THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2003, *56A* (2), 193–209

Further evidence for systematic reliability differences between explicit and implicit memory tests

Axel Buchner and Martin Brandt
Heinrich-Heine-Universität, Düsseldorf, Germany

Meier and Perrig (2000) as well as Buchner and Wippich (2000) have shown that simple dissociations between explicit and implicit memory measures need not reflect functional dissociations of hypothetical underlying memory systems. Instead, such dissociations may also result from the fact that some widely used implicit memory measures are simply less reliable than the explicit measures with which they have been compared. We extend this argument in two ways. First, we show that illusion-based memory measures such as the priming measures derived from fame and preference judgement tasks are also subject to the reliability problem. Second, we show that yes–no and two-alternative forced-choice paradigms should, and in fact do, yield virtually identical results as far as the reliability of the memory tests is concerned.

The typical functional dissociation reported in the implicit memory literature shows differences between groups or conditions when memory is measured with an explicit memory test but not when memory is measured with an implicit memory test (e.g., Buchner & Wippich, 1998; Roediger & McDermott, 1993; Schacter, 1987). For instance, when younger and older adults are compared on implicit measures of memory, age differences are generally small and significant only in a few cases, but older adults score lower than younger persons on explicit tests of memory (e.g., Light & Singh, 1987). Such dissociations have been widely interpreted to indicate that the memory processes (e.g., Roediger, 1990) or memory systems (Tulving & Schacter, 1990) underlying performance in explicit tests differ from those underlying performance in implicit memory tests.

Meier and Perrig (2000) as well as Buchner and Wippich (2000) have demonstrated empirically that many popular implicit memory tests yield less reliable memory measures than do explicit tests. It follows directly from classical test theory (cf., Gulliksen, 1950) that a

Requests for reprints should be sent to Axel Buchner, Institut für Experimentelle Psychologie, Heinrich-Heine-Universität, D-40225 Düsseldorf, Germany. Email: axel.buchner@uni-duesseldorf.de

We thank Bettina Mehl for programming the experiments, and Uwe Friese, Susanne Mayr, and Bettina Mehl for assistance with data collection. We also thank the Archiv für Kunst und Geschichte, Berlin, the Deutsche Presse-Agentur, Frankfurt, the Harenberg Kommunikation Verlags- und Mediengesellschaft, Dortmund, and the Süddeutscher Verlag, München, for their kind permission to let us use the photographs free of charge. Special thanks go to Trevor Penny for letting us use his line drawings of possible and impossible objects.

© 2003 The Experimental Psychology Society
http://www.tandf.co.uk/journals/pp/02724987.html DOI:10.1080/02724980244000260

reliable test is more likely to reflect population group differences in terms of statistically significant differences between samples than is an unreliable test. Thus, a certain proportion of dissociations between explicit and implicit memory measures is to be expected simply because of the systematic reliability difference between explicit and implicit memory tests. We cannot know how large this proportion is because the psychometric properties of memory measures are usually not reported. In fact, we would not be surprised to learn that most researchers are unaware of the psychometric properties of the memory measures that they use in their own laboratories.

The implicit tests for which a reliability deficit has already been shown by Meier and Perrig (2000) as well as by Buchner and Wippich (2000) all fall into the class of *performance-oriented* memory tests—that is, tests that are intended to reflect the operation of memory more or less directly in speed or accuracy measures or both. Examples include word-stem completion, category association, and picture naming.

Illusion-oriented implicit memory measures, in contrast, are designed to reflect the use of *memory as a tool* to accomplish a task (Jacoby & Kelley, 1987; Jacoby, Kelley, & Dywan, 1989; Whittlesea, 1993). For instance, previously read famous and nonfamous names appear more famous than previously unread names (Neely & Payne, 1983). According to Jacoby (e.g., Jacoby, Lindsay, & Toth, 1992), this so-called "false fame effect" is a memory-based illusion. It can be explained by assuming, first, that repeated information is processed more fluently, and second, that the processing fluency is automatically attributed to a source. The attribution process may be erroneous, particularly if the correct source—the episode of the previous encounter—is not readily available but another plausible source is. For instance, the fluent processing may be incorrectly attributed to the latent variable underlying a current judgement dimension such as fame.

These judgement dimensions can be manifold. For example, when previously heard and new sentences must be judged against a white noise background of varying loudness, then the background noise appears less loud when the sentences are old (and, presumably, are processed more fluently) than when they are new (Jacoby, Allan, Collins, & Larwill, 1988). Effects of prior experiences can also be misattributed to a statement being true (Begg, Anas, & Farinacci, 1992), an answer being correct (Kelley & Lindsay, 1993), a problem being easy (Jacoby & Kelley, 1987), presentation times being longer (Witherspoon & Allan, 1985), or an object being more aesthetically pleasing (the so-called "mere exposure" effect, e.g., Kunst-Wilson & Zajonc, 1980).

Of these memory-based illusions, the false fame effect has attracted considerable attention, and many facets of this effect have been investigated (e.g., Banaji & Greenwald, 1995; Bartlett, Strater, & Fulton, 1991; Buchner, Steffens, & Berry, 2000; Buchner & Wippich, 1996; Dywan & Jacoby, 1990; Jacoby, Kelley, Brown, & Jasechko, 1989; Jacoby, Woloshyn, & Kelley, 1989; Squire & McKee, 1992). Similarly, the mere exposure effect has been investigated extensively (for a review, see Bornstein, 1989; see also Bornstein & D'Agostino, 1994; Gordon & Holyoak, 1983; Seamon et al., 1995; Zajonc, 1980), and there is even interest in this phenomenon from an advertisement perspective (e.g., Baker, 1999; Perfect & Askew, 1994).

