Marker 'fatigue' and marking reliability in Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE)

Dr. Neil DRAVE
Hong Kong Examinations and Assessment Authority, Hong Kong
ndrave@hkeaa.edu.hk

Abstract
An increasing number of examination bodies around the world are replacing paper-based marking with onscreen marking (OSM), a system in which marking is done on computers. OSM has been used for the LPATE since 2008 and is being introduced for all public examinations in Hong Kong in 2012. The change to OSM has occasioned discussion of the possible consequences for marking reliability, particularly with regard to the question of whether marker 'fatigue' results from having to mark on a computer in a marking centre. This paper explores the issue of marker fatigue among markers of the written papers of the LPATE, a standards-referenced assessment which assesses the English proficiency of those who wish to teach English in Hong Kong schools, and administered by the Hong Kong Examinations and Assessment Authority (HKEAA). It presents data from the HKEAA's OSM system on the variability of marking in different marking periods. The study concludes that, while there is some evidence of marker fatigue, there is no concrete evidence of adverse fatigue effects.

Background to the research
The purpose of the current research was to investigate the issue of fatigue among makers of the composition section of Paper 2 (Writing) of the 2011 Language Proficiency Assessment for Teachers of English (LPATE), with particular reference to the effect of marking for extended periods onscreen on the reliability of marking (Meadows and Billington 2005 is a thorough review of the literature on reliability).

The LPATE is a high-stakes, standards-referenced assessment used to assess the English language proficiency of those wishing to teach English in Hong Kong schools.[1] All LPATE marking is done on computers in designated marking centres, and it is not unusual for markers to mark for extended periods, sometimes hours at a time, often without a break. Previous research has found that extended periods of concentration can be detrimental to physical and mental performance, particularly in high-pressure occupations (e.g. Petrilli et al. 2006) and when computers are used (e.g. Jensen et al. 2004). Studies of onscreen marking (OSM) have generally not explored the topic of fatigue in depth but, given the findings of research in other settings, it is reasonable to assume that marker performance might be influenced by it.

The HKEAAs OSM system records all marking times and durations so it is possible to investigate whether there is a demonstrable relationship between marking for a long period and marking reliability. Reliability in this study is defined as the degree of difference between a raw score and a statistically adjusted 'true' score. It is posited that, if a statistical relationship is found between marking duration and the scores given to compositions, this would be *de facto* evidence of marker fatigue and so any influences on scores would constitute 'fatigue effects'.

The composition component of the LPATE assessment was chosen as the focus of the research for two reasons: first, since most previous studies have concerned short

[1] Details of the structure of the LPATE assessment can be found at http://www.hkeaa.edu.hk/en/lpat/

answers, the issue of the OSM of extended writing is under-researched (Shaw 2008). Second, the LPATE composition is marked according to a set of 'Scales and Descriptors', meaning that there is greater scope for variation in marker judgement than in item-based papers. The scales used to assess candidates' compositions are as follows:

Scale 1: Organisation and Coherence
Scale 2: Grammatical and Lexical Accuracy and Range
Scale 3: Task Completion

Candidates are asked to write about 400 words, in a stated text type, on a topic of current relevance to Hong Kong. In 2011, candidates wrote a proposal for a youth event. Markers give a score of between 1 and 5 on each scale, with 5 being the highest score. All scripts are double marked and an IRT 'fair averaging' procedure (cf. Lumley and McNamara 1995) is employed to derive the final scale scores, which are reported separately. It should be noted that the present study concerns only the initial scale scores. Since there is double marking, discrepancy checking and fair averaging in the actual LPATE, any variability in raw scores is attenuated before the final marks are given to candidates.

As a preliminary to the quantitative phase of the research, LPATE markers were surveyed to see whether they felt they had suffered from fatigue during OSM and whether there had been any perceived fatigue effects. I present results of the survey in the following section before describing OSM and fatigue in more detail.

