the preloaded script and receives student contributions which are then classified and determine the next tutor contribution. The dialogue manager design has been described previously (McDonald et al., 2011).

3.3 Free-text and menu based versions

A small addition to the dialogue script and the addition of a switch to the dialogue manager code allowed us to create two systems for the price of one. The free-text entry system uses the classifiers to categorise student responses and the menu-based system simply presents students with possible options added to the script from which they can select. The addition of <menu> tags to each dialogue context in the script is shown in the following example:

```
<contribution-node id="check-hr"
parent-node="start"
default="true">

<backward class="yes">
<acknowledge/>
</backward>

<forward>

<assert>We're going to talk about
what blood pressure is....
</assert>

<info-request value="How would you
check what someone&apos;s HR is?"
define="You could take their
pulse."/>

<menu>
<item value="correct">Count the
pulse.</item>
<item value="simpler">With a
blood pressure cuff and
stethoscope.</item>
<item value="simpler">Use an
ECG.</item>
<item value="incomplete">Pulse.
</item>
<item value="dont-know">I don't
know.</item>
</menu>

</forward>
</contribution-node>
```

Note that the menu options, like the classifier training data, are derived directly from student responses to the question. There are three things to note from this. Firstly, the menu-based options tend to have a slightly different flavour to conventional multiple-choice questions (MCQs) designed by teachers. For example, the incomplete response, *'Pulse'* would probably not be included in a conventional MCQ. It is here because this was a common response from students responding with free-text and resulted in a scripted reminder to students that short-answers do require complete descriptions or explanations. Secondly, *'I don't know'* is unlikely to be found in a teacher designed MCQ; however in this context it is useful and leads to scripted remedial action as it would if the student had typed in text with a similar meaning. Finally, two different options result in the same script action, labelled *'simpler'*, being taken. This reflects the free-text student data for this question. Both are acknowledged as possible ways to check someone's heart-rate, in either

case, the student is prompted to think of a simpler method.

4 Experimental Design

Students from the 1st year Health Sciences course (N=1500) were asked to volunteer for the experiment. The first year Health Sciences course is a prerequisite for all professional health science programmes, such as Medicine, Dentistry, Pharmacy, ... Entry into these programmes is highly competitive and is dependent, amongst other things, on students achieving excellent grades in their 1st year courses. The only incentive offered to students was that taking part in the study would give them an opportunity to practice and develop their understanding of the cardiovascular section of the course by answering a series of questions related to the lectures they had received during the preceeding two weeks. Students were told that different styles of questions, short-answer and MCQ, might be used in different combinations and that not all students would receive the same style of questions. They were also told to allow 20-40 minutes to complete the questions. They could answer the questions by logging in to an online system at anytime during a three-week period which ran concurrently with their normal laboratory and self-paced study sessions on the cardiovascular system.

All student responses in the experiment were anonymised and appropriate ethics approval was obtained.

Students who logged into the system were randomly assigned to one of three conditions: A free-text condition where they completed a pre-test, then the free-text version of the tutorial dialogue, and concluded with an immediate post-test; a menu-based condition where they completed a pre-test, then the multi-choice version of the tutorial dialogue, followed by an immediate post-test, and a control condition where they simply completed pre- and post-tests.

The pre- and post-tests in each case consisted of equal numbers of MCQ and short-answers (3+3 for the pre-test and 7+7 for the post-test). The pre-test directly reflected material taught in the lectures students had just received and the post-test reflected material explicitly covered in the tutorial dialogues.

All student interactions with the system in each experimental condition were recorded and logged to a database. At the end of the experimental period only completed sessions (i.e. pre-test, post-test and tutorial condition completed) were included for analysis. The principal investigator marked all pre- and post-test short-answer questions and MCQs were scored automatically. One member of the teaching staff for the course also marked a sample of short-answer questions to check for inter-rater reliability.

Given the findings from the literature reported in Section 2, the hypotheses we wanted to test were: A. Any intervention results in better post-test performance than none; B. Free-text input results in better post-test performance overall than MCQ, because there is something special about students recalling and constructing their own response; C. Free-text tutorials lead to increased performance on short-answer questions; and D. MCQ tutorials lead to increased performance on MCQ questions.

Delayed post-tests are still to be completed and will involve correlation with short-answer and MCQ student results on the cardiovascular section of the final examination.

