Sheffield Hallam University

Partially-automated individualised assessment of higher education mathematics

ROWLETT, Peter <http://orcid.org/0000-0003-1917-7458>

Available from Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/27188/

This document is the Accepted Version [AM]

Citation:

ROWLETT, Peter (2020). Partially-automated individualised assessment of higher education mathematics. International Journal of Mathematical Education in Science and Technology. [Article]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

Partially-automated individualised assessment of higher education mathematics

Peter Rowlett[a]*

[a] *Department of Engineering and Mathematics, Sheffield Hallam University, Sheffield, U.K.*

p.rowlett@shu.ac.uk

ORCiD: https://orcid.org/0000-0003-1917-7458

Twitter: http://twitter.com/peterrowlett

LinkedIn: https://www.linkedin.com/in/peterrowlett

Partially-automated individualised assessment of higher education mathematics

A partially-automated method of assessment is proposed, in which automated question setting is used to generate individualised versions of a coursework assignment, which is completed by students and marked by hand. This is designed to be (a) comparable to a traditional written coursework assignment in validity, in that complex and open-ended tasks can be set with diverse submission formats that would not be suitable for written examination or automated marking; and, (b) comparable to e-assessment in terms of reduction of academic misconduct, with individualisation acting as a barrier to copying and collusion. This method of assessment is implemented in practice. Evaluation focuses on expert second-marking, student feedback and analysis of marks, and aims to establish that the partially-automated method can be useful in practice. The partially-automated method proposed appears to be capable of adapting a coursework assignment to make it less sensitive to copying and collusion (and therefore more reliable) while maintaining its validity, though leading to reduced efficiency for the marker. This paper therefore contributes the introduction of a novel approach to assessment which offers a way to bring automated individualisation to the assessment of higher order skills in higher education mathematics.

Keywords: partially-automated assessment; assessment; e-assessment; computer-aided assessment; skills

Introduction

There are a number of assessment methods available for higher education mathematics. One that has been popular in recent decades is automated assessment, called e-assessment, computer-aided assessment or computer-assisted assessment. There are advantages in assessing higher education mathematics via automated methods, but also limitations, which is a matter of debate in the professional and research literature (discussed later). Non-automated methods, such as coursework assignments and written examinations, also come with advantages and limitations. A model is proposed for

viewing assessment methods as offering a balance of advantages and limitations suitable for different assessment circumstances, particularly in relation to validity, reliability and efficiency.

Viewing assessment through this model, I will argue there are advantages and limitations in different circumstances when choosing one assessment methods over others. A novel method will be proposed for a partially-automated assessment approach, which might access a different balance of advantages and limitations more suitable for some circumstances. In order to demonstrate this as more than a theoretical possibility, a real module context will be described in which the partially-automated assessment method may be useful. Assessment using this method is implemented and evaluated, in order to demonstrate the proposed approach is useful in some circumstances.

The key contribution of this paper is to propose a novel assessment method and demonstrate that it can be useful in practice. The paper proceeds as follows. First, a discussion of assessment methods establishes a model for viewing assessments as offering a balance of advantages and limitations. A discussion of advantages and limitations in relation to coursework, written examination and e-assessment is detailed. Then the partially-automated method is proposed and described. The specific implementation and its evaluation are detailed. Finally, a discussion section revisits the main theme of this paper, of whether a partially-automated assessment method can offer a viable addition to the repertoire of assessment methods available to assessors of mathematics in higher education.

Summative assessment: validity, reliability and efficiency

In a learning, teaching and assessment system under constructive alignment (Biggs, 1999), importance is placed on what students do, as this has a major impact on learning. Because assessment signals to students what to focus on, desired behaviour is encoded

in learning objectives and assessment items are designed in alignment with these. Thus, summative assessment is both a prompt to students of what is to be learned and a tool for measuring whether or not learning has been successful (Biggs, 1999).

Validity is seen not a property of the assessment, but is related to the interpretation and implications of the scores that result from it. Invalidity can be introduced by an assessment task either being too narrow and failing to fully include that which it is intended to assess, or being too broad and relating to aspects not being assessed, which may cause the assessment to be too difficult or too easy, depending on the circumstances. Messick (1995) says (p. 746) that

> low scores should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected persons to display their competence. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence.

In terms of constructive alignment, a misaligned assessment task may be one at which a student can perform well without necessarily engaging in the desired learning behaviour (Biggs, 1999). If we wish to claim an assessment is aligned to certain learning objectives, we must examine whether it is possible to derive marks from it that are meaningful in relation to those outcomes. Biggs and Tang (2011) say that “the glue that holds the ILOs, the teaching/learning environment, and the assessment tasks and their interpretation together is *judgement*” (p. 219).

In a module running as part of a mathematics degree, these learning objectives should be informed by the aims of the degree programme. In the UK, the purpose of a degree programme is guided by the Benchmark Statement published by the Quality Assurance Agency for Higher Education (2015). (Similar documents exist elsewhere; for example, see Australian Council of Deans of Science, 2013.) The Benchmark

Statement outlines both subject-specific and general skills expected of graduates. The subject-specific skills include understanding and using mathematical concepts and topics, constructing and presenting mathematical arguments and problem-solving. The general skills include study skills and so-called employability or graduate skills including communication skills, group working and organisation. A well-aligned degree programme should encode this range of skills and motivate these via appropriate assessment instruments.

As well as validity, it is also important to consider the reliability of an assessment instrument, the extent to which it is objective and repeatable, including showing no bias between assessors (Gipps, 1994). An assessment with low reliability is not fair, and does not serve its purpose well because it does not necessarily present an accurate picture of a student's learning. As well as reliability and validity, real assessment activities must be considered in terms of efficiency and practicality, both resources available and time and workload for the staff and students taking part (Challis, Houston & Stirling, 2004; Cox, 2011). Efficiency here is intended to relate to the burden placed on staff or students when completing an assessment. This includes the time and effort needed to set, complete and mark a task. It is desirable that an assessment should not place more burden than is necessary for it to satisfy its goal of producing a measure of whether or not learning has been successful. In particular, for an efficient assessment the burden should be proportionate to the weighting of an assessment within the broader context.

A quality assessment instrument should reach a high level of both reliability and validity while being efficient and practical to operate. Even if we accept current assessment methods used in practice as offering the potential for reaching a suitable threshold level of both reliability and validity, there is still variety within both measures.

Indeed, these measures can be considered in tension, with measures to increase validity having an adverse effect on reliability, and vice versa (Brown, Bull & Pendlebury, 1997). Attempts to increase either validity or reliability may decrease the efficiency of the assignment, by placing additional burden on staff or students. It is possible, therefore, to conceptualise assessment methods as potentially offering different forms of balance between these factors. Thinking about assessment methods as offering different forms of balance shows there is value to different approaches in different contexts.