The important point from our perspective is that there are a number of simple dissociations between explicit memory measures on the one side and memory as expressed in fame or preference judgements on the other. A cursory overview of the relevant literature shows that these dissociations come from rather diverse areas: Dividing attention reduced recognition memory

for names relative to a full attention condition, but this manipulation appeared not to affect the false fame effect—that is, the increase in judged fame for previously seen faces over new faces (Jacoby, Woloshyn, & Kelley, 1989); recognition memory for names was reported to be much worse in amnesic patients than in healthy controls, but the false fame effect was similar for both groups of participants (Squire & McKee, 1992); recognition memory for novel melodies was worse in elderly than in younger persons, but both groups did not differ with respect to the mere exposure effect—that is, the preference for previously heard over new melodies (Halpern & O'Connor, 2000); elderly persons were impaired compared with young persons when judging the frequency with which novel, visual stimuli (Japanese ideograms) had occurred, but both groups showed a comparable mere exposure effect as a function of presentation frequency (Wiggs, 1993); patients with mild to moderate Alzheimer's disease performed worse than elderly controls in recognizing previously seen faces, but both groups did not differ significantly in their preference for old over new faces (Winograd, Goldstein, Monarch, Peluso, & Goldman, 1999); schizophrenic patients showed impairments on a recognition task when compared to controls, but both groups showed similar preferences for verbal and visual materials seen earlier as opposed to new materials (Marie et al., 2001); the memorizing of rule-based letter strings under full attention resulted in higher reproduction rates than learning under divided attention, but participants' preference of "grammatical" over "nongrammatical" letter strings did not vary as a function of the attention manipulation (Manza, Zizak, & Reber, 1998); the intentional study of full-page colour magazine advertisements led to clearly superior recognition memory for these advertisements as opposed to an incidental study condition, but on both conditions the same positive bias in attitudes toward experienced advertisements compared with novel advertisements was found (Perfect & Askew, 1994).

This list of examples may suffice to illustrate the point that a number of simple dissociations were reported in the literature that involve illusion-oriented implicit memory measures. It therefore seems important to clarify whether the substantive explanations that were offered for those dissociations might be compromised by the problem of these particular implicit memory measures being less reliable than the explicit measures with which they have been compared in past research. We decided to use recognition judgements as the explicit memory task and fame judgements (Experiments 1a and 2a) as well as preference judgements (Experiments 1b and 2b) as the implicit memory task.

Buchner and Wippich (2000) argued that a major factor that can be suspected to contribute to the reliability difference between explicit and implicit memory measures was that during the former, but not the latter memory test the performance goal is well specified and participants are constantly pushed to respond at their upper performance limits. Preference judgements obviously lack such a specific performance goal, and participants may respond as they please without any criterion available to evaluate the appropriateness of their responses. The relatively large latitude as to how to accomplish the task is likely to be accompanied by a relatively large variability of the cognitive and noncognitive processes contributing to task performance for any given person, thereby resulting in greater error variability and hence in reduced reliability on the side of the memory measure. With respect to fame judgements, one could think of criteria that may help to define true and false answers. However, these criteria are rather fuzzy (when exactly is a person famous?), and the search domain is markedly less restricted in fame judgements than in recognition judgements (cf., Meier & Perrig, 2000).

Therefore, the possible search and response strategies are likely to be relatively complex and variable. Greater variability and complexity in the cognitive processes and response strategies should again induce greater error variability, thereby reducing the reliability of the memory measure derived from the fame judgement task. Thus, both fame and preference judgements may be suspected of providing low-reliability memory measures.

EXPERIMENT 1A

Method

Participants

Participants were 197 students, 142 of whom were female. Their age ranged from 19 to 49 years (M = 22). The students were tested individually and were paid for their participation. They were assigned at random to the recognition or frame judgement task. It was also randomly determined whether a particular person would study List 1 or List 2 items, but toward the end of the experiment the particular combination of treatments was determined by the goal to end up with approximately equal sample sizes in each of the four conditions: the recognition judgement, List 1 condition (n = 48); the recognition judgement, List 2 condition (n = 49); the fame judgement, List 1 condition (n = 51); and the fame judgement, List 2 condition (n = 49).

Materials

A total of 224 greyscale photographs of faces were used. One half of the photographs showed female faces; the other half showed male faces. The persons whose faces were shown had acquired fame in the 20th century (e.g., Luciano Benetton, Italian fashion designer and manufacturer; Gwyneth Jones, Welsh soprano singer; Urho Kekkonen, Finnish president from 1956 to 1981; and Simone Veil, French politician, first president of the European parliament), but the faces themselves were only vaguely familiar to an independent sample of students (N = 71): An overall fame rating of 2.8 on a scale from 1 (no famous) to 7 (very famous) indicated that the photographed persons were considered relatively nonfamous by most students. The photographs were 4.7 cm × 6.5 cm in size. They were displayed on a computer monitor. The faces were assigned to Sets 1 and 2 with the goal to end up with roughly equal average fame ratings for the two sets. Half of the persons shown in each set were female.

Procedure

The learning phase was incidental. Participants were instructed to judge, as quickly and as accurately as possible, whether a briefly displayed photograph showed a male or a female person. Participants responded by hitting the "female" or "male" key on a keyboard. Reactions faster than 100 ms or slower than 5000 ms were counted as invalid, and the relevant photograph was repeated later during the learning phase.

Participants responded to a random sequence of 112 Set 1 or Set 2 photographs, depending on the experimental condition to which they were assigned. Each trial began with a brief warning signal after which a single photograph appeared for 300 ms. Following a post-reaction delay of 1 s, the next photograph was shown, except for every 20th trial, on which participants received feedback about the correctness and speed of their responses. The feedback was qualified depending on the performance level.

