Research Funding Proposal for Phase 1 of 3:

Combining Reliability and Validity in Assessment: Using Auditing to Create the Golden Goose in Assessment Quality

Abstract

We seek £66,375 of first phase funding for the development of a piece of research software which keeps an audit trail of activities performed during academic writing e.g. what parts of what pages are viewed in the web browser, what searches are made, what is typed, copied and pasted and where and when and so on. We believe that audit trails would be invaluable not just as a means for better understanding student and research work patterns and behaviour, but also to enable development of cross-comparative statistical testing to improve assessment reliability without impacting validity of learning objectives.

About Us

The principal researcher will be Niall Douglas. He holds two undergraduate degrees, one in Software Engineering and the other in Economics and Management. He also holds a Masters degree in Business Information Systems and is currently undertaking a further Masters degree in Educational and Social Research. He has a strong history in the development of innovative computer software, having contributed substantially to multiple projects for over fifteen years as well as participating in the ISO engineering standards process. He also has industrial in addition to his academic experience, having worked in small companies and large multinational companies. As a higher education teacher, he additionally has experience in curriculum and assessment design and implementation.

Research Questions for Phase 1 of 3

1. Can an audit trail be *usefully* (i.e. with respect to assessment) generated for electronically generated academic outputs?
2. Can such audit trails be used to compare outputs across students/authors, subjects and disciplines?

Research Questions for Phase 2 of 3 (for information only)

1. Can audit trails help identify which assessed outputs received an unusually unreliable assessment, and thereby aid quality?
2. Can audit trails empower students and staff to make better decisions and improve their learning through reflective feedback about the practice of others?
3. How common is plagiarism? And how does plagiarism actually affect long-term learning and grading?

Rationale

The quality of written assessment reliability (i.e. that grades awarded to non-numerate student written outputs are consistent across marker and institution) in higher education has been repeatedly shown by studies since the 1930s to be poor (Hartog, Rhodes and Burt, 1936;Cox, 1967;Branthwaite, Trueman and Berrisford, 1981;Elton and Johnston, 2002). Yorke, Bridges, & Woolf (2000) showed that reliability is low between coursework and examinations even with the same marker; and Briggs (1980) reported a particularly interesting finding of differences in grade between identical English papers having almost as much dependence on handwriting style as content. In

case one might think that students are not aware of this discrepancy, it has been found to feature highly in student dissatisfaction surveys (Yorke, 2000).

However, little has been done in British academia about the poor quality of reliability as it has been seen as the price to be paid for high quality assessment *validity*[1] (Elton, 1982;Elton and Johnston, 2002). In other words, if quality of reliability were to be improved, it is assumed that assessment would have to be restricted to the material on which different examiners agree – which would be biased against originality and creativity in the student output. As there is a well known "backwash" effect from the design of the assessment onto student learning outcomes, it is considered by most to be preferable to improve assessment validity at the expense of assessment reliability (Elton, 1982;Elton and Johnston, 2002).

There is of course the related problem of the effects of information technology on rates of plagiarism. The negative consequences of plagiarism upon student learning need little explanation. Using a strict definition of cheating, de Lambert, Ellen and Taylor (2002) found that 80% of New Zealand tertiary students cheat[2].

Interestingly, in the 1960s very similar problems afflicted the profession of Accounting which must also generate reliable and valid quantitative assessments of inherently qualitative value. As detailed in the history of the British accounting audit process by Matthews (2006), the introduction of systems rather than records auditing of mass market businesses in the 1960s transformed for the better what good practice is held to be. According to Matthews, this resulted from *enabling business **systems** to be compared against one another* and partial risk to be estimated, thus allowing detailed human examination of only those parts which need it. We believe that most of these auditing techniques are transferable outside the field of accounting, not least because they already have been applied to fields such as information and knowledge management where CISA has been the professional certification for information technology audit professionals since 1978.

