Speed Marking: Can intuitive skill replace conscious analysis?

Denard Lynch
University of Saskatchewan
denard.lynch@usask.ca

Abstract *This paper discusses the results of an experiment to determine the efficacy and accuracy of evaluating report-based assignments using intuitive cues versus conscious analysis.*

The experiment involves the evaluation of typewritten reports an average of four pages in length. A conscious analysis requiring twelve to fifteen minutes was performed on each report. The reports were also evaluated using an intuitive technique averaging three to four minutes each. After normalizing, the grades were compared. The data show a moderate correlation (r = .47) between the intuitive and analytical assessments. The paper concludes that the while the efficiency is attractive, the accuracy is inadequate for practical application.

Future work could include an investigation to determine ways to "train" intuition by utilizing identifiable cues .

Keywords: Student assessment, Assessment tools, summative, grading systems

1. INTRODUCTION

Many instructors, including those in engineering colleges, undoubtedly share the author's experience of having evaluated hundreds or even thousands of essay-type responses as part of the evaluation and instructional processes. This can be a very time-consuming task! When required for purely evaluative reasons (e.g. term assignments or final examinations when time is already scarce), time efficiency may be the greatest motivation for this investigation. During the term, timely feedback to students can also create "schedule challenges" which could also benefit from a more efficient way to evaluate student work.

Undoubtedly most experienced instructors have, like the author, read only the opening paragraphs or a few seconds of an essay response and already have a "sense" of what the grade will be, at least approximately. The subconscious mind is known to process information and draw conclusions much more rapidly than our conscious, analytical mind can [1].

In his book "*Blink*", Malcolm Gladwell discusses the power of expert intuition as well as how easily it can be led astray. He also acknowledges that it takes experience to "train" our intuition [4]. Klein and Kahneman mutually concluded, the requisite conditions for acquiring intuitive skill are "an environment that is sufficiently regular to be predictable" and "an opportunity to learn these regularities through prolonged practice" [1]. A general course offered every term to approximately 300 students could certainly provide such an environment and opportunity. The premise behind this investigation is that "The mind operates most efficiently by relegating a good deal of high-level, sophisticated thinking to the unconscious…" [2], and that this can be exploited to reduce the time used to evaluate students' written responses. Nalini Ambady and Robert Rosenthal concluded that college students could arrive at essentially the same rating for an instructor based on a few second of non-verbal video as those who had a semester of in-class exposure to the same instructor [3]. The objective of this investigation was to determine whether, and to what degree, the reverse is true: "Can an instructor determine the quality of an assignment submission based on a very brief inspection of a written report?"

On expert intuition, Herbert Simon states: "The situation has provided a cue; this cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition." [5].

2. THE RESEARCH QUESTION

In the language of engineering design, the main objective of this research focus is to improve the efficiency of the evaluative process while maintaining acceptable accuracy of grading results. The primary function of any solution is to grade (evaluate) essay-style submissions. Other alternative solutions may be considered to achieve this end such as automated systems or peer evaluations, but the focus of this investigation is to assess the feasibility of using experience and intuitive knowledge to achieve this goal.

Thus the question to be answered through this investigation was "Can intuitive skill be used to improve

the efficiency of evaluating essay-type responses while maintaining acceptable accuracy?"

3. EXPERIMENTAL DESIGN

The design of this experiment involved three main elements: selection of a suitable test assignment, an evaluation plan that would provide both analytical and intuitive results, and a strategy to analyze the data and answer the research question. The sub-sections below provide details for each of these elements.

3.1 Assignment Selection

The key parameters considered in selection of a "suitable assignment" were:

- availability – does the investigator have access to the assignment, and can it be used for this purpose?
- quantity – is a sufficient quantity available so the results could have statistical significance?
- experience – does the evaluator have sufficient familiarity and experience with the subject matter to qualify as an "expert"?
- size – is the assignment submission large enough that "speed marking" would be advantageous, yet small enough to allow for multiple passes in the time available?

The principle investigator had first-hand access to three potential assignments. Their key attributes are summarized in .

Table 1. Assignment alternatives

#	Available	Quantity	Experience*	Size
1	Yes	45	~22	6 pp
2	Yes	115	~700	4 pp
3	Yes	32	~75	17 pp

* approximate number of similar assignments previously marked by principle investigator.

