Moshe Kam,[1] *Ph.D.; Gabriel Fielding,*[1] *M.Sc.; and Robert Conn,*[2] *Ph.D.*

Effects of Monetary Incentives on Performance of Nonprofessionals in Document-Examination Proficiency Tests*

REFERENCE: Kam M, Fielding G, Conn R. Effects of monetary incentives on performance of nonprofessionals in document-examination proficiency tests. J Forensic Sci 1998;43(5):1000–1004.

ABSTRACT: In September 1997 we reported on a comprehensive proficiency test administered to three groups of professional document examiners (105 individuals). Each test-taker performed 144 pairwise comparisons of original handwritten documents and matched together pairs that in his/her opinion were generated by the same hand. The test was also administered to a control group of nonprofessionals (41 individuals) whose educational profile was similar to that of the tested professionals. These nonprofessionals were motivated through a monetary incentive plan ($25 gain for each correct match; $25 fine for each erroneous match; $10 fine for failure to match documents created by the same hand). In this paper, we report on a subsequent study, aimed to discover whether changes in the monetary incentive scheme would affect the performance of nonprofessionals, and whether these schemes would close the performance gap between professionals and nonprofessionals. We administered the 1997 test again, this time to four groups of nonprofessionals (132 subjects), using four different incentive schemes (including the one used originally). We found that the four sets of data obtained under different incentives were indistinguishable, in the sense that differences between the test scores were not statistically significant. We conclude that the performance of nonprofessionals in our proficiency test was relatively insensitive to the monetary incentive scheme.

KEYWORDS: forensic science, document examination, proficiency testing, writer identification, handwriting analysis, handwriting tests, questioned documents, monetary incentives

The capabilities of forensic document examiners continue to be a topic of great interest in the forensic literature and in the courts (1–6). While many articles and court briefs have been written on the subject, the only *controlled* tests on proficiency of questioned-document examiners were reported by us in 1994 (5) and 1997 (6). The 1994 test was a small-scale pilot study involving seven professional document examiners and ten nonprofessionals. The 1997 test was a full-scale comprehensive study, involving 105 professionals and a control group of 41 nonprofessionals. The present paper is an extension of the 1997 study with three interrelated objectives:

(*i*) to examine whether (and to what extent) monetary incentives affect the scores that nonprofessionals achieve on the 1997 proficiency test;

(*ii*) to respond to vigorous criticism presented in court proceedings (1–3) of the monetary incentives used in the 1997 test [e.g., (1), M Denbeaux's testimony, pp. 296–299][3]; and

(*iii*) to discover those monetary incentives that, if they exist, ''narrow the gap'' between the scores of professional document examiners and of nonprofessionals.

In order to achieve objectives (i)–(iii), we have readministered the 1997 comprehensive document-matching test to four groups of nonprofessionals, using a different monetary incentive scheme for each group. The incentives were devised to encourage different patterns of response and to span the space of awards and penalties for different decisions. We wanted to know whether changes in monetary incentives would elicit significant changes in scores of the tested individuals.

Monetary Incentives in Skill Testing

Designers of skill tests are naturally interested in employing motivated and interested subjects. This is equally true for test-takers whose skills are measured, and for the control group. Specifically, lack of motivation in the control group may cause poor performance not indicative of true skills.

Developing proper motivation for test subjects is, however, a difficult goal. It is widely recognized that ''studies on the motivation of experimental subjects, taken together, seem to suggest that human subjects are too complex and diverse to describe by simple models, presuming good, bad or anxious subjects'' (7).

Not surprisingly, there is no consensus in the research-methodology and applied-psychology literature on the ''correct'' way to assign rewards and punishments in skill tests and visual tests (e.g., (8) and its references). While money is widely recognized as a good motivator, lack of monetary resources led many experimental psychologists to use unpaid student volunteers or assign skill tests to the ''captive audience'' of students in their classes (9,10). An

[1] Data Fusion Laboratory, Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA 19104.

[2] Applied Physics Laboratory, The Johns Hopkins University, Baltimore, MD 20723.

* The study was supported by the Federal Bureau of Investigation.

Received 23 Sept. 1997; and in revised form 12 Jan. 1998; accepted 3 Feb. 1998.