In our opinion, present academic assessment techniques continue to ignore the opportunities that the near universal computerisation of non-exam student writing enables. In particular, there is no reason why the computer cannot record the student's activities, thus keeping an *audit trail* of how an assessed output was generated over time[3]. These audit trails have four main uses: (i) they allow the assessor to make use of the audit trail when a grade is on a boundary (ii) they allow statistical comparison and categorisation of audit trails, thus illuminating which student outputs are more likely to be problematic and therefore deserving of more intense human inspection (iii) they enable summative assessment to become formative by allowing a student to compare their work patterns with those of others (Sly and Rennie, 1999) and (iv) they reduce the potential for abuse and uncertainty by providing a method for independent vetting of both student and marker.

[1] i.e. that the assessment is seen to properly support the desired learning outcomes which in HE generally require an element of original thought.

[2] The most common offences being (in order): paraphrasing from a web site, book or periodical without referencing (42.1%); padding out a bibliography with references that were not actually used (38.4%); copying information directly from a web site, book or periodical with reference to the source but without quotation marks (36.8%); one student allowing another student to copy their assignment (34.7%); and copying from a web site, book or periodical without referencing (25.8%).

[3] If this sounds vaguely familiar as an idea in education and teaching methodology, it may be because the old Computer Managed Learning systems of the 1970s and 1980s also tracked students' progress through their (very tightly defined) syllabus and made use of that metadata for all sorts of feedback purposes. This idea fell out in favour of a more open, dynamic and participatory approach to the use of technology in Education since the 1990s – what we now call "e-Learning" – which pushed out the tracking feature (Reiser, 2001;Reiser, 2001), mostly due to the expense of storing large amounts of per user tracking data in a form which can be easily cross-referenced *en masse*. However, the very recent development of the *internet cloud* has reduced these costs to just US$0.15/Gb/month (Amazon S3).

Methods and Analysis

The principle goal of phase 1 of the research intends to discover if correlations can be found between academic output work patterns and assessment grades achieved, so the methodological framework to be used shall be observational whereby we observe existing practice. If phase 1 is successful, phase 2 will be mostly based on participatory and interventionist Action Research to see how we can positively affect quality. This document covers phase 1 only.

Pre-Experiment

Firstly software is written which captures everything a user does during the generation of an academic output, including activity *preceding* the beginning of the writing of the assessed output – see Appendix 2 for details. All data, including all data viewed and created, is stored *in its entirety* on a commodity cloud server. Each user of the software is securely identified using Janrain Engage[4] to ensure that their audit trail is genuine and to collect information on the user's social networking activities and that of their friendship networks. Before submitting a piece of coursework to their assessor, the user is presented with an audit trail generated from the data collected by the software. The audit trail permits records to be deleted but not otherwise modified before submission. This allows the user to filter out unnecessary or spurious records before submitting them to the assessor[5]. Our system retains the full audit trail however, including what the user decides to delete.

For the degrees at the School of Education, University of Wales Institute, Cardiff (UWIC) (there are approximately 2000 students, and we assume approximately 150 will opt in), an optional ten credit (5 ECTS) module is made available which students may utilise to fulfil their optional credit requirements. This module would require them to utilise this software during all their other coursework for that year, to attend classes which train them in the use of the software as well as establishing bidirectional dialogue about data the software is collecting, and to successfully do both to obtain the top grade for this module. The participants will be fully informed about the research before deciding to partake – this introduces some selection bias, but this is preferable to the alternatives for phase 1 because we are testing for *feasibility* before all else.

The staff have already agreed to mark coursework electronically by adding comments to submitted works. This allows their assessment actions to be audited as well. They also have agreed to add all supplied notes and suggested readings to the auditing system so these can be used for statistical analysis.

The Experiment

We don't know what factors might correlate with academic performance, so we have opted to collect everything possible during this initial phase and distil from there.