The assignment represented as #2 was selected, with available quantity and the investigator's previous experience with this assignment over the previous 16 years heavily influencing the selection. The assignment rubric is included in App. A.

3.2 Evaluation Plan

In order to meet the study objectives and provide grade data, the following evaluation strategy was followed:

- first pass (quick):
  - review the rubric and assignment objectives and expectations prior to evaluation;
  - "speed mark" the assignment submissions by scanning or "speed-reading" only until an intuitive sense of a grade is reached;
  - record the grade and proceed.
- second pass (analytical):
  - set-up a rubric grid to assess individual components of each submission;
  - consciously analyze the submission against each assessment category and record marks;
  - review result and proceed.

The strategy of performing the "quick" pass prior to the "analytical" evaluation was to mitigate the effect of familiarization that would occur if the longer analysis was done first. To further minimize the possible effects of familiarity (memory), approximately two weeks were allowed to pass between the evaluations.
In addition to the principle investigator, a junior colleague also evaluated all submissions against the rubric, although not thoroughly. A comparison of these results are included in the analysis.

3.2 Analysis Plan

The final component of the experimental set-up is a plan to analyze the data generated in relation to the research question.

The raw data resulting from the evaluation outlined in sub-section 3.2 consists of three sets of final scores for 114 separate assignment submissions. The data was prepared for analysis by:

- calculating the z-score for each grade
- standardizing the range of each score set to approximately equal values
- determining the correlation between the scores
- determining the grade error that would result from "speed marking" versus the traditional analytical approach.

The z-score for each was determined using the formula:

$$\text{z-score}_i = \frac{\left(\bar{X} - n_i\right)}{\sigma},$$

where $\bar{X}$ is the population mean, n_i is an individual score, and σ is the standard deviation of the data set. Z-scores were converted to percentage grades (i.e. max = 100) by using the inverse formula:

$$n_i = \bar{X} + \left(\text{z-score}_i\right)\sigma,$$

where the mean and standard deviation were adjusted to provide an approximately equal range of grades for each method. No adjustments were made to force the distribution to conform to a standard Normal distribution, although a visual inspection of a histogram of the data revealed an approximately Normal distribution, leading to the assumption that any variation from Normal would not

significantly affect the correlation. Table 2 shows the statistics for the adjusted datasets where the 'Analytical' column represents the scores obtained with a detailed, analytical evaluation, 'Quick Senior' the intuitive evaluation by the principle investigator, and 'Quick Junior' that of the junior colleague.

Table 2. Data set statistics

	Analytical	Quick_Senior	Quick_Junior
Avg	70.00	70.08	72.00
Min	47.63	47.16	47.22
Max	91.36	91.36	90.40
Range	43.73	44.20	43.17
StDev	7.48	7.99	7.79
set StDv	10.00	8.00	10.00
set Avg	70.00	70.00	72.00

The 'set StDv' and 'set Avg' rows show the values used to convert the z-scores for each set to the percentage scores used in the subsequent analysis.

The resulting dataset values were used in the correlation analysis which is described in section 4. RESULTS.

4. RESULTS

3.2 Correlation

The correlation coefficients between the three pairs of datasets, adjusted to have approximately the same minimum and maximum, were determined using the Pearson product moment correlation coefficient (PMCC) Table 3 summarizes the correlation between the analytical results and those obtained by "speed marking" (based on intuitive assessment).

Table 3. Correlation coefficient matrix

		Analytical	Quick_Senior
Analytical	R	1	
	R Std Err		
	t		
	p-value		
	H0 (5%)		
Quick_Senior	R	0.47	1
	R Std Err	6.94E-03	
	t	5.67	
	p-value	1.12E-07	
	H0 (5%)	rejected	
Quick_Junior	R	0.46	0.29
	R Std Err	7.06E-03	8.18E-03
	t	5.45	3.21
	p-value	3.06E-07	1.72E-03
	H0 (5%)	rejected	rejected

Limiting observations to the 'Quick_Senior versus Analytical' and 'Quick_Junior versus Analytical' comparisons, it is noted that the R standard error is very low, the *t* statistic is high compared to R, and the p-value is very, very low leading to a sound rejection of the null hypothesis. In aggregate, these indicators give a high degree of confidence that there is a statistically significant correlation between the datasets.