Immediately after the learning phase, participants received their instructions for the test phase. The recognition groups were told to decide for each individual photograph whether they had seen it during the immediately preceding gender judgement task. Participants in the fame judgement group were told

to indicate whether the photograph showed a famous or a nonfamous person. Participants responded by clicking, with the computer mouse, the button below the photograph that corresponded to their choice. The next randomly selected photograph was shown 800 ms after the response to the previous one. There was no feedback about the correctness of the response.

Design

The main independent variable was the test condition (recognition vs. fame judgement). The dependent variables were the sensitivity measures reflecting memory performance on the recognition and the fame judgement tasks. However, the variable of most interest was the reliability of the memory measures. Split-half correlations were used as reliability estimates. We expected this correlation ρ to be larger for the explicit than for the implicit test. The test of H_1: $\rho_{explicit} > \rho_{implicit}$ against H_0: $\rho_{explicit} \leq \rho_{implicit}$ is a one-tailed test problem. Given $\alpha = \beta = .05$ and the goal to detect "large" differences between correlations in the population ($q = 0.50$, cf., Cohen, 1977), an a priori power analysis suggested that at least $N = 88$ participants were needed in each test condition. We were able to collect data from $n_1 = 97$ and $n_2 = 100$ participants in the recognition and the fame judgement conditions, respectively. All other things being equal, the power of our test was therefore $1 - \beta = .97$ and thus even slightly larger than had been planned.

The level of α was set to .05 for all analyses reported in this article so that we reject H_0 whenever $p < .05$. Individual p values are nevertheless reported for completeness.

Results and discussion

We decided to use both d' and P_r as sensitivity measures. The former denotes the sensitivity measure defined within signal detection theory (assuming normal, equal-variance signal and noise distributions of the evidence variable). P_r denotes the sensitivity measure within the two-high threshold model, which has been evaluated favourably by Snodgrass and Corwin (1988) and is used quite frequently in recognition memory studies, although usually without reference to the underlying measurement model. Both measures were used for the recognition and the fame judgement tasks. To explicate, for both the recognition and the fame judgement tasks the hit rate was defined as the proportion of "old" List 1 (or List 2, depending on the experimental condition) faces that received and "old" or a "famous" judgement. Analogously, the false alarm rate was defined as the proportion of "new" List 2 (or List 1) faces that received an "old" or "famous" judgement.

Two halves of each test were created using the odd–even method—that is, items with odd and even ordinal numbers were assigned to the first and second test halves, respectively. In that way, two summary scores could be computed for every participant. The split-half correlations were estimated as the correlations between these summary scores.

The descriptive and inferential statistics pertaining to the performance measures are summarized in Table 1 (see Table A1 in the Appendix for the variables underlying the performance measures). The sensitivity measures differed significantly from zero for both the recognition and the fame judgement tasks. We therefore concluded that performance was above chance on both memory measures, which justified the reliability-related analyses.

When looking ahead at the data pattern obtained across all experiments it is clear that the priming effects (the false fame and the mere exposure effects) were generally much smaller than the recognition effects. The latter aspect is particularly obvious from the sample effect size measures in the last column of Table 1. The inspection of the effect size measures also reveals another interesting feature of the results: In general, the sensitivity measure P_r of the

TABLE 1
Sample means of the performance measures for Experiments 1a, 1b, 2a, and 2b and results of the tests of the H_0 positing that the performance measures do not differ from zero

		List 1		*List 2*		*All*					
								H_0: $M = 0$			
Experiment		*M*	*SE*	*M*	*SE*	*M*	*SE*	*t*	*df*	*p*	d_3'[a]
1a	Face recognition P_r	.180	.013	.161	.013	.171	.009	18.788	96	<.0001	1.921
	Fame judgement P_r	.037	.008	.028	.009	.033	.018	5.428	99	<.0001	0.550
	Face recognition d'	.482	.035	.446	.036	.464	.025	18.478	96	<.0001	1.885
	Fame judgement d'	.114	.029	.085	.024	.100	.019	5.256	99	<.0001	0.529
1b	Object recognition P_r	.113	.009	.086	.010	.100	.007	13.908	99	<.0001	1.409
	Preference judgement P_r	.018	.006	.024	.045	.021	.005	4.616	98	<.0001	.0467
	Object recognition d'	.324	.029	.232	.028	.278	.021	13.522	99	<.0001	1.359
	Preference judgement d'	.047	.017	.062	.017	.054	.012	4.529	98	<.0001	0.452
2a	Face recognition P_r	.231	.025	.240	.024	.236	.017	13.525	82	<.0001	1.494
	Fame judgement P_r	.061	.013	.085	.015	.072	.010	12.726	85	<.0001	0.783
	Face recognition d'	.439	.050	.455	.050	.447	.035	7.272	82	<.0001	1.406
	Fame judgement d'	.112	.024	.155	.029	.132	.018	7.135	85	<.0001	.0777
2b	Object recognition P_r	.189	.019	.158	.015	.173	.012	14.422	79	<.0001	1.617
	Preference judgement P_r	.033	.014	.036	.013	.034	.010	3.508	79	<.0004	0.395
	Object recognition d'	.347	.034	.288	.027	.318	.022	14.277	79	<.0001	1.606
	Preference judgement d'	.059	.026	.065	.024	.062	.018	3.504	79	<.0004	0.391

Note: See Table A1 for the variables underlying these measures.
[a]See Cohen (1977).

two-high threshold model appears to be slightly more "sensitive" than the d' measure of the normal-distribution, equal-variance signal detection theory. This is parallel to the results of Snodgrass and Corwin (1988) that the measures derived from the two-high threshold model were more sensitive to the experimental manipulations than were the signal detection theory measures.

