Journal of Behavioral Decision Making, Vol. 11, 235–249 (1998)

Seek and Ye Shall Find: Test Results Are What You Hypothesize They Are

GERSHON BEN-SHAKHAR, MAYA BAR-HILLEL*, YORAM BILU and GABY SHEFLER
The Hebrew University, Israel

ABSTRACT

Expert clinicians were given batteries of psychodiagnostic test results (Rorschach, TAT, Draw-A-Person, Bender-Gestalt, Wechsler) to analyze. For half, a battery came along with a suggestion that the person suffers from Borderline Personality disorder, and for half, that battery was accompanied by a suggestion that he suffers from Paranoid Personality disorder. In Study 1, the suggestion was made indirectly, through a background story that preceded the test results. In Study 2, the suggestion was made directly, by the instructions given. The experts saw in the tests what they hypothesized to be there. In particular, the target diagnoses were rated higher when they were hypothesized than when they were not. © 1998 John Wiley & Sons, Ltd.

Journal of Behavioral Decision Making, **11:** 235–249 (1998)

KEY WORDS: Baye's theorem; confirmation bias; cognitive confirmation; clinical judgment; hypothesis testing

Several years ago, the TV program *60 Minutes* did an exposé of polygraphers. It enlisted the help of a commercial firm in staging a mock crime. The firm called upon three polygraphers, whom we shall call A, B and C, to assist in discovering which of three suspected employees, whom we shall call X, Y and Z, was stealing from the firm. A was told that X was suspected, B was told that Y was suspected, and C was told that Z was suspected. No reasons were given for this suspicion. The three polygraphers, who were unaware of each other's existence, conducted a polygraph investigation of the three suspects (using the standard Control Question Technique), and handed in their conclusions. A found X to be 'guilty' (i.e. deceptive), B put the blame on Y, and C pointed to Z as the culprit.

A cynic might assume that the polygraphers were knowingly selecting as guilty the employee that had been selected by the firm as most likely to be the offender. Psychologists can be less cynical. They know that a polygraph investigation, like many data-gathering enterprises, can be conducted, even in good faith, in a manner that yields biased results. For example, if the tone of the interrogation is relatively more cold or hostile to some suspect, this suspect in turn may be relatively more uneasy,

* Correspondence to: Professor M. Bar-Hillel, Department of Psychology, The Hebrew University, Jerusalem 91905, Israel. E-mail: msmaya@pluto.mscc.huji.ac.il

Contract grant sponsor: F. A. Schonbrunn Research Foundation, Center for the Study of Rationality and Interactive Decision Making, Sturman Center for Human Development at the Hebrew University.

CCC 0894–3257/98/040235–15$17.50

Accepted 11 January 1998

which could result in a relatively more incriminating chart. This is what Snyder and Swann (1978) called 'the behavioral confirmation effect', whereby the 'perceiver's behaviors . . . channel the course of the interaction such that expectancy-confirming behaviors are elicited' (Darley and Gross, 1983, p. 20). In such a case, the unwitting bias is in the evidence-elicitation procedure, but not necessarily in its interpretation. However, evidence interpretation can also be biased, as in the following imaginary scenario.

A committee is considering two candidates for a single academic position. Professor X prefers Dr A to Dr B, and Professor Y prefers Dr B to Dr A. Each of them has prepared a statement in support of the candidate they prefer. Suppose that A's list of publications consists largely of joint publications, and they are almost all focused on a single phenomenon. B's list of publications consists largely of single-author publications, which appeared in a wide range of journals and cover many areas. It is not far-fetched to expect X to point out that A exhibits an ability to cooperate with others and work in a team which is lacking in the more idiosyncratic B, while B tends to spread out too thinly, lacking the ability to pursue a topic at depth and to occupy a well-defined scientific niche. Y, on the other hand, might point out that A has not established an ability to do independent research, which is clearly characteristic of B, and moreover, A is a narrow, one-issue person, lacking the breadth and intellectual diversity so clearly exhibited by B. Unlike the polygraphers, the committee members did not, in this case, elicit and observe different data from the two candidates, but perhaps interpreted the same two CVs differently, in accordance with the prior opinions these members held regarding the relative merits of the two candidates.

The committee example is of what Darley and Gross (1983) have called a 'cognitive confirmation effect', which occurs 'in the absence of any interaction between the perceiver and the target persons. In these cases, the perceivers simply selectively interpret, attribute, or recall aspects of the target person's actions in ways that are consistent with their expectations' (p. 20).[1]

Bias in evidence elicitation, and bias in evidence interpretation, are but two psychological mechanisms that contribute to a meta bias which has been called confirmation bias. A confirmation bias is the advantage that the hypotheses or beliefs which one holds have over alternative or competing hypotheses or beliefs, simply by virtue of the fact that they *are* the hypotheses or beliefs which one holds.

The present study is a study of the cognitive confirmation effect. Its novelty is twofold. First, unlike the typical participants in studies of the confirmation bias carried out by social psychologists and cognitive psychologists, the participants were not students doing a novel task but rather professionals, who were rendering a professionally familiar kind of judgment (but see Koehler, 1993). Second, the information at their disposal was a battery of five psychodiagnostic test results, which had been collected by professionals like themselves. Thus, although the target case being judged is a person, as it is in the social psychological studies, the information is not social in nature but rather strictly clinical, of the kind that psychodiagnosticians are trained to interpret with the help of manuals allegedly developed from extensive statistical evidence and psychodynamic theory.

In this study, all the information was available to the participants all the time, and they did not have to retrieve any of it from memory. They worked at home, unsupervised and unconstrained. In this respect, this study resembles Chapman and Chapman's classical series of studies on illusory correlation (1967, 1969, 1982). The subjects in those studies, too, were trained psychodiagnosticians, working with psychodiagnostic tests (Draw-A-Person and Rorschach). Their 1982 title 'Test results are what you think they are' inspired the title for the present paper. Whereas Chapman and Chapman focused on the malleability of test results to selective interpretation (and recall) guided by pre-existing associative links, in the present case, we focus on test results that are biased by the question one addresses to them.

