Anonymous examination marking at University of Cape Town: The quest for an "agonising-free zone"

Suellen Shay and Barbara Jones

A/Prof. Suellen Shay
Centre for Higher Education Development
University of Cape Town
Rondebosch 7700
sshay@ched.uct.ac.za

B. Jones
Freelance Educational Researcher
bemjones@telkom.net

ABSTRACT 2

In 2003 the University of Cape Town introduced an anonymous examination policy. This paper reports on a study of the impact of the implementation of this policy on student performance. Comparisons of student results pre- and post policy implementation showed no evidence of negative or positive discrimination of students in the examination marking. Interviews with course conveners suggested however, that, irrespective of the policy, markers infer student identity from examinations and that these inferences can influence their assessment. The most commonly cited example was 'sympathetic marking', that is, assessors marking more generously if they infer a student to be educationally under-prepared. The paper concludes that the implementation of this policy has had a limited impact on strengthening the validity of assessment results, but is likely to be retained given both staff and students' perceptions that the policy 'objectifies' the marking process.

Introduction

Anonymous marking of examinations, that is, concealing the name of the student from the marker, has become standard good practice in many universities around the world. At the University of Cape Town (UCT) calls for anonymous examination came from students in the early 1990's – a politically turbulent time in South Africa's history -- who argued that anonymous marking would minimize the possibility of "passive discrimination" or "unconscious (gender and racial) bias" in the final written examinations[1]. Responses from the academic community to this policy proposal ranged from general support to strong opposition. In 2003 after more than a decade of institutional debate, UCT's Senate finally approved an anonymous examination policy.

[1] SRC letter entitled Anonymity of Evaluation Scripts, dated 28 February 1991.

The policy claims that anonymous marking will strengthen the validity of UCT's assessment in two ways[2]. Firstly, anonymous marking minimizes the possibility of subconscious[3] irrelevant inferencing which may discriminate for or against students, in particular inferences based on gender, race and any other kind of information that can be made on the basis of a student's name. Secondly, anonymous marking strengthens UCT's assessment system by addressing students' *perceptions* that the marking could be biased. Even where discrimination cannot be conclusively confirmed, it is important that students perceive the assessment to be free of bias.

While gaining Senate approval for this policy was challenging, what proved to be almost as difficult was reaching consensus on a system of implementation. The system finally adopted, after much discussion, was to conceal the students' name and student number under a sealed corner flap. For practical administrative purposes, some form of identification was needed on the front cover of the script. It was argued that until such time as UCT was able to implement an alternative identification system (e.g. use of bar codes), the student number must be written on the front cover of the script. The student number at UCT is an alpha-numeric code consisting of the first three consonants of the surname, followed by the first three letters of the first name, followed by a three-digit number. Objectors argued that this method of implementation only ensured partial anonymity since it is possible from the student number to infer some aspects of identity, for example, a student number MJLXOL001 is likely to be that of an African student given the consonant combinations and the first name beginning with 'X' which are commonly found in Xhosa names. Despite these objections, the use of the student number was accepted as the most administratively efficient mechanism until a new student identification system was available.

Given the sensitivities around this policy, the Examinations and Assessment Committee of Senate called for an impact study into whether the policy had indeed

[2] The first author was responsible for drafting the policy on behalf of UCT's Examinations and Assessment sub-committee of Senate.
[3] The policy acknowledges that anonymous marking only addresses subconscious discrimination. No policy can stop a marker from discriminating for or against a student if they intentionally set out to do so.

strengthened the validity of UCT's examination assessment system in the two years of its implementation.

Literature Review

Educational and psychological assessment scholars have come to understand validity as the adequacy and appropriateness of the *inferences* from assessment and the decisions and consequences which emerge from these inferences (Messick 1989). Putting the emphasis of validity on the inferences (rather than the assessment instruments or scores) highlights assessment as an interpretive process which is "social as much as rational" (Cronbach 1989). Messick (1989) and Cronbach (1989) argue that there are no value-free inferences. The challenge lies in the rigour of the practices and processes which support the assessment interpretations, ensuring inferences that are relevant to the knowledge, skills and attitudes being assessed.

The influence of irrelevant inferences, or bias, on the assessment of student performance is a long-standing concern and subject of research in the field of educational assessment. More recently quality assurance imperatives in higher education have cast a spotlight on the validity of assessment systems and in particular the reliability of examination results – a crucial issue since the whole edifice of quality education rests on the overall validity of the assessment system (Knight 2002). Concerns about reliability include, among others, the appropriateness of examination assessment methods, the prevalence of cheating, and the reliability of marking. Numerous studies have documented the problem of consistency between markers particularly at the higher levels of the education system where tasks become more complex and open to interpretation (Newstead and Dennis 1994, Scharaschkin and Baird 1999, Shay 2004, 2005, Spear 1997). Inconsistencies that emerge are often attributed to assessor bias and there is an extensive body of research in educational assessment on the influence of bias on test design, administration and assessment (Fleming 1999, Gipps and Murphy 1994, Greatorex and Bell 2004).

While the practice of anonymous examination marking is fairly prevalent in higher education, there are few studies on the impact of these policies. Two studies conducted in the U.K mid-80's (Bradley 1984 and Belsey 1988) provided evidence

that the introduction of 'blind' marking resulted in the improved performance of women. Another study, however, where gender and ethnic bias were investigated found no evidence of discrimination (Dorsey and Colliver 1995). Newstead and Dennis (1993) question whether studies such as those conducted by Belsey and Bradley can provide conclusive answers regarding the existence of sex or gender bias in marking. They conclude that in the area of gender bias it has been very difficult to disentangle effects of bias from true differences in performance and that it is necessary to broaden the types of bias investigated to include things such as effects of social class, race and prior knowledge of the student.