Consider an open-ended piece of mathematics coursework. This might loosely-specify a problem or task with which students must engage, with some quite wide limits for the format of submission of work completed. Such a task might be used to require students to select and apply appropriate mathematical techniques and make choices about how to communicate their findings, rather than simply performing well-specified computations on request, and so perform a valid function in relation to the aims of mathematics undergraduate teaching. However, many open-ended and complex tasks are assessed via professional judgement rather than a tightly-specified mark scheme, and it is possible that multiple markers may struggle to agree on the standard of such a piece of work, affecting reliability (Bloxham, 2009). Another issue for reliability is academic misconduct, either copying work from another in a way which cannot be fixed by attribution (including plagiarism), or collusion, which is cooperating with another person in a way which obscures the origin of a piece of work (Seaton, 2019). Iannone and Simpson (2012) report some university mathematics departments moving away from coursework due to concerns around copying and collusion. A mathematics lecturer interviewed by Thomlinson, Robinson and Challis (2010) said that it is ‘not clear what the real benefit is’ of coursework, given that copying is a particular problem among weaker students. Copying and collusion affect reliability, because another assessment

completed by the same student on the same learning objectives may come to a different rating if it is more effective at combating such academic misconduct. For example, a timed examination might reveal that a student did not know a topic so well as a take-home piece of coursework suggested they did if they completed the latter with the help of others.

In order to address concerns about copying and collusion in a piece of coursework, there are two natural approaches a mathematics assessor might take. First, changing the assessment task to be a written examination virtually guarantees that the work is the student's own. However, this move to increase reliability may come at the cost of reduced validity. An examination is naturally a more tightly-specified series of tasks with closed-form questions, which may mean students have less agency in decision-making about their work and so less opportunity to demonstrate their broader skills base. Timed exam conditions also prioritise speed and memory to some extent, which may not be the learning objectives associated with the task.

An alternative approach may be to replace the coursework by an individualised, automated assessment. This may be called e-assessment or computer-aided assessment and can be implemented in a number of ways, with corresponding effects on validity and reliability.

A significant advantage of e-assessment in relation to copying and collusion is that each student can be given a unique set of questions via pseudo-randomisation, either through *random selection* of items from a question bank or *random generation* – the use of pseudo-randomised parameters in a question template. Individualisation can reduce opportunities for copying and collusion considerably because a student who may consciously or unconsciously engage in these kinds of academic misconduct cannot find another student with the same questions from whom to copy or with whom to

collaborate. Two or more students with similar but not identical questions discussing approaches but not actually working together on questions could be viewed as being engaged in the far more positive and constructive behaviour of collaboration (Seaton, 2019). It should be noted that as well as copying and collusion, Seaton (2019) also includes contract cheating in her definition of academic misconduct. This is when someone asks another person to complete a piece of work and submits as their own, and cannot be stopped by individualised approaches such as take-home e-assessment.

Although e-assessment with individualisation may act to reduce academic misconduct, it carries with it certain limitations. Random selection may offer limited range of questions (Broughton, Robinson and Hernandez-Martinez, 2013) and writing questions using random generation is difficult, requiring expertise unlike the setting of paper tests (Greenhow, 2015; Sangwin, 2015). The challenge is technical, because of the need to understand the minutiae of how an automated marking system will handle a response (Sangwin, 2007), and pedagogic, because this requires much clearer specification of what is assumed and tested along with knowledge of typical student mistakes (Greenhow, 2015). It is also possible to introduce mathematically impossible questions (Sangwin, 2004). Question authors must take care to avoid introducing alternative or additional learning requirements while being ‘creative in findings ways around’ the ‘limitations’ of e-assessment (Lawson, 2002; pp. 4–5). These issues affect the efficiency of the assessment for staff and students, as well as potentially impacting on validity by reducing the range of what can be asked.

Validity is also affected by difficulty communicating answers to the computer. An e-assessment system might provide responses in some form for the student to select, such as with multiple-choice questions, giving a hint or the opportunity of guessing or working backwards from the answer (Lawson, 2002; Sangwin, 2007). If an e-

assessment does not provide a response, it must allow input of answers. Numeric input is of limited use for mathematics (Sangwin, 2007), and simple string-matching is inadequate (Klai, Kolokolnikov & Van den Bergh, 2000). A more sophisticated possibility is free-text input which is tested for algebraic equivalence by a computer algebra system (Sangwin, 2007). Such input requires practice for students to use correctly (Sangwin, 2015) and may add additional learning requirements unrelated to the assessment objectives (Lawson, 2002). Alternatively, menu-based interfaces may be used, also requiring additional learning to use. One possibility may be hand-writing recognition of mathematics, a technology under development but not yet in common use (Pacheco-Venegas, Lopez & Andrade-Aréchiga, 2015).

Automated marking avoids human error and lack of objectivity (Ferrão, 2010; Sangwin, 2004), potentially improving reliability provided systematic marking errors can be avoided (Ferrão, 2010). However, the need to ask questions that can be marked by computer leads to reduction in validity, as this tends to focus questions on procedural aspects (Broughton, Hernandez-Martinez and Robinson, 2017). The final answer input into a computer is not necessarily enough to establish partial credit (Rønning, 2017), a normal part of assessment in mathematics (Genemo, Miah & McAndrew, 2016), leading to a solution with a minor error potentially being marked as completely wrong (Greenhow, 2015). Some systems attempt to address this by breaking questions into parts, steps or sub-questions, prioritising procedural aspects and so reducing validity (Quinney, 2010). Students are reported as preferring human-marked work because of the ability for a human marker to act flexibly to award partial credit marks (Cigdem & Oncu, 2015). Marking extended and open-ended work may be impossible (Beevers & Paterson, 2003; Greenhow, 2015; Sangwin, 2015).

Proposal for a partially-automated assessment method

It is possible, to a large extent, to unpick the advantages and limitations of summative e-assessment in relation to copying and collusion. A significant advantage is that automated question generation enables individualisation of work, which can reduce opportunities for copying and collusion. However, the range and depth that can be assessed is limited by the capabilities of automated marking, leading to difficulties assessing complex, open-ended work and testing conceptual understanding. Setting questions is technically and pedagogically challenging because of the limitations of automated marking. The need for students to input mathematics into computers may cause inefficiencies for students, or lead to reduced validity via additional learning requirements or more structured questions.

