Can graphology predict occupational success? Two empirical studies and some theoretical ruminations

Article *in* Journal of Applied Psychology · November 1986

DOI: 10.1037/0021-9010.71.4.645

Gershon Ben-Shakhar
Hebrew University of Jerusalem
163 PUBLICATIONS **4,478** CITATIONS

Maya Bar-Hillel
Hebrew University of Jerusalem
76 PUBLICATIONS **2,582** CITATIONS

Journal of Applied Psychology
1986, Vol. 71, No. 4, 645–653

Can Graphology Predict Occupational Success? Two Empirical Studies and Some Methodological Ruminations

Gershon Ben-Shakhar, Maya Bar-Hillel, Yoram Bilu, Edor Ben-Abba, and Anat Flug
The Hebrew University
Jerusalem, Israel

Two empirical studies for testing the validity of graphological predictions are reported. In the first, the graphologists rated bank employees on several job relevant traits, based on handwritten biographies. The scripts were also rated on the same traits by a clinical psychologist with no knowledge of graphology. The criterion was the ratings on the same traits by the employees' supervisors. The graphologists' and the clinician's correlations with the criterion were typically between 0.2 and 0.3. To test whether these validities might be attributable to the scripts' content, we developed a third method of prediction. The information in the texts (e.g., education) was systematically extracted and combined in a linear model. This model outperformed the human judges. In the second study, graphologists were asked only to judge the profession, out of 8 possibilities, of 40 successful professionals. This was done on the basis of rich (e.g., containing numbers and Latin script as well as Hebrew text), though uniform, scripts. The graphologists did not perform significantly better than a chance model.

The measurement and prediction of personality traits present a major obstacle for personnel selection. Traits such as honesty, responsibility, independence, sociability, and so forth, seem to be desirable and even necessary for many occupations, yet traditional psychological testing devices typically fail to predict associated job behavior with anything approaching satisfactory rigor. The increasing demand for better personnel selection, combined with the weakness of standard personality tests, has led many firms to turn to alternative prediction methods—most notably, graphology. Levy (1979) reported that graphology is routinely used in the hiring of personnel by 85% of firms in Europe. Rafaeli and Klimoski (1983) estimated that 3,000 American firms use this tool, and the number appears to be growing. In Israel, graphology is more widespread than any other single personality test.

In view of this trend, it is surprising to note the paucity of serious research efforts to assess the validity of graphology in predicting job performance. Such research as is available typically suffers from one or more of the following methodological problems.

1. Nonstandardized assessments. The typical graphological output is a free-style overall qualitative personality description. This kind of material is hard to correlate with any independent criterion. There are two ways to deal with this problem—by standardizing the assessments, or by choosing an evaluation method that can handle the nonstandard assessments. To achieve the first, the graphologists may be requested to dispense with the overall description and merely to rate the writers on several predefined traits. Unfortunately, most graphologists find it unnatural to work this way, and for reasons that are at least partially valid (see Bem & Allen, 1974). To achieve the second, Crumbaugh and Stockholm (1977) developed what they called a holistic technique. Judges who are acquainted with all the writers in a sample match the graphologists' free-style descriptions (from which direct references to the writers are deleted) against the names of the writers. Correct matches in excess of chance expectation is taken as evidence that the descriptions carry at least some valid information about the writers. This method, however, has only limited applicability. For example, it cannot be used to corroborate the claim that graphological signs can indicate whether a person is honest or not, suitable for some job or not, and so forth.

2. Flawed criteria. When graphologists attribute to writers traits such as honesty and responsibility, satisfactory validation criteria are hard to come by, as these traits are not directly observable and can seldom be independently ascertained with sufficient certitude. The most commonly used criteria are the predictions of other, more standard, personality tests, or the subjective evaluations of people who are well acquainted with the writers. The problem with these criteria is that personality tests have notoriously low validities themselves, and subjective

We wish to thank the following people for their assistance in various phases of this project: Avishai Margalit for alerting us to the proliferation of graphology for personnel selection in Israel; Hagit Benziman for giving us her valuable time and the benefit of her expertise as a psychodiagnostician; Micha Weiss and Vered Kozachok, who assisted in the data gathering stages of Experiment 1; Arie Naftali for advice in collecting the graphological material in Experiment 2; Yossi Rinot for help in analyzing the data of Experiment 2; the graphologists who participated in the studies and requested to remain anonymous; and the Sturman Center for Human Development for supporting this study. Experiment 1 is also described in Ben-Shakhar, Bar-Hillel, and Flug (1986).

Correspondence concerning this article should be sent to Gershon Ben-Shakar, Department of Psychology, Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905.

evaluations are often unreliable. This makes it hard to identify the culprit if a mismatch is found between the graphologists' predictions and the criteria.