The sample split-half correlations for the recognition and the fame judgements are displayed in Table 2, separately for the List 1 (first column) and List 2 (second column) samples. These split-half correlation estimates were averaged on the assumption that the List 1 and List 2 samples were independently drawn from the same population. Compatible with this assumption, the List 1 and List 2 split-half correlations did not differ significantly from each other in this experiment and in any of the other experiments reported in this article. In that way, the power of the critical comparison between test types was maximized. Averaging was achieved by first transforming the correlations to Fisher's Z and then back-transforming the averaged Z value to r (cf., Silver & Dunlap, 1987). The averaged correlations are displayed in the third column of Table 2. The results of the relevant statistical tests are also shown in Table 2.

The split-half correlations that we used as reliability indices were significantly lower for the implicit fame judgement measure than for the explicit recognition judgement measure when

TABLE 2
Sample split-half correlations of the performance measures for Experiments 1a, 1b, 2a, and 2b, and results of the tests of the H_0 positing equality of the reliabilities of the explicit and of the implicit memory measures

		List 1	*List 2*	*All*			
					H_0: $r_{recognition} = r_{fame\ or\ preference}$		
Experiment		$r_{split\text{-}half}$	$r_{split\text{-}half}$	$r_{split\text{-}half}$	z	p	q^a
1a	Face recognition P_r	.522	.303	.417	2.607	.005	0.377
	Fame judgement P_r	.018	.118	.067			
	Face recognition d'	.525	.271	.410	1.555	.060	0.225
	Fame judgement d'	.284	.140	.208			
1b	Object recognition P_r	.431	.476	.454	3.068	.001	0.442
	Preference judgement P_r	.001	.096	.048			
	Object recognition d'	.277	.471	.378	2.422	.008	0.349
	Preference judgement d'	.011	.088	.049			
2a	Face recognition P_r	.527	.644	.592	3.850	<.001	0.603
	Fame judgement P_r	–.089	.227	.078			
	Face recognition d'	.550	.698	.634	4.325	<.001	0.678
	Fame judgement d'	–.101	.265	.071			
2b	Object recognition P_r	.566	.286	.437	2.075	.019	0.334
	Preference judgement P_r	.205	.059	.133			
	Object recognition d'	.546	.286	.384	1.685	.046	0.272
	Preference judgement d'	.195	.059	.133			

[a]See Cohen (1977).

the sensitivity measure P_r was used. This difference just missed out preset criterion for statistical significance when d' was used. The slight difference between the P_r-based and the d'-based results is parallel to that obtained for the performance measures per se: The P_r statistic appear to be slightly more sensitive. Nevertheless, the general pattern of the results is compatible with the expectation that the implicit fame judgement measure of memory is less reliable than the explicit recognition measure. Experiment 1b extends this result to preference judgements.

EXPERIMENT 1B

Method

Participants

Participants were 199 students, 151 of whom were female. Their age ranged from 18 to 45 years (M = 23). The students were tested individually and were paid for their participation. They were assigned at random to the recognition or preference judgement task. It was also randomly determined whether a particular person would study List 1 or List 2 items, but toward the end of the experiment the particular combination of treatments was determined by the goal to end up with approximately equal sample sizes

in each of the four conditions: the recognition judgement, List 1 condition (n = 50); the recognition judgement, List 2 condition (n = 50); the preference judgement, List 1 condition (n = 50); and the preference judgement, List 2 condition (n = 49).

Materials

A total of 224 line drawings were used. One half of the line drawings showed objects that were possible in three-dimensional space; the other half showed impossible objects. The line drawings were scaled so that they occupied about the same area as the faces in Experiment 1a. They were displayed on a computer monitor. The line drawings were assigned to Sets 1 and 2 at random with the restriction that half of the objects shown in each set were possible.

Procedure

The learning phase was identical to that of Experiment 1a with the exception that participants were instructed to judge, as quickly and as accurately as possible, whether a randomly selected, briefly displayed line drawing showed a possible or an impossible object. Participants responded by hitting the "possible" or "impossible" key on a keyboard.

The procedure for the test phase was also identical to that of Experiment 1a, except that participants in the preference judgement group were told to indicate for each individual line drawing whether it appeared aesthetically pleasing to them or not.

Design

The design considerations were identical to those underlying Experiment 1a. Given $\alpha = \beta = 0.5$ and the goal to detect "large" differences between correlations in the population (q = 0.50, cf., Cohen, 1977), an a priori power analysis suggested that at least N = 88 participants were needed in each test condition. We were able to collect data from n_1 = 100 and n_2 = 99 participants in the recognition and the preference judgement conditions, respectively. As in Experiment 1a, the power of our test was therefore $1 - \beta = .97$ and thus even slightly larger than had been planned.

Results and discussion

The results pertaining to the performance measures are very similar to those of Experiment 1a (see Table 1), and so are our conclusions: Performance was above chance on both memory measures, which justified the reliability-related analyses.

The averaged split-half correlations (see the third column in Table 2) that we used as reliability indices were significantly lower for the implicit preference judgement measure than for the explicit recognition judgement measure, and this was so independently of which sensitivity measure was used. However, this pattern seemed again slightly more pronounced when P_r rather than d' was used as the sensitivity measure. Nevertheless, the pattern of results of Experiment 1b is compatible with the expectation that the implicit preference judgement measure of memory is less reliable than the explicit recognition measure.