If we separate automated question generation from automated marking, we can see the advantages are linked to the former and the limitations arise principally from the latter. A partially-automated approach is thus proposed, in which questions are set via an automated question generator but completed by students and marked by hand as if it were a non-automated piece of coursework. This could access the chief advantage of individualisation while avoiding the major limitations of computer input and automated marking. Because assessors setting questions would not have to adapt their practice to suit the limitations of automated marking, tasks could be more open-ended and the difficulty of setting questions becomes comparable to setting questions for coursework, though care would still need to be taken that randomisation generates comparable tasks for each student. A partially-automated method would lose the advantage of algorithmic objectivity in marking, but should be no worse than traditional coursework in this regard. One fresh limitation is that individualised work will be less efficient to mark

than traditional coursework, because each student has a different set of answers that cannot be memorised by the marker.

Viewing an assessment method as offering a balance of reliability and validity, we can view a piece of coursework as offering high potential validity, because it can be used to assess more complex, open-ended tasks, with reduced reliability, in part due to the increased risk of copying and collusion. The reliability can be improved by obstructing opportunities for copying and collusion via either a written examination or e-assessment, but these methods reduce validity potential in different ways. By individualising the assessment through the proposed partially-automated approach, we might decrease the risk of copying and collusion, and so increase reliability, without reduction in validity. In fact, this method could be less open to academic misconduct than e-assessment individualisation, since the students could be asked to submit an extended piece of work for a human marker to read, whereas an e-assessment system typically only examines the final answer. The proposed method has the potential to maintain validity and increase reliability with respect to copying and collusion, compared with a traditional piece of open-ended coursework, at a cost of decreased efficiency since marking will be more time-consuming. This would make this method an unusual and potentially useful addition to the assessment methods in common use.

Since making this proposal, I have become aware of three approaches with similarities to the proposed approach, although these differ in significant ways.

The first is from my experience of teaching. I taught part of a computational methods module in which students were given coursework to solve using MATLAB. Each question contained a randomised parameter, r, and students were required to compute the question in MATLAB and then answer it. For example, r might be used in one question as a coefficient in a differential equation and in another as a term in a

matrix. Each student was given the same questions but a different value of r. A similar approach is taken by Blyth and Labovic (2009), who use automation via Maple worksheets. These approaches are different to the proposed approach here because the individualisation is a collaboration between the question author, who must carefully specify the questions, and the student, who must actually vary the questions themselves through software. This is acceptable because students are demonstrating their ability to meet learning objectives around using software, but this approach would not work outside of a computing context because of the imposition of additional learning objectives. The marking, being by hand, was not limited by automation.

The second approach is in statistics and is described by Hunt (2007). This approach uses Microsoft Excel to draw a randomised data sample from a larger data set. Each student uses a five digit PIN as a seed which generates the data sample, and the marker uses this PIN to populate an answer sheet. A similar, more automated approach is taken by Fawcett, Foster and Youd (2008) via an e-assessment system. They provide statistics assessment with each student being presented with a randomly generated unique dataset to analyse, with marking by hand. These approaches involve individualised work which is marked by hand, so there are some similarities. However, the individualisation is achieved by drawing a random data sample from a larger database, meaning this approach is only applicable to topics involving data. The system used by Fawcett et al. has limitations caused by computer input, as it uses multiple-choice questions for ‘more descriptive parts’ (p. 46).

Combined tests are proposed as a possibility by Sangwin (2015), in which a ‘routine calculation within a longer proof’ is ‘checked automatically... before the whole piece of work is submitted to an intelligent human marker’ (p. 712). Such a system may have some efficiency advantages over what is proposed here, because of the use of

automatic marking for some parts, but it will still suffer the limitations caused by the difficulty of inputting mathematics to computers.

The proposed partially-automated approach to assessment, as theorised, has potentially a different set of advantages and limitations over other types of assessment currently in use, and therefore could make a useful addition to the assessment methods used in higher education mathematics. I take a sceptical approach to technology innovation, in which developments are not implemented simply because the technology makes them possible, but in response to a recognised educational need (for an exploration of this, see Rowlett, 2013). To move this proposal beyond a theoretical possibility, it is therefore necessary to consider whether there is a context in which this approach could be more useful than existing assessment methods, and this becomes the main question of this research. A live teaching context is sought, following Kounin's (1970) view that a classroom has 'its own ecology', meaning that a well-controlled experiment may not be sufficient to demonstrate that this method would work in the reality of a teaching context (p. 59). If such a context can be found and a partially-automated approach shown to have the advantages theorised above, then a useful, novel assessment approach will have been developed. This research does not seek to demonstrate that this is the only way in which a partially-automated approach might be useful, simply to demonstrate that such an approach can be useful in some way.

Teaching and assessment context

The teaching context used was a final year, optional module in undergraduate mathematics which aimed to develop skills needed in employment which may not be developed by traditional mathematics teaching. These skills were working in depth on a problem over an extended period, writing reports, communicating mathematical results to different audiences, working in collaboration with others and articulation of graduate

skills. The module was project-based, so that the main activities designed to drive learning were a series of student-led, summative group projects rather than, say, delivery of content via lectures.

One issue of group work is that of uneven contribution, which can lead to student perceptions that assessment is unfair if all students get the same mark regardless of their contribution (MacBean, Graham & Sangwin, 2004). To increase the amount of the module mark which reflected individual ability, alongside other methods to address uneven contribution, individual work was set alongside the group projects. This individual work contributed to individual marks but not to the group mark, which was based on the group report.

One group project saw students spend three weeks answering a brief from a (fictional) client. Specifically, students were to investigate ‘Art Gallery Problems’, which are concerned with determining the minimum number of point ‘guards’ necessary for all points in a polygon (the ‘art gallery’) to be connected by a straight line (line of sight) to at least one guard (O’Rourke, 1987). The brief gave three art gallery floor plans and asked groups to propose the size of a staff which must be hired to guard each of these, justifying their findings and giving discussion of real world considerations in a short report. Art Gallery Problems are a curiosity of combinatorial geometry and not really intended to be applied in this way, leading to plenty of opportunity for students to demonstrate their awareness of the limitations of the approach taken. The individual coursework gave a single art gallery floor plan and asked the same question (a sample piece of individualised work is given in the appendix). These tasks were deliberately similar because the intention was to examine individuals on similar work to the main group project. This was designed so that students would be advantaged in the individual work by contributing well to the main group project.

The similarity of the individual and group tasks meant I viewed the risk of in-team copying or collusion, for example the temptation to view the group task as answering the question for four floor plans, as high. Approaches to reduce this risk could be exam conditions or individualised work. As the work was quite open-ended, I felt that timed exam conditions would not be suitable for reasons discussed earlier in this paper. As the work to be submitted was a diagram showing positions of guards and an extended report, this was beyond the limits of automated marking. The production of a diagram by computer in a specific format was not part of the learning on the module, so requiring this would introduce additional learning requirements compared with allowing hand-written answers. I could have attempted to write a different assessment on the topic of Art Gallery Problems that would have been suitable for a timed examination or e-assessment, but this would have been at the cost of reduced validity with respect to the assessment goals. The need to produce individualised work via randomisation, lack of suitability of automated marking and the need for students to be able to hand-write their answers suggests that the proposed partially-automated approach may be appropriate.