[1] Of course, in order to be sure that A and B don't simply value different characteristics in CVs, a properly controlled study would have to be done, but the anecdote serves to illustrate our point.

Chapman and Chapman dealt with the effect of prior expectations about the *tests* on test interpretation, whereas we deal with the effect of prior expectations about the *testee* on test interpretation.

The present study shares with previous studies a characteristic that facilitates cognitive confirmation bias: the evidence producing it is rich and ambiguous. By 'ambiguous' one can refer either to a property of the individual items, such as vagueness, or to a property of the ensemble, such as inconsistency. For example, a single poem may provide ambiguous or unambiguous evidence about the writer's talents. A portfolio of poems could be ambiguous even if no single poem were, provided it contained both good and bad poems. Information which is sufficiently 'rich' — namely, as plentiful and varied as a candidate's folder, a suspect's polygraph chart, or a patient's test battery results — is almost bound to be ambiguous.

Klayman (1995) drew a distinction between confirmation bias, which he defines as 'an inclination to retain, or a disinclination to abandon, a currently favored hypothesis' (p. 386), and what he calls positive hypothesis testing, which is looking for 'features that are expected to be present if the hypothesis is true' (p. 399). The latter may lead to the former if people are unaware of, or make insufficient allowances for, the fact that the presence of such features, even in quantity, does not by itself support an hypothesis. The former refers to end results, the latter, to the process that produces them. Our study is less about the process than about its end result.

STUDY 1

Method

Participants

The participants in this study were all fully accredited Israeli psychodiagnosticians. They were initially recruited through a personal approach.[2] Later we relied on the mailing list of the Israeli Association of Psychologists, which in addition to names and addresses classifies members into categories such as 'clinical', 'educational', etc. We sent a letter to all Israeli psychodiagnosticians listed as such (several hundred at the time, and growing), asking them to volunteer as participants in 'a study of psychodiagnostics'. The exact purpose of the study was not disclosed, and anonymity (though not from us) was promised. Participants were promised full payment at going rates. Some letters were returned with 'address unknown' stamps (about 5%), and some people never answered at all (about 25%). Of those who wrote back, on prepaid postcards which we provided, about one third disqualified themselves, either for lack of expertise in analyzing projective tests, especially the Rorschach, or for lack of time.

We ended up with almost two hundred consenting psychodiagnosticians (almost a half of the initial mailing) and sent out the materials roughly in the order in which they responded. We didn't send materials to all of them, because we ran out of money, and because some positive responses came in after long delays, too late to be of use. Although those that responded positively are clearly not a random sample, they cover a wide range of professionals. We had men and women, young and old, from large cities and kibbutzim, Israeli trained and immigrants, people with or without PhDs[3], and people with private practice as well as people working in hospitals and public clinics. However, all we know about our participants (beyond the fact that they are registered expert psychodiagnosticians, and felt confident enough of their diagnostic skills to be willing to participate in our study) is what can be derived from names and addresses. We did not ask them for any personal or professional information about themselves.

[2] Israel being a small country, which up to the 1960s had only one psychology department, many of the Israeli-trained psychologists know each other, and so some of the participants were colleagues and acquaintances.
[3] A PhD is not a requirement for clinical psychologists in Israel.

Instruments

Clinical test materials were drawn from real clinical files, from the records of a Jerusalem mental hospital. These consisted of Rorschach, TAT, Draw-A-Person, Bender-Gestalt and Wechsler profile results. We constructed two batteries of test results, labelled I and II, mixing real tests from three different patients. This was intended to produce an essentially non-diagnostic battery. Of course, all identifying information was removed from the test materials.

Two brief fictitious life histories were written. They were meant to be suggestive of either a paranoid personality (PP) disorder or a borderline personality (BP) disorder.

Design

Four experimental groups were constructed. Each expert evaluated two cases. One group received only the two brief life histories, PP and BP. One group received only the two test batteries, I and II. Two groups received two complete files, each consisting of both a life history and a test battery. One of these groups received the test batteries coupled with the life histories one way (I + PP; II + BP) and the other received the test batteries coupled with the life histories the other way (II + PP; I + BP).

The experts were assigned into the last three groups at random. However, the first group (life history only) consisted largely of participants who expressed a willingness to participate, but were either unfamiliar with the Rorschach or pleaded they did not have the time required for analyzing full test batteries, so that this group was not formed at random. This first group required but a short time for each case, whereas the other three groups required several hours of work. The first group participated on an unpaid voluntary basis, whereas the other three groups were paid real wages at a fixed rate of 120 NS an hour, plus 17% VAT, for four hours of diagnostic work. This brought fees to close to $200 per participant (at then current exchange rates).

Task

The experts gave two kinds of responses. First, the three groups who received test materials were instructed as follows: 'Please diagnose two cases on the basis of psychodiagnostic materials: [listed]. The end product should be a diagnostic report of up to one page. It is very important that you record the findings and their diagnostic interpretations. The more the diagnostic conclusions are based on the test materials, the better'.

Second, all the experts received a list of eight diagnostic categories, in the following order: schizophrenic paranoia; obsessive–compulsive neurosis; borderline personality disorder; hysteria; paranoid personality disorder; bipolar disorder; narcissistic personality disorder; schizophrenia simplex. They were asked to write down a number between 1 and 10 next to each diagnostic category, according to the degree to which it was plausible on the basis of the test battery (1 — not at all; 10 — perfectly). The group who received a story only, judged the plausibility of these diagnoses in light of the story.[4]

[4] Shefler, a licensed practising clinician and an expert in psychodiagnosis, notes:

> Diagnosis based on Psychodiagnostic test results only is not the standard or most frequent form of diagnosis in clinical practice, but it is certainly routine and commonly done. For example, the Israeli Army uses this form for screening large numbers of candidates for sensitive positions. Although most of these tests precede the introduction of the Borderline Personality Disorder as a diagnostic category in the DSM, there is a vast literature that spells out direct and clear signs for this disorder. Its main diagnostic features, which are instability and the existence of primitive defense mechanisms (e.g., projection and splitting), can be identified in the psychological tests comprising our batteries (see, e.g., Kwaker, J. S., Lerner, H. D., Lerner, P. M. & Sugarman, A. (eds) *Borderline Phenomena and the Rorschach Test*. 1985, NY: International Universities Press). The situation is slightly harder for Paranoid Personality Disorder, since many of its psychodiagnostic aspects resemble in nature, if not in intensity, those of normal people. However, clinical signs such as suspiciousness and extreme caution appear primarily in the projective techniques — Rorschach, Draw-A-Person, and TAT (see, e.g., Schafer, R. *The Clinical Application of Psychological Tests: Diagnostic Summaries and Case Studies*. 1970, 13th printing. NY: International Universities Press).