[3] M. Denbeaux and others offered the conjecture that the monetary incentives used in (6) somehow encouraged ''betting'' by nonprofessional test-takers. No empirical evidence was presented to substantiate this conjecture. We provide in the Appendix an excerpt of Prof. Denbeaux's testimony on this subject in (1).

extensive review of relevant journals[4] shows that test-per-course-credit is the most popular method used by experimentalists. When pay was used, several different pay incentive schemes were employed, quite often involving token sums ($5 for responding to a survey (11), $20 for one day of work (12), $75 for training sessions over 15 days (15)). Paying methods varied significantly from study to study, and experimenters seldom explained in their written articles the rationale behind the chosen payment scheme. Most studies that used monetary incentives paid a flat fee (e.g., (11,12)) or an hourly fee (e.g., $5/hour in (13)). Some studies paid a flat fee plus performance-based bonuses. In (14) participants in a cognitive ability test were paid a $25 flat fee plus a fee of $15 for good performance. In (15) trainees received $75 each for 15 days of training, and competed for three bonuses of $100, $60, and $40. In (16) participants in a traffic information test received $5/hour plus performance-related bonuses and penalties in the range of $15.

The new tests described below are much more substantive than the ones described in the literature in that the monetary incentives are much more substantial, and that several different incentives are experimentally compared.

The Task

As in (6), the test was to associate original handwritten documents from a package of six documents (the *unknown* package), with original handwritten documents in a separate package of 24 documents (the *database*). A total of $6 \times 24 = 144$ comparisons were needed in order to complete a test. A test-taker was to declare a ''match'' between two documents only if in his/her opinion both documents were created by the same writer. With the exception of monetary incentives, we have used the same documents, methods, and instructions that were used in (6).

The Test Takers

We have selected a homogeneous group of students of ages 19 to 21, recruited primarily from the sophomore class of Drexel University's College of Engineering. About 87% of these students require financial aid during their studies at the college. Members of this group are therefore likely to be highly sensitive to monetary incentives of the kind that our tests offered.

Method

We have tested 132 individuals in four groups (of 33, 33, 32, and 34 individuals). There were two testing sessions, one on a ''leisurely'' Saturday morning in the middle of the winter academic term, and one on a ''busy'' Thursday evening before a Friday deadline for two major homework assignments. Four distinct sets of test instructions were used, describing four different incentive schemes. The decision on which incentive each test-taker would receive was made at random a few minutes before the test started. As in (6), test-takers received explicit oral explanations of the task and the test, and they had an opportunity to ask questions. No a priori information about the distribution of the documents, the tests, or their writers was revealed.

[4] Journal of Experimental Psychology: Learning, Memory, and Cognition; Journal of Experimental Psychology: Human Perception and Performance; Journal of Applied Psychology: Memory and Language; Human Factors.

The Incentives

Four incentive schemes were provided. We use the following notation:

- H_0: null hypothesis, the document in the *database* and the document in the *unknown set* do not match;
- H_1: alternative hypothesis, the document in the *database* and the document in the *unknown set* do match; and
- C_{ij}: penalty in US dollars for accepting H_i when H_j is true (negative penalties are monetary gains for the participants).

Incentive I $C_{00} = 0$ $C_{01} = 10$ $C_{10} = 25$ $C_{11} = -25$
Incentive II $C_{00} = 0$ $C_{01} = 25$ $C_{10} = 25$ $C_{11} = -25$
Incentive III $C_{00} = 0$ $C_{01} = 5$ $C_{10} = 50$ $C_{11} = -25$
Incentive IV Participant receives $100 at the beginning of the test, then
$C_{00} = 0$ $C_{01} = 25$ $C_{10} = 25$ $C_{11} = 0$

All incentives had a gain floor of $25—if the overall reward based on these penalties was less than $25, the participant received $25.

Incentive 1—Exactly the one used in (6). It is intended to encourage *correct matches* (gain of $25), discourage *wrong matches* (penalty of $25), and discourage *failures to match* (penalty of $10). It therefore discourages *failures to match* in a milder way than it discourages *wrong matches.*

Incentive 2—Provides the same monetary penalty for a *wrong match* ($25) as for a *failure to match* ($25); it rewards correct matches to the same extent ($25) that it penalizes for an error.

Incentive 3—Provides a high penalty ($50) for a *wrong match,* and a light penalty ($5) for a *failure to match,* with a medium reward ($25) for a *correct match.* It is intended to strongly discourage wrong matches and encourage participants to be cautious.

Incentive 4—Provides a significant monetary incentive ''up front,'' then penalizes equally for both types of error. The idea was to see whether the risk of losing the initial reward will change the behavior of the test-taker.