The first problem to be solved will be construction of the proper timeline and source interaction graphs. For example, if a user has a web browser tab showing a Wikipedia page, another tab showing JSTOR and a Microsoft Word document containing their partially completed essay, then it is straightforward to sequence operations per-web page and per-essay and cross link them in time appropriately into a highly reliable audit trail of how the work was gestated. However, if a user does a lot of pre-reading mixed in with other non-study web browsing a number of

[4] There have been a number of recent advances in *federated identity* which allows us to verify a person's identity using one or all of a person's social networking login accounts. For example, using Janrain's Engage service (http://www.janrain.com/products/engage), when linked one knows not just the contents of a person's Facebook, Google, Twitter, Windows Live, Yahoo, AOL, LinkedIn, Paypal and blogging profiles but also all their friends, all the activity of the person and all their friends, their birthday, home address and all other demographic information including any photos and movies of them as well as the contents and location history of their mobile phone. As universities increasingly deploy Google Apps and Email to their students and staff, this means that students and staff simply log into our service using their university credentials and if without a Google login, simply using their Facebook credentials.

[5] If using Microsoft Office 2007 or later, the audit trail is actually stored internally in the document.

weeks before even starting their essay, the problem of graph building is rather harder because determining where a student got an idea from is much more complex. One can see how it could be possible that the most conscientious and widely read students are the least easily representable in our data set, and we shall try to account for this by including what activities of each user we couldn't categorise or graph.

Answering Research Question 1:

Having formed gestation graphs per assessed work for each user, one must now try to elicit use patterns. One would assume that each person is likely to keep similar use patterns over time, so longitudinal analyses cohorted by year and module would search for repetition in use patterns. For example, student A might consistently prefer term searches over drilling down links in order to find suitable supporting journal article citations by say 20% over student B, and student C might consistently search for citations rather than terms. Finding such needles in haystacks can be brute forced by calculating all possible metadata from each person's graphs, and performing a Kullback-Leibler divergence test to determine which contain unique information (Cover, Thomas and Wiley, 1991) in order to reduce the set of variables to something manageable. One can then use an appropriate mixture of cross-correlation detection and removal, rank correlation coefficient analysis and Pearson's product-moment coefficient analysis as appropriate to elicit use patterns. A novel idea which we will test (where we have sufficient data) is to induct strange attractors via phase space analysis to elicit use patterns not findable using probabilistic analysis (Kyrtsou, 2005;Kyrtsou and Labys, 2006).

At this stage one ought to be able to group and categorise users according to common use patterns. We now bring assessments and grades awarded into the picture, so we try to regress assessment (the dependent variable) from use patterns (the independent variables) bearing in mind that many of the factors may also be cross-correlated, have non-linear relations or indeed substantial stochastic noise. This ought to illuminate which common use patterns have the most effect on assessed grades, and if we succeed we answer research question 1.

Answering Research Question 2:

We now need to compare across users and modules, so taking our set of use patterns, assessment grades and dependent factors we perform a series of cross-sectional studies across all assessed works looking for what factors other than use patterns might also have a correlation with grade. For example, one might hypothesise that high achieving individuals are more likely to be in a social group together, so one could discover to what extent this hypothesis might be true.

For obvious reasons, it is hard to be more specific without knowing the results of research question 1. However, given the much more limited search space in a cross-sectional study, it ought to be computationally feasible to employ automated searchers for testable hypotheses e.g. a genetic algorithm automated solver (De Rooij and Vitányi, 2011), a Kolmogorov structure finder (Vereshchagin and Vitányi, 2004), or even self-generating memetic algorithms which find contact boundaries between information structures (Krasnogor, 2004).

We plan to employ a segmented approach to the use of automated solver, so after determining a structure for a given assessed work we can then look for correlations between these structures. For example, one might hypothesise that there ought to be a strong consistency in structure of assessment per marker and a weaker one per department (where one would assume there is peer marking and other such assessment quality measures).

Time frame

June 2011 – September 2011: Writing of the software.

September 2011 – June 2012: Deployment of the software and ongoing collection of data.

June 2012 – September 2012: Refinement of data analysis and writing of phase one project report.

Ethics

We supply an example End User Licensing Agreement in Appendix 3 to which participants must agree in order to take part in this research. Crucially, we guarantee that any data collected can be edited before submission to assessors, that personal data can be deleted, and that we will not allow individuals to be identified without their prior permission.