Individualised worksheets were generated using the system Numbas, principally a mathematically-aware e-assessment system (Foster, Perfect & Youd, 2012) that can also provide printable question sheets and corresponding answer sheets (where generating answers is possible). Questions included a diagram selected from a bank of diagrams and the insertion of randomised parameters into question templates. Producing this was much like writing questions for an e-assessment system, without the requirement to comply with the limits of automated marking. (Systems other than Numbas could presumably be adapted for a similar approach.) When marking, answers

could not be learned and student submissions needed to be matched to an appropriate answer sheet using an ID number, which added to the time taken for marking.

For the main group project, students were expected to gather information from multiple literature sources and work for an extended period (3 weeks) on problems, in groups, and communicate their solutions clearly via reports. This aimed to address learning objectives around problem-solving in a real-world context, working in depth over an extended period, communication via reports and group working. The related individual assignment, to be individualised via partially-automated assessment, assessed the same learning objectives except group working.

Evaluation method

The main question of this research is: Can a context be found in which the proposed partially-automated approach more appropriately aligns an assessment task to its associated learning objectives than existing assessment methods? The partially-automated approach was proposed as having potential to maintain the validity and reliability (with respect to marker consistency) of an open-ended piece of coursework while increasing reliability (with respect to academic misconduct), so these aspects will be examined. To answer this, a series of sub-questions are asked:

(1) Are the marks particularly sensitive to who is doing the marking (marker consistency)?
(2) Is the assignment assessing the learning objectives it was intended to assess (validity)?
(3) Does the individualised nature of the assignment work to reduce copying and collusion (academic misconduct)?

The purpose of evaluation in this project is a combination of Chelimsky's (1997) 'evaluation for development' and 'evaluation for knowledge' (p. 100). It is evaluation for development because it aims to evaluate a teaching innovation in a particular context to decide whether this has been effective and to provide formative information for an innovation process, i.e. it is worthwhile to use this innovative approach in the circumstances described? There is also an element of evaluation for knowledge because the potential exists for this research to establish whether the proposed partially-automated assessment approach can be put to effective use in higher education mathematics in general.

As discussed, in order to draw authentic and useful conclusions, this evaluation took place in a live teaching situation, raising ethical issues around impact on student grades. For this reason, the main driver of individual variation in marks for group work was the better-established peer assessment of contribution (see, e.g. Earl, 1986), using an approach similar to that used elsewhere in the degree course, and the partially-automated assessment was used to only contribute a small proportion of the overall mark (4% of the module) in order to keep the untested new approach from having an undue influence on the final grade. Student feedback was collected anonymously and, because student marks and the process of assigning those marks are discussed in detail, the identities of the universities involved are kept confidential. The research design was approved by a faculty research ethics process.

When selecting statistical tests, assessment marks are considered as being on an interval scale and normality is not assumed. For example, considering marks from 44 students submitted for the partially-automated assessment, the Shapiro-Wilk test for normality gives statistically significant evidence to reject the null hypothesis that the marks arise from a normal distribution ($p=0.0068$).

This evaluation comprised a second-marker experiment, student feedback and evaluation of marks. These stages are described in detail in the remainder of this section and summarised in Table 1.

[Table 1 near here]

Second-marker experiment

Marker consistency may be evaluated by having different markers scoring the same piece of work (Gipps, 1994). We should not expect complete agreement between multiple markers for this more subjective form of assessment. Also, Bloxham (2009) criticises the inherent assumptions that higher education work can be awarded an accurate and reliable mark and that academics share common views regarding academic standards. Since we cannot expect complete agreement, conclusions about whether the level of agreement found between multiple markers is reasonable or not require context. In order to calibrate expectations and provide reference information, the level of agreement for multiple markers of two more established assessment methods was examined. This used: a class test under examination conditions, a method of assessment recognised as being highly reliable; and, an open-ended piece of coursework, a method reported as having problems with consistency of marking (Iannone & Simpson, 2012).

Each second-marking experiment had a piece of work which was marked by an original marker and at least one second-marker. To assist with interpretation, comments on differences in the marks and the intraclass correlation coefficient (ICC) are presented. The intraclass correlation coefficient (ICC) is used to assess measurement error in judgements made by humans. A two way model on single score data ICC which considers the agreement between raters is computed. This takes no account of any ‘true’ value of the mark, if such a thing exists, but only considers the level of agreement between multiple markers. This means the ICC rating given depends on the reliability of

the markers. For this reason, only people who have professionally marked student work in universities will be used as markers, for a reasonable expectation of reliability.

Evaluating validity requires professional judgement (Brown et al., 1997). A simple test of validity was to ask the second-markers what they thought the coursework was assessing. If the second-markers' views on what was being assessed matches the intended learning objectives, then we can say the work was viewed as assessing what it was intended to assess. Care was taken to control the information given to second-markers. They were shown student work and given a mark scheme with which to mark this, but they were not given a broader context for the assessment or told the intended learning objectives.

Written test reference experiment

The work for this experiment arose from an open-book test, taken under examination conditions, during a basic mathematical methods module for first year mathematics students. The test comprised five well-focused, short problem questions for which 50 marks were available. A 10% sample of all scripts was checked by a moderator, with reference to the original marks, as part of the usual institutional process. The moderator agreed with the marks awarded in all cases.

I marked a sample of ten scripts without reference to the marks assigned by the original marker but using the same mark scheme (blind second-marking). The mark scheme was a set of worked solutions with individual marks indicated for components of answers and for working. The original marker was working at the same university as me so was used to marking work from similar students.

Coursework reference experiment

The work for this reference experiment arose from a coursework task to write an 800-

1000 word review of a popular book or textbook on mathematics or the history of mathematics. The marking criteria specified those pieces of information that each review should contain, as well as some general subjective criteria around the quality of the writing and level of critical understanding. Marks were a simple percentage. A sample of work had previously been approved via an institutional moderation procedure, conducted with reference to the original marks.

I marked a sample of 14 scripts via blind second-marking. The original marker was working at a different university with a similar entry requirement to my own.

Second marking of the individualised coursework

Three second-marker volunteers were recruited opportunistically from personal contacts. Each had experience of marking work at university; one as a senior academic, one as a junior academic and one as a PhD student. One had experience of marking at a university with a similar entry requirement to where the work was submitted, one with a lower entry requirement and with a higher entry requirement.

