Can training improve marker accuracy at detecting contract cheating? A multi-disciplinary pre-post study

Journal:	*Assessment and Evaluation in Higher Education*
Manuscript ID	CAEH-2018-0347.R1
Manuscript Type:	Original Paper
Keywords:	academic integrity, assessment, marking, contract cheating, cheating

Can training improve marker accuracy at detecting contract cheating? A multi-disciplinary pre-post study

<blind>

Contract cheating occurs when students outsource assessed work. In this study, we asked experienced markers from four disciplines to detect contract cheating in a set of 20 discipline-specific assignments. We then conducted a training workshop to improve their detection accuracy, and afterwards asked them to detect contract cheating in 20 new assignments. We analysed the data in terms of sensitivity (the rate at which markers spotted contract cheating) and specificity (the rate at which markers spotted real student work). Pre-workshop marker sensitivity was 58% and specificity was 83%. Post-workshop marker sensitivity was 82% and specificity was 87%. The increase in sensitivity was statistically significant, however the increase in specificity was not. These results indicate that markers can often detect contract cheating when asked to do so, and that training may be helpful in improving their accuracy. We suggest that markers' suspicions may be crucial in addressing contract cheating.

Keywords: academic integrity; assessment; marking; contract cheating; cheating

Contract cheating is the commissioning of bespoke university assignments with the intention of submitting them for assessment (Lancaster and Clarke 2007). Although contract cheating can include parents or friends doing work for free on behalf of a student, the term has become synonymous with online essay mills and assignment writing services. For a fee, thousands of these websites offer to produce custom-made assignments in as little as a few hours. Contract cheating websites are sophisticated online businesses, and they make a range of promises, including guaranteeing confidentiality and customer satisfaction (Rowland et al. 2017). While contract cheating includes the outsourcing of nearly any task type, including examinations, this paper is focused on contract cheating of take-home tasks through commercial contract cheating sites. Analysis of self-report studies published since 2014 place the prevalence of students admitting to this sort of contract cheating at around 15.7% of students (Newton 2018).

Students who succeed at contract cheating appear to have met learning outcomes they have not demonstrated themselves. This creates significant problems. Community confidence in higher education suffers when students appear to be able to buy their way through degrees. Most critically, public safety is endangered when students cheat to gain accreditation into professions with significant responsibility.

We support the global academic integrity movement's positive agendas and its focus on integrity, education and the promotion of ethical behavior (Bertram-Gallant 2015, Davies and Howard 2016). However, we also believe that universities are responsible to their communities to have measures in place to detect academic dishonesty, and that the existence of these measures may have a significant deterrent effect. While the dramatic fall in copy-paste plagiarism since the mid-2000s can be partially attributed to educational interventions, there is significant evidence that the widespread introduction of text-matching tools like Turnitin played a significant role as well (Bertram-Gallant 2015, Li 2012). More recently, there is evidence that cheating detection in online exams via remote proctoring – such as the monitoring of students through biometrics and webcams – also has a deterrent effect (Brothen and Peterson 2012). Detecting cheating and improving academic integrity are complementary, not contradictory; we believe universities are responsible to their students and to the broader community to make evidence-based efforts to detect cheating.

Detecting contract cheating

University teachers are at the front lines of detecting contract cheating, as they are the ones who undertake routine marking of student work. Recent large-scale survey research by Harper et al. (2018) suggests that most academics have encountered what they suspect to be contract cheating. Two thirds of their 916 academic staff participants had identified what they thought was contract cheating one or more times. Slightly less than 40 per cent of their participants had spotted contract cheating five times or more. It is difficult to make judgements about academics' ability at detecting contract cheating based on this type of research because (a) it is based on self-report from academics who volunteered to fill in a survey about contract cheating, (b) we do not know if the academics' suspicions were correct and (c) we do not know how much contract cheating they did not spot. However, given a 15.7% prevalence rate since 2014 in terms of students admitting to having contract cheated (Newton 2018), and that each academic teaches many students over their career, it is likely that most contract cheating goes undetected.