Two more recent studies conducted on the assessment of clinical skills, although not directly studies of anonymous marking, are relevant in the South African context where race, social class and educational performance are intertwined. McManus et al (1996) in the UK investigated the claim that much higher failure rates of minority medical students in clinical examinations was evidence of bias. She found that while UK ethnic minority students tended to perform less well in clinical exams than UK white students, they also tended to perform less well in all examinations, including MCQ's marked by machines. She concluded that disparities in marking could not be accounted for by racial discrimination. Wass et al (2003) conducted a quantitative and qualitative study investigating the effect of ethnicity on performance in clinical examinations. Like McManus et al, their study concluded that disparities in performance could not be accounted for by explicit racial bias. Their study suggests that the poorer performance of ethnic students may be due to these students' particular 'styles of communication' with the patients being examined, being deemed as inappropriate by the assessors (what Gee 1996 would call 'discourse' differences).

These studies all point to the difficulty of disentangling a variety of issues which impact on students' performance, as well as the range of construct-relevant and construct-irrelevant inferences which influence markers' assessment**.**

The premise of this study is that the assessment of complex tasks, is a socially situated, interpretive act (Shay 2004), and that all judgments of student performance (irrespective of who the markers are) involve a tangled complexity of inferences, some deemed relevant and others deemed irrelevant. There will be at in any given

historical period and institutional context more or less heated contestation about which is which. The aim of this research is to explore, on the basis of student results, whether there are any shifts in the patterns of student performance in examinations which might be attributed to the implementation of the anonymous examination policy and to explore through interviews the course conveners' interpretations and explanations for these patterns[4].

Thus, the study set out to answer the following questions:
1) Are there any significant differences in the patterns of student performance in examinations prior to and following the implementation of the anonymous examination policy?
2) What do markers perceive to be the major reasons for these differences (or lack of differences)?

The study

The study was conducted on a sample of five undergraduate courses, one from each of UCT's six faculty, with the exception of the Law Faculty where anonymous marking has been in place for many years. The courses were selected on the basis of the following criteria:

- Curriculum stability -- Courses with relative stability over the period under investigation with respect to course content, staff, student intake, assessment system in order to control for the number of variables contributing to possible changes in student performance;
- Size -- Courses with large student numbers (>200 students) where anonymity is more likely to be achieved;
- Diversity of students -- Courses with diverse student body in terms of population group, educational background, language since students' concerns about racial discrimination were central to the study.

[4] It was the intention of this study to include students' interpretations of the performance patterns. While students are represented on various fora where this study has been presented, it has not been possible by the time of this publication to canvass their views in any systematic way.

- Varied assessment items -- Courses where the examination included high inference items (e.g. essays), low inference items (e.g. short answer questions or mathematical problems) and no-inference items (e.g. computer-marked multiple choice questions). This would enable us to better isolate changes in performance patterns which could be attributed to the policy vs. other contributing variables. For example, changes in performance on computer-marked items could not be attributed to the anonymity policy.

In two of the faculties -- Engineering and Science -- it was not possible to find courses with sufficiently large student numbers that met the stability criterion, therefore courses with fewer students were selected. Only Humanities had a course which had both high and no-inference items, which also met the other criteria.

Both quantitative and qualitative data were collected. The quantitative data comprised mean course marks, mean examination marks and mark distributions for two years prior to the implementation of the policy (2002-2003) and two years following implementation (2004-2005)[5]. The data were analyzed for the overall student cohort for each course, and disaggregated by gender, population group[6], language[7] and educational background[8]. In the Humanities course the data were also disaggregated by high and no-inference items. Analyses of the quantitative data were then discussed in interviews with each of the course conveners. The course conveners were all

[5] Examination results were only available from 2003-2005 for the Health Science course and the anonymous examination policy was only implemented in 2005 in the Engineering course.

[6] Population group in the South African context refers to the 'race' classifications used under apartheid: Black (meaning African), Coloured, Indian and White. These classifications are still required in South Africa for the purposes of monitoring equity.

[7] Language refers to home or first language. For the purposes of this research students were classified as English or 'other' since the interest was whether students were English first language or English as an additional language.

[8] Educational background refers to the type of school from which the student matriculated. Although South African national education has been housed under one department since 1994, the legacy of apartheid education continues to result in uneven quality of education to this day. Students are classified as coming from four groups of schools: House of Assembly (HOA) which refers to schools which were located within the historically White schooling system; House of Representatives (HOR) which refers to schools from the historically Indian schooling system; Department of Education and Culture (DEC) which refers to schools which were located within the historically Coloured schooling system, and Department of Education and Training (DET) which refers to schools from the historically Black African schooling system.

relatively experienced academics (> 5 years in academe). Four were males and all were white. This latter information is relevant given the primary concern about racial and gender discrimination which initiated the policy.

Findings & Discussion

Findings: Student Performance

Given the history of racial discrimination in South Africa and the students' particular concerns about the potential for such discrimination in examination marking, the primary interest of the study was to establish whether there was any evidence of negative discrimination prior to the implementation of the policy on the basis of population group, language, educational background and gender. We speculate that, if there had been significant levels of discrimination on these grounds, and if the implementation of the policy succeeded in masking these aspects of student identity, then, as was the case in the Belsey (1988) study, the findings should reveal changed performance for certain categories of students (for example, Black African students) following the policy implementation.