3. Contaminated texts. This factor refers to the confounding of graphological information with other sources of information. Contamination is most apparent when the handwritten text is a brief autobiography of the writer, as it typically is in personnel screening contexts. Clearly, such texts contain a great deal of information about the writer that is relevant for predicting job performance criteria, (e.g., education, previous work record). Moreover, nonbiographical but spontaneous text is also contaminated, most notably by the writer's verbal abilities, such as vocabulary, articulateness, and clarity of expression. These are correlated with successful performance in many jobs. Because graphological validity refers to the form, rather than the content, of written material, the confounding of the two makes it difficult to assign the appropriate weight to the one versus the other.

Contamination is hard to eliminate, because many graphologists insist on analyzing only spontaneously produced text, claiming that copying a text changes the graphological characteristics of the written material. Graphologists insist that they attend only to the graphological features of the text, ignoring its contents. However, besides the a priori implausibility of this claim, studies typically find that nongraphologists who read the same texts achieve the same (low) validities as do graphologists (e.g., Jansen, 1973; Rafaeli & Klimoski, 1983), or even outperform them (e.g., Frederick, 1965). Such results clearly shift the burden of proof (that their validities are not due to content) to the graphologists.

In view of these various methodological problems, it should come as no surprise that validation studies of graphology have shown a mixed bag of results. In general, the methodologically tighter a study, the less impressive the graphologists' performance. On the other hand, the tightening of the method is usually disadvantageous to the graphologists, who are often thereby barred from working in their natural mode with their preferred material.

We conducted two validity studies, each addressing these methodological difficulties differently. The first, conducted in a typical personnel selection situation, used a typical criterion—supervisor evaluation. The graphologists were given their favored material—authentic handwritten autobiographical sketches written by people who were unaware that their text would be graphologically analyzed. Control for text content was exercised by predicting from the texts' nongraphological information. In addition to recruiting a human judge for this task, we derived a simple linear model based on this information. The latter control constitutes a novel contribution of this study. All predictions by all predictors were made on standardized quantitative scales.

The second study required concurrent prediction of occupation. To overcome the contamination problem, the texts—contributed by eminently successful representatives of several professions—were identical in content. The graphologists' task was confined to guessing the writer's profession out of a given list. This relieved them from the need to use artificial scales, while relieving us of the difficulties of freely written personality descriptions, and providing a clearcut criterion.

Experiment 1

Method

Writers and Material

The handwritten texts were taken from the files of employees of two large Israeli banks. These files had been compiled by a reputable professional Israeli firm specializing in preemployment screening and personnel evaluation (we call the firm PT). The handwritten samples had been requested at the job application time (one to three years prior to our study) by candidates for employment at the banks. Almost all of them were brief (up to one page) autobiographical sketches. The search for material was stopped when 80 scripts were found that also had criterion information (see below) in their files. No selection was exercised. The candidates included equal numbers of men and women. Most were between 19 and 27 years old, with a few older. A short questionnaire attached to their scripts stated their age, as well as their army health classification and rank.

Besides the scripts, the files contained the predictions, on similar scales, that had been made at the preemployment screening time by the firm's psychologists on the basis of an entire battery of aptitude and personality tests and other observations (e.g., small-group interactions and interviews). Some analyses were extended to these evaluations.

Graphologists

The three graphologists who participated in this study were recruited via PT, on whose payroll they are. The purpose and design of the study was explained to them, as well as the nature of the criterion. They consented to make their predictions on the evaluation form described below.

Evaluation Form

The evaluation form was a subset of the form actually filled out by the employees' supervisors for purposes of performance evaluation and promotion. It consisted of three job-related areas: (a) nine items related to level of performance and ability (e.g., "ability to learn from mistakes and make appropriate adjustments"), (b) six items related to interpersonal relations (e.g., "willingness to help fellow workers"), and (c) nine items related to loyalty to the job and compliance with job requirements (e.g., "shows up on time to work"). There was also a single summary item called "overall evaluation." The ratings of all the items were on a scale from 1 to 6.

Criterion

The criterion was provided by the assessments of the employees by their direct supervisors, as found in their files (PT had already standardized these to control for individual differences among different supervisors in using the rating scales). So the task was one of postdiction.

Procedure

The scripts were assessed four times—once by each of the four judges. Each one worked independently, in his or her own free time and order. They were permitted to refrain from judging texts that they saw as unfit or baffling.