Early work on contract cheating detection relied on publicly available data, such as students negotiating the purchase of work in open online forums (Clarke and Lancaster 2007). While monitoring these forums may have initially been an effective approach, in recent years contract cheating has become much more sophisticated and clandestine. Contract cheating transactions are now conducted privately and confidentially, with guarantees of anonymity. While there will still be a small number of contract cheating cases detected due to carelessness on the part of students, a reliance on this is likely to leave most contract cheating undetected.

There has been limited research conducted in controlled settings that gathered data about markers' accuracy at detecting known contract cheating assignments. To date, there have been three small studies involving the purchase and detection of contract cheating assignments by markers that we are aware of. The first, by Lines (2016), involved markers marking 26 history assignments bought from contract cheating websites. Contract cheating was not mentioned to these markers and they did not suspect it; most scored passing grades and 23 appeared acceptable when put through Turnitin's text-matching software. In a later study, Dawson and Sutherland-Smith (2018) specifically asked markers to detect contract cheating, and found they were able to detect it most of the time. In that pilot study, they asked seven markers to mark the same bundle of 20 second-year psychology assignments, including 14 real assignments and 6 purchased assignments. Their reporting focused on two measures: sensitivity and specificity. Sensitivity is the true positive rate, or the proportion of contract cheating detected. They reported a sensitivity rate of 62% (95% CI: 0.46–0.76), meaning that 62% of the time markers accurately detected contract cheating. Specificity is the true negative rate, or the proportion of legitimate student work that is not incorrectly suspected of contract cheating. They reported a specificity rate (true negative) of 96% (95% CI: 0.89–0.99). This means that 96% of the time markers accurately identified real student work and did not flag it as contract cheating. The third study, by Medway, Roper, and Gillooly (2018), involved two contract cheating assignments being marked by ten markers who were not aware they were involved in a study about contract cheating. As in Lines' study, these markers were not primed to look for contract cheating and they did not detect it. Taken together, these three studies reveal two promising possibilities for contract cheating research. Firstly, when markers are not alerted to the possibility of contract cheating they might not detect it. Secondly, when markers are told of the possibility of contract cheating they might spot it.

However, this existing empirical research on contract cheating detection is somewhat problematic from a practical perspective. The generalisability of all three studies is poor, as they are small-scale and from single disciplines. From a statistical perspective, the confidence interval for contract cheating detection sensitivity in Dawson and Sutherland-Smith (2018) is very broad (95% CI: 0.46–0.76), meaning that the true detection rate may have really been anywhere from 46% to 76%. No attempts to replicate these studies have been published, which although not uncommon in educational research (Makel and Plucker 2014), does present problems in terms of reliability. Also, while they represent a useful baseline for contract cheating detection, they do not offer empirically-tested approaches to improve detection, with the possible exception of increasing marker awareness of contract cheating.

This paper provides larger-scale evidence of marker accuracy at detecting contract cheating, through a multi-disciplinary study with markers before and after a workshop on detecting contract cheating. In particular it addresses the questions:

RQ1: “Can markers detect contract cheating?”

RQ2: “Can training improve marker accuracy at detecting contract cheating?”

Method

This study used a pre-post design to compare the accuracy of markers at detecting contract cheating before and after a training workshop. Markers were each given a bundle of 20 assignments, consisting of a mixture of 6 contract cheating assignments and 14 assignments provided by students. This prevalence was necessary to achieve sufficient statistical power for sensitivity, which we consider the primary outcome variable for this study, within the resourcing available. The contract cheating work had been purchased from a range of contract cheating websites by our team, and the real student work was provided voluntarily by students. Markers then attended a training workshop, and afterwards they were given a new bundle of assignments of the same size procured in the same manner.