Before turning to the research questions, it needs to be noted that the legacy of apartheid's discriminatory educational policy continues to persist in racially-distributed academic performance. Across all five courses there is a consistent pattern of differentials in performance across population groups, educational backgrounds and language. (There are also gender differentials, with the general trend that women outperform males by small margins. See C-4, E-4, H-4, HS-4, S-4) In other words, there is a general pattern that students who are White outperform students who are Indian, Coloured and Black (See C-2, E-2, H-2, HS-2, S-2). Students who are first language English-speakers outperform speakers of English as an additional language (See tables C-1, E-1, H-1, HS-1, S-1); and students from former HoA schools tend to outperform students from HOR and DET (See tables C-3, E-3, H-3, HS-3, S-3). There is evidence of this pattern in other performance data collected at UCT (Cliff, Yeld and Hanslo under review; Visser and Hanslo 2006; Scott, Hendry, Jawitz and Gibbon

2005). As increasing numbers of black[9] students emerge from historically HoA schools, there is a perception among academic staff that these patterns are beginning to shift. However the data from these courses is evidence of a persistent reality of racially differentiated performance where race, educational background and language continue to serve as proxies for disadvantage. Against this backdrop the interest of this paper is whether there are any significant *changes* in these performances patterns after the policy implementation. We now turn to address this question.

In relation to differences in performances patterns pre- and post- policy implementation, there were no significant differences in any of the courses in performance patterns disaggregated by population group, education background, language and gender, in other words the racially differentiated performance pattern described above do not shift post-policy implementation (See C 1-4, E 1-4, H 1-4, HS 1-4, S 1-4) . Thus in relation to the students' concerns which initiated calls for the policy, there is no evidence from the performance data of negative discrimination on the basis of population group, language, gender or ex-education background. Nor is there any evidence of positive discrimination being applied to particular groups of students.

[9] The term 'black' is used here inclusively of Black, Indian and Coloured population groups.

In terms of differences in the performance of the overall cohort pre- and post-policy, the data reveals no significant differences in means in two courses (Science and Health Science with p-values of 0.03 and 0.003 respectively) (see S-5 and HS-5), nor is there any significant difference in distribution of scores for these two courses. In the other three courses (Commerce, Engineering and Humanities) there is a significant difference in the mean scores pre and post policy. In Commerce there is a significant difference ($p<0.0001$) between the examination means for each year (2002-2005) with a marked significant decrease of 7% between 2003 and 2004, the year of policy implementation (see C-5). There is also a shift in the distribution patterns with a wider distribution of scores post-policy. In the Engineering course there is a statistically significant increase in mean scores for 2002, 2003 and 2004 and a statistically significant decrease in 2005 ($p<0.0001$), the year of policy implementation (see E-5). In the Humanities course there is no significant difference between examination means for 2002 and 2003 but there is a statistically significant decrease ($p<0.0001$) of 7% between 2003 and 2004, the year of policy implementation(see H-5). The examination means disaggregated by item type reveals that this decrease is consistent for both the high inference items (i.e. essays) and the no-inference items (i.e. multiple choice questions which are computer marked).

Although it is tempting to attribute these decreases in mean average and widening of distribution to the policy implementation, the fact that the mean average on the Humanities multiple choice items also decreased, and by the same percentage as the high inference questions in that examination, suggests that we must be cautious about any causal links between the changes in performance patterns and the policy. If the implementation of the policy contributed to the decrease in marks in some of the courses and the wider distribution in one of the courses, it was only one of multiple contributing factors.

Findings: Conveners' Explanations

The student performance data seems to suggest that students' concerns about discrimination in examination marking are unfounded. We felt however that it was

important to explore other possible interpretations of these performance patterns through interviews with the course conveners. Where there were no differences in performance post-policy implementation (as in Health Science and Science), where there were small decreases (as in Engineering), or where there were significant decreases (as in Commerce and Humanities), what explanations did conveners have to offer for these patterns? The interviews would also provide an opportunity for conveners to offer any other insights on this sensitive issue.

The Health Science course convener's explanation for no difference in the performance patterns was that the course, since its inception in 2001, has always had a rigorous internal moderation process in place. She argued that discrimination on the basis of irrelevant inferences would be picked up through their own moderation process. Thus in relation to their course, the anonymous examination policy was redundant.

The Science and Engineering course conveners speculated that the small decrease in examination mean, although statistically significant in the case of the latter course, was not particularly significant from their perspective. They both argued that this was most likely a result of variation in examination difficulty from one year to the next.

In the Commerce course, the course convener felt strongly that the policy was in part responsible for the marked decrease in examination mean as well as the wider distribution of marks. His argument was that the policy had minimized positive discrimination, or "sympathetic marking"- that is, assessors marking more generously if they perceive a student to be educationally "disadvantaged". This theme of "sympathetic marking" came up in all the interviews and we return to it below.

In Humanities, the course convener was deeply puzzled by the decrease and was unable to offer any satisfactory explanation. Given that the decrease was evidenced on both the essay items as well as the MCQ items, he argued it was not possible to attribute the decrease to the policy alone. He proposed that perhaps the students were academically weaker in that year, although to counter this argument he noted that admissions criteria for this programme had in fact been tightened in 2004.

In addition to these explanations, further probing revealed a general scepticism from all the conveners about attempts to "anonymize" examination marking as well as an acknowledgment that, with or without the policy, markers do make inferences about student identity and these inferences can influence their judgments.