The correlation coefficient of 0.47 (or 0.46) indicates a *moderate* positive correlation between the 'Analytical' assessment and the "speed marking" assessments [6]. However, the correlation coefficient alone does not give sufficient indication of the practical implications of "speed marking" from an evaluative or a student's perspective. To further analyze these data, the absolute value of the *difference* in percentage grade (i.e. out of 100) for the 'Analytical' versus 'speedmarking' grade was plotted in a histogram (Figure 1). The associated numerical data is shown in Table 4.

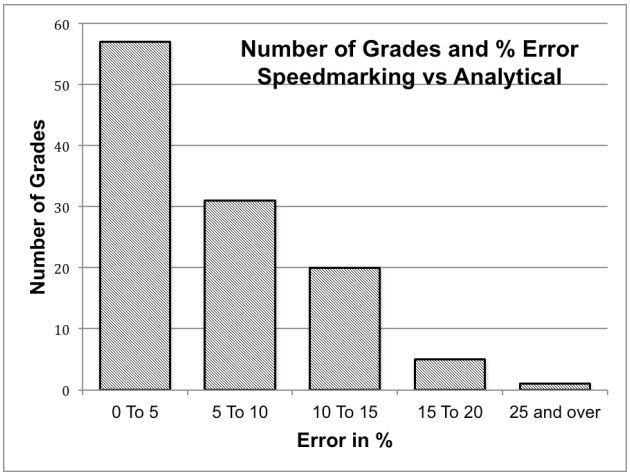

Figure 1. Speedmarking grade error

Table 4. Grade error vs analytical

Grade Error	Count	Percent	Cumulative %
0 To 5	57	50.0%	50%
5 To 10	31	27.0%	77%
10 To 15	20	18.0%	95%
15 To 20	5	4.0%	99%
25 and over	1	1.0%	100%

Making the assumption that the 'Analytical' evaluation is accurate, this result would show that 77% (88 grades) would be, on average, within one letter-grade of their "true" value. However, the remaining 23% (26 grades) would be in error by more than 10 points with one being approximately 25% in error.

3.2 Time Savings

It is a simple exercise to analyze the time savings for the "speedmarking" methodology. Table 5 shows the average time recorded for evaluation for each method as well as the time savings estimated for the 114 assignment submissions used in this investigation.

Table 5. Time savings

	per paper (min)	per 114 (hrs)	savings (hrs)
Analytical	13.5	25.65	
Quick Senior	3.4	6.46	19.19
Quick Junior	6.5	12.35	13.3

5. CONCLUSIONS

To properly frame any conclusions from this experiment and analysis, reconsider the original question: "Can intuitive skill be used to improve the efficiency of evaluating essay-type responses while maintaining acceptable accuracy?".

Addressing the simpler question first, yes, there is undoubtedly an improvement in time efficiency – the mathematics is simple and clear: if one spends one quarter the time on a given evaluation task, one can provide four times as much formative feedback for students or at least make more time available for other pedagogical needs. In a practical sense, however, the advantage is less compelling. Although the "burden" of marking for purely evaluative or ranking purposes can seem intrusive at times, its share of the institutional or even one's personal budget is relatively insignificant in the larger picture. This certainly does not mean that any efficiency should be sought after, provided the side effects are absorbable.

The second part of this question "…while maintaining acceptable accuracy…" is more complex and more challenging to answer with confidence. Based on the results of this experiment and analysis, the answer is clearly "no". Most would agree that having one quarter of students receiving a result that is more than ten grade points off would generally be unacceptable. Obviously one could not ignore a ten point error on any other examination, so it seems equally unacceptable even if the subject material is somewhat subjective.

However, this "inaccuracy" is based on the default assumption is that the analytical approach will produce a more accurate result in grading, but there is some evidence that where the decision requires consideration of a large number of factors, the intuitive method may actually produce a more accurate result. In addition to Ambady and Rosenthal's work, Gottman and more recently Carrère, Dijksterhuis and even Sigmund Freud have suggested that engaging our subconscious mind can provide more accurate results in some circumstances [4]. There is also some evidence that the accuracy of "quick assessments" can be improved through training [7]. Based on this literature, the concept may deserve further investigation.

6. FUTURE WORK

The results of this experiment have suggested there is an opportunity for further investigation to determine if and how intuitive techniques can be exploited to improve on current common evaluative methodologies.