Course and assessment contexts

There were four compulsory units across three faculties involved in this study: one first year marketing unit; one first year unit in genetics; one second year unit in psychology and one third year unit from nutrition. These units were chosen because the Associate Deans of Teaching and Learning in each faculty recommended them as being the most suitable to meet our criteria, which were: large enrolments with a minimum of 400 students, a minimum of two written tasks in each unit, offered in both face-to-face and online modes and with at least 5 sessional marking staff in each unit. The teaching staff co-ordinating and administering these units were willing to participate, as were the marking staff.

The first year marketing unit focusses on working with an online environmentally friendly company’s launch of green products. The unit’s first written task of 750 words was to outline an approach to segment the market and provide evidence that their approach targeted different consumers in the specific marketplace. The second task relied on feedback from the first and asked students to provide a full marketing mix strategy for the company in 1,000 words.

The first year genetics unit focused on understanding molecular cell biology and its role in the control of gene expression and the principles of DNA technologies. The first task required students to write a 2,000 word scientific report, following the standard scientific IMRaD (Introduction, Method, Results and Discussion) report format. The second task asked students to write a full practical report on DNA extraction in relation to laboratory experiments.

The second year psychology unit focusses on the processes of cognition and the research methods psychologists use to study these cognitive processes. The unit's first assessment task was a 900 word written introduction to laboratory report. The second task was to use the feedback from the first task and write a complete 1,800 word laboratory report, following the APA conventions.

The third year nutrition unit's first task was to prepare a fact sheet reviewing the evidence linking obesity with systemic inflammation and potential inflammation treatment using vitamin D. It was 1,000 words. The second task required students to construct a mock research article in 2,000 words using a set of results provided on the topic of FODMAPs (Fermentable, Oligo-, Di-, Mono-saccharides And Polyols, which are a type of carbohydrate that is poorly absorbed by the small intestine) and obesity.

Participants and recruitment

All students enrolled in the units were invited to participate in this study by emailing their consent to the research team. Student participation was voluntary and entailed them giving us permission to use anonymised copies of their assignments in the study. Students were not approached until after the teaching period was completed, and they were assured that participation in the study could not affect their grades or result in them being accused of any academic integrity breach. All experienced markers on each unit were invited to participate. Markers were paid for their time at the usual marking rate. The typical marker in our study had been marking on that unit of study for multiple semesters and was currently undertaking a doctorate in the same discipline. The teaching staff on the units were not members of the research team, however they did assist in recruiting marker and student participants.

Data collection – assignments

Student assignments were obtained from the university learning management system, and were anonymised by a research assistant. A total of 48 contract cheating assignments were purchased from 13 different websites by a research assistant. Contract cheating sites were given the same instructions that students in the units were given, including the assignment specifications, any required materials, and marking criteria. Transactions were conducted by PayPal.

Data collection – marker decisions

Markers were provided with a bundle of assignments and asked to identify which assignments they thought were contract cheating, and which assignments they thought were not contract cheating. Then, after the workshop, this process was repeated with a new batch of assignments. Markers returned their decisions via email on a spreadsheet.

Workshop

After the first round of marking had been collated, the researchers analysed the data to find which assignments had proven most challenging for markers; that is, the assignments that led to the most 'false positives' (incorrect detection of contract cheating) and 'false negatives' (incorrectly labelling contract cheating as legitimate work). A three-hour workshop was developed where markers worked with their peers on the same unit and discussed the four most problematic of these assignments, including at least one legitimate assignment and at least one contract cheated assignment. In brief, markers were provided with an assignment, they debated if they thought it was contract cheating, then they were told if it was contract cheated or not. At that point they discussed and wrote down implications that this may have for contract cheating detection, and then the same process was repeated for all the other selected assignments. Markers were not told that these were the most problematic assignments. At the end of the workshop the markers produced a list of 'indicators of contract cheating' for their course, which the researchers collected, compiled and circulated to the markers by email on completion of the workshop and prior to the markers commencing the second round of marking. The workshop facilitators had access to the pre-workshop assignments before the workshop, but they did not have access to the post-workshop assignments until after the workshop. We piloted this workshop design with a different set of markers and adapted it, based on their feedback, prior to implementing it in this study. An outline and agenda for the workshop is provided in Appendix A.