As noted in the introduction, one of the concerns in the implementation of this policy was the use of the student number which encodes aspects of student identity. This issue re-emerged as a key theme across all the interviews. The limits of anonymity were exemplified in three ways: Firstly, the Engineering course convenor noted that in smaller courses such as his (fewer than 100 students), where academic staff often teach and assess their own students, staff are likely to know students by their student numbers. If this is the case, it suggests that at UCT there is no anonymity in examination marking in the majority of courses at the upper undergraduate and postgraduate levels where numbers are relatively small. Secondly, in large courses where staff are unlikely to know students personally, as noted before, inferences about student identity can sometimes be made on the basis of the student number. One lecturer noted, "If UCT wants to eliminate any bias from marking, which I think is a good thing, they should go to the system they have at X university where… a student number is just a number". Thirdly, irrespective of the student number, staff testified to making inferences about identity on the basis of the actual student performance, for example, grammar, handwriting, or word usage. One marker noted, "I would be willing to bet if you were to give me thirty or forty exams I can pick them (out)…it's not hard, the way they use language. I can say – he's Black or he's Coloured or he is Afrikaans or if he's English. It's not that I am looking for it…". What these accounts suggest is that with or without the student number assessors consciously or unconsciously infer student identity from the examination script.

The question is, do these inferences about identity influence assessor's judgments? If so, how? A couple of interviewees initially dismissed the suggestion that markers are influenced by these inferences. With reference to student identity one noted, 'I don't pay any attention to it…'. Another noted, "It's just not even possible, not really. Anyone faced with a pile of marking doesn't give a hoot or damn about who the person is: you're just getting through it." But with further probing a consistent theme

across all the interviews was that knowledge about the student does influence assessors' interpretations – whether the assessor has first-hand knowledge about the student or is inferring this on the basis of the student number or script.

For example, one convenor noted the prevalence of the 'halo effect', that is, students who "catch your eye…whom you come to like because they are visible". He noted, "all the research in education shows that once you've got that you can do no wrong, your marks are going to be better because people expect you to be better". However the most common example of how inferences about student identity influence the assessment was the case of 'sympathetic marking'. These are instances where a name, a student number or the actual performance evokes in the marker a stereotype of a 'disadvantaged' student which may generate sympathy on the part of the marker, particularly if the student's performance is on the pass/fail borderline. One convener admitted, "You note the name and think the language isn't going to be good. And with that you'd have an element of, you know, how would I do in a second language?... Here's somebody carrying two bags of cement on their shoulders, not one…And so you go a bit easy….If it's obvious to me that a student is not a first language English speaker and there's really something I'm struggling to understand, I would tend to give them the benefit of the doubt. I would sort of say to myself, 'Could he mean that?'"

As noted above, one of the course conveners initially argued for a causal relationship between the policy and the decreased performance. His explanation for the decreased marks was that the policy had minimized sympathetic marking. He argued that since the implementation of the policy markers could now assess on the basis of "merit" rather than accounting for extenuating circumstances – "a poor answer now gets a low mark". He describes the reasoning as follows: "If you know the race of the student and…you make an assumption about the quality of their schooling that in general (it) was better than the majority of black students…one can almost say that there's a feeling that it's their fault. (It's) the student's fault if they haven't answered the question well, because they haven't worked hard enough or they haven't paid attention or whatever it might be. Whereas if you find that with a student that you know is a black student, it's not that easy to make that judgment. You think well, they were a victim of bad schooling and that principle of charity would come in, where you

would give them the benefit of the doubt." By contrast the new policy "liberated" markers. "I don't need to care any more…and it's easier to mark…there's no agonising, there used to be agonising. We are in an agonising-free zone now." Although the failure rate was notably higher, he argued that since the policy implementation the examination results were a "fair assessment of the actual stuff before us". On the other hand, "if we think fairness requires preferential treatment, then it's not fair…on my version of fairness, it's more fair in 2004 (post-policy) because I don't think it's the university's job to redress the social norms."

Thus while the student performance data suggests no positive or negative discrimination of particular groups of students, the interviews point to another reality – that of white academics caught between a university's contending discourses of equity and excellence, of redress and success. Not surprisingly, despite the administrative burden of implementing the policy, and despite their skepticism about its ability to ensure anonymity, the policy is welcomed. They all agreed that if the intention of the policy is to separate the person (the student) from the product (the performance), in the current political climate this is a good thing. As one convener noted, "We should be treating the scripts as objective products as much as possible, and if (the policy) contributes to that, then by all means."

Conclusion

In summary, on the basis of this sample, comparisons of pre- and post-policy examination results indicate no evidence of *negative* discrimination against students on the basis of population group, language, gender or ex-education background. While there is a strong perception on the part of one convener that the effect of the policy may have been to minimize 'sympathetic marking', there is no evidence in the performance data to support this. Where there is a significant decrease in examination results post-policy as well as a widening in the distribution of scores, these patterns are consistent across all groups. To the extent that the anonymity policy has contributed to this decrease, it has affected the whole cohort. We speculate that one possible reason for this decrease is that de-coupling 'person' from 'product' results in more conservative marking, at least in the initial stages of implementation. This hypothesis needs further exploration however. While the conveners were generally

sceptical about the degree to which UCT's system ensured anonymity, they noted that the policy signalled the importance of assessing examination products on their merit without contextual considerations. This move to 'objectify' examination marking was welcomed.

The intention of the policy was to strengthen the validity of UCT's examination assessment. What conclusions can we make on the basis of this study about the impact of this policy? As previously discussed, validity is the degree to which the inferences made about student performance are relevant to the knowledge, skills and attitudes being assessed. The intention of the policy is that by removing the students' name, aspects of student identity such as race and gender, which are deemed irrelevant, are thereby removed. In South Africa, however, race, language, and educational opportunity remain inextricably intertwined with performance for the foreseeable future. As all assessors of complex performances know stripping 'subject' from 'object' is very challenging, often not possible or even desirable. This is not an argument in support of 'sympathetic marking'. Attempting to compensate for poor schooling by awarding extra points is indefensible. What it does suggest, however, is that, while we might retain the anonymous examination policy for political reasons, its contribution to strengthening validity is limited. Valid assessment results will emerge from rigorous assessment and moderation practices which require the community of interpreters – academics, tutors, external examiners – to articulate what we really value. Practically, this supports the need for clear marking criteria, model answers, marker training, and moderation meetings. These are not to be seen as technical or bureaucratic managerial requirements but rather opportunities for rigorous collegial debate about the value-basis of our assessment judgements, what is relevant and what is not, and under what circumstances. It would appear that there is in fact no "agonising-free zone" after all.