Criteria for Comparison

We use same scoring criteria used in (6) (with the exception of the *earning ratio* criterion, which became irrelevant once different monetary incentives were introduced.)

Criterion 1: Number of Wrongly Associated Documents

We have assigned to each test-taker a score based on the number of *unknown* documents that he/she associated wrongly with *database* documents. The score 0 was given to examiners that associated no *unknown* document wrongly with a *database* document. The score 6 was given to examiners that wrongly associated all six *unknown* documents with *database* documents.

Criterion 2: Hit Rate

We scored individuals by *hit rate* $= m/n$, where m = number of correct matches declared by the individual, and n = number of matches that existed in the individual's test.

TABLE 1—*The P-rank method for assigning grades for performance.*

Incorrect Matches	Missed Detections	Grade
0	0	1
0	1	2
0	2	3
0	more than 2	4
1	0	5
1	1 or more	6
2	0	7
2	1 or more	8
3	0	9
All other combinations of errors		10

TABLE 2—*Values of the KS statistic, D.*

	1 vs. 2	1 vs. 3	1 vs. 4	2 vs. 3	2 vs. 4	3 vs. 4
wrong association rate	0.1515	0.1619	0.0642	0.1600	0.1248	0.2261
Hit Rate	0.1515	0.1515	0.1230	0.1515	0.0980	0.1489
$P(H_1\|D_1)$	0.1515	0.1638	0.0695	0.1600	0.1337	0.1710
P-rank	0.1818	0.2396	0.1025	0.1335	0.1319	0.1710
Critical Value	0.3348	0.3374	0.3323	0.3374	0.3323	0.3350

Criterion 3: Probability that an Association Declaration is Correct

We have scored all test-takers in terms of the ratio of correct-matching decisions to the total number of matching decisions [$P(H_1$ is the correct hypothesis | H_1 was the declared correct) or $P(H_1|D_1)$].

Criterion 4: P-rank

We developed a grading scheme, called a *performance rank* (or P-rank) based on the different types of errors observed in our test. This scheme divides the test-takers into ten different sub-groups, based on the severity and number of errors observed. The grading scheme is described in Table 1.

Hypotheses Tested

Using these four criteria, we tested seven hypotheses.

Hypothesis Tests 1–6

(Corresponding to incentive pairs $(i, j) = (1, 2), (1, 3), (1, 4), (2, 3), (2, 4)$, and $(3, 4)$).

We test

- H: there is no difference between the scores collected from the group that had incentive *i* and those collected from the group that had incentive *j;* against the hypothesis that
- K: there is a difference in the scores collected from the group that had incentive *i* and the group that had incentive *j*.

Hypothesis Test 7

We test

- H: there is no difference in the scores collected from the four groups of nonprofessionals; against the hypothesis that
- K: there is a difference in the scores collected from the four groups of nonprofessionals.

We have used the *Kolmogorov-Smirnov (KS) two-sample test* for hypotheses 1–6 (17, p. 127, 18, section 3.9.3). For hypothesis 7 we have used both the *Birnbaum-Hall (BH) k-sample test* (19, p. 382, 20) and the *Kruskal-Wallis (KW) one-way analysis of variance by ranks* (17, chapter 8, 18, section 3.9.5). In presenting results from these tests we follow the notation in (6).

Results of Hypotheses Testing

Hypotheses Tests 1–6

Table 2 shows the KS statistic, D. The critical value of D for a significance level of 0.05 is shown in the last row of Table 2. Since none of the values in Table 2 exceed the critical value, we accept H for all criteria and all pairs of incentive schemes. An alternative way to analyze the results is to examine the ''*p*-values'' associated with scores Table 3; these are the probabilities of obtaining, under hypothesis H, larger values than the observed scores). Both approaches lead to the same conclusion. **The differences between the data produced by any two test-taker groups that had different incentive schemes are not statistically significant.**

Hypothesis Test 7

Table 4 shows the BH statistic T, as well as the probability *p* of obtaining, under hypothesis H, a larger value of the statistic than the observed value (calculated by the iterative scheme proposed in (20)). We clearly accept H. **The data produced by the four groups that were given different incentives exhibited no statistically significant differences.**

The main conclusion of our study is therefore that the incentives did not induce statistically significant differences in the performance of the non-experts. Additional testing found no significant differences between the data from the ''Saturday group'' and the ''Thursday group.''