Research ethics

All student and cheated work was anonymised. Researchers have previously acknowledged that there are ethical concerns relating to the purchasing of contract cheating assignments for research (Dawson and Sutherland-Smith 2018, Medway, Roper, and Gillooly 2018). In particular, we were concerned that we were providing financial support to businesses that we think are unethical. However, given the size of the global contract cheating industry, which is estimated to exceed £200m (Adams 2015), our support is relatively minor. As with other research, and in consultation with relevant institutional, national and learned society guidelines (American Educational Research Association 2011, Hammersley and Traianou 2012, National Health and Medical Research Council, Australian Research Council, and Australian Vice-Chancellors' Committee 2015), the decision to engage in this work considered the potential harms and benefits of conducting the work, as well as their likelihood and magnitude. This study was approved by the relevant ethics committee at our university (approval number HEAG-H 136:2016) as low risk research.

Results

Fifteen markers participated in this study in total: four from psychology, four from nutrition, two from marketing, and five from biology. Each marker marked 20 assignments from their unit before the workshop, and 20 assignments after the workshop, resulting in a total of 600 instances of marking. Tables 1 and 2 below provide descriptive statistics about the decisions markers made in each unit. In these tables, a 'true positive' is a correct detection of contract cheating; a 'true negative' is correctly not detecting contract cheating (ie identifying legitimate student work); a 'false positive' is flagging real student work as contract cheating; and a 'false negative' is not detecting a piece of contract cheating work.

<Tables 1 and 2>

From the data in tables 1 and 2 it is possible to calculate sensitivity and specificity statistics for the pre- and post-workshop marking. Using the method described by Newcombe (2001) it is also possible to test if there is a statistically significant difference between the pre- and post-workshop marking detection rates. These statistics are presented in Table 3 below:

<Table 3>

The pre-workshop scores for sensitivity and specificity are lower (therefore worse) than the post-workshop scores: prior to the workshops markers detected 58% of contract cheating and 83% of legitimate work, and after the workshops markers detected 82% of contract cheating and 87% of students' legitimate work. However, only the difference in sensitivity was statistically significant.

A pilot study on marker accuracy at detecting contract cheating (Dawson and Sutherland-Smith 2018) also used untrained markers, and had better results for both sensitivity and specificity. Using Newcombe's (2001) method it is also possible to determine if the difference in sensitivity and specificity between the untrained markers in the pilot and this larger-scale study is significant. This comparison is presented in Table 4 below. While the difference in sensitivity was not significant, the difference in specificity was significant.

<Table 4>

The DAG_Stats package (Mackinnon 2000) was used to calculate further statistics on both the pre-workshop and post-workshop marking results for this study independently, which are presented in Table 5. These further analyses are alternative representations of essentially the same results, which are useful in addressing different practical questions, like 'when markers thought they spotted contract cheating, how likely is it they were correct'? (predictive value of positive test), or 'what was the overall accuracy of markers'? (correct classification rate).

<Table 5>

The correct classification rate is the overall accuracy rate of markers in this study, and it can be interpreted as the likelihood of any given decision by a marker as being correct. Prior to the workshop, markers made the right decision 75% of the time; after the workshop they were right 86% of the time. The incorrect classification rate is the inverse of these numbers.

The predictive value of positive and negative tests has substantial use in practice, as it represents the likelihood that a given positive or negative decision was correct. Before the workshop, when markers thought they had detected contract cheating they were correct 59% of the time; after the workshop they were correct 73% of the time. When markers thought they were looking at legitimate student work they were correct 82% of the time before the workshop and 92% of the time after the workshop.