Survey of LPATE markers

Markers of all LPATE 2011 written papers were surveyed after the marking period had concluded to find out whether they felt they had suffered from fatigue and what its effects had been (responses n=30, 79% of all markers). Most questions were open to all markers to answer but three were directed exclusively at composition markers (responses n=10). The key findings were as follows:

Questions answered by all markers

- Most markers felt that they had experienced fatigue during the marking period.
- Fatigue manifested itself as tired eyes, difficulty concentrating, sleepiness and muscle pain.
- Markers attributed the fatigue to OSM and to marking for long periods.
- Markers felt that fatigue had made them mark more slowly than normal.
- Sixteen respondents stated that there had been no effect on the reliability of their marking, with the remainder unsure.
- Four markers said that they were more willing to give the benefit of the doubt to candidates when fatigued, 17 said that they were not, while the remainder were undecided.

Questions answered by Composition markers only

- Two markers said that they had been less willing to give high scores to compositions when fatigued but most said fatigue had had no influence on the scores they had given.
- Three markers expressed uncertainty about whether fatigue had influenced their willingness to give low scores to compositions; all but one of the others said there had been no influence.
- Two markers said that they had given similar scores to compositions when fatigued while 3 others said they didn't know whether they had or not.

The findings suggest that LPATE markers felt they had suffered from fatigue and some attributed this to OSM as well as marking duration. Some 'fatigue effects' were mentioned, although more respondents either rejected the notion that fatigue had influenced their marking or were unsure about this.

Given these findings, it was decided to look in more depth at the issue of fatigue and to explore statistically whether marking for long periods onscreen did, in fact, have an impact on the reliability and range of scores given. In the sections that follow, I first review the OSM system, then relevant research on fatigue, before describing data and presenting findings.

Onscreen Marking (OSM)

OSM of examinations is becoming increasingly common around the world (Drasgow et al. 2006, Shaw 2008) and Hong Kong is following this trend, progressively implementing OSM for all subjects offered in its public examinations.[2] Current HKEAA practice is that candidates write their answers on paper, in Question-Answer booklets which are bar-coded. These booklets are scanned and the anonymised script images stored on a server, ready for distribution to markers via a computer intranet. The actual marking is done in one of the HKEAA's designated marking centres. These are premises with hundreds of computers running the OSM interface, as well as facilities for meetings and areas where markers can relax. The marking centres are open all day (usually until 10pm) and so markers have great flexibility in when they mark, and for how long.

Before marking begins, markers are trained and standardised using scripts which have been pre-marked by the Chief Examiner (see Drave 2010 for details of standard-setting procedures for item-based papers). Throughout the training and marking process, scripts are delivered automatically and randomly, and markers read them onscreen before clicking a radio button to indicate whether an item is correct or incorrect (in the case of item-based papers) or clicking on a number between 1 and 5 to enter a scale score (for the composition part of the Writing paper). The OSM system also allows for monitoring of marker performance by randomly allocating 'Control' scripts, which have been pre-marked. For practical and legal reasons, we do not allow annotations of scripts, although research has found that markers find this beneficial (Shaw 2008).

OSM Reliability

Previous research has demonstrated that the move to OSM has certain consequences for how marking is conducted, how reliable/valid it is and for marker comfort and wellbeing (Shaw 2008, Geranpayeh 2011, Whitehouse 2010). In general, OSM marking studies seem to fall into three types: those which compare the reliability of OSM marking to paper-based marking (of the same scripts); those which examine the reliability of marking by using internal reliability measures such as those based on IRT models; and those which survey subjective attitudinal factors.