Performance of Professional and Nonprofessional Test-Takers

The primary objective of the present study was to quantify the effect of incentives on the performance of nonprofessionals in document examination. However, we are afforded an opportunity to

TABLE 3—p-*values for the KS statistic, D, as given in Table 2.*

	1 vs. 2	1 vs. 3	1 vs. 4	2 vs. 3	2 vs. 4	3 vs. 4
wrong association rate	0.8017	0.7498	1.0000	0.7626	0.9424	0.3259
Hit Rate	0.8107	0.8178	0.9486	0.8178	0.9952	0.8275
$P(H_1\|D_1)$	0.8107	0.7369	1.0000	0.7626	0.9050	0.6790
P-rank	0.6015	0.2690	0.9916	0.9147	0.9133	0.6790

TABLE 4—*BH statistics for the four groups of nonprofessional test-takers.*

	BH statistic (T)	p	Decision
wrong association rate	0.2261	0.768131	DO NOT REJECT H
Hit Rate	0.1515	0.996453	DO NOT REJECT H
$P(H_1\|D_1)$	0.1710	0.977039	DO NOT REJECT H
P-rank	0.2396	0.677348	DO NOT REJECT H

TABLE 5—*Performance of three tested groups.*

	Hit Rate	Wrong Association Rate	Average Time, min	Std. Dev. Time, min
''New'' nonprofessionals	0.811	0.227	0:49	0:19
''Old'' nonprofessionals (6)	0.877	0.383	0:58	0:24
Professionals (6)	0.879	0.065	1:32	0:30
Ideal	1.0	0.000		

compare the present test-takers (''new'' nonprofessionals) to the test-takers that we have examined before in (6) (''old'' nonprofessionals), and to compare both groups to the professionals from (6).

Table 5 shows the absolute performance of three groups along with the average and standard deviation of the time they have spent on the tests.

Performance is shown in terms of the **pair** of probabilities:

- Hit Rate: probability that a match was declared given that a match existed (i.e., P (accept $H_1|H_1$ is true), ideally 1.0, higher scores signify better performance); and
- Wrong Association Rate: probability that an *unknown* document was wrongly matched to a *database* document, ideally 0.0, lower scores signify better performance.

As we have explained in (6) these probabilities are *linked* and they should be *always* considered *together.*

We came to the following conclusions:

1. Both the ''old'' nonprofessionals and the ''new'' nonprofessionals perform significantly worse than the professionals; and
2. the ''new'' nonprofessionals have a lower (i.e., worse) hit rate, and lower (i.e., better) w.a.r. when compared to the ''old'' nonprofessionals.

The second conclusion is consistent with the general tendency of the ''new'' nonprofessionals to make fewer matching declarations of *any* kind (right or wrong) compared to the ''old'' nonprofessionals. With fewer matching declarations, the wrong association rate of the ''new'' nonprofessionals improved, but the hit rate worsened.

Conclusion

We have tested data produced by four groups of nonprofessionals who took the proficiency test described in (6) under four different monetary incentives. There were no statistically significant differences between the data produced by the four groups, leading us to conclude that the details of monetary penalty/reward scheme did not play a significant role in affecting the proficiency scores of nonprofessionals. The nonprofessionals performed the same regardless of incentive scheme, and exhibited markedly inferior performance (in terms of the linked pair *hit rate* and *wrong association rate*) compared with that of the professional document examiners tested in (6).

Appendix

Criticism of Monetary Incentives in (6)

The following is an excerpt from testimony by M. Denbeaux in US vs. Martin (1, pp. 296–299). Questions (Q) are by Ms. L. C. Matucci, Assistant US Attorney. Answers (A) are by Professor M. Denbeaux.

A: . . . we also know that there are some methodological questions I have with the incentive system dealing with the college students, and I think that could lead to some—

Q: Based upon your experience as a law professor?

A: No. Actually it's dealing with my experience as a college student myself and the parent of three others. That if my children were told here is the test and here's how you make money, you get paid $25 if you are right on each match, you lose $25 if you are wrong, and if every time you get in equipoise, it's very close, you're not sure, but you know that if you say I'm not sure you are going to lose $10, and if you say—and you gamble and you say you are right, you have a chance to make $25, so it's a guaranteed loss of $10, it's a possible loss of $15 more and it's a possible winning of $25, I have at least two of my children who would take that bet every time.

Q: Dr. Kam was very clear about the fact that there was a bottom line as far as the amount of money that the participants in the examination could lose.