Table 1
Nine Items of Nongraphological Information Contained in the Scripts, and Their a priori Assigned Values

Variable	Value
V1. Education	
University graduate	V1 = 4
College studies without degree	V1 = 3
High school graduate with matriculation	V1 = 2
12 years of school without matriculation	V1 = 0
Less than 12 years of school	V1 = −2
V2. Army rank	
Officer	V2 = 5
Enlisted man in commanding position	V2 = 0
No army service, or enlisted man	V2 = −1
V3. Arrival in Israel	
Prior to 1967 and Israeli born	V3 = 0
Between 1967 and 1971	V3 = −1
Later than 1971	V3 = −2
V4. Marital status	
Married	V4 = 1
Single	V4 = 0
V5. Vocational interests	
Theoretical, commercial	V5 = 2
Other or none	V5 = 0
V6. Overall quality of written essay	
Very good	V6 = 3
Good	V6 = 1
Fair	V6 = 0
Poor	V6 = −1
Very poor	V6 = −2
V7. Aesthetic evaluation of script	
Beautiful or nice	V7 = 1
Fair or poor	V7 = 0
V8. Grammatical or spelling errors	
None	V8 = 1
Almost none	V8 = 0
V9. Overall impression of writer	
Good	V9 = 2
Fair	V9 = 0
Poor	V9 = −1

Linear model

The attempt to build a linear model of the information contained in the scripts was guided by informal and intuitive considerations, based largely on the authors' judgments and the clinician's introspection as to what variables she felt had influenced her judgments when reading the scripts. Nine variables and their values were defined, as summarized in Table 1. Each script was scored on each variable independently by two graduate students.

Results

Data Reduction

Because the reliabilities of single-item ratings are lower than those of combined scores, we arbitrarily reduced the entire set of scores given to each script to just five scores. For each judge, the specific scales within each of the three areas (level of performance, interpersonal relations, and job compliance) were averaged to arrive at a single score for that area. In addition, a simple average of the three area scores was computed to arrive at a grand average. Along with the original scale of "general evaluation," this yields five scores per handwriting—three specific ones and two general ones.

Validities

The five scores given by each judge (the three graphologists, the clinician, and the firm test battery) were correlated with the corresponding five supervisor ratings. The resulting Pearson product-moment coefficients are presented on the left-hand side of Table 2. Where the table reports fewer cases than 80, this is due to a failure on the part of the corresponding judge to rate some script(s) (e.g., the graphologists were reluctant to judge scripts not written by native Hebrew speakers). To facilitate meaningful comparisons of the correlations, these were recomputed for the 58 scripts that were evaluated by *all* five judges, and these correlations are displayed in the right-hand side of Table 2. Although all these correlations are positive, indicating some predictive ability, almost none exceed .4, and only about half are significant (at these sample sizes, significance at the .05 level occurs at about $r = .25$).

The clinician outperformed (though not significantly) all three graphologists on the two global criteria, and also compared favorably with the firm's test battery. With respect to the three specific areas, her performance was about on a par with the graphologists'—better than some, worse than others.

To see whether the five different predictors (the four human judges and the test battery) tapped different sources of variance, a regression analysis was conducted on the criterion of "general evaluation." The firm test battery accounted for the most criterion variance—11.2%; none of the other judges added significantly to the percentage of explained variance, although the total explained variance of the judges was 20.25%.

Linear Model

Each script was scored on each of the nine variables in Table 1 by the average of the two scores given independently by our two research assistants. Their interjudge reliabilities on the nine variables were measured by the Pearson correlation between their ratings (see Table 3). Note that some of the variables are strictly objective (e.g., V1, V2, V3, V4), whereas others involve some subjective judgment (e.g., V6, V7, V9). Not surprisingly, the reliabilities were higher for the former (.95 and up) than for the latter (between .20 for V9 and .53 for V7). Nonetheless, the validities for V6 through V9 were, on the whole, higher than for V1, V2, or V3. The correlations between the assessors' average rating and the two general criteria (i.e., the supervisors' general evaluations, and the computed average of the three area scores) are displayed in Table 3. Because some texts failed to include information about some of the variables, Table 3 is based on those 52 scripts for which *all* nine variables could be scored.

For the sake of simplicity, we decided to combine all nine variables into a single predictor by computing a simple sum of their a priori values, and use that as our a priori linear model. This is clearly an "improper" linear model (Dawes, 1979), as no effort—and no pretense—was made to select weights optimally. We did not seek optimal weights because, with a sample of 52, the results would not have been robust enough.

Table 2
Correlations Between the Predictions of Five Judges and Supervisor Ratings

	Supervisor ratings									
	Based on all available information					Based on the common core (n = 58)				
Judge	General evaluation	Level of job performance	Job compliance	Human relations	Average	General evaluation	Level of job performance	Job compliance	Human relations	Average
Graphologist A (n = 66)	0.24	0.31*	0.17	0.11	0.22	0.21	0.34*	0.18	0.07	0.21
Graphologist B (n = 72)	0.21	0.37*	0.26*	0.11	0.27*	0.21	0.42*	0.25	0.10	0.29*
Graphologist C (n = 64)	0.25*	0.07	0.33*	0.11	0.29*	0.21	0.05	0.39*	0.06	0.25*
Clinical psychologist (n = 75)	0.21	0.33*	−0.13	0.27*	0.24*	0.28*	0.34*	−0.08	0.42*	0.34*
Firm test battery (n = 79)	0.33*	0.28*	0.12	0.12	0.37*	0.33*	0.32*	0.13	0.07	0.34*

* Significant at the .05 level, two-tailed test.