We describe our early results and analysis of the immediate post-test data in the next section.

5 Results

720 students logged into the experimental system during the 3 week period in which it was available. Of these, 578 students completed the session through to the end of the post-test and these were relatively evenly distributed across the three conditions suggesting that there are no sampling bias effects across conditions. We report here the results from the first 338 of the 578 completed tutorials/tests. Short-answer sections of both pre- and post-tests were checked for inter-rater reliability. A Cohen's kappa of 0.93 (p=0) confirmed very high agreement between 2 markers on pre- and post-test questions for a sample of 30 students.

Table 1 summarises the descriptive statistics for the three experimental conditions. Across all three conditions students performed well in the pre-test with a mean normalised score of 0.83. In the post-test, which was inherently harder, student scores dropped across all three conditions but the mean scores were higher in both the tutorial conditions compared to the control (0.75 and 0.77 c.f. 0.69).

	Control n=119		Free-text n=101		Menu-based n=118	
	mean	*sd*	*mean*	*sd*	*mean*	*sd*
Pre-test	0.83	0.15	0.83	0.14	0.84	0.14
Post-test	0.69	0.19	0.75	0.16	0.77	0.17

Table 1: Descriptive Statistics

The dependent variable to test our first hypothesis was taken as the difference between pre- and post-test performance for each student with the pre-test result serving as a common baseline in each case. The differences between pre- and post-test scores were normally distributed which allowed us to use parametric tests to see if there were differences between the means in each condition. A between subjects Anova gave an F value of 4.95 and a post-hoc Tukey multiple comparison of means at 95% confidence level showed a significant difference when compared with the control for both the free-text tutorial condition (p=0.03) and the menu-based tutorial condition (p=0.01) .

However, there was no support for our second hypothesis that free-text input results in better post-test performance overall than menu-based input; comparison between the mean scores for free-text condition and menu-based condition was not significant (p=0.94). Given this result it was also clear that there was no demonstrated benefit for free-text tutorials improving scores on free-text questions in the post-test nor multiple-choice questions improving post-test performance on the MCQs.

We discuss the implications of these early results in the final section and also outline our plan for further detailed analysis of the data obtained.

6 Discussion

Several features stand out from the results. The most striking initial feature is the much higher tutorial completion rate (80%) for this system compared with the original tutorial system (23%) which was fielded in order to collect student responses ((McDonald et al., 2011)) as discussed in Section 3. Formal evaluation of the free-text version classifier performance is currently underway and will be reported separately but the overall higher completion rate and only slightly lower numbers completing the dialogue tutorial (29% of the 578 completions) compared with the multi-choice tutorial (34% of the 578 completions) is suggestive of a considerable improvement.

On average, students performed better in the pre-test than they did in the post-test. This was expected: the pre-test was designed to measure the degree to which students had comprehended key points from the lectures they had just attended, while the post-test was designed to be more challenging. It is worth noting that in real in-class settings it is not uncommon for students to perform well in initial tests and subsequently perform less well as they work to make sense and meaning of the subject under study (see for example, Cree and Macaulay (2000)). However in this specific context, given that the pre-test was designed to measure the degree to which students had comprehended key points from the lectures they had just attended it is not too surprising that they did uniformly well in the pre-test. The post-test was designed to be more challenging and required an ability to demonstrate understanding as well as the ability to manipulate key cardiovascular variables and understand whether and how these relate to each other. These skills and abilities are developed through experience in the laboratory teaching sessions and with self-directed study materials; they are also directly covered in each of the tutorial conditions. Certainly the results confirmed that students in each condition started at a similar level and support our hypothesis that post-test performance is significantly improved through exposure to either tutorial condition when compared to the control condition.

In a practical sense it is important to see not only whether there are actual differences in performance but also whether these differences are large enough to be worth the additional effort for both teaching staff and students. Effect sizes are commonly reported in the educational literature and we believe it is worth doing so here. The standardised effect size is relatively small in each tutorial condition (0.17-0.22). Hattie (2008) and

many others make the point that in general an effect size larger than 0.40 is suggestive of an intervention worth pursuing but that this also depends on local context. In the context of this study, for the 'price' of a single relatively brief intervention, an average effect size increase of between 6 to 8 percentage points on the immediate post-test suggests that engagement with either tutorial, particularly in a high stakes course, where every percentage point counts, does produce a gain worth having. With such a brief one-off intervention it would be surprising indeed to have found much larger effect sizes.