Research has found that OSM compares favourably to paper-based marking on reliability and consistency measures (Whitehouse 2010, Shaw 2008, Coniam 2009). Whitehouse (2010) notes that OSM was found to cause no change in marker leniency/severity in a GCSE Literature exam (n=180 scripts) but there was greater variation in marking. A similar tendency towards greater variation was found by Fowles

[2] For details of the HKEAA's OSM system, please refer to www.hkeaa.edu.hk/DocLibrary/Media/Leaflets/ea-osm-eng.pdf

(cited by Whitehouse 2010: 2) with regard to GCSE English, with markers being either more severe or more lenient onscreen. This was a very small study, however (n=40 electronic scripts). Johnson et al. (2010) studied 12 experienced English literature assessors who marked two matched samples of 90 essay exam scripts on screen and on paper. A variety of statistical methods were used to compare the reliability of the essay marks given and no differences were found. Coniam (2009) studied the 2007 Hong Kong CEE English Language Writing Paper. Thirty markers marked on paper 100 scripts they had marked onscreen nine months before. There were no differences between the marking of the two modes.

Some studies do not compare marking modes but instead seek to establish the internal reliability/consistency of OSM using the same techniques as traditional/paper-based rater variability studies. Whitehouse's (2010) study of an essay-based unit of the UK's AS Level Geography found that the marking reliability was acceptable, as judged by comparing OSM scores with a 'true' mark given by experts (n=173 scripts). A study of the Primary 6 Territory-wide Systems Assessment in Hong Kong found that there were high correlations between the 'true' marks given by expert markers and those given by markers marking onscreen (Cheung and Chang 2009).

Coniam's attitudinal survey of markers of the Liberal Studies examination in Hong Kong found that there was considerable marker resistance to moving from paper-based marking to OSM, and some markers felt that there were negative consequences to doing so (Coniam 2010, Coniam and Yeung 2010).

The two Hong Kong studies cited above both touch on the issue of marker fatigue. Cheung and Chang (2009: 2) claim that "rater fatigue may be a factor weakening the reliability of marking and raters tend to rate more severely over time." Fatigue is used as a rationale for not giving the markers in their study too many scripts to mark. Coniam and Yeung (2010) report that markers said they suffered from tired eyes when using OSM. The issue of fatigue is not explored in detail in either study, however.

Fatigue

According to Theander (2007), there are more than 250 ways of measuring fatigue but no agreed definition of it. There have been many attempts to classify and categorise fatigue symptoms in medical settings, the most widely used scale for measuring fatigue being the Multidimensional Assessment of Fatigue Scale (MAF). However, scales such as this – and there are dozens of them in use (Whitehead 2009, Hjollund et al. 2004) – are normally used to measure self-reported fatigue in conditions of (usually chronic) illness, over a number of days (Belza 2010), so are not directly relevant to the present study.

In non-medical studies, there is a tendency to regard fatigue as a loose set of deleterious physical, emotional, behavioural and cognitive symptoms which negatively impact human performance (Theander 2007). Meadows and Billington (2005) categorise fatigue as a 'transient examiner trait' and thus recognise its relevance to marking, but their literature review covers only two studies, neither of which is particularly relevant to the current research.

Of more relevance to this study is research on prolonged computer use, which has been found to give rise to ergonomic consequences for users, such as physical discomfort and mental strain (Sonnea et al. 2010, Müllera et al. 2010). Geranpayeh (2011) reports that markers in his OSM survey felt strain from the fact that they were using a mouse when marking for long periods. Markers also felt that scrolling had an impact on their

marking because this made them forget information. Simply reading onscreen presented no problems for Gernapayeh's respondents, although other research has found that it necessitates different practices from reading on paper, and so causes some mental strain for readers (Shaw 2008).

The issue of fatigue is particularly important for occupations in which errors of judgement can be costly in terms of human life. A study of surgeons who had been on-call for a long period, for example, found that fatigue caused an increase in cognitive errors, a decrease in attention and in the ability to carry out certain psychomotor tasks (Kahol et al. 2008). Nurses who work for more than 12 consecutive hours, or work when they have not had sufficient sleep, put their patients' health at risk (Rogers 2008). Studies of various occupations have found that engaging in a task for long periods results in negative cognitive-behavioural effects: air traffic controllers (Signal et al. 2009), lorry drivers (Moore and Brooks 2000) and pilots (Petrilli et al. 2006, Stewart et al. 2006) have all been found to suffer fatigue effects.