To see how the validities of variables V1 through V9 compared with the validities of the five judges and the validity of the linear models, we computed the latter on the same subsample of 52. The results are presented in the lower part of Table 3. Note that the subjective variables (V6 through V9) predicted the two criteria better—albeit not significantly better—than any of the three graphologists (one exception is graphologist B's correlation with the average score, .29). Indeed, some single variables (e.g., "script's aesthetics," V7) had higher observed validity coefficients than any of the three graphologists.

Sadly, our a priori guess failed to predict even the correct direction for scoring marital status (V4) and vocational interests (V5). This apriori formula, though containing the two variables that were scored in opposition to the empirically appropriate direction—thus detracting from the model's validity—did better (though not significantly) than any of the four human judges (the graphologists and the clinician); deleting V4 and V5 from the formula resulted in higher correlations (again not significantly so) than obtained by the firm's test battery as well.

Table 3
Interjudge Reliabilities and Predictive Validities for Variables Extracted From 52 Scripts

		Criterion	
Variable	Interjudge reliability	General evaluation	Average rating
Education: VI	0.95*	0.20	0.21
Army rank: V2	1.00*	0.17	0.12
Arrival in Israel: V3	1.00*	0.17	0.12
Marital status: V4	0.95*	−0.11	−0.15
Vocational interests: V5	0.77*	−0.07	−0.03
Quality of essay: V6	0.50*	0.27	0.28*
Aesthetics of script: V7	0.53*	0.24	0.28*
Language errors: V8	0.53*	0.24	0.27
Overall impression: V9	0.20	0.22	0.25
Graphologist A		0.16	0.21
Graphologist B		0.21	0.29*
Graphologist C		0.19	0.08
Clinical psychologist		0.22	0.34*
Firm test battery		0.31*	0.34*
Sum of V1–V9		0.28*	0.30*
Sum of V1–V3 with V6–V9		0.35*	0.35*

* Significantly different from 0 at $p < .05$.

Discussion

The results of this study lead to the conclusion, shared with previous studies, that when graphologists base their judgments on spontaneously produced text, such as autobiographical sketches, they can achieve positive, if small, validities. However, when nongraphologists analyze the same data, they achieve similar validities. So does a naive and clearly nonoptimal linear model of the information in these texts.

In light of some recent literature that recommends the use of biographical information in personnel selection (e.g., Cascio, 1978), it is of some interest that such information, at least when extracted from freehand autobiographical sketches, had validities that were inferior to the "softer" properties of the written text. As shown in Table 3, nonbiographical properties of the writers, such as the quality of their writing in terms of grammar, aesthetics, or articulateness, produce validity coefficients comparable to those achieved by professional graphologists. A possible reason for the advantage of "script variables" over biographical variables is that the former are more directly related to ability and intelligence than the latter.

This does not account, however, for the superiority of the "script aesthetics" variable (V7), which—unlike writing quality and linguistic errors—does not seem related to ability or intelligence. We speculate that, because the aesthetic features of one's handwriting are, at least to some extent, under one's voluntary control, this variable might predict supervisor satisfaction to the extent that it reflects willingness to please, to "do well," to match up to some standard. This result is particularly striking in view of the low reliability of this variable. Presumably, if this reliability could be increased (e.g., by adding judges), its validity might grow higher than even the firm's test battery.

An important implication of these results is that even in studies that use standard scripts it is still worthwhile to use a control group of nongraphologists. The aesthetic features of a handwritten script may be a graphological variable, but our result suggests that it is the kind of variable that is intuitively accessible to nongraphologists as well. We caution against generalizing injudiciously from this finding: Personality characteristics such as honesty, leadership, and sense of humor, may bear no relationship to this variable, even if success on the job does.

Experiment 2

Method

Writers

Handwritten scripts were collected from 40 adult males, 33 of whom were Israeli born, and the rest had immigrated to Israel before the age of six, so that Hebrew was their major writing language. The writers who were chosen (on the basis of personal acquintance with the authors or nationwide renown) had all worked their entire career in one profession, had been working in it for at least 10 years, expressed great satisfaction in their own professional choice, and are considered successes by their colleagues and by those who benefit from their services. Indeed, most are noted representatives of no little repute in their professions. Thus, it can be safely stated, by any reasonable criterion, that they are suited to their professions. The forty writers included 9 mathematicians, 5 clinical psychologists, 5 philosophers, 5 artists (all painters), 4 middle- to top-level executives in the chemical engineering industries, 4 architects, 4 physicians (all surgeons), and 4 jurists.

Materials

All writers contributed the same handwritten materials, written with a standard pencil on a standard, unlined, sheet of paper according to detailed instructions we provided. The choice of materials was made in consultation with one of Israel's leading graphologists. The scripts, all in Hebrew, included: (a) a full page of text copied from a book by Ephraim Kishon (a noted Israeli humorist), (b) an adage that was copied out three times consecutively, (c) a well-known children's song that was written down from memory, (d) five simple arithmetic problems, that were first copied down and then solved, (e) a couple of rows of the Latin letters M and O intermittently, (f) three repetitions of the numbers 1 through 10, and (g) the same fictitious name, which the writers were asked to write down "the way you would sign it if it were your name."