YES–NO VERSUS FORCED-CHOICE DESIGNS

Experiments 2a and 2b were designed as conceptual replications of Experiments 1a and 1b with a different design. More precisely, Experiments 2a and 2b used a two-alternative

forced-choice (2AFC) task rather than a yes–no task. This should not alter the basic results although, superficially, the two tasks may appear to differ. For instance, the performance measure derived from yes–no tasks typically is a difference score (d' or P_r). Difference scores are known to cause reliability concerns (cf., Meier & Perrig, 2000). In contrast, the proportion of correct trials, $p(c)_{2AFC}$, is usually taken as the performance measure in 2AFC tasks under the assumption that response biases are negligible. However, although $p(c)_{2AFC}$ does not look like a difference score on the surface, it can be shown to be formally equivalent to the sensitivity measure P_r of the two-high threshold model. For the purpose of this illustration we assume that two stimuli are presented simultaneously for recognition in a 2AFC design, one to the right and one to the left. The hit rate (H) may then be defined as the probability of correctly selecting the left stimulus on ⟨old, new⟩ trials,

$$H = p(\text{“left”} \mid \langle \text{old, new} \rangle),$$

and the false alarm rate (FA) may be defined as the probability of incorrectly selecting the left stimulus on <new,old> trials (cf., Macmillan & Creelman, 1991),

$$FA = p(\text{“left”} \mid \langle \text{new, old} \rangle).$$

For simplicity we assume that both types of trial are equally likely. Because $p(c)_{2AFC}$ is simply the overall proportion of trials on which the old stimulus was selected correctly, we find that

$$p(c)_{2AFC} = (H + (1 - FA))/2,$$

and hence

$$2p(c)_{2AFC} - 1 = H - FA$$

The left side of this equation is simply the typical proportion correct corrected for guessing that is often used in 2AFC tasks (cf., Macmillan & Creelman, 1991), whereas the right side of the equation is equivalent to P_r. Following these considerations the pattern of reliability estimates in Experiments 2a and 2b should closely resemble the pattern observed in Experiments 1a and 1b.

According to signal detection theory, the theoretical decision spaces underlying yes–no and 2AFC tasks differ, as a result of which 2AFC tasks are predicted to be easier than yes–no tasks—that is, the expected proportion correct given the same sensitivity is larger for 2AFC than for yes–no tasks. The decision-space differences are taken into account by the signal-detection models of the tasks so that the values of the sensitivity measure d' are predicted to be at the same level across yes–no and 2AFC tasks for individuals with equal abilities in old–new discrimination. However, past research has shown that the $d'_{2AFC}/d'_{yes–no}$ ratios may be greater than 1 (e.g., Creelman & Macmillan, 1979; Jesteadt & Bilger, 1974), indicating a certain degree of misspecification of the underlying measurement model for a given task. By comparing the d' statistics of Experiments 2a and 2b to those of Experiments 1a and 1b we may test whether the signal detection measurement model is adequate for the tasks used here.

Experiment 2a contrasted recognition and fame judgements and was thus parallel to Experiment 1a. Experiment 2b contrasted recognition and preference judgements and was thus parallel to Experiment 1b.

EXPERIMENT 2A

Method

Participants

Participants were 167 students, 112 of whom were female. Their age ranged from 17 to 45 years (M = 25). The students were tested individually and were paid for their participation. They were assigned at random to the recognition or fame judgement task. It was also randomly determined whether a particular person would study List 1 or List 2 items, but toward the end of the experiment the particular combination of treatments was determined by the goal to end up with approximately equal sample sizes in each of the four conditions: the recognition judgement, List 1 condition (n = 39); the recognition judgement, List 2 condition (n = 44); the fame judgement, List 1 condition (n = 46); and the fame judgement, List 2 condition (n = 40).

Materials

The materials were identical to those of Experiment 1a.

Procedure

The learning phase was identical to that of Experiment 1a. The same was true for the test phase except that participants saw two photographs of faces at a time, one of which had been presented during the gender judgement phase. Participants in the recognition judgement condition indicated which of the two photographs they had seen during the immediately preceding gender judgement task. Participants in the fame judgement condition indicated which photograph showed the more famous person. Participants responded by clicking with the computer mouse, on the photograph that corresponded to their choice.

Design

The design considerations were identical to those underlying Experiment 1a. Given $\alpha = \beta = .05$ and the goal to detect "large" differences between correlations in the population (q = 0.50, cf., Cohen, 1977), an a priori power analysis suggested that at least N = 88 participants were needed in each test condition. We were able to collect data from n_1 = 83 and n_2 = 86 participants in the recognition and the fame judgement conditions, respectively. All other things being equal, the power of our test was therefore $1 - \beta = .94$ and thus slightly smaller than had been planned.

Results and discussion

The results pertaining to the performance measures are similar to those of Experiment 1a (see Table 1) in that performance was above chance on both memory measures, which justified the reliability-related analyses. Also note that performance was better in the present 2AFC task than in the yes–no task of Experiment 1a. This is discussed briefly in the General Discussion section.

The averaged split-half correlations (see the third column in Table 2) that we used as reliability indices were significantly lower for the implicit fame judgement measure than for the explicit recognition judgement measure, and this was so independently of which sensitivity measure was used. This time the pattern seems slightly more pronounced when d' rather than P_r is used as the sensitivity measure, but the reason for the deviation from the general pattern

observed in all other experiments is unclear. In any case, the pattern of the results of Experiment 2a is compatible with the expectation that the implicit fame judgement measure of memory is less reliable than the explicit recognition measure.