It is difficult to derive any definitive conclusions from these studies, however, because fatigue is defined in many different ways, with different aspects in focus in different studies, and it is often confounded with sleep deprivation. What seems clear, however, is that fatigue is essentially a time-based concept: in contexts of computer use and when undertaking activities requiring concentration, the longer one undertakes a task for, the more fatigue there is. This does not mean that there will necessarily be fatigue effects, of course, but the above research suggests that this is a likely outcome. The key research question to ask in relation to OSM, therefore, is: 'Does a marker suffer fatigue effects (become less reliable) the longer they mark for? The specific research questions addressed in the study were:

Q1 Were LPATE markers equally reliable (i.e. severe/lenient) in different marking periods?
Q2 Did fatigue effects occur in particular scales?
Q3 Did fatigue effects occur for particular markers?

Method[3]

In 2011, 1369 candidates sat the LPATE Writing paper. Each composition script was double marked over a period of approximately one month. There were 12 markers, plus a Chief Examiner (CE) and Assistant Chief Examiner, and each marker marked approximately 220 scripts each (including double marking). The scores given were then fair averaged using Rasch analysis to take into account marker leniency/strictness tendencies. The final scores were then checked against the original mean scores on each scale and anomalies dealt with by the CE. This adjusted fair average score is regarded as the 'true' score of the candidate on each scale.

The raw data for this study comprised an Excel file extracted from the OSM system with all the original scores for the 2011 composition paper on Scale 1, 2, and 3. The data had been subdivided by the marking date, hour (with '1' indicating the first hour of the day in which a marker marked scripts) and candidate number. A list of the fair averaged scores of each candidate was provided in a separate Excel file. The fair scores which corresponded to the original scores were then located for each hour of the markers' marking. The fair scores given in each time slot were then averaged so that the markers' leniency could be calculated. For each time slot, the leniency degree of a marker was:

[3] I would like to thank my colleague Dr. Eric Fung, who did the statistical analysis.

Leniency degree = Average of the raw marks – Average of the corresponding fair marks
The hour of marking (i.e., hr. 1, 2, 3, etc.) was treated as an indicator of the degree of marker fatigue. This is not ideal, but is a convenient assumption in view of the difficulty of applying any independent fatigue measures. By calculating the leniency degree and fatigue degree values for a marker in different timeslots, one can examine whether there is any linear relationship between them. In view of the small number of markers involved, further sub-division of the data by marker ID was possible only for three of the markers, for whom sufficient hourly data were available.

Findings
For each scale, the summary statistics on leniency degree in different hours are provided in Appendix 1 (Scale 1), Appendix 2 (Scale 2) and Appendix 3 (Scale 3).

The summary statistics and box plots (not provided) show that there are no major differences in leniency for different marking hours. In other words, there is no evidence to support the claim that, overall, there is a linear relationship between marker fatigue and leniency degree. This observation was further tested by carrying out a one-way analysis of variance (ANOVA), which was used to test whether there were any significant differences in the leniency degree means between different marking hours.

As shown in the summary statistics tables (Table 1-3), the number of observations drops dramatically after marking hour 8. Because a small sample size can violate the basic assumptions for ANOVA, such as the normality assumption (which assumes that the leniency degree within each marking hour is approximately normally distributed), the data for marking hours above 8 were excluded from statistical testing.

The results of the ANOVA for each scale are as follows:

Table 1: ANOVA test results for Scale 1

	Sum of Squares	df	Mean Square	F-statistics	Significance level
Between-Group	.155	7	.022	.221	**.980**
Within-Group	33.665	337	.100		
Total	33.820	344			

Table 2: ANOVA test results for Scale 2

	Sum of Squares	df	Mean Square	F-statistics	Significance level
Between-Group	.674	7	.096	1.144	**.335**
Within-Group	28.389	337	.084		
Total	29.064	344			

Table 3: ANOVA test results for Scale 3

	Sum of Squares	df	Mean Square	F-statistics	Significance level
Between-Group	.694	7	.099	.880	**.522**
Within-Group	37.969	337	.113		
Total	38.663	344			

The tables show that the significance levels are all greater than 0.05, which means that there are no statistically significant differences in terms of leniency degree between different marking hours for each of these three scales.