The materials were solicited by mail, with the purpose of the study clearly disclosed. Close to 50% of those approached responded, and of those, all but a few (those who indicated in the short questionnaire attached that Hebrew was not their major or mother tongue), were ultimately used.

Graphologists

Five Israeli graphologists took part in this study, at least three of whom are quite famous (having written popular books or appeared on popular media shows featuring their skills). Over a dozen more were approached by us and asked to participate. Some declined right away, but others withdrew only after actually starting on the assignment or viewing the materials. Hence, among those approached, our group of five was self-selected.

Procedure

Each graphologist received the 40 scripts along with information regarding the age, handedness, country of birth of the writer, what major illnesses he had ever suffered from, and whether he wore glasses.

The graphologists were told that each writer belonged to one and only one of the eight professions but not to assume an equal number in each profession. They were asked to indicate which profession(s) the writer was best suited to. They were allowed to check more than one profession as suitable to each writer, to group several professions together if they found them hard to distinguish, to eliminate altogether professions that they considered unpredictable on the basis of handwriting, and to abstain from making any predictions about scripts they found unanalyzable.

Three graphologists, labeled A, B, and C, worked independently, whereas D and E consulted each other while analyzing the scripts.

Results

Validity

The manner in which this study was carried out precludes a presentation of the results in the customary terms of correlation coefficients. Relying upon the number of correct predictions made by each graphologist is also troublesome, because the graphologists used an unrestricted—and often different—number of guesses for each script (e.g., A gave 52 guesses for the 40 scripts, whereas C gave 92 guesses to 39 scripts, and up to 5 guesses to a single script). The way we evaluated the validity of the graphologists was to compare the number of correct guesses made by each graphologist with the number expected by a chance model (i.e., a model that assumes that guesses are made at random) based on his or her own base rate of guesses (i.e., the total actual number of guesses divided by the total possible number, which is eight times the number of judged scripts).

A graphologist's guess was defined as "correct" if the writer's true profession was included among the professions assigned to him by the graphologist; otherwise (unless the true profession happened to be one that the graphologist declared unpredictable, in which case that script was deleted from the analysis altogether) it was defined as "incorrect."

The chance model was based on a binomial distribution, conditioned by the number of guesses given to each script assuming independence between scripts. The statistical analysis compared the observed percentage of correct guesses made by each graphologist, P_o, with the percentage expected by the chance model, P_c, using a normal approximation to the distribution of the total number of correct responses, y. A measure Δ was defined, denoting the marginal increment to the base rate (i.e., the difference between a graphologist's P_o and his P_c. To overcome the differences in chance expectations between graphologists, we also computed $\Delta/(1 - P_c)$. All these measures appear in Table 4. The observed number of correct guesses always fell within the 95% range of expected correct guesses, hence was not significant at $p = .05$. One graphologist, E, came close to the criti-

Table 4
Comparison of Number of Correct Guesses by Five Graphologists With Number Expected by a Chance Model

Graphologist	No. of scripts	No. of guesses given	No. of possible guesses	% expected correct guesses: P_c	% observed correct guesses: P_o	No. of observed correct guesses: N_o	Marginal increase (Δ) in percent guesses	$\frac{\Delta}{1-P_c}$	95% expected range of correct guesses
A	40	52	320	16.25	17.50	7	1.25	.0149	1.93–11.07
B	40	69	320	21.56	27.50	11	5.94	.0757	3.53–13.72
C	39[a]	92	312	29.49	30.77	12	1.28	.0182	5.92–17.08
D	32[a]	44	222[b]	19.82	28.13	9	8.31	.1036	1.92–10.76
E	39[a]	73	297[b]	25.25	38.46	15	13.21	.1767	4.53–15.17

[a] When the writer's true profession was classified as "unknown" by a graphologist, that script was omitted from the analysis for that graphologist, resulting in number of scripts less than 40.
[b] When a graphologist didn't classify a certain profession as either suitable or unsuitable for some writer, the "number of possible guesses" was lower than 8 per script.

cal value. But note that had we considered the labeling of the writer's actual profession "unknown" as an error, rather than conservatively deleting that script from the sample, the proportion of correct guesses would have gone down (for C, D, and E).

Power

The present results do not support a claim of graphological validity in predicting professional suitability. Because this is based on the failure to reject a null hypothesis, the question of power may well be raised. Power is a function of size of effect, which in the present case is measured in terms of Δ. We computed statistical power for five values of Δ ranging from 0.1 to 0.3 in increments of .05. The computations were based on a normal approximation to the distribution of y. The results are displayed in Table 5.