A 10% sample of student work was anonymised (5 pieces from 44 submitted). This was provided along with grade descriptions, a mark scheme and a sample piece of marked work (written to be correct on the non-subjective parts of the mark scheme) as a reference piece since the second-markers were not familiar with the topic. Second-markers were asked to assign a mark to each piece of work and to provide some comment on what they thought the assessment was assessing, which could be by guessing at learning objectives or writing a general statement of what a student who gets full marks will have demonstrated their ability to do. These tasks were designed to test reliability and validity, respectively. Second-markers were also asked for any comments on the marking process.

Student feedback

The approach taken to individualised assessment was intended to reduce the possibility of students discussing answers and copying. Students volunteers from the target cohort were asked via a questionnaire to express their views, anonymously, on the role of individualised work and how this affected interaction with other students. Some professional sources question whether concern about copying and collusion is overblown (e.g., Cox, 2011), so questions were included also about copying in this assignment and other work. Details of questions asked are included in the results section below. This was completed by the cohort taking the assessment task described above (group A) and by a group at a different university (group B) to provide input from an independent cohort of students with which I had not interacted. The lecturer for group B had also used the technique developed for this project via Numbas for an individualised formative in-class question sheet in a final year digital signal processing module. I helped the lecturer with the implementation, but had no contact with the students in group B. For both groups, questionnaires were administered via Google Forms. No questions were compulsory. For group A, due to practical considerations this was during the final session of the module, six weeks after the group project had been submitted. For group B, this was at the end of the session in which the individualised assessment was used.

Where interval or ordinal data have been collected from two independent groups, Fisher's Exact Test can be used to test whether the distribution between categories is independent of group membership, i.e. that the two samples are drawn from equivalent populations. Evidence to reject the null hypothesis would mean that the distribution of responses (to, say, a Likert-style question) is significantly different for the two groups. This will be used to see if there is evidence that the responses by the

group I had interacted with are different from a group with which I had no contact, i.e. to attempt to control for my influence as potentially an enthusiast innovator.

Comparison of marks

The academic misconduct risk was around intra-group copying or collusion, since group members were working together on similar problems. Inter-group copying or collusion may be a risk also, but this seemed less likely because groups were partly self-selected and group work encourages a sense of inter-group competition. Individual marks from within groups were therefore examined for differences. If there was wide variety of marks within groups, we might conclude that intra-group copying and collusion was not a large problem. A lack of variety of marks, however, could indicate copying or collusion, or perhaps just that students had been learning the topic together and so have similar levels of understanding. If group members colluded on the individual work, we might expect to see similarity between individual and group marks, since they certainly (properly) collaborated on the latter.

The correlation of raw group project marks and rankings (prior to scaling due to peer assessment of contribution) with the individualised coursework is presented via Pearson's ρ and Kendall's τ. Pearson's ρ does not assume data are normal, is appropriate for data taken from an interval scale and gives a measure of linear correlation between variables. Similarly, Kendall's τ provides a measure of association between ordinal variables. The dispersion of marks for the individual assignment is examined via the range and standard deviation of the marks within each group.

Results

Second-marker experiment

Written examination reference experiment

Table 2 contains the original marks and those which I ('PR') assigned during the blind second-marking. There were five discrepancies of one or two marks (2% or 4% of the total) in ten scripts. The ICC for the two sets of marks is 0.992. This value is regarded by Landis and Koch (1977) as an 'almost perfect' level of agreement (p. 165).

[Table 2 near here]

Coursework reference experiment

Table 3 contains the original marks and those which I assigned during blind second-marking. There were differences in all fourteen pieces of work. Six were differences of around 5% or less, a further six were around 10% and two were greater differences. The ICC for the two sets of marks is 0.586. This value is regarded by Landis and Koch (1977) as a 'moderate' level of agreement (p. 165).

[Table 3 near here]

Second marking of the individualised coursework

The marks from each marker are given in Table 4. The ICC for the four sets of marks is 0.635. This value is regarded by Landis and Koch (1977) as a 'substantial' level of agreement (p. 165).

[Table 4 near here]

Comments on learning objectives

The three learning objectives intended to be addressed by this individual piece of work

were:

(1) problem-solving in a real-world context;

(2) working in depth over an extended period; and,

(3) communication via reports.

Second-marker A suggested the following as the learning objectives that were addressed by this assignment:

- Ability to solve unfamiliar problems (or unfamiliar variants of problems discussed in class);
- ability to use literature;
- ability to present mathematical work clearly.

I would suggest that the first of these is problem-solving, the second is part of working in depth and the third refers to communicating results. These three statements do not quite cover all aspects of the three intended learning objectives, but neither do they represent additional, unplanned or extraneous requirements.

Second-marker B gave the following description of what the work was assessing:

> The exercise seemed to be designed to assess a student's ability to apply a piece of mathematics, in this case an aspect of computational geometry, and interpret the real world viability of their solution. Particular emphasis was given in the mark scheme for rewarding the students' awareness of the mathematical and legal literature as well as communication skills - suggesting that you wanted the students to actually take seriously how one uses mathematics outside of the classroom.

In this description, second-marker B identified problem-solving in a real-world context, awareness of background information and communication skills. Again, these fit within

and do not extend the actual intended learning objectives.

Second-marker C suggested the following intended learning objectives:

- understanding of the theory;
- ability to apply the theory;
- understanding that theory doesn't always apply perfectly to the real world;
- understanding of the difference between 'necessary' and 'sufficient', and that there's more than one possible answer.

Second-marker C identified aspects of problem-solving, depth of understanding and the process of relating a solution to a real-world context, which match the first two learning objectives fairly well. They did not identify communication skills, nor put forward any additional learning objectives.

Comments on process

Marker B gave no comments. Marker A said:

> I don't think I did this very well. In particular I am unsure about the marks awarded for quality of exposition. To some extent these may have (subconscioulsy [sic]) overlapped with marks for showing familiarity with literature etc. In general I felt my sample didn't present the work very well. I am a little worried about the marks awarded for consideration of uniqueness and possibility of triangulation, which none of my sample mentioned - was it clear they were expected to address this? Does creating a triangulation show that a triangulation is possible?
> I gave a mark to one student in 2 for familiarity with literature although the answer was wrong because there was sufficient similarity with the correct answer to suggest they might have seen it. I was perhaps over-generous. I was probably too harsh elsewhere!

Marker C said:

> I had to look up the theory first to understand the question, as I've never come across this problem before some of the students seem to miss the point somewhat, and apply the method without really understanding it (i.e. drawing the triangulation but failing to notice by inspection that they have far more guards than necessary) none of the students considered that guards are human and might get ill or want to go on holiday!