Acknowledgements **16**

We would like to thank Alvin Visser, Monique Hanslo and Kutlwano Ramaboa from the Alternative Admissions Research Project at UCT for assistance with the analysis of the quantitative data.

References

Baird, J. 1998. What's in a Name? Experiments with blind marking in A-level examinations. *Educational Research*, 40(2): 191-202.

Belsey, C. 1988. Marking by Number. *Association of University Teachers Woman,* 15: 1-2.

Bradley, C. 1984. Sex Bias in the Evaluation of Students. *British Journal of Social Psychology*, 23: 147-153.

Cliff, A.F., N. Yeld and M. Hanslo (under review). Assessing the academic literacy skills of entry-level students, using the Placement Test in English for Educational Purposes (PTEEP). *Assessment in Education*.

Cronbach, L. 1989. Construct validation after thirty years. In R. Linn (Ed), *Intelligence: Measurement, theory, and publich policy: Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 146-71). Urbana: University of Illinois Press.

Dennis, I., & Newstead, S. 1996. A new approach to exploring biases in educational assessment. *British Journal of Psychology*, 87: 515-534.

Dorsey, J., and J. Colliver. 1995. Effect of Anonymous Test Grading on Passing Rates as Related to Gender and Race. *Academic Medicine*, 70(4): 321-323.

Fleming, N. 1999. Biases in Marking Students' Written Work: Quality? In S. Brown & A. Glasner (Eds.), *Assessment Matters in HE: Choosing and Using Diverse Approaches* (pp. p. 83-92). Society of Research in Higher Education & Open University Press.

Gee, J. 1996. *Social Linguistics and Literacies: Ideology in Discourses*. Basingstoke: Falmer Press.

Gipps, C. and Murphy, X. 1994. A Fair Test? Assessment, achievement and equity. Buckingham: Open University Press.

Greatorex, J., & Bell, J. 2004. Does the gender of examiners influence their marking. *Research in Education*, 71: 25-36.

Knight, P. 2002. Summative assessment in Higher Education: practices in disarray, *Studies in Higher Education*, 27(3): 275-286.

McManus IC, Richards P, Winder BC, and KA Sproston. 1996. Final Examination Performance of Medical Students from Ethnic Minorities. *Medical Education* 30: 195-200.

Messick, S. 1989. Meaning and Values in Text Validation: The Science and Ethics of Assessment. *Educational Researcher*, 18(2): 5-11.

Newstead, S. 1996. The psychology of student assessment. *The Psychologist,* 543-547.

Newstead, S. and I. Dennis. 1993. Bias in Student Assessment. *The Psychologist,* 451-452.

Newstead, S. and I. Dennis. 1994. Examiners Examined. *The Psychologist,* 216-219.

Scharaschkin A, and J. Baird. 2000. The Effects of Consistency of Performance on A Level Examiners' Judgments of Standards. *British Educational Research Journal* 26(3): 343 – 357.

Scott, I., Hendry, J., Jawitz, J. & Gibon, F. CHED/IPD Equity and Efficiency ('Throughput') Project. Unpublished report to the UCT Senate Executive Committee, November 2005.

Shay, S. 2004. The assessment of complex performances: A socially-situated interpretive act, *Harvard Educational Review,* 74(3): 307-329.

Shay, S. 2005. The assessment of complex tasks: A double reading *Studies in Higher Education,* 30(6): 663-679.

Spear M. Summer 1997. The Influence of Contrast Effects upon Teachers' Marks. *Educational Research* 39(2): 229-233.

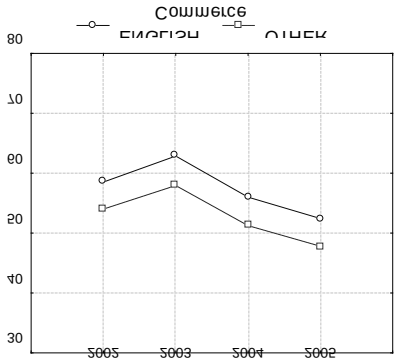
Visser, A. and M. Hanslo, M. (in press). Approaches to predictive studies: Possibilities and challenges. *South African Journal of Higher Education*, 19(6).

Wass V, Roberts C, Hoogenboom R, Jones R, Van der Vleuten C. (2003) Effect of Ethnicity on Performance in a Final Objective Structured Clinical Examination: Qualitative and Quantitative Study. *British Medical Journal*, 326: 800-803.

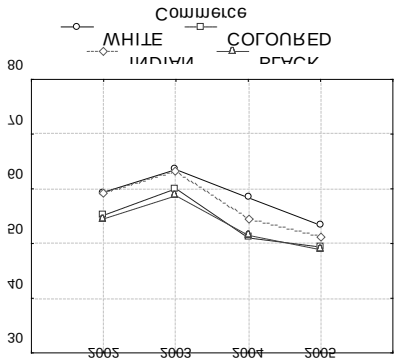
COMMERCE

Year	Student No/ Mean Score	Cohort	Lang		Pop Group				Ex-Education			Gender	
			Eng	O	W	B	I	C	HoA	HoR/DEC	DET	F	M
2002	**N**	976	715	257	465	303	89	114	669	22	30	423	549
	M	57	59	54	59	55	59	55	59	55	52	59	56
2003	**N**	1035	752	282	436	315	130	152	877	57	78	430	604
	M	62	63	58	64	59	63	60	62	60	54	62	61
2004	**N**	958	692	266	405	298	114	132	830	40	62	411	547
	M	55	56	51	58	52	54	51	56	52	46	55	54
2005	**N**	872	592	280	328	287	110	124	711	51	71	339	533
	M	51	52	48	53	49	51	49	52	46	43	52	50