It could be argued that Δ does not carry the same meaning for every graphologist, because they differed markedly in their base rate performance. Hence, we computed two additional power values for each graphologist, based on an increase of 50% and of 100% in their initial base rate probability of guessing correctly. These are also displayed in Table 5.

The results of the power computations indicate them to have been rather low for detecting small effects (i.e., an increment of 0.1, or an increase of 50%) but satisfactory for increments of 0.2 at least, or for a doubling of the base rate probability. We leave it to the reader to judge what is an "important" effect size in the present case. If a Δ of 0.1 is already such an effect, then our study needs to be replicated using more scripts, but it can stand on its own for the purpose of detecting larger effects.

Reliability

Besides looking at the validities of our graphologists, we also looked at their pairwise rates of agreement, which would indicate the extent to which they are employing a unified—whether valid or not—approach. A nonstandard measure of reliability was called for (as it was for the validities), because of the different base-rates characterizing each graphologist and because the graphologists were allowed to classify into more than one category. We settled on the difference between the observed number of agreements between a pair of graphologists, and the number of agreements expected on the basis of the chance model that takes their base rates into account. A script was scored as an "agreement" if both graphologists included its writer's true profession among their guesses, or both didn't. The number of agreements could, thus, assume any value between 0 and 8 per script. Clearly, its import depends on the total number of professions guessed for the script.

The observed number of guesses was compared with its expected value, and the percentage of cases where the former was

Table 5
Calculations of Statistical Power of the Tests Checking Marginal Increases in Probability of Correct Guesses

Graphologist	Power of statistical test for Δ = 0.10	Power of statistical test for Δ = 0.15	Power of statistical test for Δ = 0.20	Power of statistical test for Δ = 0.25	Power of statistical test for Δ = 0.30	Power of statistical test for increasing base rate by 50%	Power of statistical test for increasing base rate by 100%
A	.419	.687	.870	.959	.991	.315	.742
B	.355	.616	.824	.940	.983	.394	.870
C	.290	.535	.762	.910	.977	.520	.973
D	.319	.556	.763	.899	.966	.314	.755
E	.317	.569	.788	.922	.980	.448	.927

Table 6
Comparison of Extent of Agreement Between Pairs of Graphologists as Expected by Chance Model Versus as Actually Observed

Pair of graphologists	No. of common scripts	Average expected no. of agreements by chance model	Average observed no. of agreements	Average of observed minus expected agreements	% cases where observed exceeded expected no. of agreements	95% range of agreements expected by chance model[a]
A-B	40	5.5152	5.5250	0.0125	30.00	5.21–5.82
A-C	39	5.1154	5.3590	0.2436	48.72	4.74–5.49
A-D	32	5.7578	5.8438	0.0859	36.36	5.47–6.04
A-E	39	5.3974	5.3333	−0.0644	23.68	5.11–5.68
B-C	39	5.0064	5.4103	0.4035	64.10	4.81–5.19*
B-D	32	5.4375	5.5000	0.0625	36.36	5.19–5.68
B-E	39	5.1859	5.3077	0.1218	36.84	4.88–5.50
C-D	31	5.0500	5.4516	0.4016	61.90	4.67–5.43*
C-E	38	4.8616	5.1842	0.3224	54.05	4.49–5.23
D-E[b]	32	5.5703	6.9688	1.3984	86.36	5.05–6.09*

[a] In some cases the number of agreements expected by the chance model was exactly the same as that actually observed.
[b] This pair of graphologists worked in collaboration.
* The observed number of agreements fell outside of the 95% range expected by the chance model, so that it significantly exceeded the chance expectation, $p < .05$.

larger is shown in Table 6, along with a *t* test for matched groups that was performed for each pair of graphologists.

In 7 of 10 cases the extent of agreement between pairs of graphologists failed to reach significance. In addition, one of the three pairs whose agreement was significant worked by mutual consultation, and hence cannot be regarded as independent. Moreover, only in three pairs did the observed number of matches even exceed the expected number more than half the time.

Discussion

None of the graphologists who participated in Experiment 2 was able to predict a writer's profession from a standard handwritten script to a significant degree. Indeed, their mean probability of making a correct prediction only exceeded that of a chance prediction model by 0.06, and the largest probability increment was 0.13 (for E). Because five graphologists took part in the study, the probability that at least one of them would achieve a significant level of correct predictions purely by chance is approximately 0.25. Thus, even a higher rate of correct guesses by one graphologist out of the five would not have constituted sufficient evidence to establish even that graphologist's prediction ability without further replication and cross validation, and certainly our results fail to support a correspondence between graphological signs and professional suitability.

Even the low accuracies that we found may have been artificially inflated by extraneous clues. For example, one of our graphologists pointed out that architects often adopt a style of writing certain digits in the manner favored in schools of architecture (e.g., an 8 made of a small oval on a large one). A casual inspection of our scripts revealed that in at least one case an architect indeed used this special style, and was correctly identified by four of our graphologists. Such clues, however, merely reflect acquired habits, not personality characteristics.