Procedure

Psychodiagnosticians who volunteered in writing to participate in the study received an envelope in the mail, containing all the materials needed for their experimental group, as well as the two preprepared diagnostic rating forms described above, and written instructions. They worked unsupervised at home, at their own pace, for the preagreed fixed fee. Some took weeks to send in the results, others took months, and required prompting on our part.

Results

The average rating for each of the eight diagnostic categories was computed across participants. These means, with the corresponding standard deviations, are displayed in Exhibit 1. In addition, the exhibit shows the number of experts who gave that diagnosis their highest rating (these numbers sometimes sum to more than the total number of experts, because for some experts two or more diagnoses tied for the highest rating). The two target categories, Paranoid Personality Disorder and Borderline Personality Disorder, are displayed in the first two columns.

The first two rows in the exhibit, belonging to a single group of psychodiagnosticians, provide a manipulation check on the two stories we wrote. It is apparent that one story is highly suggestive of a

Exhibit 1. Mean ratings and SDs for each of eight diagnostic categories, and number of experts who gave the diagnosis their maximal rating[a]

	Target categories		Non-target categories						
	Paranoid pers.	Border. pers.	Hysteria	Paranoid schiz.	BiPolar	Narciss. pers.	Schiz. simplex	Compul. neurosis	N
1. PP alone	**8.05**	1.95	1.79	3.22	1.63	3.11	1.53	3.94	19
	2.20	1.43	1.03	2.31	1.16	1.97	1.39	2.73	
	18	0	0	1	0	0	0	1	
1. BP alone	1.95	**8.21**	2.17	2.68	3.50	5.17	2.37	1.53	19
	1.84	1.75	1.79	2.38	2.20	2.87	1.92	1.84	
	0	16	0	0	0	2	2	1	
2. I alone	4.71	5.33	2.79	5.63	**6.08**	3.26	2.13	1.63	24
	2.18	2.57	1.91	3.08	3.03	1.57	1.48	1.13	
	1	4	1	9	10	0	1	0	
2. II alone	3.52	**6.87**	4.57	2.43	3.04	6.41	1.70	3.00	23[b]
	2.56	2.14	2.02	1.53	2.20	2.50	1.61	2.43	
	1	11	2	0	1	8	0	2	
3. I + PP	**7.56**	3.53	2.25	5.21	2.71	3.81	1.72	3.06	19
	2.38	2.03	2.14	2.99	2.14	2.81	1.45	2.29	
	12	0	0	6	1	2	0	0	
3. II + BP	2.87	**7.89**	3.00	2.87	3.00	4.83	2.87	2.06	18[b]
	1.88	2.17	2.00	2.92	2.50	2.66	2.33	1.34	
	0	14	0	2	1	1	0	0	
4. II + PP	**7.38**	4.63	2.13	5.38	3.00	2.63	2.06	3.13	16
	2.99	2.55	2.13	2.42	2.92	1.93	2.11	2.58	
	9	1	1	1	2	1	0	1	
4. I + BP	2.94	**9.19**	2.38	3.88	3.19	5.19	1.81	1.56	16
	2.38	1.72	2.00	2.53	2.29	2.74	2.26	1.26	
	0	14	0	1	0	1	1	0	

[a]The sum of these numbers exceeded the number of experts when experts gave a maximal rating to more than one category
[b]One participant sent back an evaluation of only one of the two cases.

paranoid personality disorder, and another of a borderline personality disorder, as intended. For PP alone, the paranoid personality disorder diagnosis received a mean rating of 8.05; for BP alone, the borderline personality disorder diagnosis received a mean rating of 8.21. Nothing else came even close: the second highest mean rating was less than 4.

The next two rows, again belonging to a single group of psychodiagnosticians, are a manipulation check on our psychodiagnostic test batteries — or, if you will, a control. Ideally, we would have wanted a totally non-diagnostic battery. The batteries we actually obtained were not quite that, but neither had a single dominant diagnostic category on which most experts agreed. In battery I, nine and ten experts, respectively, favored Paranoid Schizophrenia and Bipolar disorder, and in battery II, eleven and eight experts, respectively, favored Borderline Personality disorder and Narcissistic Personality disorder. No diagnostic category received a mean rating higher than 7, and five received mean ratings between 5 and 7. This group provides a base line against which the other two experimental groups can be compared. If they were obeying our instructions to analyze the test batteries alone, there should be little difference between these three groups.

Although we were hoping that the test batteries wouldn't elicit a high mean rating for any of the eight diagnostic categories, we were curious whether the low mean ratings we obtained resulted from low individual ratings, or from lack of consensus. Low individual ratings mean that judges don't regard any diagnosis as highly plausible on the basis of the tests, whereas lack of consensus simply means that individual judges don't agree on which are the highly plausible diagnoses. To see which was the case, we analyzed individual ratings for the test batteries, and compared them with individual ratings for the stories — both of which yielded just one diagnosis with a high mean rating. First, we looked at the frequency of the two highest ratings for plausibility, 9 and 10. These ratings of 9 or 10 constituted 9% of all (all = number of judges × number of diagnostic categories) ratings given to batteries alone, and 8% of all ratings given to stories alone. Next, we considered number of judges who used a 9 or 10 rating at least once. This happened to 60% of the judges who were given the stories alone, and 53% of the judges who were given the batteries alone. Finally, we looked at the average individual rating (i.e. the average of all the numbers in our raw data). It was 3.92 for batteries, 3.02 for stories. By and large, it seems that judges were as willing to use high ratings in either group, so the fact that no single category received a high rating in the battery-only group does not reflect more reluctance to give confident diagnoses on the basis of the batteries than on the basis of the stories.