A: Well, actually he said they could never lose money, they could only make—but they could make $25. But he did say that in his incentive system in his advertisements he projected that they could make as much as $100 to $200 for an hour's work. Well, you can't make $100 to $200 for an hour's work unless you put some bets down and try to make some matches.

Q: So you are saying that you think these people were just betting on Dr. Kam's tests or guessing to make some money?

A: No, I didn't say guessing. I am very sure that where it was clearly—knowing my children, and at least two of them are bettors and my daughter is probably not—

Q: Well, regardless of what you—

A: Well, excuse me. Let me just—without relating to my children, I think it's very likely that they would divide it up with a triage system. Where they are sure it's a match, they'd say it's a match. Where they are sure it's not a match, they would say it wasn't a match. But in that middle area where it's very close and they are not sure, the economic incentives of that system—and Dr. Kam himself said that he had never taken into account the psychology of the people doing it—would lead students to make bets in favor of trying to make a match. That seems to me—

Q: Am I understanding you are saying that Dr. Kam's study is flawed because of the incentive system that was used here?

A: Yes, although I think it's very hard to come up—

Q: Thank you.

A: —with a matching incentive system.

The court: Let him finish his answer, please.

The witness: I think it's very hard to come up—as Dr. Kam and I have discussed, it's very hard to come up with a perfect incentive system for nonprofessionals.

References

1. United States vs. James Martin. United States District Court, Northern District of Georgia, Case No. 1:96-CR-287(s).
2. United States vs. Douglas Lambert. United States District Court, Middle District of Florida, Case No. 93-139-Cr-Orl-19.
3. United States vs. Curtis Evans, United States District Court, Western District of Pennsylvania, Case No. 96–167.
4. Hansen M. Handwriting Analysis Under Fire. Am Bar Assoc J, 1997;83(5)76–8.

5. Kam M, Wetstein J, and Conn R. Proficiency of professional document examiners in writer identification. J Forensic Sci 1994;39: 5–14.
6. Kam M, Fielding G, Conn R. Writer identification by professional document examiners. J Forensic Sci 1997;42(5):778–86.
7. Singleton SA, Straits BC, Straits MM. Approaches to social research. New York: Oxford University Press, 1993;200.
8. Smith RL, Lucaccini LF, Epstein MH. Effects of monetary rewards and punishments on vigilance performance. J Appl Psychol 1967; 51(5):411–6.
9. Lupker SJ, Brown P, Colombo L. Strategic control in naming tasks: changing routes or changing deadlines? J Exp Psychol Learn Mem Cogn 1997;23(3):570–90.
10. Chan D, Schmitt N. Video-based versus paper-and-pencil method of assessment in situational judgement tests: subgroup differences in test performance and face validity perceptions. J Appl Psychol 1997;82(1):143–59.
11. Wanberg CR, Andecedents and outcomes of coping behaviors among unemployed and reemployed individuals. J Appl Psychol 1997;82(5):731–44.
12. Horowitz IA, ForsterLee L, Brolly I. Effects of trial complexity on decision making. J Appl Psychol 1996;81(6):757–68.
13. Downing PE, Treisman AM. A line-motion illusion: attention of impletion? J Exp Psychol Hum Percept Perform 1997;23(3): 768–79.
14. Chan D, Schmitt N, DeShon RP, Clause CS, Delbridge K. Reactions to cognitive ability tests: the relationships between race, test performance, face validity perceptions, and test-taking motivation. J Appl Psychol 1997;82(2):300–10.
15. Arthur W, Day EA, Bennett W, McNelly TL, Jordan JA. Dyadic versus individual training protocols: loss and reaquisition of a complex skill. J Appl Psychol 1997;82(5):783–91.
16. Kantowitz BH, Hanowski RJ, Kantowitz SC. Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. Hum Factors 1997;39(2):164–76.
17. Siegel S. Nonparametric statistics for the behavioral sciences. New York: McGraw Hill, 1956;127–36,184–94.
18. Sachs L. Applied statistics—a handbook of techniques. New York: Springer Verlag, 1984.
19. Conover WJ. Practical non-parametric statistics. New York: Wiley, 1980.
20. Birnbaum ZW, Hall RA. Small sample distributions for multi-sample statistics of the Smirnov type. Ann Math Stat 1960;31:710–20.

Additional information and reprint requests:
Moshe Kam
Data Fusion Laboratory
Electrical and Computer Engineering Department
Drexel University
Philadelphia, PA 19104