Both markers who made comments expressed concern about their unfamiliarity with the topic. Marker B also appears to be asking the kinds of questions that would be answered by a more detailed set of marking criteria. Marker B mentions mathematical rigour, while Marker C is concerned with real-world aspects. Again, more detail about the aims of the task and a more detailed mark scheme could have addressed these concerns, though these were omitted as part of the research design around validity.

Student feedback

Students in groups A and B were asked to indicate their level of agreement with each of four statements, listed with numbers of responses in Table 5. Also in Table 5 are the p-values obtained for each Likert-type question when comparing the two groups via Fisher's Exact Test. In each case, there is no evidence at the 5% level to reject the null hypothesis that the distribution of answers is independent of the group answering. Responses to two questions about copying, which were accompanied by a reminder that the questionnaire was anonymous, are given in Table 6. Again, p-values from Fisher's Exact Test are listed in Table 6 and do not give evidence at the 5% level to reject the same null hypothesis.

[Table 5 near here]

[Table 6 near here]

Comparison of marks

There were five groups. The raw group project marks and rankings do not correlate well with the marks and rankings for the individualised coursework (ρ=0.230; τ=0.229). The range and standard deviation of the individual marks within each group are presented in Table 7. Individual marks for each group represent a range of at least 23 marks and up to 31 marks, and have a standard deviation of at least 8.216 and up to 11.411.

[Table 7 near here]

Conclusions and discussion

I proposed a novel partially-automated approach to assessment, in which the tools of e-assessment are used to set a piece of individualised coursework that is marked by hand. I argued that this would offer the opportunity to improve the reliability of a piece of coursework while maintaining validity, and set this against converting the coursework into a written examination or e-assessment, which would improve reliability but reduce validity in different ways. The proposed partially-automated method would have reduced efficiency of marking compared to written examination or e-assessment. As such, I argued that this proposed assessment method offered a different balance of advantages and limitations over other established methods.

In order to determine whether such a balance of advantages and limitations could be useful in practice, the proposed method was implemented in a real assessment context. This was a piece of individual coursework taking place alongside a group project, where the risk of copying or collusion was felt to be high but a written examination or e-assessment would have been less suitable due to reduced validity.

Evaluation of the implementation examined whether the partially-automated approach could be used to set an assessment that was of comparable validity and

reliability (with respect to who is doing the marking) to a piece of open-ended coursework, while contributing to an increase in reliability (with respect to academic misconduct) via reduction in opportunity for copying and collusion.

Second-marker reference experiments showed a high level of agreement between markers for an open-book written examination and a moderate level of agreement for an open-ended piece of coursework. For the individualised coursework, a group of four markers showed a level of agreement that was between the two reference experiments, and close to the open-ended piece of work. This suggests that the individualised coursework was, despite being set by the partially-automated approach, not unduly unreliable with respect to who was doing the marking.

The three second-markers identified the intended learning objectives with a fair degree of accuracy and did not recognise unintended learning objectives being assessed. This provides evidence that the assignment was assessing what it was intended to assess, and no more.

Questionnaire responses from 42 students saw 22 saying they had copied work from another student at university and 35 saying another student had copied from them at university. Clearly these questions carried risks of inaccurate responses, because students might be unwilling to confess to this undesirable behaviour, but that risk is around deflating the numbers, not inflating them, so this does not weaken the finding that copying and collusion appears to be risks, which refutes the suggestion that concern about academic misconduct is overblown (see, e.g., Cox, 2011). Students generally reported appreciating being able to discuss individualised work with no risk of plagiarism and reported concerns about copying, including that if identical work had been set then they believed some students would have copied from others. The responses from an independent reference group of students at another university are

apparently similar, suggesting that my influence as the person who initiated the innovation was not a factor.

Individual marks were not well correlated with group marks and dispersion of individual marks in each group was high. This indicates that intra-group copying or collusion, which would tend to produce homogeneity of work, was not a big problem. Since student feedback indicated a high risk of copying or collusion and none was detected, this suggests that the individualised nature of the coursework did contribute to a reduction in copying and collusion.

Conclusions from the evaluation are that: the partially-automated assessment was of comparable reliability (with respect to who did the marking) to an open-ended piece of coursework; the assessment was valid, in the sense of assessing what it was intended to assess and no more; copying and collusion were confirmed as risks and found to not have been a large problem, suggesting that the individualised nature of the assessment contributed to a reduction in academic misconduct.

Some sources report a devaluing of coursework in the eyes of assessors, due to concerns about academic misconduct (Iannone & Simpson, 2012; Thomlinson et al., 2010). The partially-automated approach proposed here appears to be capable of adapting a coursework assignment to make it less sensitive to copying and collusion (and therefore more reliable) while maintaining its validity, though it led to reduced efficiency for the marker.

It should be noted that one of the second-markers reported concern about his consistency and another reported unfamiliarity with the topic being examined. These issues limit the reliability of the findings, and having markers who are more familiar with the topic or a screening test of markers could have improved this and perhaps led to the marks being more consistent. Training and consultation between markers was

avoided, though, in order to give information on how the work would be marked without my influence or that of other markers. Having more markers or comparing with more pieces of assessed work would improve the robustness of the findings here, but comparing three assessments and involving five other markers with me was the limit of practicality. Student questionnaire data was taken six weeks after the partially-automated assessment was completed, as part of an end of module survey activity. It should be understood, then, that the student reflections reported were six weeks distant from the experience that had taken place.

The main research question asked whether the proposed partially-automated approach to assessment can be implemented in a way that offers more appropriate alignment of an assessment task to its associated learning objectives. The answer is yes, so this paper has proposed a useful, novel approach to the assessment of mathematics in higher education which offers a way to bring automated individualisation to the assessment of higher order skills. This approach differs from typical e-assessment in mathematics, where the limitations of computer input and automated marking tend to lead to a focus on mathematical techniques and procedural approaches.

This research provides one context in which the partially-automated approach is apparently suitable and more advantageous than other methods which could be used. The partially-automated approach is therefore recommended as an addition to the repertoire of higher education mathematics assessment methods, particularly in the case where an assessment carries a high risk of copying or collusion but issues such as validity make an examination or automated marking e-assessment sub-optimal. Future work is recommended to identify and evaluate other contexts in which this approach may be useful.

Word count: 7,705.

Acknowledgements

An early version of this paper was published in the proceedings of the 8th British Congress of Mathematics Education. I am grateful to Christian Lawson-Perfect (School of Mathematics and Statistics, Newcastle University) for adapting the Numbas e-assessment system for my experiment. The 'worksheet' theme he created to enable my partially-automated approach remains part of the free system at numbas.mathcentre.ac.uk. I am also grateful for the anonymous contributions to this work by the students who gave feedback, the second-markers and the other lecturer who administered the survey with group B.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Australian Council of Deans of Science. (2013). *Mathematical sciences standards statement and threshold learning outcomes*. Melbourne, VIC: Deakin University.