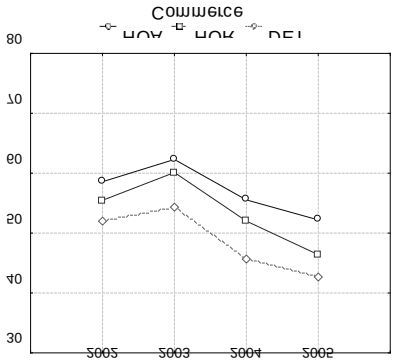
C-1



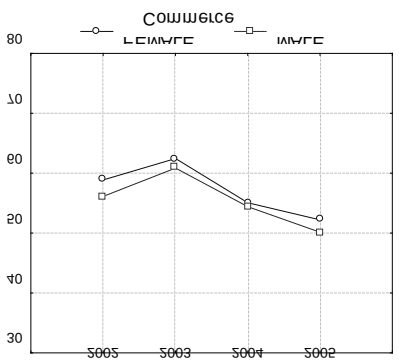
C-2



C-3



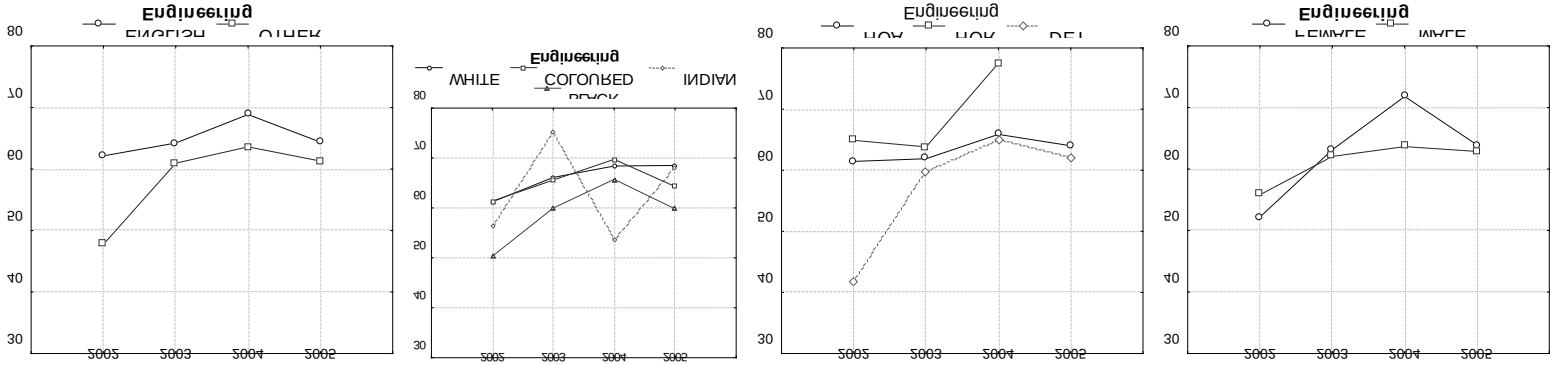
C-4



ENGINEERING

Year	Student No/ Mean Score	Cohort	Lang		Pop Group				Ex-Education			Gender	
			Eng	O	W	B	I	C	HoA	HoR/DEC	DET	F	M
2002	**N**	90	46	44	25	49	4	12	47	7	23	18	72
	M	55	62	48	61	50	56	61	61	65	42	52	56
2003	**N**	84	34	50	22	55	2	5	43	3	28	18	66
	M	62	64	61	66	60	75	66	62	64	60	63	62
2004	**N**	93	47	46	25	52	5	11	57	4	29	30	63
	M	66	69	64	68	66	53	70	66	78	65	72	63
2005	**N**	89	51	36	21	49	7	10	63	0	12	20	67
	M	63	64	61	68	60	68	64	64		62	64	63

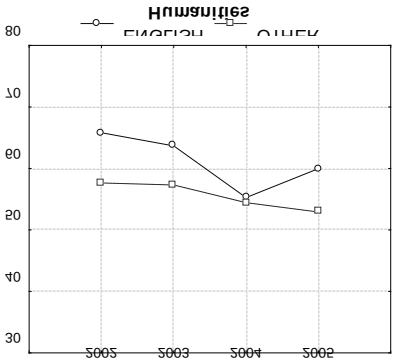
E-1 **E-2** **E-3** **E-4**



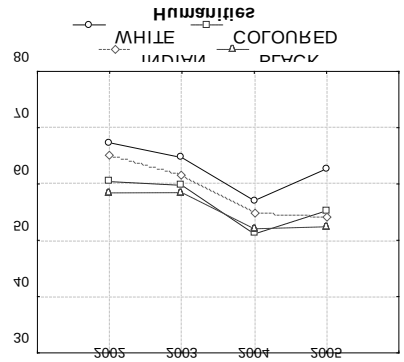
HUMANITIES

Year	Student No/ Mean Score	Cohort	Lang		Pop Group				Ex-Education			Gender	
			Eng	O	W	B	I	C	HoA	HoR/DEC	DET	F	M
2002	**N**	712	572	139	408	149	35	119	620	37	24	531	180
	M	64	66	58	67	58	65	60	65	59	53	65	62
2003	**N**	651	560	90	416	101	26	106	596	27	18	505	146
	M	63	64	57	65	58	62	60	63	58	52	64	60
2004	**N**	574	477	96	352	107	29	79	516	26	15	460	114
	M	55	55	54	57	52	55	51	55	53	47	56	53
2005	**N**	628	526	102	353	128	40	104	418	17	7	495	133
	M	59	60	53	63	52	54	55	60	57	37	59	58