The false positive rate represents the proportion of real student work that was incorrectly flagged as contract cheating, and the false negative rate represents the proportion of contract cheating work that was classified as real student work by markers. The false positive rate and the false negative rate are the inverse of specificity and sensitivity respectively. Prior to the workshop markers incorrectly classified 17% of legitimate work as contract cheating and 42% of contract cheated work as legitimate; after the workshop this decreased to 13% of legitimate work flagged as cheating and 18% of cheated work going undetected.

Discussion and conclusions

This paper provides the largest study to date of marker accuracy at detecting contract cheating, and tests a workshop design to improve detection. It also provides more sophisticated statistical evidence than previous studies into contract cheating detection by markers.

Compared with a pilot study by Dawson and Sutherland-Smith (2018), these new results for untrained marker accuracy at detecting contract cheating are lower in terms of sensitivity and specificity, however this is only statistically significant for specificity. The new specificity score presented in this paper of 83%, should lead to greater caution in handling marker suspicions of contract cheating than Dawson and Sutherland-Smith's (2018) pilot score of 96%. The broad confidence interval of this new result (95% CI: 0.77 - 0.88) should lead to even greater caution: it is possible that the true rate of incorrect suspicions of contract cheating was actually as high as 23%, if the lower bound of the confidence interval (77%) represents reality. Of similar concern is the low score for the predictive value of a positive test of 0.59 (95% CI: 0.48 - 0.69), which suggests that more than 40% of the time when pre-workshop markers came to the conclusion that a piece of work was contract cheating, they were wrong. Taken together, this high false positive rate and low predictive value for a positive test are a reminder to use appropriate caution when untrained markers think they have detected contract cheating. We used a 30% prevalence rate for contract cheating in this study in order to provide the necessary statistical power for our primary measure, sensitivity, given the resources available. In circumstances where prevalence is lower, the predictive value of a positive test may be even lower. However, none of this suggests that marker suspicions of contract cheating should be dismissed. Our markers were right most of the time, and marker suspicion is a crucial first step in identifying contract cheating.

This paper demonstrates that the same markers, when taken through a workshop on detecting contract cheating, can show improved sensitivity without lowering their specificity. For our markers this represented a statistically significant improvement of 24% in terms of their detection rate. While a 24% increase may sound small, this represents a shift from 58% to 82% detection of contract cheated work. Expressed differently, untrained markers let contract cheating slip by 42% of the time, which is more than twice as often as trained markers, who only missed 18% of contract cheated assignments. However, it is worth acknowledging that alerting untrained markers to the potential of contract cheating and asking them to detect it resulted in an increase from Lines' (2016) and Medway, et al.'s (2018) result of 0% to our pre-workshop result of 58%. Both awareness raising and marker training could play a role in addressing contract cheating.

As has been reported previously (Lancaster and Clarke 2017) it is possible to contract cheat on almost any assessment task in almost any discipline. Our results show that for the range of tasks we considered, which spanned a range of disciplines, it was also possible to detect contract cheating some of the time. However, the detection rates varied substantially – from 100% sensitivity and specificity for post-workshop markers in marketing, to 33% sensitivity for pre-workshop marketing and 79% specificity for pre-workshop psychology markers. This significant diversity suggests that generalising from our results to any specific real-world task could be hazardous.

Given the detection rates reported in this article, it is tempting for us to speculate on what made contract cheating detectable. We have chosen not to share our own internal views on that question in this article for a number of reasons. Firstly, on looking at the 'Indicators of contract cheating' our markers identified there is more variation than similarity across disciplines and task types. It is unclear how much of these lists will be generalizable, and how much is unique to particular tasks, units or disciplines. Secondly, we do not think we have the level of evidence required to make claims of this type. We think that making unfounded claims about the indicators of contract cheating, even tentatively, could have unintended consequences. Thirdly, and possibly most importantly, the dissemination of potential indicators of contract cheating is likely to lead to contract cheating sites adapting and improving their products. While we think the identification of features of contract cheating is a worthwhile endeavor, however it is one that needs to be undertaken carefully, through studies that are designed specifically with this purpose.