Differences between markers

The present study considered whether there is a relationship between leniency degree and fatigue degree for individual markers. Due to the small number of cases (scripts) being marked in some of the sessions, the focus was on the following three markers (of the 12 total) for whom there were sufficient data in each hour: Marker 010 (68 scripts), Marker 002 (33 scripts) and Marker 005 (29 scripts).

There were no significance differences in marking on any of the scales for Marker 002 and Marker 005. Marker 010, however, tended to give lower scores than the fair averaged score in Hour 2 and higher scores than the fair averaged score in Hour 4 on Scale 1 (Table 4). Marking on the other scales followed the expected pattern.

Table 4 Marker 10 ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Scale 1 Leniency degree	Between Groups	1.443	3	.481	3.251	.030
	Within Groups	6.955	47	.148		
	Total	8.399	50			

Since the significance level is 0.03, which is below 0.05, there is a statistically significant difference in the mean leniency degree between different marking hours. Gabriel's pair wise test was used to conduct post-hoc tests on the one-way ANOVA for Marker 10 (Table 5).

Table 5 Marker 010 Multiple Comparisons

Dependent Variable		(I) Hour	(J) Hour	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Scale 1 Leniency degree	Gabriel	1	2	.185	.128	.624	-.17	.54
			3	-.158	.157	.886	-.58	.27
			4	-.317	.181	.371	-.80	.16
		2	1	-.185	.128	.624	-.54	.17
			3	-.343	.157	.171	-.77	.08
			4	-.502	.181	**.036**	-.98	-.02
		3	1	.158	.157	.886	-.27	.58
			2	.343	.157	.171	-.08	.77
			4	-.159	.203	.964	-.71	.39
		4	1	.317	.181	.371	-.16	.80
			2	.502	.181	**.036**	.02	.98
			3	.159	.203	.964	-.39	.71

We can see that significant differences can be found only between Hour 2 and Hour 4. It is difficult to interpret these data however without exploring in detail the marking behaviour in these hours (which is beyond the scope of this research).

Overall, the current data set does not provide any evidence that marking reliability is affected by marking duration for the three markers for which adequate data are available.

Conclusion

This paper has focused on the interaction between marking duration and scores awarded. I am pleased to note that, for the paper under consideration, marking for extended periods does not seem to have influenced the reliability of marking. Markers may indeed suffer from fatigue but there is no evidence in these data of fatigue *effects*.

References

Belza, B. (2010). Multidimensional Assessment of Fatigue (MAF) User's Guide. Available online: http://www.son.washington.edu/research/maf/users-guide.asp

Cheung, K. M. and R. Chang. (2009). Investigating reliability and validity in rating scripts for standardisation purposes in onscreen marking. Paper presented at the IAEA conference, Brisbane, September 2009.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. Educational Research and Evaluation, 15(3), 243–263.

Coniam, D. (2010). A qualitative examination of the attitudes of Liberal Studies markers towards onscreen marking in Hong Kong. British Journal of Educational Technology. Available online: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8535.2010.01136.x/pdf

Coniam, D. and A. Yeung. (2010). Markers' perceptions regarding the onscreen marking of Liberal Studies in the Hong Kong public examination system. Asia Pacific Journal of Education, 30(3), 249–271.

Drasgow, F., R. Luecht and R. Bennett. (2006). Technology and testing. Educational Measurement, ed. R. Brennan. Westport (CT): ACE/Praeger, 471–515.