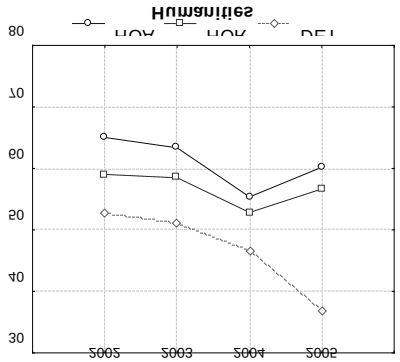
H-1



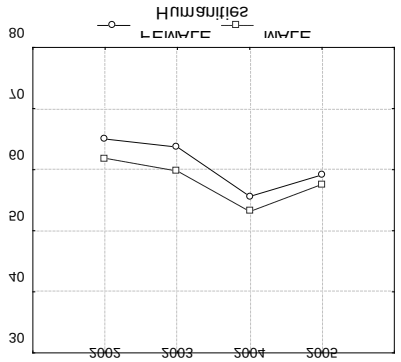
H-2



H-3



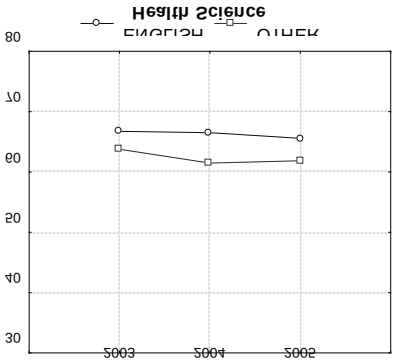
H-4



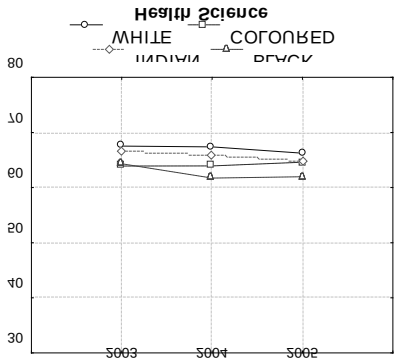
HEALTH SCIENCE

Year	Student No/ Mean Score	Cohort	Lang		Pop Group				Ex-Education			Gender	
			Eng	O	W	B	I	C	HoA	HoR/DEC	DET	F	M
2003	**N**	335	253	80	154	77	42	60	284	26	18	255	78
	M	66	67	64	68	64	67	64	66	64	62	67	64
2004	**N**	351	239	112	152	111	34	51	286	17	38	260	91
	M	65	67	61	67	62	66	64	66	62	61	66	63
2005	**N**	345	265	80	152	84	40	63	286	20	29	250	95
	M	65	66	62	66	62	65	65	65	64	61	65	63