EXPERIMENT 2B

Method

Participants

Participants were 160 students, 103 of whom were female. Their age ranged from 17 to 45 years (M = 24). The students were tested individually and were paid for their participation. They were assigned at random to the recognition or preference judgement task. It was also randomly determined whether a particular person would study List 1 or List 2 items, but toward the end of the experiment the particular combination of treatments was determined by the goal and end up with approximately equal sample sizes in each of the four conditions: the recognition judgement, List 1 condition (n = 40); the recognition judgement, List 2 condition (n = 40); the preference judgement, List 1 condition (n = 40); and the preference judgement, List 2 condition (n = 40).

Materials

The materials were identical to those of Experiment 1b.

Procedure

The learning phase was identical to that of Experiment 1b. The test phase was identical to that of Experiment 2a, except that participants in the preference judgement group were told to indicate which of two simultaneously presented line drawings appeared more aesthetically pleasing to them.

Design

The design considerations were identical to those underlying Experiment 1a. Given $\alpha = \beta = .05$ and the goal to detect "large" differences between correlations in the population (q = 0.50, cf., Cohen, 1977), an a priori power analysis suggested that at least N = 88 participants were needed in each test condition. We were able to collect data from n_1 = 80 and n_2 = 80 participants in the recognition and the preference judgement conditions, respectively. All other things being equal, the power of our test was therefore $1 - \beta = .93$ and thus slightly smaller than had been planned.

Results and discussion

The results pertaining to the performance measures are very similar to those of Experiment 2a (see Table 1) in that, first, performance was above chance on both memory measures, which justified the reliability-related analyses, and, second, performance was better than with the same stimuli in Experiment 1b.

The averaged split-half correlations (see the third column in Table 2) that we used as reliability indices were significantly lower for the implicit preference judgement measure than for the explicit recognition judgement measure, and this was so independently of which sensitivity measure was used. Again, this pattern seemed slightly more pronounced when P_r rather than d' was used as the sensitivity measure. In essence, however, the pattern of results of

Experiment 2b is compatible with the expectation that the implicit preference judgement measure of memory is less reliable than the explicit recognition measure.

GENERAL DISCUSSION

The present experiments aimed at demonstrating a reliability deficit of illusion-oriented implicit memory measures such as those derived from fame and preference judgement tasks. Independently of whether fame or preference judgements were used, and independently of whether the memory test was implemented as a yes–no or a 2AFC design, the reliability of the implicit memory measure was significantly lower than the reliability of the explicit memory measure. These results systematically extend previous results presented by Meier and Perrig (2000) and by Buchner and Wippich (2000). As a consequence, it is now clear that simple dissociations between explicit and implicit, illusion-oriented memory measures such as those discussed in the Introduction may not need a substantive explanation in terms of different memory processes or systems being involved in the two types of task. Rather, such dissociations could also be due to the plain methodological artefact of the implicit memory measure being less reliable than its explicit counterpart.

However, this must not be taken to mean that all such dissociations are necessarily artifacts. In principle, it is still possible that the simple dissociations of the sort considered here (a difference between groups or conditions on the explicit measure, no difference on the implicit measure) have a substantive basis in terms of differences in the underlying memory processes or systems, but that the probability of finding a dissociation in a sample is also augmented by the reliability difference. The fundamental problem is that we do not know whether it is justified to give a substantive interpretation to a particular dissociation pattern (e.g., that between recognition and fame judgements as discussed in the Introduction) as long as the reliabilities of the memory measures are unknown. It should also be noted that, as Buchner and Wippich (2000) have already pointed out, the reliability problem does not concern systematic patterns of double dissociations in which an experimental manipulation affects the explicit memory measure in one direction and the implicit memory measure in the other. Buchner and Wippich have also shown that implicit memory measures are not unreliable per se. Rather, it seems that the reliability problem occurs because implicit memory measures typically, but not necessarily, allow considerable latitude as to how to accomplish the task in which memory performance is measured. As a result of this greater latitude, the variability of the cognitive and noncognitive processes contributing to task performance for any given person may be considerable, which, in turn, reduces the reliability of the memory measure derived rom the task. In support of this hypothesis, Buchner and Wippich found that memory measures derived from speeded word and picture identification tasks (i.e., tasks that have a well-defined performance goal and push participants towards their upper performance limits) may be about as reliable as recognition measures using the same materials. Along the same line, Meier (2001) reported that both picture identification and recognition were affected by a depth-of-processing manipulation so that there was no dissociation between the implicit and explicit memory measures. This is interesting because implicit memory measures have been reported to be unaffected by depth-of-processing manipulations that have robust effects on explicit memory measures (e.g., Chiu & Schacter, 1995; Meier & Perrig, 2000; Roediger, Weldon, Stadler, & Riegler, 1992), although often such effects seem to have been present at a descriptive level

without reaching statistical significance (Challis & Brodbeck, 1992). However, Meier's picture identification measure was as reliable as his recognition measure, suggesting that prior failures to find statistically significant depth-of-processing effects on implicit memory measures may have been due to the low reliability of those measures.

Before discussing possible consequences for future research we would like to address two aspects that concern the present data. A first aspect of the data that may seem notable is the apparently low absolute performance level on both the explicit and implicit memory tasks. However, the sample effect size measures reported in the last column of Table 1 show that the sizes of the effects were actually quite large. To illustrate, let us accept for the moment that the average of the sample effect sizes for the implicit memory measures of the present four experiments is the best possible estimate of the population effect size ($d_3' = .549$ for P_r). Then the sample size needed given $\alpha = .05$ and a desired power $(1 - \beta)$ of .80 (which Cohen, 1977, recommends) or .95 (which we prefer when, as in the present case, α and β errors are equally serious) is only $N = 22$ or $N = 37$, respectively. Thus, false fame or mere exposure effects of the sort reported here can be called very robust in that they can be expected to be replicated with a very high probability even if the sample sizes are as small as they often are in psychological research (way too small, that is, see, e.g., Cohen, 1962; Rossi, 1990; Sedlmeier & Grigerenzer, 1989). Also, the absolute performance level is not an issue as long as performance is above chance so that the reliability-related analyses are meaningful. Finally, we were interested only in the properties of the implicit memory measure *relative* to the properties of the explicit memory measure under otherwise identical conditions because this is the situation of the typical simple-dissociation experiment (see the Introduction of this article for a list of examples).