The next four rows of Exhibit 1 are pertinent to our research hypothesis. I + PP and II + BP belong to the same group of psychodiagnosticians, and likewise II + PP and I + BP both belong to another group of psychodiagnosticians. Notice the striking similarity between I + PP and II + PP, obtained from different experts, and likewise the similarity between I + BP and II + BP, also obtained from different experts. Furthermore, I + PP and II + PP are both similar to PP alone, and I + BP and II + BP are both similar to BP alone. These impressions will be quantified systematically below.

Inspection of Exhibit 1 reveals that fairly large mean values were obtained for each of the two target diagnoses whenever there was background information which was compatible with it (7.56, 7.89, 7.38 and 9.91). In contrast, the highest rating for a non-target diagnostic category in any group was 6.41 (Narcissistic Personality in the II alone group). In almost all other cases the means were much smaller than that.

Two types of analyses were carried out on the data in Exhibit 1 — one on the target categories (first two columns) alone and one on all eight categories. The analysis done on the target categories alone consisted of *t*-tests (for independent samples) conducted over the mean difference between the mean ratings they received from the full information (i.e. battery following background) groups versus under the partial information groups (background alone, or battery alone). These *t*-tests are shown in Exhibit 2. The analysis done on all eight diagnostic categories was correlational. For each pair of conditions (e.g. I alone with I + PP), a Pearson correlation was computed across the profiles of the

Exhibit 2. *t*-Tests between diagnoses made on the basis of full information versus diagnoses made on the basis of partial information

	Partial information	
Full information	Battery alone	Background alone
Battery I + Paranoid background story	$t_{40df} = 4.03$[a]	$t_{35df} = -0.66$
Battery II + Paranoid background story	$t_{37df} = 4.32$[a]	$t_{33df} = -0.77$
Battery I + Borderline background story	$t_{38df} = 5.26$[a]	$t_{33df} = 1.66$
Battery II + Borderline background story	$t_{39df} = 1.51$	$t_{35df} = -0.50$

[a]Significant at the 0.05 level.

eight mean ratings given to the eight diagnostic categories. These correlations are displayed in Exhibit 3. In the main diagonal are the average interjudge reliabilities, obtained separately for each condition.

Exhibit 2 shows that when comparing a test battery only condition (I alone or II alone) with its two full-information conditions (i.e. that test battery following one story or the other), the difference between the mean ratings of the target diagnosis (i.e. PP when the background story was Paranoid, or BP when it was Borderline) is significant in three of the comparisons, and falls just short of significance in the fourth (which is II versus II + BP). On the other hand, none of the four comparisons between a full information condition and its background component alone produce a significant difference on the target diagnoses.

The correlations displayed in Exhibit 3 form three groups:

(1) Shared background information, preceding either different test batteries, or none. The six correlations in this group range between 0.80 and 0.98, with a median at 0.96.
(2) Shared test battery, following either different background stories, or none. The six correlations in this group range between −0.01 and 0.83, with a median at 0.29.
(3) No common component. The 16 correlations in this group range between −0.25 and 0.80, with a median at 0.04.

A comparison of groups 1 and 2 shows that in 35 out of the $6 \times 6 = 36$ possible pairwise comparisons, the correlation from group 1 exceeded that from group 2 (statistically significant, using the Mann–Whitney test). Moreover, when comparing correlations between conditions with a shared story to correlations between conditions with a shared battery, the first was significantly larger than the second in all comparisons (t_{5df} for correlated samples, see McNemar, 1955). In addition, in 67 out of 96

Exhibit 3. Pearson correlations between eightfold diagnostic profiles made on the basis of different sets of information. The main diagonal (italicized) gives the average interjudge reliability for that condition. Group numbers correspond to Exhibit 1

	PP	BP	I	II	I + PP	II + PP	I + BP	II + BP
1. PP	*0.73*							
1. BP	−0.33	*0.57*						
2. I	0.06	0.41	*0.37*					
2. II	−0.08	0.80[a]	0.14	*0.42*				
3. I + PP	0.89[a]	−0.08	0.44	−0.04	*0.41*			
4. II + PP	0.80[a]	0.02	0.55	−0.01	0.94[a]	*0.37*		
4. I + BP	−0.17	0.97[a]	0.51	0.80[a]	0.13	0.24	*0.58*	
3. II + BP	−0.25	0.98[a]	0.34	0.83[a]	−0.03	0.09	0.97[a]	*0.34*

[a]With 6 degrees of freedom, correlations in excess of 0.7 are significant at the 0.05 level.

possible comparisons, the correlations in group 2 exceeded those of group 3, but this is not statistically significant (Mann–Whitney test).

This analysis provides further support for our hypothesis. According to a weak version of our hypothesis, positive correlations were expected between the test battery preceded by a background story and that story alone, in spite of the instructions to only interpret the battery, and judge only the battery's diagnostic implications. But the correlation analysis supports an even stronger hypothesis, according to which the similarity between the test battery following a background story and that story alone is even larger than between the test battery following a background story and that battery alone: in all four cases, the correlation of the full information with its story component is significantly higher than with its battery component.

Without the hypothesized influence of the story on the battery interpretation, the above-mentioned 0.97 and 0.94 coefficients — between full-information conditions with a shared story but different test batteries — should have resembled the correlation between the diagnostic profiles of the two test batteries alone (which happened to be 0.14); but in fact, they are significantly higher (z-test for independent samples equals 3.09 and 2.53, respectively). Furthermore, without the hypothesized influence of the story on the battery interpretation, the 0.13 and 0.09 coefficients — between full-information conditions with a shared test battery but different stories — should have been about as high as the interjudge reliabilities for those batteries (see main diagonal of Exhibit 3), but in fact they were not significantly larger than zero.