After being confronted with their own poor performance, some of our graphologists protested that our criterion was not really a fair one, because one's actual occupation does not necessarily reflect one's professional tendencies. Although in general this may be true to some extent, we believe it is not a valid objection in the present study. If our writers were not suited to their occupation (and recall how they were selected), then we fail to understand what professional suitability is. We concede, of course, that our writers may have also been suited to other professions, but we allowed the graphologists to make that judgment without penalty.

General Discussion

Generalizability

We have often heard the objection that the fact that some graphologists may be incompetent does not invalidate the graphological enterprise itself. This may well be true, but until such a time as graphological analysis can be objectified, there is no way of testing the enterprise independently of its practitioners. Our studies join many previous ones that—though conducted by different investigators in different countries, using different settings, criteria, graphologists, texts, and so forth—gave rise to the same general picture. In the few cases where significant levels of accuracy were reported, the effects were no bigger than those we found, and significance was achieved largely on the strength of the sample size, rather than the effect size (e.g., Crumbaugh & Stockholm, 1977). Hence, we find ourselves compelled to conclude that it is graphology, rather than just our small sample of graphologists, that is invalid. This conclusion is based not only on empirical results but on a theoretical analysis as well. A detailed analysis appears in Bar-Hillel and Ben-Shakhar (1986). Here we shall only summarize some points of this analysis.

On the face of it, handwriting analysis looks like an excellent candidate for personality assessment. It seems to have all of the right characteristics from a substantive point of view: The analysis relies on a sample of self-generated and expressive individual behavior (Allport & Vernon, 1933). Handwriting is rich enough in features and attributes to afford it the requisite scope for expressing the richness of personalities. It is as unique as personalities, and—like them—exhibits both individual differences and shared structure. Handwriting is a stable characteristic of the individual (Fluckinger, Tripp, & Weinberg, 1961). It is more or less enduring over short times, and yet shows development over longer times.

Closer scrutiny of these features reveals them to be flawed.

1. Although it would not be surprising if it were found that sloppy handwriting characterized sloppy writers, stylized calligraphy indicated some artistic flair, and bold, energetic people had bold, energetic handwriting, there is no reason to believe that traits such as honesty, insight, leadership, responsibility, warmth, and promiscuity, find any kind of expression in graphological features. Some may have no somatic expression at all. Indeed, if a correspondence were to be empirically found between graphological features and such traits, it would be a major theoretical challenge to account for it.

2. There are not enough constraints in graphological analysis, and the very richness of handwriting can be its downfall. Unless the graphologist makes firm commitments to the nature of the correspondence between handwriting and personality, one can find ad hoc corroboration for any claim.

3. The a priori intuitions supporting graphology listed above operate on a much wider range of texts than those graphologists find acceptable. As graphologists practice their craft, it appears that from a graphological viewpoint, handwriting—rather than being a robust and stable form of expressive behavior—is actually extremely sensitive to extraneous influences, that have nothing to do with personality (e.g., whether the script is copied or not, or the paper lined or not).

4. It is noteworthy that most graphologists decline to predict the sex of the writer from handwriting, although even lay people can diagnose a writer's sex from handwriting correctly about 70% of the time (e.g., Goodenough, 1945). They explain this by insisting that handwriting only reveals psychological, rather than biological, gender (e.g., Crepieux-Jamin, 1926). Although common sense would agree that some women are masculine and some men are effeminate, it would be somewhat perverse to argue against the presumption that most women must be feminine and most men masculine. Could the graphologists simply be reluctant to predict so readily verifiable—or falsifiable—a variable?

The graphological enterprise also must face the difficulties attendant on the predicted variables, namely behavior and personality. Graphological analysis is an attempt to infer from how people behave in a single context what kind of people they really are. It relies on a supreme article of faith that the characteristics of such behavior, as they are expressed in handwriting features, are indicative of the personality as a whole, and therefore of the entire range of an individual's behavior. This, however, is a strongly holographic notion of personality, and flies in the face of much of the evidence in the field (e.g., Mischel, 1976; Nisbett & Ross, 1980). Although the person reading a graphological character analysis has a distinct sense that an integrated, whole personality has been put together, and that he or she now actually knows the person described, the sense of being now able to predict that person's behavior is not supported by the facts.

Why Does Graphology Appear to Work?

If both empirical validation studies and methodological considerations are so unflattering to graphology, why are its users and clients so satisfied with it and so often prepared to swear by it? It appears that though graphological predictions have no empirical validity, they have two other kinds of compelling "validities"—face validity and personal validity. Face validity refers to the fact that handwriting *appears* to have the right kind of properties for reflecting personality. Personal validation refers to the subjective feeling imparted by exposure to a graphological analysis that it is accurate and right on the button, that it managed to capture the true core of one's personality. Unfortunately, neither of these two types of validation can substitute for straightforward empirical validation. The latter particularly is very vulnerable to manipulation.