A significant caveat to this study is that while we have demonstrated that markers can often detect contract cheating, and that detection accuracy might be improved through training, marker detection alone is not necessarily sufficient evidence to satisfy the burden of proof for a contract cheating allegation. Our markers were required to make a judgement, not build a case to evidence their suspicions of contract cheating. Building a case to allege an instance of contract cheating can be almost impossible in some instances, even when the marker is sure of contract cheating. The difficulty in proving cases of contract cheating is likely the major reason why academics sometimes do not pursue cases of suspected contract cheating (Harper et al. 2018). We suggest that marker detection be used in conjunction with other strategies, and suggest that vivas or other interactive approaches may be used in cases of suspected contract cheating, although we are aware of logistical issues and concerns about performance by English as Other Language speakers in vivas.

Like other pre-post designs, there is a chance that the ‘pre’ component of this study was actually the active ingredient that led to the change observed in the ‘post’ component of this study. In other words, practicing making decisions about contract cheating may have led to improvements in marker accuracy at actually detecting contract cheating. It is also possible that differences between the batches of assignments, peculiarities of our markers or the disciplines chosen, or some other unknown factor is the cause of the observed changes. These are limitations of our study that we have not been able to isolate due to resourcing constraints. Another potential limitation of the study is that we only tested one workshop design. It is possible that more effective workshop designs exist. While our design took a social constructivist approach in building a shared understanding of detecting contract cheating, there may be justification for providing more direct instruction in what contract cheating work typically looks like (Kirschner, Sweller, and Clark 2006). The study is also a ‘simulation’ of marking rather than a real-life marking experience; it is possible that if we had conducted this study on live student work, with marks that mattered, that the results would differ. A further limitation to this study is that in collecting ‘legitimate’ student work we have assumed that the work with which we have been provided by students is not, itself, contract cheated.

Returning to the research questions posed in the introduction to this paper, we have demonstrated that markers can sometimes detect contract cheating – 58% of the time in our pre-workshop study. However, their detection rates are not perfect, they vary across disciplines, and detection is accompanied by false positives. We have also demonstrated that markers who completed a workshop on detecting contract cheating had improved detection rates compared to pre-workshop rates. Based on our results, we recommend universities inform all teaching staff about what contract cheating is, and we suggest that in areas of particular concern that it may be useful to provide markers with specialist training on detecting contract cheating. However, further research into the efficacy of training and other approaches to improving marker accuracy at detecting contract cheating is still needed.

Contract cheating website operators make sophisticated sales pitches to potential cheating students, involving money-back guarantees, 24-hour online support, and even copies of Turnitin similarity reports (Rowland et al. 2017). While we cannot stop most of these services, our study provides strong evidence against one of their common promises: that contract cheating is undetectable. We have shown across a multitude of disciplines and task types that when our markers were looking for contract cheating, they usually found it, and that when they were trained they were even more accurate at spotting contract cheating. This is an important finding for the field of academic integrity, and one that could usefully be integrated in student-focused campaigns to reduce contract cheating.

Acknowledgements

<blind>

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

<blind>

Notes on contributors

<blind>

References

Adams, R. 2015. "Cheating found to be rife in British schools and universities." accessed 14 June 2015. http://www.theguardian.com/education/2015/jun/15/cheating-rife-in-uk-education-system-dispatches-investigation-shows.

American Educational Research Association. 2011. "AERA Code of Ethics, Approved by the AERA Council February 2011." *Educational Researcher* 40 (3):145-156. doi: 10.3102/0013189X11410403.