Interjudge reliabilities were estimated as follows. Within every experimental condition, we computed the Pearson correlation between the eightfold diagnostic ratings profile of every pair of experts. The mean of these correlations was taken to be the interjudge reliability of that condition. The main diagonal of Exhibit 3 shows these interjudge reliabilities. They range from 0.34 (II + BP) to 0.73 (PP). Note that even the lowest interjudge reliability is considerably higher than the 0.13 and 0.09 reported above — and even the highest is lower than standard benchmark interjudgment reliability requirements in psychological testing (which are upwards of 0.85). We cannot do a statistical significance test for this difference, however, because our interjudge reliability measure is a mean of correlations between individual ratings, whereas the 0.13 and 0.09 figures are the correlations between mean ratings, so they are not really comparable. The apparent difference is thereby probably an underestimate of the real difference, because correlations between mean ratings are expected to be larger than between individual ratings.

Correlation coefficients may not be the best way to characterize interjudge reliability in a task like the present one, because even if two judges agree on the relative plausibility of the diagnostic categories, there is a big difference between giving high ratings (say, 8 and above) or low ratings (say, 3 and below) to the most likely diagnosis, since only the latter acknowledges the objective weakness or problematics of our batteries — or perhaps even of all projective tests. It might be enlightening, therefore, to simply consider the variance among the psychodiagnosticians who were in the same experimental condition (who numbered between 16 and 24 per group).

Exhibit 1 gives the standard deviations of the ratings that our experts gave each diagnosis when they all received identical information. Note that in each of the eight conditions, at least one diagnostic category had an SD of 2.6 or more. For comparative purposes, we note that 2.6 is the SD of a uniform distribution between 1 and 10! Our raw data also show that in every condition, most of the diagnostic categories elicited ratings ranging from 1 to 10. Whereas not all the experts disagreed this much all of the time, some of the experts certainly disagreed rather alarmingly with each other some of the time.

We had hoped to report analyses of the one-page free-style diagnostic reports as well. Primarily, we were interested in the question whether clinicians who were interpreting a given battery with a different background in mind were focusing on different features of the battery test results, or were interpreting the same features differently. Unfortunately, we discovered that the reports do now allow this kind of

analysis. Although the reports often made reference to the test materials, as we asked, this was typically done merely by saying: 'see Bender test', 'as indicated by the TAT', etc.

Discussion

The results of Study 1 are consistent with our hypothesis: background information provided to psychodiagnostic experts had an influence on their subsequent diagnostic evaluation of test materials. Indeed, our correlational analysis shows that it had a whopping influence; much greater, in fact, than the influence of the information contained in the batteries themselves. The ratings assigned to a target diagnostic category when participants were given a background story which suggested it, were highly similar regardless of which test battery, if any, was combined with it. This, although the task was to say what diagnosis was implied by the battery (when there was one), not by the story.

The correlation analysis indicates that the background story actually affected the entire diagnostic profile, and not just the target diagnosis. The correlations between the profiles given on the basis of the background stories alone and those given on the basis of batteries that were preceded by these stories were impressively high. Furthermore, while the ratings given on the basis of the two test batteries alone were hardly correlated at all, the assessments made on the basis of these selfsame test batteries were almost perfectly correlated when the two batteries followed an identical background story.

Although these results are totally compatible with our hypothesis, there is an alternative account for them. It could be argued that in spite of being instructed to evaluate the diagnostic implications of the batteries only, our psychodiagnosticians actually integrated the information they derived from the batteries with what they derived from the background information. If so — the argument continues — two groups that are integrating the same test battery with different background stories (e.g. I + PP and I + BP, or II + PP and II + BP) could well end up with different final ratings.

This account raises a new question: why were final ratings so much closer to those based on stories alone than to those based on tests alone? Could this reflect the experts' feeling that the background information is simply more compelling than the psychodiagnostic test results? After all, such test batteries have notoriously low validities, and the ones used here were even patched up from more than one patient. However, our analysis of the individual ratings reported in the Results section belies this possibility. Recall that mean ratings, as well as proportion of extreme ratings, given by the story-only group were no higher than those given by the batteries-only group. So even if the experts were integrating all the evidence at their disposal, it is unclear *prima facie* why they would weigh the background stories more than the test batteries.

Our research hypothesis, however, provides an explanation for the greater weight put on the background story, even under the interpretation that experts are explicitly integrating all the evidence they were given. Suppose the experts first interpret the test results in a manner influenced by the background story — as our research hypothesis suggests. They then integrate their analysis of the test results with their clinical reading of the background story — as the alternative account presently under consideration suggests. This gives the background story 'double counting' in the integrated rating: once indirectly through its influence on the test interpretation and once directly, thereby accounting for why observed mean ratings of the full information groups were closer to those based on only the respective story than to those based on only the respective test component. So there is a bias in favor of the diagnostic implication of the story even under the alternative account of integration.

Comparison of some of the correlations displayed in Exhibit 3 considerably strengthens this conclusion. The correlations between conditions with shared story (namely between I + PP and II + PP and between I + BP and II + BP) were 0.94 and 0.97, respectively. That they exceeded those between the conditions without any shared component (namely between I + PP and II + BP, −0.03, and between II + PP and I + BP, 0.24), shows that the story influenced the final diagnoses. That they

also exceeded those between conditions with shared battery (namely between I + PP and I + BP, 0.13, and between II + PP and II + BP, 0.09), shows that the story's influence was stronger than the test battery's influence (all differences significant, t_{5df} for correlated samples, see McNemar, 1955). Most telling, though the fact that neither the correlation between I + PP and I + BP, 0.13, nor between II + PP and II + BP, 0.09, were larger than either −0.03 or 0.24 shows that the test battery's influence on the final diagnoses was virtually nil. Thus, it seems that even if the ratings we received were based on the totality of evidence presented, they were influenced only by the story, and not by the battery test results.

It is easy to fall into the misunderstanding that our instructions to evaluate only the diagnostic implications of the test batteries, rather than the entire body of evidence, is tantamount to asking that the background stories be ignored or forgotten. That, however, is not the case. To borrow a convenient notation from another area (but without claiming any equivalence): let A and B be two events from a probability space, and suppose one knows that A, and is asked to give the probability that B. Formally, the required probability is P(B/A). It is not (generally) the same as P(B&A). But neither is it the same as P(B). Giving A and asking for an evaluation of B is not to be confused either with asking that A be ignored or with asking that A&B be evaluated. Insofar as we gave judges background stories and instructed them to evaluate test batteries, we did not ask them to do that which we realize to be both difficult psychologically and inferior normatively — ignore given evidence. But insofar as they evaluated both, rather than evaluating the one while keeping the other in mind, they violated the instructions. Either way, they exhibit a confirmation bias.