Beevers, C.E., & Paterson, J.S. (2003). Automatic assessment of problem-solving skills in mathematics. *Active Learning in Higher Education*, 4(2), 127–144, https://doi.org/10.1177/1469787403004002002

Biggs, J. (1999). What the Student Does: teaching for enhanced learning. *Higher Education Research & Development*, *18*(1), 57–75. https://doi.org/10.1080/0729436990180105

Biggs, J., and Tang, C. (2011). *Teaching for Quality Learning at University: What the Student Does* (4th ed.). Maidenhead: Open University Press.

Bloxham, S. (2009). Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209–220. https://doi.org/10.1080/02602930801955978

Blyth, B., & Labovic, A. (2009). Using Maple to implement eLearning integrated with computer aided assessment. *International Journal of Mathematical Education in*

Science and Technology, 40(7), 975–988, https://doi.org/10.1080/00207390903226856

Broughton, S.J., Hernandez-Martinez, P., & Robinson, C.L. (2017). The effectiveness of computer-aided assessment for purposes of a mathematical sciences lecturer. In M. Ramirez-Montoya (Ed.), *Handbook of Research on Driving STEM Learning with Educational Technologies* (pp. 427–443). Hershey, PA, U.S.A.: IGI Global.

Broughton, S.J., Robinson, C.L., & Hernandez-Martinez, P. (2013). Lecturers' perspectives on the use of a mathematics-based computer-aided assessment system. *Teaching Mathematics and its Applications*, 32(2), 88–94, https://doi.org/10.1093/teamat/hrt008

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing Student Learning in Higher Education*. Oxon, U.K.: Routledge.

Challis, N., Houston, K., & Stirling, D. (2004). S*upporting Good Practice in Assessment*. Birmingham, U.K.: Maths, Stats and OR Network.

Chelimsky, E. (1997). Thoughts for a New Evaluation Society. *Evaluation*, 3(1), 97–118, https://doi.org/10.1177/135638909700300107

Cigdem, H., & Oncu, S. (2015). E-Assessment Adaptation at a Military Vocational College: Student Perceptions. *Eurasia Journal of Mathematics, Science & Technology Education*, 11(5), 971–988.

Cox, B. (2011). *Teaching Mathematics in Higher Education – the basics and beyond*. Birmingham, U.K.: Maths, Stats and OR Network.

Earl, S.E. (1986). Staff and peer assessment – measuring an individual's contribution to group performance. *Assessment & Evaluation in Higher Education*, 11(1), 60–69, https://doi.org/10.1080/0260293860110105

Fawcett, L., Foster, B., & Youd, A. (2008). Using computer based assessments in a large statistics service course. *MSOR Connections*, 8(3), 45–48, https://doi.org/10.11120/msor.2008.08030045

Ferrão, M. (2010). E-assessment within the Bologna paradigm: evidence from Portugal. *Assessment & Evaluation in Higher Education*, 35(7), 819–830, https://doi.org/10.1080/02602930903060990

Foster, B., Perfect, C., & Youd, A. (2012). A completely client-side approach to e-assessment and e-learning of mathematics and statistics. *International Journal of*

e-Assessment, 2(2). Retrieved from http://journals.sfu.ca/ijea/index.php/journal/article/viewFile/35/37

Genemo, H., Miah, S.J., & McAndrew, A. (2016). A design science research methodology for developing a computer-aided assessment approach using method marking concept. *Education and Information Technologies*, 6, 1769–1784, https://doi.org/10.1007/s10639-015-9417-1.

Gipps, C. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London, U.K.: RoutledgeFalmer.

Greenhow, M. (2015). Effective computer-aided assessment of mathematics; principles, practice and results. *Teaching Mathematics and its Applications*, 34(3), 117–137, https://doi.org/10.1093/teamat/hrv012

Hunt, N. (2007). Individualized Statistics Coursework Using Spreadsheets. *Teaching Statistics*, 29(2), 38–43, https://doi.org/10.1111/j.1467-9639.2007.00254.x

Iannone, P., & Simpson, A. (2012). A Survey of Current Assessment Practices. In P. Iannone & A. Simpson (Eds.), *Mapping University Mathematics Assessment Practices* (pp. 3–15). Norwick, U.K.: University of East Anglia.

Klai, S., Kolokolnikov, T., & Van den Bergh, N. (2000). Using Maple and the Web to Grade Mathematics Tests. In J.C. Kinshuk & T. Okamoto (Eds.), *Advanced Learning Technology: Design and Development Issues* (pp. 89–92). Los Alamitos, CA, U.S.A.: IEEE Computer Society.

Kounin, J.S. (1970). *Discipline and Group Management in Classrooms*. London: Holt, Rinehart and Winston.

Landis, J.R., & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174, https://doi.org/10.2307/2529310

Lawson, D. (2002). Computer-aided assessment in mathematics: Panacea or propaganda? *International Journal of Innovation in Science and Mathematics Education*, 9(1).

MacBean, J., Graham, T., & Sangwin, C. (2004). Group work in mathematics: a survey of students' experiences and attitudes. *Teaching Mathematics and its Applications*, 23(2), 49–68, https://doi.org/10.1093/teamat/23.2.49

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist*, 50(9), 741–749.

O'Rourke, J. (1987). *Art gallery theorems and algorithms*. New York, NY, U.S.A.: Oxford University Press.

Pacheco-Venegas, N.B., Lopez, G., & Andrade-Aréchiga, M. (2015). Conceptualization, development and implementation of a web-based system for automatic evaluation of mathematical expressions. *Computers & Education*, 88, 15–28, https://doi.org/10.1016/j.compedu.2015.03.021

The Quality Assurance Agency for Higher Education. (2015). *Subject benchmark statement: Mathematics, statistics and operational research*. Gloucester, U.K.