Drave, N. (2010). Keeping up appearances: maintaining standards in Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE). Paper presented at the IAEA conference Bangkok, Thailand, August 2010.

Geranpayeh, A. (2011 January). The impact of online marking on examiners' behaviour. Cambridge ESOL Research Notes, 43, 15–21.

Hjollund, N., J. Andersen and P. Bech. (2007). Health and quality of assessment of fatigue in chronic disease: a bibliographic study of fatigue measurement scales. Life Outcomes, 5(12). Available online: http://www.hqlo.com/content/5/1/12

Jensen, C., L. Finsen, K. Søgaard and H. Christensen. (2004). Musculoskeletal symptoms and duration of computer and mouse use. Computers and Industrial Engineering, 46(3), 399–411.

Johnson, M., R. Nadas, and J. Bell. (2010). Marking essays on screen: an investigation into the reliability of marking extended subjective texts. British Journal of Educational Technology, 41(5), 814–826.

Kahol, K., M. Leyba, M. Deka, V. Deka, S. Meyers, M. Smith J. Ferrara and S. Panachanathan. (2008). Effect of fatigue on psychomotor and cognitive skills. American Journal of Surgery, 195(2), 195–204.

Lumley, T. and T. F. McNamara. (1995). Rater characteristics and rater bias: implications for training. Language Testing, 12, 54–71.

Meadows, M. and L. Billington. (2005). A review of literature on marking reliability. Report produced for the National Assessment Agency.

Moore, B. and C. Brooks. (2000). Heavy vehicle driver fatigue: a policy adviser's perspective. Proceedings of the 4th Int. Conference on Fatigue and Transportation, Fremantle, March 2000.

Müllera, C., L. Tomatisa and T. Läublia. (2010). Muscular load and performance compared between a pen and a computer mouse as input devices. International Journal of Industrial Ergonomics, 40(6), 607–617.

Petrilli, R., G. Roach, D. Dawson and N. Lamond. (2006). The sleep, subjective fatigue, and sustained attention of commercial airline pilots during an international pattern. Chronobiol Int., 23(6), 1357–62.

Rogers, A. E. (2008). The effects of fatigue and sleepiness on nurse performance and patient safety. Patient Safety and Quality: an Evidence-Based Handbook for Nurses, ed. R. Hughes. Rockville (MD): AHRQ (US), 509–545.

Shaw, S. (2008). Essay marking on-screen: implications for assessment validity. E–learning, 5(3). Available online: http://www.wwwords.co.uk/ELEA

Signal, T., P. Gander, H. Anderson and S. Brash. (2009). Scheduled napping as a countermeasure to sleepiness in air traffic controllers. Journal of Sleep Research 18(1), 11–9.

Sonnea, M., D. Villaltab and D. Andrews. (2010). Development and evaluation of an office ergonomic risk checklist: ROSA. Applied Ergonomics, 42(1), 29–36.

Stewart, S., A. Holmes, P. Jackson and R. Abboudh. (2006). An integrated system for managing fatigue risk within a low cost carrier. Paper presented at the 159th Annual International Air Safety Seminar (IASS), Paris, October 2006.

Theander, K. (2007). Fatigue, functional status, health and pulmonary rehabilitation in patients with chronic obstructive pulmonary disease. Linköping University Medical Dissertations No. 980.

Whitehead, L. (2009). The measurement of fatigue in chronic illness: a systematic review of unidimensional and multidimensional fatigue measures. Journal of Pain and Symptom Management, 37(1), 107–128.

Whitehouse, C. (2010). Reliability of on-screen marking of essays. AQA Paper RPA_10_CW_RP_012.