Study 2 was designed specifically to make it difficult, if not impossible, to do anything but interpret a psychodiagnostic test battery, by the simple expedient of not giving the experts any information but the psychodiagnostic test battery.

STUDY 2

The major difference between Study 1 and Study 2 is that in Study 2, almost no background information was provided, so of course, there was naught to be used in combination with the test battery results. There were two patients, Carlos and Oren. The manner in which a hypothesis was suggested for the experts to test was by a single sentence, prefacing the response sheet, as in the following:

> Carlos (or Oren) is 41 (or 17) years old. As a boy, he immigrated to Israel from Argentina (or he is Israeli-born). Following an hypothesis that he is suffering from a borderline personality disorder (or paranoid personality disorder), he underwent a series of psychodiagnostic tests, whose results are attached. We refrain deliberately from giving you any further information about Carlos (Oren), because we are interested only in the diagnostic implications of these tests. Is the hypothesis supported by the tests? We thank you for interpreting them for us.

I believe that the tests point to a diagnosis of _____.

In addition, we ask you to rank the plausibility of each of the following diagnoses. The ranking should be done on a scale from 1 to 10, where 1 = no plausibility, and 10 = very strong plausibility. (Please do not use 0).

Clearly, we had gone all the way from (possibly) strong prior evidence to no prior evidence. We intended the offered hypothesis just to guide, via the question frame, the search for diagnostic evidence in the tests. This manipulation seemed so weak, that some of us were not sure it could uphold the hypothesis. In addition, the instructions were explicit about wanting no outside influences on the test interpretation.

Exhibit 4. Mean ratings and SDs for each of eight diagnostic categories, and number of experts who gave the diagnosis their maximal rating.[a] The score applies to the individual named in the Task Condition

	Target categories		Non-target categories						
Condition	Paranoid pers.	Border. pers.	Hysteria	Paranoid schiz.	BiPolar	Narciss. pers.	Schiz. simplex	Compul. neurosis	N
1. Oren B	2.89	**5.67**	2.56	1.44	2.22	4.44	2.33	3.67	9
	2	4	0	0	0	2	1	1	
1. Carlos P	**5.78**	5.00	2.78	2.33	1.44	3.11	1.78	2.44	9
	5	3	0	0	0	0	1	1	
2. Carlos B	3.11	**8.11**	3.89	2.44	2.44	4.89	2.78	2.44	9
	0	7	0	0	0	2	0	1	
2. Oren P	5.44	**7.00**	2.89	3.00	2.22	3.78	2.22	3.11	9
	4	5	1	0	0	0	0	0	

[a]The sum of these numbers exceeded the number of experts when experts gave a maximal rating to more than one category.

Results

As in Study 1, the average rating for each of the eight diagnostic categories was computed across participants. These means, with the corresponding standard deviations, are displayed in Exhibit 4 (which is formatted like Exhibit 1). In addition, the exhibit shows the number of experts who gave that diagnosis their highest rating (these numbers sometimes sum to more than the total number of experts, because some experts gave a highest rating to more than one diagnosis). The two target categories, Paranoid Personality Disorder and Borderline Personality Disorder, are displayed in the first two columns.

In this study, we had to analyze the results differently from those in Study 1 because we no longer had a base line of diagnoses, based on a background story or a test battery only, to compare with. We could only test here for the weak version of our hypothesis, namely, that interpreting a battery under a given hypothesis would raise the judged plausibility of the corresponding diagnosis relative to what that diagnosis receives when it does not correspond to the initial hypothesis.

Each expert was given a single score, D, defined as: $D = (B_h + P_h) - (B_{nh} + P_{nh})$, where B_h and P_h are the ratings the expert gave the Borderline and Paranoid diagnoses, respectively, when they were hypothesized, while B_{nh} and P_{nh} are the ratings the expert gave these diagnoses when they were not hypothesized. Take, for example, one of our participants, who was in the Oren B–Carlos P condition. This expert gave Oren a rating of 6 for the Borderline diagnosis and 1 for the Paranoid diagnosis, and Carlos a rating of 2 for Borderline and 6 for Paranoid. The expert's D score is therefore $(6 + 6) - (2 + 1) = 9$. If our hypothesis is correct, then D scores should be positive. For 12 of our 18 experts, it was. In addition, the mean D score was a positive 3.5. Against a null hypothesis that D would not be positive, a one tailed *t*-test showed D to be statistically significant ($t_{17} = 2.44$, $p = 0.05$).

Our results also allowed us to make four separate comparisons, as follows. For each of the two target diagnoses, we compared the mean rating it got when it was compatible with the question frame (e.g. Borderline for Oren B) versus the mean rating it got when it was not (e.g. Borderline for Oren P). This can be done for both batteries, generating four possible comparisons. Of the four (see the first two columns of Exhibit 4), three were in the predicted direction, and moreover, were statistically significant ($8.11 > 5.00$, $t_{16} = 2.55$; $5.78 > 3.11$, $t_{16} = 2.26$; $5.44 > 2.89$, $t_{16} = 2.27$, *t*-test for independent samples), whereas one was in the opposite direction ($5.67 < 7.00$, not statistically significant). The effect size of the latter comparison was 0.48 (considered a medium effect by Cohen, 1988), whereas the effect sizes for the three other comparisons were 1.2, 1.06, and 1.07, respectively, which are considered large effects.

Exhibit 5. Pearson correlations between eightfold diagnostic profiles made on the basis of different sets of information. The main diagonal (italicized) gives the average interjudge reliability for that condition. Group numbers correspond to Exhibit 4

	Carlos + BP	Carlos + PP	Oren + BP	Oren + PP
1. Carlos + BP	*0.26*			
2. Carlos + PP	0.57	*0.44*		
3. Oren + BP	0.86[a]	0.54	*0.22*	
4. Oren + PP	0.79[a]	0.92[a]	0.74[a]	*0.36*

[a]With 6 degrees of freedom, correlations in excess of 0.7 are significant at the 0.05 level.