Bibliography

Branthwaite, A., Trueman, M. and Berrisford, T. (1981) 'Unreliability of Marking: further evidence and a possible explanation', *Educational Review*, vol. 33, pp. 41--46.

Briggs, D. (1980) 'A study of the influence of handwriting upon grades using examination scripts', *Educational Review*, vol. 32, pp. 186--193.

Cover, T.M., Thomas, J.A. and Wiley, J. (1991) *Elements of information theory*, Wiley Online Library.

Cox, R. (1967) 'Examinations and higher education: a survey of the literature', *Higher Education Quarterly*, vol. 21, pp. 292--340.

de Lambert, K., Ellen, N. and Taylor, L. (2002) 'Prevalence of academic dishonesty in tertiary institutions: The New Zealand story', *New Zealand Journal of Applied Business Research*, vol. 1.

De Rooij, S. and Vitányi, P. (2011) 'Approximating rate-distortion graphs of individual data: Experiments in lossy compression and denoising', *IEEE Transactions on Computers*.

Elton, L. (1982) 'Assessment for learning', *Professionalism and flexibility in learning*, pp. 106--135.

Elton, L. and Johnston, B. (2002) 'Assessment in universities: a critical review of research'.

Gerrish, S.M. and Blei, D.M. (2010) 'A language-based approach to measuring scholarly impact'.

Hartog, S.P., Rhodes, E.C. and Burt, S.C. (1936) *The marks of examiners: being a comparison of marks allotted to examination scripts by independent examiners and boards of examiners, together with a section on a viva voce examination*, Macmillan.

Krasnogor, N. (2004) 'Self generating metaheuristics in bioinformatics: The proteins structure comparison case', *Genetic Programming and Evolvable Machines*, vol. 5, pp. 181--201.

Kyrtsou, C. (2005) 'Evidence for neglected linearity in noisy chaotic models', *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, vol. 15, p. 3391.

Kyrtsou, C. and Labys, W.C. (2006) 'Evidence for chaotic dependence between US inflation and commodity prices', *Journal of Macroeconomics*, vol. 28, pp. 256--266.

Matthews, D. (2006) *A history of auditing: the changing audit process in Britain from the nineteenth century to the present day*, Taylor & Francis.

Reiser, R.A. (2001a) 'A history of instructional design and technology: Part I: A history of instructional media', *Educational technology research and development*, vol. 49, pp. 53--64.

Reiser, R.A. (2001b) 'A history of instructional design and technology: Part II: A history of instructional design', *Educational Technology Research & Development*, vol. 49, pp. 57--67.

Sly, L. and Rennie, L.J. (1999) 'Computer managed learning: Its use in formative as well as summative assessment'.

Vereshchagin, N.K. and Vitányi, P.M. (2004) 'Kolmogorovs structure functions and model selection', *Information Theory, IEEE Transactions on*, vol. 50, pp. 3265--3290.

Yorke, M. (2000) 'The quality of the student experience: what can institutions learn from data relating to non-completion?', *Quality in Higher Education*, vol. 6, pp. 61--75.

Yorke, M., Bridges, P. and Woolf, H. (2000) 'Mark distributions and marking practices in UK higher education', *Active Learning in Higher Education*, vol. 1, p. 7.

Appendix 1: Estimated Costs

Stipend for Principal Researcher (15 months):	@ £35,000 p/a	£43,750
+ standard 46% UK HE research overhead		£20,125
		£63,875
Conference and publication fees		£1,000
Rental of cloud space for 15 months	@ £100 p/m	£1,500
		£66,375

Appendix 2: Raw Data Collected

For each event:

Time	Location	Event Type	Additional Information Stored	Rationale
Securely signed monotonic nano-second timestamp	URL File Window Web page view Social network contact	Window state change e.g. size, order, position, focus, scroll position, window open or close etc.	Screen coverage	Lets you determine what parts of which files are visible to user at which time. What files are used to compose academic work output.
		Mouse movements	Mouse pointer location	Commands issued by user. Left or right handedness. Whether user uses menus or key shortcuts. How familiar is user with computer functionality.
		Keyboard button state change	Key in question and its new state	What is typed into computer. Movement of text cursor.
		Text cursor position change	Where to in text	What items or text user selects. Where user intends to type next.
		All Microsoft Office operations e.g. application of italics/bold, choose/modify styles, spell check, word count etc.	The operation in question and what it was applied to	Changes to academic work output. Familiarity with Microsoft Office functionality.
		Copy/Cut/Paste text/items	From where to where	What data is copied in part or in full from where to where.
		All web browser operations e.g. change current tab or window, web page link click, page load, page display refresh, page end load, page scroll, page search etc.	Web page/PDF source data in full	Records data and which parts of data viewed during production of academic work output.
		All social networking activity	Activity in question	Records social networking activity.

Events undergo a certain amount of processing before being sent to internet cloud repository: Mouse movements are parabolised into vectors, typing into search boxes and clicking are converted into search terms, web page data/PDFs are checked to see if they are already on the cloud to avoid having to transfer them, and texts and changes to texts have their Kolmogorov complexities and Levenshtein distances calculated. Software permits additional plugin processing modules to be added remotely.

Appendix 3: Proposed End User Licence Agreement

This service is at a very early stage. It will contain bugs, misbehaviour, misoperation and may at any stage fail to function completely including the protection of the privacy or integrity of your data. We utilise state of the art testing methodology to prevent this from happening, but early stage concept testing software is simply like this and we must protect ourselves from legal liability.

We recognise the European Data Protection Directive, and we believe that we meet its requirements.

By using this service or any software connected to or associated with this service, you automatically agree to following:

1. You agree to testing software which is still being developed, and which may or may not function correctly.
2. You give us permission to store personally identifying data on your computer and any other computer or device which you use to access this service. We use these data to ensure that you are actually you.
3. For content that is covered by intellectual property (IP) rights, you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you access on any other media from the moment that you first access this service, including data about how and when you access any data. We use these data to provide the service.
4. You permit us to access and make any use of third party relationship and/or personal data derived from any logins supplied by you to us. We use these data to provide the service.
5. You permit us to make these data available to third parties as we see fit. We shall try to not reveal personal data, or personally identifying data, without your permission.
6. You permit us to transfer and/or store these data to anywhere on the planet as we see fit. We require this ability as we make use of third party providers of cloud processing services which may not reside in your geographical region.
7. There is no warranty of any kind for this service, so you agree to assume all risk when using this service. You shall not hold us liable for any data loss, misappropriation of data, damage to data or any other thing including any form of tangible or intangible good or service. This is early stage software, and it will likely fail at some point.
8. You agree to not assert, or cause others to assert, any of your essential patent claims against us.
9. You will not use this service to do anything unlawful, misleading, malicious or discriminatory.
10. You will not do anything that could disable, overburden, or impair the proper working of this service.
11. You will not facilitate or encourage any violations of this agreement.
12. You agree that any violation of this agreement permits us to arbitrarily delete any data of yours stored on our servers. This is to protect us in case you transfer illegal material onto our servers.

Here is what we shall try (but do not guarantee) to do:

1. We shall try our best to notify you when personal or personally identifying data is about to be transferred away from your machine. Our software will try its best to respect your decision if you refuse to permit this, however be aware that it will then likely fail to operate and no longer be fit for purpose.
2. We shall try our best to delete all personal and personally identifying data when requested by you. This is a legal requirement in many jurisdictions, and we shall try our best to meet it. However, this service is early stage software, and bugs may arise which cause us to inadvertently fail part or all of this requirement.
3. We shall try our best to not permit others to access your personal data without your permission. They may, however, access data about your data (i.e. metadata) freely.
4. We shall try our best to not permit others to identify you without your permission. They may, however, be able to access statistics generated by us aggregated from your personal and personally identifying data.