In a paper entitled "Cold reading: How to make strangers believe that you know all about them," Hyman (1977) listed many of the tricks and methods whereby "cold readers" (e.g., palmists, astrologers, and crystal ball gazers) practice their trade. One powerful trick is the notorious Barnum Effect (e.g., Forer, 1949; Snyder, 1974; Snyder & Shenkel, 1975). The power of this effect can be underestimated only by those who have not witnessed it in action.

Graphologists, however, enjoy advantages above and beyond those of the cold reader that help ensure them of satisfied clients. The firms and organizations who seek their services are seldom in a position to evaluate the reading given by the graphologists against the truth. In other words, a criterion may be completely unavailable for many reasons: Some candidates are simply rejected at the graphologist's recommendations, other predictions have no clear observable correlates, and so forth. Yet even here, clients can get a sense of personal validation, based simply on the fact that a proficient graphologist can give a reading that just sounds good. The character sketch may be rich or credible or familiar enough that it checks not with any specific piece of reality, merely with one's notion of what people are like. Graphologists are happy to step in where psychologists have left a void that begs to be filled—the prediction of such personality traits as honesty, reliability, trustworthiness, and so forth. Their clients are all too pleased to pass the responsibility for making such important judgments to people who claim to have the professional capability of making them.

In light of the present results, the sense of security imparted by passing the responsibility on to the graphologists hardly seems warranted. This is not to say, however, that graphological services should be dispensed with altogether. It should be recalled that with contaminated material—which is the kind they customarily work with—graphologists achieve validities not much lower than those achieved by entire batteries of psychological tests, and—it should be noted—at far less an investment in both time and money. Graphological predictions thus would seem to play a role akin to that played by placebos in medicine:

not completely ineffective, but for reasons others than those that make the real thing effective.

References

Allport, G. W., & Vernon, P. E. (1933). *Studies in expressive movement.* New York: The Macmillan Company.

Bar-Hillel, M., & Ben-Shakhar, G. (1986, in press). The a priori case against graphology: Methodological and conceptual issues. In B. Nevo (Ed.), *Handbook of scientific aspects of graphology.*

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross situational consistencies in behavior. *Psychological Review, 81,* 506–520.

Ben-Shakhar, G., Bar-Hillel, M., & Flug, A. (1986, in press). A validation study of graphology in personnel selection. In B. Nevo (Ed.), *Handbook of scientific aspects of graphology.* Illinois: C. C. Thomas.

Cascio, W. F. (1978). *Applied psychology in personnel management.* Virginia: Prentice Hall.

Crepieux-Jamin, J. (1926). *The psychology of the movements of handwriting.* (trans. and arranged by L. K. Given-Wilson). London: Routledge.

Crumbaugh, J. C., & Stockholm, E. (1977). Validation of graphoanalysis by "global" or "holistic" method. *Perceptual and Motor Skills, 44,* 403–410.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Fluckinger, F. A., Tripp, C. A., & Weinberg, G. H. (1961). A review of experimental research in graphology, 1933–1960. *Perceptual and Motor Skills, 12,* 67–90.

Forer, B. R. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology, 44,* 118–123.

Frederick, C. J. (1965). Some phenomena affecting handwriting analysis. *Perceptual and Motor Skills, 20,* 211–218.

Goodenough, F. L. (1945). Sex differences in judging the sex of handwriting. *Journal of Social Psychology, 22,* 61–68.

Hyman, R. (1977). "Cold reading": How to convince strangers that you know all about them. *The Zetetic,* p. 18–37.

Jansen, A. (1973). *Validation of graphological judgements: An experimental study.* The Hague, Netherlands: Mouton.

Levy, L. (1979). Handwriting and hiring. *Dun's Review, 113,* 72–79.

Mischel, W. (1976). *Introduction to personality.* New York: Holt, Rinehart and Winston.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Rafaeli, A., & Klimoski, R. J. (1983). Predicting sales success through handwriting analysis: An evaluation of the effects of training and handwriting sample context. *Journal of Applied Psychology, 68,* 212–217.

Snyder, C. R. (1974). Why horoscopes are true: The effects of specificity on acceptance of astrological interpretations. *Journal of Clinical Psychology, 30,* 577–580.

Snyder, C. R., & Shenkel, R. J. (1975). The P. T. Barnum effect. *Psychology Today, 8*(3), 52–54.

Received October 18, 1985
Revision received February 19, 1986 ■

Change in Distribution of APA Convention "Call for Programs"

In an effort to facilitate distribution of the APA "Call for Programs" for the annual convention, the "Call" for the 1987 convention will appear in the December issue of the *APA Monitor* instead of being a separate mailing to APA members. The 1987 convention will be in New York from August 28 to September 1. Deadline for submission of program and presentation proposals is January 20, 1987. Additional copies of the "Call" will be available from the APA Convention Office in December.