A second interesting side-aspect of the present data is that the sample d' values of the 2AFC tasks in Experiments 2a and 2b were in the order of magnitude of the corresponding performance levels observed for the yes–no tasks in Experiments 1a and 1b. Considering Experiments 1a and 2a, the d'_{2AFC}/d'_{yes-no} ratio was 0.963 for the recognition judgements and 1.32 for the fame judgements. Considering Experiments 1b and 2b, the d'_{2AFC}/d'_{yes-no} ratio was 1.144 for the recognition judgements and 1.148 for the preference judgements. Considering that these values were obtained using samples that were randomly drawn but may still have differed in terms of their "memory ability", the d'_{2AFC}/d'_{yes-no} ratios seem close enough to 1 to conclude that signal detection theory took into account adequately the differences in decision space between the present yes–no and 2AFC memory tasks.

Returning to the central issue of this article, it is quite obvious that reliability estimates obtained for the memory measures in the present experiments are relatively low compared to what we would expect from a reasonably well-designed psychometric measurement instrument. Unfortunately, the present data cannot be regarded as anomalies, because reliability estimates for the explicit and implicit memory measures were all in the order of magnitude of what has been reported recently for comparable tasks (Buchner & Wippich, 2000; Meier & Perrig, 2000).

What can be done to remedy the problem? In discussing the issue of difference scores causing reliability concerns, Meier and Perrig (2000) suggested that base-rate corrections could be dropped once the population base rates could be estimated well enough. Meier and Perrig had word-stem completion tasks in mind. Extending their idea to the present yes–no and 2AFC tasks, this would be essentially equivalent to using the hit rate alone. Along the same line, one could be tempted to calculate the priming scores for implicit memory measures by simply

subtracting the overall sample base rate (as an estimate of the population base rate) from the hit rate of each individual. Because associativity holds, the overall group-wise priming score would not be affected. Thus, tests of the hypothesis that priming has been observed (i.e., that the priming score is different from zero) would be insensitive to the way the base rate is taken into account. Further, as a constant is subtracted from the hit rate, a reliability measure calculated from this priming score would be equivalent to the reliability measure computed from the hit rate directly. As the hit rate is not a difference score, reliability measures could be expected to reach larger values. It does indeed turn out that reliability estimates based on hit rates alone are routinely higher than those based on the sensitivity measures. For instance, the split-half correlations for the recognition and fame judgement hit rates of Experiment 1 were .916 and .835, respectively (still a significant difference, $z = 2.483$, $p < .007$).

At first sight, such a procedure may sound like an ideal solution of the low-reliability problem, but it is not. Hit rates (and hit rates minus a constant) are contaminated by response strategies (such as preferring yes over no responses or stimuli presented on the left over those presented on the right side). An increase in a reliability estimate for hit rates over the corresponding reliability estimate for the sensitivity measures is thus very likely to result from a consistency in response strategies across items and measurement occasions.

Currently, we see three strategies to deal with the low-reliability problem. First, the number of items could be increased in the implicit relative to the explicit memory test. Tests with more items tend to yield more reliable measures. Second, as the reliability problem translates directly into a statistical power problem (Buchner & Wippich, 2000), more participants could be run in conditions in which memory is measured with a low-reliability memory test. Third, only implicit memory tests with psychometric properties that are equivalent to their explicit counterparts could be used. Such measures exist (Buchner & Wippich, 2000), and, as discussed earlier, previously observed simple dissociations may disappear when they are used (Meier, 2001). In addition, it may be possible to increase the reliability of existing implicit memory measures (e.g., by imposing additional restrictions such as time pressure). It may also be possible that implicit memory measures other than those analysed so far turn out to be as reliable as the explicit measures with which they have been compared in the past. At any rate, the general presupposition for all of this is that the psychometric properties of the memory measures become known.

REFERENCES

Baker, W.E. (1999). When can affective conditioning and mere exposure directly influence brand choice? *Journal of Advertising*, *28*, 31–46.

Banaji, M.R., & Greenwald, A.G. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology*, *68*, 181–198.

Bartlett, J.C., Strater, L., & Fulton, A. (1991). False recency and false fame of faces in young adulthood and old age. *Memory & Cognition*, *19*, 177–188.

Begg, I.M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*, 446–458.

Bornstein, R.F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Bornstein, R.F., & D'Agostino, P.R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition*, *12*, 103–128.

Buchner, A., Steffens, M.C., & Berry, D.C. (2000). Gender stereotyping and decision processes: Extending and reversing the gender bias in fame judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1215–1227.

Buchner, A., & Wippich, W. (1996). Unconscious gender bias in fame judgments? *Consciousness and Cognition*, *5*, 197–220.

Buchner, A., & Wippich, W. (1998). Differences and commonalities between implicit learning and implicit memory. In M.A. Stadler & P.A. Frensch (Eds.), *Handbook of implicit learning* (pp. 3–46). Thousand Oaks, CA: Sage Publications.

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*, 227–259.

Challis, B.H., & Brodbeck, D.R. (1992). Level of processing affects priming in word fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 595–607.