In addition to this analysis, we also did a correlational analysis, similar to the one reported for Study 1 (see Exhibit 5, formatted like Exhibit 3). However, unlike in Study 1, in this study, rather high correlations are found between the eightfold diagnostic profiles even when the assessment conditions had nothing in common (namely, they shared neither test batteries *nor* task frames). For example, the correlation between the mean ratings made on the basis of Carlos' battery under 'check for Borderline', and those made on the basis of Oren's battery under 'check for Paranoid', is 0.79, and when the frames and the batteries are switched over, the correlation is 0.54. These unexpected correlations are not problematic in themselves, since they might simply reflect some similarity between the two genuine patients (remember: we didn't tamper with the batteries in this study!), but it makes the correlational analysis somewhat problematic: a ceiling effect makes it difficult to obtain a statistically significant effect of the experimental manipulation.

Nevertheless, it should be noted that the two correlations between the ratings made for the same test battery with different frames were 0.57 for Carlos's battery and 0.74 for Oren's battery (comparable to the 0.54 and 0.79 correlations reported above), whereas the correlations between the ratings made with the same frame for two different batteries were 0.86 for the Borderline frame and 0.92 for the Paranoid frame. Although the higher of the first pair of correlations, 0.74, is lower than the lower of the second pair of correlations, 0.86, the only significant difference between the pairs is that between 0.57 in the first pair, and 0.92 in the second. As in Exhibit 3, the main diagonal of Exhibit 5 shows the mean pairwise interjudge reliabilities in each of the four conditions.

Discussion

The results of Study 2 many not be as striking as those of Study 1, but then, a very minimal manipulation was used. Yet here, too, there is little if any evidence that some objective diagnostic implications of the batteries were determining the diagnoses (rather than the question frame). But for one case, where the suggested diagnosis was favored by seven of the nine experts (Carlos B), there was little consensus among the experts (see Exhibit 4), except insofar as rare diagnoses, such as Paranoid schizophrenia and Schizophrenia simplex, were rarely given.

Recall that in Study 2 we used integral batteries, whereas in Study 1 the batteries were a composite of several patients. What effect, if any, did this have upon the interjudge reliabilities? In other words, can the low reliabilities of Study 1 be attributed to the fact that the batteries were not integral? The interjudge reliabilities (see main diagonal of Exhibit 5) allow us, by comparison with those of Study 1 (see main diagonal of Exhibit 2), to answer this question. The relevant reliabilities to look at in Study 1 are clearly those for battery I and II without stories, which were 0.37 and 0.42, respectively. Even though in Study 2 the interjudge reliabilities were computed for batteries that shared a question frame, they were no higher than this — and in two cases, lower. The same conclusion was reached from a comparison of the SDs of the ratings given to the non-target categories in the two studies. Exhibit 4

shows little consensus among our experts on the appropriate diagnosis. Among the six non-target diagnostic categories, we count six SDs of 2.6 and up (out of a possible $6 \times 4 = 24$), as compared with ten such in Exhibit 1 (out of a possible $6 \times 8 = 48$).

We have heard some argue that the very words 'Following an hypothesis that he is suffering from ...' provide some kind of evidence in favor of that hypothesis, even though nothing is said about who hypothesized it, on what basis it was hypothesized, etc. That is like saying that a criminal court judge can believe the defendant to be guilty even before the first piece of evidence is presented, on the (valid) premise that there would not have been a trial otherwise. We find it as unacceptable for a psychodiagnostician to rely on an anonymous, unsupported, and vague 'hypothesis' as for one of *60 Minutes'* hapless polygraphers to have relied on the hiring firm's suspicions. Such practice and attitudes undermine the entire rationale for testing, and we hope few (if any) of our experts reasoned in this unprofessional way.

GENERAL DISCUSSION

The present study, both in its conception and in its results, extends an earlier paper by Elaad, Ginton and Ben-Shakhar (1994). In that study, expert polygraph examiners were given inconclusive polygraph charts for interpretation and scoring. They were told in advance whether each examinee had subsequently confessed (i.e. was Guilty), or had been exonerated by another's confession (i.e. was Not-Guilty). This information was reversed for half of the polygraphers. Examiners who believed they were scoring the chart of a Guilty examinee gave scores that were more indicative of a Guilty finding than that chart received when the examiners thought that it belonged to a Not-Guilty examinee.

Though participants in both studies were experts in their respective fields, and were analyzing familiar materials of the kind that they routinely analyze on the job — albeit not materials that they gathered themselves — results resembled those found in the typical cognitive confirmation bias studies with students performing an artificial task. Some researchers (e.g. Smith and Kida, 1991) have suggested that people can be effective judges when operating in natural familiar contexts. This does not, however, appear to protect them from the present bias.

It has also been suggested (e.g. Klayman, 1995) that the consideration of alternative hypotheses mitigates confirmation biases. Indeed, the facilitated consideration of alternative hypotheses is supposed to be one of the advantages brought about by expertise: 'There is reason to believe that training ... can facilitate the consideration of alternatives. In familiar situations, people may learn certain natural sets of competing hypotheses that must be distinguished ...' (p. 405). Our study required the participants explicitly to evaluate a set of eight diagnostic categories, not just a single target diagnosis. This did not seem to prevent the hypothesis-biased judgment.

The results of these studies join those of Chapman and Chapman in contributing to our understanding of why 'Most clinicians know about the research showing that [projective tests] are invalid, yet ... continue to use the test[s] regularly because they claim they have seen the signs work in their own clinical practice' (Chapman and Chapman, 1982, p. 240) — a state of affairs which still holds true 25 years after the Chapmans' studies. They also join those of Einhorn and Hogarth (1978) in explaining the persistence of the illusion of validity which human judges in general, and clinicians in particular, attribute to their own intuitions. When test results confirm a clinician's hunch or initial hypothesis, whether it is formed on the basis of presenting symptoms or arbitrarily, the 'validity' of the tests is simultaneously confirmed — and so is the psychodiagnostician's clinical intuition.