One key area to explore is the development of a framework or test that can provide an acceptable standard method of determining an accurate evaluation for partially subjective material. The comparison of the two evaluators in this experiment ($R = 0.29$) suggests that there is often a difference between instructors, and it would be presumptuous to assume that the "experienced" evaluator is more accurate. Options to consider in this area may include: an exhaustive and exceptionally detailed rubric, an outcome-based assessment that doesn't include a written response, and multiple evaluations (either by different evaluators or the same evaluator over a broader timeframe).

Another avenue of investigation may be to identify "cues" that trigger intuitive decision making and evaluate them to use as conscious indicators, or use them to improve intuitive skills through training.

Should subsequent investigations conclude there is a viable use for intuitive techniques, the effects of varying the time spent evaluating could be explored.

Acknowledgements

The author wishes to acknowledge the support of the University of Saskatchewan for providing the means and opportunity for its instructors to explore and enhance their pedagogical skill and knowledge and contribute to the body of knowledge in this area. The author would also like to thank Dr. Sean Maw and the other members of the Innovative Teaching and Research in Engineering Education Group at the University of Saskatchewan for their support and suggestions. Finally, the author thanks the students in the subject course for their patience and understanding of the delays in grading related to this research.

References

[1] Daniel Kahneman, *Thinking, Fast and Slow*. N/A: Anchor Canada (Penguin Random House Canada), 2013, eISBN: 978-0-385-67652-6.

[2] Timothy D. Wilson, *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA, USA: Belknap Press of Harvard University Press, 2002, ISBN: 0-674-01382-4.

[3] N. Ambady and R. Rosenthal, "Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness," *Journal of Personality and Social Psychology*, vol. 64, no. 3, pp. 431-441, 1993.

[4] Malcolm Gladwell, *Blink: the power of thinking without thinking*. New York, NY, USA: Bay Back Books / Little, Brown and Company, 2005, ISBN 978-0-316-01066-5 /pb.

[5] Herbert A. Simon, "What is an Explnation of Behavior?," *Psychological Science*, vol. 3, pp. 150-161, 1992.

[6] James D. Evans, *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove, CA, USA: Brooks/Cole, 1996.

[7] HealthScoutNews. (2002, July) www.lifeclinic.com. [Online]. http://www.lifeclinic.com/fullpage.aspx?prid=508121&type=1

APPENDIX A: ASSIGNMENT RUBRIC

Report Format:
The report is to be submitted electronically in response to the corresponding assignment posted in Blackboard Learn, in MS Word format (please name your submitted document with the filename: lname1_lname2_Section.doc[x]; name1_name 2 in alphabetic order, e.g. fondstad_lynch_S04.docx, where 'S04' is the Section you attend. For split groups, you can use '…_S0402'. The recommended format for the written report is *semi-formal* (ref: MacLennan p.146). (also, 1-1/2 spaced, standard fonts of Arial, Times New Roman for the body text - font size 12, 2.5 cm margins, etc.). The *maximum* length for the body of the report is four (4) pages; (note that a cover page, abstract, references, interview questions and appendices are not included in the page count).

Report Rubric:

- Selection of subject
  - 2 Clear justification of relevant subject selection related to assignment objectives
  - 1 Some suggestion that relevance was considered in subject selection
  - 0 No indication that suitability of subject was considered
- Questions
  - 2 Interview questions were designed to stimulate conversation (i.e. open-ended) and focus on assignment objectives; adaptable to flow of conversation
  - 1 Some questions were open, evidence of some adaptation (if required)
  - 0 Questioning did not draw out relevant conversation or adapt to responses.
- Analysis and interpretation
  - 3 Clear evidence of analysis or interpretation beyond direct response of subject
  - 2 Some instances of evidence of analysis or interpretation beyond direct response of subject
  - 0 Very limited or no evidence of analysis or interpretation beyond direct response of subject
- Lessons learned
  - 3 3 or more interpretations of results that could be applied to improve engineering practice, or corroboration of principles considered in class
  - 1.5 1 or 2 interpretation of results applied to engineering
  - 0 No clear interpretation of results applied to engineering
- Formatting, grammar and spelling
  - 5 Well formatted/organized to accommodate reader; only very minor or no grammatical or spelling errors; very readable.
  - 3 Some grammatical or spelling errors, but did not substantially detract from readability
  - 0 Noticeable formatting, grammatical or spelling errors that significantly affected readability