Examination of the variability of pre- and post-test results in each of the three conditions shows a highly consistent distribution of marks in all three conditions on the pre-test but a wider variation in results in the post-test control group (sd=0.19) than in either of the tutorial groups (sd=0.16 in menu-based condition and sd=0.17 in free-text condition). Again, given that the post-test was specifically testing material taught in the tutorial this is perhaps not suprising. You would hope that in any teaching situation student marks would start to converge in a positive direction! Nevertheless, once the complete set of student results is marked we will investigate this further. Of particular interest is to see whether poorer performing students benefit more from the tutorial than others.

Finally, the lack of difference between the two tutorial conditions, free-text and menu-based, was consistent with indications from existing literature. However, we found no advantage for free-text entry over menu-based choice overall, nor indeed did either condition confer any advantage in performance when post-testing was in the same mode. However, given previous research results we are keen to explore this further. In particular we want to examine specific questions from the post-test and see whether there is a difference between conditions on questions which required simple recall and those which required further analysis or description by the student. We also intend to look at several other factors: whether the average length of written responses to the tutorial in the free-text condition has any bearing on performance in either condition, time on task relative to performance and the stage at which the student logged in to the experimental system. (For example, did the group which took the tutorial later, once they had more laboratory work and self-study time under their belts, perform better in either condition than those who took the tutorial earlier?)

Additional work still to do includes correlating these experimental results with student performance on relevant questions in the final course examination (short-answer and MCQ); this will effectively provide delayed post-test data. Also, we will be gathering student feedback on their experience and perceptions of the tutorial systems via a course evaluation questionnaire.

Developing a deeper understanding of the potential role of natural language tutorial dialogue systems in improving student performance has been the focus of this paper. Nevertheless a striking side-effect from undertaking this research has been realising the role dialogue systems like this may be able to play in providing feedback to teachers on the conceptions held by students in large classes about the material they are being taught. The range and depth of large numbers of student free-text responses provide important clues about student conceptions. The ability to describe these conceptions is invaluable for teaching (Marton and Saljo, 1976). The facility to do this in an automated or semi-automated way for large classes, presumably, is even more so. Teaching staff who have had some involvement in the project have commented on the usefulness of being able to see student responses to questions grouped into categories: this grouping provides a powerful way for teachers to gauge the range of responses which they receive to their questions.

References

Vincent Aleven, Amy Ogan, Octav Popescu, Cristen Torrey, and Kenneth Koedinger. 2004. Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In *In*, pages 443–454. Springer.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michelene T. H. Chi. 2009. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1):73–105.

Albert Corbett, Angela Wagner, Sharon Lesgold, Harry Ulrich, and Scott Stevens. 2006. The impact on learning of generating vs. selecting descrip-

tions in analyzing algebra example solutions. In *Proceedings of the 7th international conference on Learning sciences*, ICLS '06, pages 99–105. International Society of the Learning Sciences.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA, November.

Viviene Cree and Cathlin Macaulay. 2000. *Transfer of Learning in Professional and Vocational Education*. Routledge, London: Psychology Press.

Terence J. Crooks. 1988. The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4):pp. 438–481.

Norman Frederiksen. 1984. The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39(3):193–202.

Lorraine R. Gay. 1980. The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, 17(1):45–50.

Caroline V. Gipps *. 2005. What is the role for ict-based assessment in universities? *Studies in Higher Education*, 30(2):171–180.

J. Hattie. 2008. *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement*. Taylor & Francis.

N. Japkowicz. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI2000)*, volume 1, pages 111–117.

F. Marton and R. Saljo. 1976. On qualitative differences in learning: Ioutcome and process*. *British Journal of Educational Psychology*, 46(1):4–11.

Mark A. McDaniel, Janis L. Anderson, Mary H. Derbish, and Nova Morrisette. 2007. Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19(4-5):494–513.

J. McDonald, A. Knott, R. Zeng, and A. Cohen. 2011. Learning from student responses: A domain-independent natural language tutor. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, volume 148, page 156.

Tom Murray. 1999. Authoring intelligent tutoring systems: An analysis of state of the art. *International Journal of Artificial Intelligence in Education*, 10:98–129.

Michael C. Rodriguez. 2003. Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40(2):163–184.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.