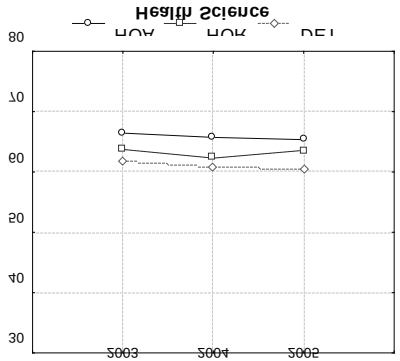
HS-1



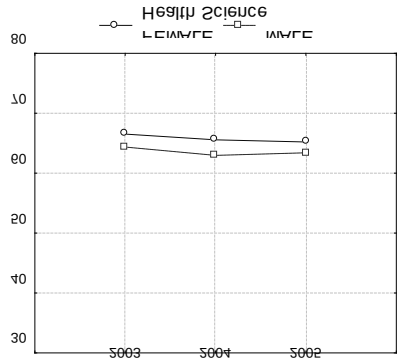
HS-2



HS-3



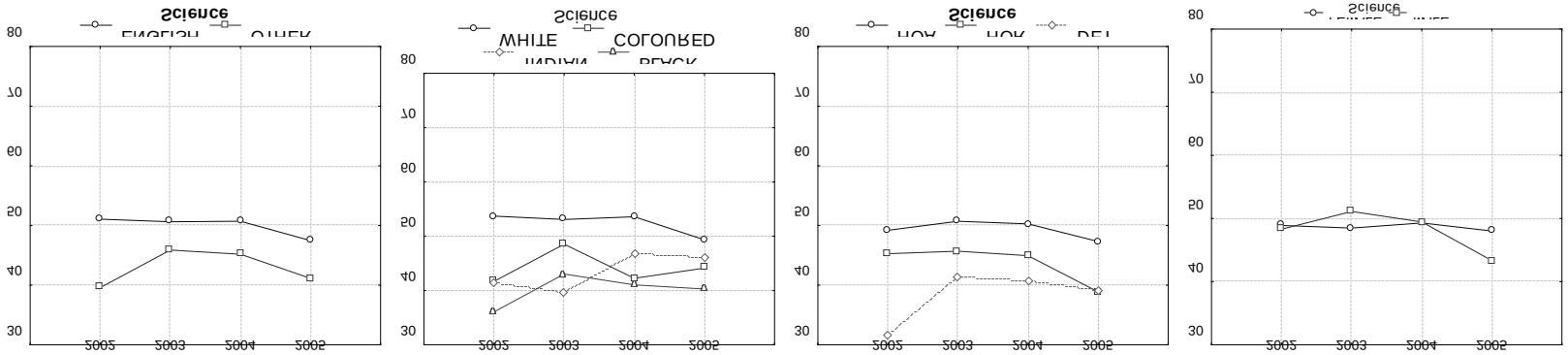
HS-4



SCIENCE

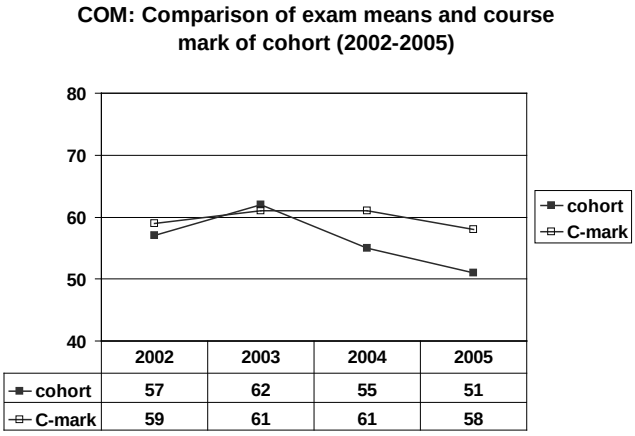
Year	Student No/ Mean Score	Cohort	Lang		Pop Group				Ex-Education			Gender	
			Eng	O	W	B	I	C	HoA	HoR/DEC	DET	F	M
2002	**N**	195	146	40	122	34	11	18	171	9	12	89	97
	M	48	51	40	54	36	42	42	49	45	32	50	48
2003	**N**	203	158	43	116	43	9	32	171	15	14	115	86
	M	50	51	46	53	43	40	49	51	46	41	48	51
2004	**N**	210	161	45	131	42	9	23	183	10	13	103	104
	M	49	51	45	54	41	47	42	50	45	41	49	49
2005	**N**	180	130	49	92	57	11	17	131	4	18	92	87
	M	46	47	41	49	40	46	44	47	39	39	48	43

S-1 **S-2** **S-3** **S-4**



COMPARISON OF MEANS FOR OVERALL COHORT

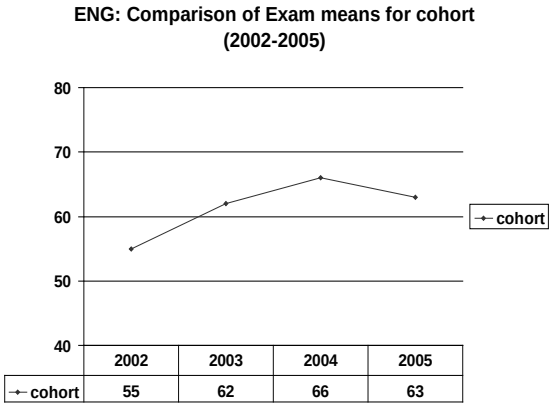
C-5 COMMERCE



	2002	2003	2004	2005
cohort	57	62	55	51
C-mark	59	61	61	58

$p<0.0001$

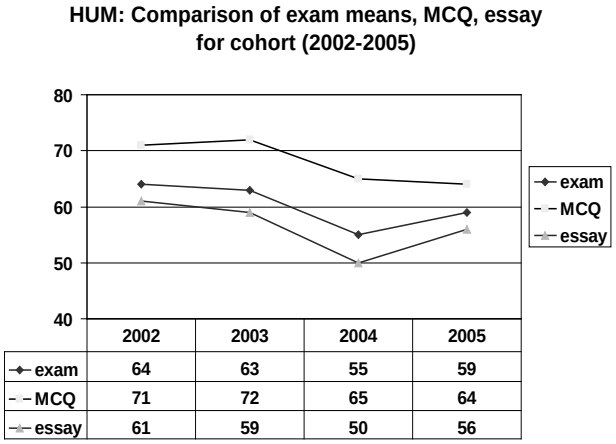
E-5 ENGINEERING



	2002	2003	2004	2005
cohort	55	62	66	63

$p<0.0001$

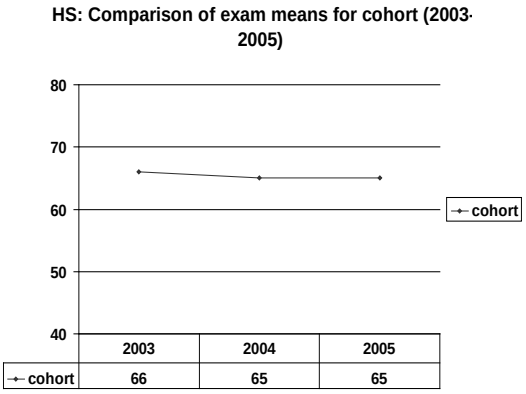
H-5 HUMANITIES



	2002	2003	2004	2005
exam	64	63	55	59
MCQ	71	72	65	64
essay	61	59	50	56

P<0.0001

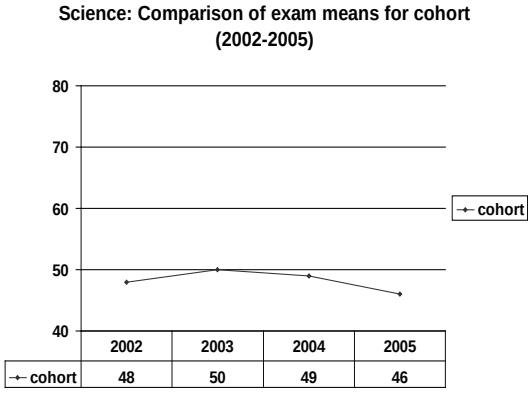
HS- HEALTH SCIENCE



	2003	2004	2005
cohort	66	65	65

P< 0.003

S-5 SCIENCE



$p<0.03$