Bertram-Gallant, T. 2015. "Leveraging institutional integrity for the betterment of education." In *Handbook for academic integrity*, edited by Tracey Bretag, pp.979-994. Singapore: Springer.

Brothen, T., and G. Peterson. 2012. "Online Exam Cheating: A Natural Experiment." *International Journal of Instructional Technology and Distance Learning* 9 (2):15-20.

Clarke, R., and T. Lancaster. 2007. "Establishing a Systematic Six-Stage Process for Detecting Contract Cheating." 2007 2nd International Conference on Pervasive Computing and Applications, 26-27 July 2007.

Davies, L., and R. M. Howard. 2016. "Plagiarism for the Internet: Fears, Facts and Pedagogies." In *Handbook of Academic Integrity*, edited by Tracey Bretag, 591-606. Singapore: Springer.

Dawson, P., and W. Sutherland-Smith. 2018. "Can markers detect contract cheating? Results from a pilot study." *Assessment & Evaluation in Higher Education* 43 (2):286-293. doi: 10.1080/02602938.2017.1336746.

Hammersley, M., and A. Traianou. 2012. "Ethics and Educational Research, British Educational Research Association on-line resource.". British Educational Research Association, accessed 10th September. http://www.bera.ac.uk/resources/ethics-and-educational-research.

Harper, R., T. Bretag, C. Ellis, P. Newton, P. Rozenberg, S. Saddiqui, and K. van Haeringen. 2018. "Contract cheating: a survey of Australian university staff." *Studies in Higher Education*:1-17. doi: 10.1080/03075079.2018.1462789.

Kirschner, P. A., J. Sweller, and R. E. Clark. 2006. "Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching." *Educational Psychologist* 41 (2):75-86. doi: 10.1207/s15326985ep4102_1.

Lancaster, T., and R. Clarke. 2007. "The phenomena of contract cheating." In *Student plagiarism in an online world: Problems and solutions*, edited by T. Roberts, 144-158. Hershey, USA: Idea Group Inc.

Lancaster, T., and R. Clarke. 2017. "Rethinking assessment by examination in the age of contract cheating." Plagiarism across Europe and Beyond, Brno, Czech Republic, May 24-26.

Li, Y. 2012. "Studying the Issue of Plagiarism at a University in Hong Kong: An Exploration of the Teaching and Learning Processes." *Hong Kong Journal of Applied Linguistics* 13 (2):22-32.

Lines, L. 2016. "Ghostwriters guaranteeing grades? The quality of online ghostwriting services available to tertiary students in Australia." *Teaching in Higher Education*:1-26. doi: 10.1080/13562517.2016.1198759.

Mackinnon, A. 2000. "A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement." *Computers in Biology and Medicine* 30 (3):127-134. doi: https://doi.org/10.1016/S0010-4825(00)00006-8.

Makel, M. C., and J. A. Plucker. 2014. "Facts Are More Important Than Novelty: Replication in the Education Sciences." *Educational Researcher* 43 (6):304-316. doi: 10.3102/0013189x14545513.

Medway, D., S. Roper, and L. Gillooly. 2018. "Contract cheating in UK higher education: A covert investigation of essay mills." *British Educational Research Journal* 44 (3):393-418. doi: 10.1002/berj.3335.

National Health and Medical Research Council, Australian Research Council, and Australian Vice-Chancellors' Committee. 2015. National Statement on Ethical Conduct in Human Research Research. edited by Commonwealth of Australia. Canberra.

Newcombe, R. G. 2001. "Simultaneous comparison of sensitivity and specificity of two tests in the paired design: a straightforward graphical approach." *Statistics in Medicine* 20 (6):907-915. doi: 10.1002/sim.906.

Newton, P. M. 2018. "How Common Is Commercial Contract Cheating in Higher Education and Is It Increasing? A Systematic Review." *Frontiers in Education* 3 (67). doi: 10.3389/feduc.2018.00067.