Quinney, D. (2010). The Role of E-Assessment in Mathematics. In P. Bogacki (Ed.), *Electronic Proceedings of the Twenty-second Annual International Conference on Technology in Collegiate Mathematics*, Chicago, Illinois (pp. 279–288). Retrieved from http://archives.math.utk.edu/ICTCM/VOL22/S093/paper.pdf

Rønning, F. (2017). Influence of computer-aided assessment on ways of working with mathematics. *Teaching Mathematics and its Applications*, 36(2), 94–107, https://doi.org/10.1093/teamat/hrx001

Rowlett, P.J. (2013). Developing a Healthy Scepticism About Technology in Mathematics Teaching. *Journal of Humanistic Mathematics*, 3(1), 136–149, https://doi.org/10.5642/jhummath.201301.11

Sangwin, C. (2004). Assessing mathematics automatically using computer algebra and the internet. *Teaching Mathematics and its Applications*, 23(1), 1–14, https://doi.org/10.1093/teamat/23.1.1

Sangwin, C.J. (2007). Assessing elementary algebra with STACK. *International Journal of Mathematical Education in Science and Technology*, 38(8), 987–1002, https://doi.org/10.1080/00207390601002906

Sangwin, C. (2015). Computer Aided Assessment of Mathematics Using STACK. In S.J. Cho (Ed.), *Selected Regular Lectures from the 12th International Congress on Mathematical Education* (pp. 698–713). Cham, Switzerland: Springer, https://doi.org/10.1007/978-3-319-17187-6_39

Seaton, K.A. (2019). Laying groundwork for an understanding of academic integrity in mathematics tasks. *International Journal of Mathematical Education in Science and Technology*, *50*(7), 1063–1072. https://doi.org/10.1080/0020739X.2019.1640399

Thomlinson, M.M., Robinson, M., & Challis, N.V. (2010). Coursework, what should be its nature and assessment weight? In M. Robinson, N. Challis & M. Thomlinson

(Eds.), *Maths at University: Reflections on experience, practice and provision* (pp. 122–126). Birmingham, U.K.: More Maths Grads.

Appendix – Sample piece of individualised work

A sample individualised question sheet is shown in figure 1. The diagram and various numbers in the questions have been randomised. The ID number allows matching question and answer sheets.

A sample individualised answer sheet is shown in figure 2. The answer sheet gives information where it is possible to calculate this, to assist marking. This might be the answer or a short-hand reminder of the details of the question asked.

Table 1. Summary of evaluation process.

Phase	**Details**		**Evaluation target**
Second marker experiment	Reference	Written test, exam conditions	Calibrating expectations re. marker consistency (ICC).
		Coursework, open-ended	
	Main	Second-marking of individualised coursework	Marker consistency (ICC). Validity (professional judgement). Efficiency for staff (marker comments).
Student feedback	Questionnaire	Group A, who took the summative individualised coursework described	Student experience. Risk of copying/collusion.
		Group B, who used the same individualised assessment technique for formative work at a different institution	Comparison group to detect innovator influence.
Comparison of marks	Intra-group individual marks variation		Occurrence of copying/collusion.
	Raw group marks and individualised coursework marks		

Table 2. Original and second marks for ten pieces of work blind-second-marked for the written examination reference experiment.

Student	**Original marker (%)**	**PR (%)**
1	72	72
2	88	86
3	80	80
4	94	94
5	54	54
6	72	72
7	78	76
8	52	48
9	60	58
10	60	62

Table 3. Original and second marks for fourteen pieces of work blind-second-marked for the coursework reference experiment.

Student	**Original marker (%)**	**PR (%)**
1	58	56
2	65	55
3	48	43
4	80	71
5	68	45
6	75	74
7	70	58
8	60	63
9	80	68
10	70	58
11	75	61
12	62	58
13	55	51
14	82	72

Table 4. Original and second marks for five pieces of work submitted for the individual coursework.

Student	**PR (%)**	**Second-marker A (%)**	**Second-marker B (%)**	**Second-marker C (%)**
1	56	31	38	49
2	74	64	59	72
3	67	72	74	77
4	67	46	51	51
5	74	59	54	69

Table 5. Number of students indicating level of agreement with four statements about individualised work.

Group	‘Strongly disagree’ 1	2	3	4	5 ‘Strongly agree’	p-value
‘I disliked having different questions because I wanted to work together with another student on our answers.’						
A	12	16	13	1	0	0.08851
B	3	8	2	3	0	
‘I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism.’						
A	0	1	10	22	9	0.6193
B	0	0	4	6	6	
‘I liked having different questions because it meant that no one could copy from me.’						
A	0	4	14	17	7	0.1366
B	2	2	5	3	4	
‘If we had been set identical questions, (members of our group [group A]/some students [group B]) would have copied answers from other students.’						
A	2	5	11	15	9	0.3132
B	2	1	6	2	5	

Table 6. Number of students answering yes and no to two questions about copying.

Group	**Yes**	**No**	**p-value**
‘While at university, I have copied work from other students’			
A	22	19	0.1513
B	5	11	
‘While at university, other students have copied work from me’			
A	35	7	0.2811
B	11	5	

Table 7. Marks range and standard deviation for the individualised coursework within each group.

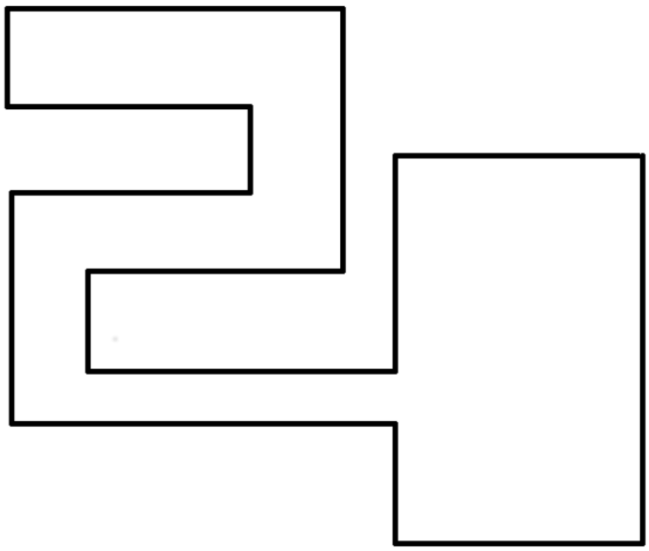
Group	**Individualised coursework marks range for group members**	**Individualised coursework standard deviation for group members (3 d.p.)**
1	31	11.411
2	30	10.706
3	23	8.216
4	28	9.584
5	30	9.513

Figure 1. Sample individualised question sheet.

ID: 1

Individual assignment 2

1.



a) Show, by triangulating and three-colouring the polygon, how many guards are necessary to guard every point in the museum shown above at any one time.

b) In reality, given a staff of 11 guards, could you arrange for every point in the museum shown to be guarded 24 hours a day and seven days a week?
If so, how would you arrange this?
If not, why not and how many staff would you require?

2. Draw a polygon using 16 vertices for which 5 guards are necessary to guard every point at any one time.

Figure 2. Sample individualised answer sheet.

ID: 1

Individual assignment 2

1.
a) 3 guards (variant 1).

b) asking about 11 guards on staff.

2.
using 16 vertices for which 5 guards are necessary