At the end of the introduction to this article, we stated that our studies were less about the process underlying the cognitive confirmation bias than about its end result, namely, the enhanced probability enjoyed by an hypothesis merely by virtue of being the one tested. Nonetheless, we are prepared at this

point to offer an account of the nature of the cognitive process that might yield this bias, which is based on an extension of the principle of compatibility (Fitts and Seeger, 1953). In its generalized form, 'This principle states that when stimuli and responses are mentally represented, the weight of a stimulus attribute is enhanced to the extent that it is compatible with the required response' (Shafir, 1995, p. 248). In a comprehensive survey, Shafir (1995) discusses how this principle has 'recently been evoked by cognitive psychologists and by students of judgment and decision making to account for a series of surprising yet systematic patterns of behavior' (p. 248), in areas ranging beyond the original ones of perception and motor skills to choice behavior and social judgment.

In particular, Shafir suggests, 'Selective focusing on features that are compatible with a currently held hypothesis or with the given instructions may be seen to underlie numerous studies reporting ... confirmatory biases' (p. 267). Thus, in a task such as our own, the compatibility principle predicts that borderline features will be weighted more when one is testing for a borderline diagnosis, and paranoid features will be weighted more when one is testing for a paranoid diagnosis. Whether the diagnosis is suggested by background information (as in Study 1) or simply by the experimenter's stipulation (as in Study 2) is irrelevant for the principle. Moreover, the compatibility principle does not even require that one actually believes the hypothesis being tested. The fact that an hypothesis directs the search through the test materials suffices to bestow extra weight on diagnostic indicators that are compatible with it. In that sense, the principle of compatibility is more general than the confirmation bias, which is often spoken of in terms of self-perpetuating beliefs.

It is possible, of course, that as investigators testing a confirmation bias research hypothesis, we ourselves are more likely to believe our hypothesis to have been confirmed by our data than had we held a more skeptical and open-minded, or contrary, position. We leave this possibility for the unbiased reader to judge.

ACKNOWLEDGEMENTS

We thank the F.A. Schonbrunn Research Foundation, the Center for the Study of Rationality and Interactive Decision Making, and Sturman Center for Human Development at The Hebrew University for supporting this study. We thank David Budescu, Ilan Yaniv, Jonathan Evans and our thoughtful referees for valuable comments on an earlier version.

REFERENCES

Chapman, L. J. and Chapman, J. P. 'Genesis of popular but erroneous diagnostic observations', *Journal of Abnormal Psychology*, **72** (1967), 193–204.

Chapman, L. J. and Chapman, J. P. 'Illusory correlation as an obstacle to the use of valid psychodiagnostic signs', *Journal of Abnormal Psychology*, **74** (1969), 271–280.

Chapman, L. J. and Chapman, J. P. 'Test results are what you think they are', in Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, 1982, Ch. 17, 239–248.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (2nd edn.), Hillsdale, NJ: Erlbaum, 1988.

Darley, J. M. and Gross, P. H. 'A hypothesis-confirming bias in labelling effects', *Journal of Personality and Social Psychology*, **44** (1983), 20–33.

Einhorn, H. J. and Hogarth, R. M. 'Confidence in judgment: Persistence in the illusion of validity'. *Psychological Review*, **85**, (1978), 395–416.

Elaad, E., Ginton, A. and Ben-Shakhar, G. 'The effects of prior expectations and outcome knowledge on polygraph examiners' decisions', *Journal of Behavioral Decision Making*, **7** (1994), 279–292.

Fitts, P. M. and Seeger, C. M. 'S-R compatibility: Spatial characteristics of stimulus and response codes', *Journal of Experimental Psychology*, **46** (1953), 199–210.

Klayman, J. 'Varieties of confirmation bias', in Busemeyer, J. R., Hastie, R. and Medin, D. L. (eds), *Decision Making from the Perspective of Cognitive Psychology (The Psychology of Learning and Motivation* Vol. 32), New York: Academic Press, 1995, 385–418.

Koehler, J. J. 'The influence of prior beliefs on scientific judgments of evidence quality', *Organizational Behavior and Human Decision Processes*, **56** (1993), 28–55.

McNemar, Q. *Psychological Statistics* (2nd edn), New York: John Wiley, 1955.

Shafir, E. 'Compatibility in cognition and decision', in Busemeyer, J. R., Hastie, R. and Medin, D. L. (eds), *Decision Making from the Perspective of Cognitive Psychology (The Psychology of Learning and Motivation*, Vol. 32), New York: Academic Press, 1995, 247–274.

Smith, J. F. and Kida, T. 'Heuristics and biases: Expertise and task realism in auditing', *Psychological Bulletin*, **109** (1991), 472–489.

Snyder, M. and Swann, W. B. Jr. 'Behavioral confirmation in social interaction: From social perception to social reality', *Journal of Experimental Psychology*, **14** (1978), 148–162.

Authors' biographies:

Gershon Ben-Shakhar (Ph.D., The Hebrew University of Jerusalem) is a professor of psychology at the Hebrew University of Jerusalem. His main research is focused on human psychophysiology (orienting and habituation processes, and psychophysiological detection of information). He also studied psychological testing and its applications to personnel selection.

Maya Bar-Hillel, Ph.D. is professor of psychology, and member of the Center for the Study of Rationality. She studies probabilistic reasoning and rational choice. A recent project was a debunking of the so-called Bible Code. She has collaborated in the past with Ben-Shakhar on JDM aspects of polygraph use, and with Bilu and Ben-Shakhar on a debunking of graphology.

Yoram Bilu (Ph.D. from the Hebrew University, 1979) is a professor of psychology and anthropology at the Hebrew U. His research interests include psychiatry and culture, and the anthropology of religion. He has studied various topics in these fields among Moroccan Jews and Ashkenazi ultraorthodox Jews in Israel.

Gaby Shefler, Ph.D. Senior clinical psychologist, psychoanalyst. Chief psychologist of ''Herzog-Ezrat Nashim'' Hospital, Jerusalem. Clinical senior lecturer, Department of Psychology, The Hebrew University Jerusalem. Main research interests: Time limited psychotherapy outcome and process research, personal and professional development of psychotherapists, ethical issues in psychologists work.