Rowland, S., C. Slade, K.-S. Wong, and B. Whiting. 2017. "'Just turn to us': the persuasive features of contract cheating websites." *Assessment & Evaluation in Higher Education*:1-14. doi: 10.1080/02602938.2017.1391948.

	Psychology	**Nutrition**	**Marketing**	**Biology**	**Total**
True positives	21	14	4	13	52
True negatives	44	48	24	58	174
False positives	12	8	4	12	36
False negatives	3	10	8	17	38
Total	80	80	40	100	300

Table 1. Pre-workshop marking results.

	Psychology	**Nutrition**	**Marketing**	**Biology**	**Total**
True positives	22	23	12	17	74
True negatives	50	52	28	53	183
False positives	6	4	0	17	27
False negatives	2	1	0	13	16
Total	80	80	40	100	300

Table 2. Post-workshop marking results.

	Pre-workshop	Post-workshop	Difference
Sensitivity	0.58 [0.47, 0.68]	0.82 [0.73, 0.89]	0.24 [0.11, 0.37]
Specificity	0.83 [0.77, 0.88]	0.87 [0.82, 0.91]	0.04 [-0.03, 0.11]

Table 3. Comparing pre- and post-workshop sensitivity and specificity scores. Numbers in brackets are 95% CIs.

	Pilot study	**Pre-workshop markers in this study**	**Difference**
Sensitivity	0.62 [0.46, 0.76]	0.58 [0.47, 0.68]	-0.04 [-0.21, 0.14])
Specificity	0.96 [0.89, 0.99]	0.83 [0.77, 0.88]	-0.13 [-0.19, -0.06]

Table 4. Comparing pre- and post-workshop sensitivity and specificity scores. Numbers in brackets are 95% CIs.

	Pre-workshop	**Post-workshop**
Correct classification rate	0.75 [0.70, 0.80]	0.86 [0.81, 0.89]
Incorrect classification rate	0.25 [0.20, 0.30]	0.14 [0.11, 0.15]
Predictive value of positive test	0.59 [0.48, 0.69]	0.73 [0.64, 0.82]
Predictive value of negative test	0.82 [0.76, 0.87]	0.92 [0.87, 0.95]
False positive rate	0.17 [0.12, 0.23]	0.13 [0.09, 0.18]
False negative rate	0.42 [0.32, 0.53]	0.18 [0.11, 0.27]

Table 5. Further statistical analyses of pre- and post-workshop results. Numbers in brackets are 95% CIs.

Appendix A: Marker contract cheating detection workshop outline

Intended learning outcome: on completion of this workshop, markers should be better able to detect contract cheating in their particular unit/subject

Time: three hours

Participants: experienced markers

Space: workshop-style room, with markers sharing a table with other markers from the same unit/subject

Required materials: four assignments per unit/subject, including at least one contract cheating assignment and at least one legitimate student assignment. Assignments are selected based on them being difficult to tell if they are contract cheating or not.

Pre-work: markers to read each assignment and make a preliminary decision about if each is contract cheating or not

Workshop agenda

15 minutes: Introducing the facilitators, the markers and the workshop

15 minutes: Markers reflect in groups on the experience of marking contract cheating assignments

4x20 minutes: Each table group discusses each of their assignments for 20 minutes per assignment, focused on deciding if each one is contract cheating or not, and why. Each marker takes notes about potential ‘Indicators of contract cheating’. After a decision has been made, markers are told if that assignment was contract cheating or not. Table groups then discuss and revise their potential ‘Indicators of contract cheating’. Then they move on to the next assignment.

60 minutes: Marker groups construct a shared ‘Indicators of contract cheating’ list.

10 minutes: Workshop close; brief reflection and Q&A session; thank participants; collect all copies of assignments and each group’s ‘Indicators of contract cheating’ list.

Post-workshop: facilitators type up each group’s ‘Indicators of contract cheating’ list and email it to that marker group.