Appendix

Table 1: Summary statistics on leniency degree by different marking hours for Scale 1

Marking Hour	N	Mean	Median	SD	Minimum	Maximum	25th Percentile	50th Percentile	75th Percentile
Hour 1	99	.09	.10	.315	-0.75	1.00	-.04	.10	.31
Hour 2	95	.08	.07	.355	-0.80	1.50	-.17	.07	.27
Hour 3	57	.09	.13	.284	-0.50	0.75	-.09	.13	.29
Hour 4	36	.09	.09	.297	-0.50	0.80	-.08	.09	.25
Hour 5	23	.08	.09	.298	-0.75	0.50	-.10	.09	.36
Hour 6	16	.02	.00	.229	-0.38	0.33	-.14	.00	.26
Hour 7	11	.01	.00	.336	-0.50	0.50	-.20	.00	.35
Hour 8	8	.04	.04	.302	-0.30	0.63	-.25	.04	.21
Hour 9	3	-.25	-.13	.696	-1.00	0.38	-1.00	-.13	.
Hour 10	2	-.11	-.11	.734	-0.63	0.41	-.63	-.11	.
Hour 11	3	.55	.50	.079	0.50	0.64	.50	.50	.
Hour 12	2	-.12	-.12	.357	-0.38	0.13	-.38	-.12	.

Table 2: Summary statistics on leniency degree by different marking hours for Scale 2

Marking Hour	N	Mean	Median	SD	Minimum	Maximum	25th Percentile	50th Percentile	75th Percentile
Hour 1	99	-0.08	0.00	0.309	-1.00	0.50	-0.19	0.00	0.10
Hour 2	95	-0.01	0.00	0.293	-1.00	1.00	-0.17	0.00	0.17
Hour 3	57	-0.04	0.00	0.312	-1.00	0.65	-0.17	0.00	0.14
Hour 4	36	-0.04	0.00	0.289	-0.75	0.50	-0.17	0.00	0.13
Hour 5	23	-0.01	0.00	0.204	-0.46	0.40	-0.17	0.00	0.17
Hour 6	16	0.06	0.06	0.281	-0.50	0.63	-0.11	0.06	0.28
Hour 7	11	0.04	0.00	0.149	-0.14	0.31	-0.08	0.00	0.17
Hour 8	8	0.13	0.12	0.201	-0.10	0.50	-0.06	0.12	0.26
Hour 9	3	-0.05	-0.03	0.065	-0.13	0.00	-0.13	-0.03	.
Hour 10	2	0.08	0.08	0.058	0.04	0.13	0.04	0.08	.
Hour 11	3	-0.01	0.00	0.152	-0.17	0.14	-0.17	0.00	.
Hour 12	2	-0.22	-0.22	0.038	-0.25	-0.20	-0.25	-0.22	.

Table 3: Summary statistics on leniency degree by different marking hours for Scale 3

Marking Hour	N	Mean	Median	SD	Minimum	Maximum	25th Percentile	50th Percentile	75th Percentile
Hour 1	99	-0.07	0.00	0.357	-1.50	0.57	-0.17	0.00	0.08
Hour 2	95	0.03	0.00	0.364	-0.80	1.25	-0.17	0.00	0.20
Hour 3	57	-0.05	-0.07	0.284	-1.00	0.67	-0.20	-0.07	0.11
Hour 4	36	-0.04	-0.12	0.329	-1.00	0.58	-0.20	-0.12	0.16
Hour 5	23	-0.03	-0.05	0.353	-0.75	0.90	-0.20	-0.05	0.20
Hour 6	16	0.04	0.00	0.220	-0.28	0.50	-0.12	0.00	0.21
Hour 7	11	0.00	0.00	0.269	-0.50	0.50	-0.17	0.00	0.08
Hour 8	8	0.02	-0.08	0.268	-0.21	0.50	-0.19	-0.08	0.28
Hour 9	3	0.01	0.00	0.266	-0.25	0.28	-0.25	0.00	.
Hour 10	2	-0.12	-0.12	0.538	-0.50	0.26	-0.50	-0.12	.
Hour 11	3	0.19	0.25	0.168	0.00	0.32	0.00	0.25	.
Hour 12	2	-0.53	-0.53	0.315	-0.75	-0.30	-0.75	-0.53	.