Chiu, C.Y.P., & Schacter, D.L. (1995). Auditory priming for nonverbal information: Implicit and explicit memory for environmental sounds. *Consciousness and Cognition*, *4*, 440–458.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Creelman, C.D., & Macmillan, N.A. (1979). Auditory phase and frequency discrimination: A comparison of nine procedures. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 146–156.

Dywan, J., & Jacoby, L.L. (1990). Effects of aging on source monitoring: Differences in susceptibility to false fame. *Psychology and Aging*, *5*, 379–387.

Gordon, P.C., & Holyoak, K.J. (1983). Implicit learning and generalization of the "mere exposure" effect. *Journal of Personality and Social Psychology*, *45*, 492–500.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.

Halpern, A.R., & O'Connor, M.G. (2000). Implicit memory for music in Alzheimer's disease. *Neuropsychology*, *14*, 391–397.

Jacoby, L.L., Allan, L.G., Collins, J.C., & Larwill, L.K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 240–247.

Jacoby, L.L., & Kelley, C.M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin*, *13*, 314–336.

Jacoby, L.L., Kelley, C.M., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*, 326–338.

Jacoby, L.L., Kelley, C.M., & Dywan, J. (1989). Memory attributions. In H.L. Roediger & F.I.M. Craik (Eds.), *Varieties of memory and consciousness. Essays in honour of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Jacoby, L.L., Lindsay, D.S., & Toth, J.P. (1992). Unconscious influences revealed: Attention, awareness, and control. *American Psychologist*, *47*, 802–809.

Jacoby, L.L., Woloshyn, V., & Kelley, C.M. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, *118*, 115–125.

Jesteadt, W., & Bilger, R.C. (1974). Intensity and frequency discrimination in one-and two-interval paradigms. *Journal of the Acoustical Society of America*, *55*, 1266–1276.

Kelley, C.M., & Lindsay, D.S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*, 1–24.

Kunst-Wilson, W.R., & Zajonc, R.B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*, 557–558.

Light, L.L., & Singh, A. (1987). Implicit and explicit memory in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 531–541.

Macmillan, N.A., & Creelman, C.D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Manza, L., Zizak, D., & Reber, A.S. (1998). Artificial grammar learning and the mere exposure effect: Emotional preference tasks and the implicit learning process. In M.A. Stadler & P.A. Frensch (Eds.), *Handbook of implicit learning* (pp. 201–222). Thousand Oaks, CA: Sage Publications.

Marie, A., Gabrieli, J.D.E., Vaidya, C., Brown, B., Pratto, F., Zajonc, R.B., & Shaw, R.J. (2001). The mere exposure effect in patients with schizophrenia. *Schizophrenia Bulletin*, *27*, 297–303.

Meier, B. (2001). Verschwinden Dissoziationen zwischen implizitem und explizitem Gedaechtnis, wenn die Reliabilitaet der Tests vergleichbar is? Ein Beispiel [Do dissociations between implicit and explicit memory measures disappear when their reliability is considered? An example]. *Zeitschrift fuer Experimentalle Psychologie*, *48*, 207–313.

Meier, B., & Perrig, W.J. (2000). Low reliability of perceptual priming: Its impact on experimental and individual difference findings. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *53A*, 211–233.

Neely, J.H., & Payne, D.G. (1983). A direct comparison of recognition failure rates for recallable names in episodic and semantic memory tests. *Memory & Cognition*, *11*, 161–171.

Perfect, T.J., & Askew, C. (1994). Print adverts: Not remembered by memorable. *Applied Cognitive Psychology*, *8*, 693–703.

Roediger, H.L. (1990). Implicit memory. Retention without awareness. *American Psychologist*, *45*, 1043–1056.

Roediger, H.L., & McDermott, K.B. (1993). Implicit memory in normal human subjects. In H. Spinnler & F. Boller (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63–131). Amsterdam, Netherlands: Elsevier Science Publishers.

Roediger, H.L., Weldon, M.S., Stadler, M.L., & Riegler, G.L. (1992). Direct comparison of two implicit memory tests: Word fragment and word stem completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1251–1269.

Rossi, J.S. (1990). Statistical power off psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646–656.

Schacter, D.L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 501–518.

Seamon, J.G., Williams, P.C., Crowley, M.J., Kim, I.J., Langer, S.A., Orne, P.J., & Wishengard, D.L. (1995). The mere exposure effect is based on implicit memory: Effects of stimulus type, encoding conditions, and number of exposures on recognition and affect judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 711–721.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Silver, N.C., & Dunlap, W.P. (1987). Averaging correlation coefficients: Should Fishers' z transformation be used? *Journal of Applied Psychology*, *72*, 146–148.

Snodgrass, J.G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

Squire, L.R., & McKee, R. (1992). Influence of prior events on cognitive judgments in amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 106–115.

Tulving, E., & Schacter, D.L. (1990). Priming and human memory systems. *Science*, *247*, 301–306.

Whittlesea, B.W.A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1235–1353.

Wiggs, C.L. (1993). Aging and memory for frequency of occurrence of novel, visual stimuli: Direct and indirect measures. *Psychology and Aging*, *8*, 400–410.

Winograd, E., Goldstein, F.C., Monarch, E.S., Peluso, J.P., & Goldman, W.P. (1999). The mere exposure effect in patients with Alzheimer's disease. *Neuropsychology*, *13*, 41–46.

Witherspoon, D., & Allan, L.G. (1985). The effect of prior presentation on time estimation in a perceptual identification task. *Memory & Cognition*, *13*, 101–111.

Zajonc, R.B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*, 151–175.

Original manuscript received 28 August 2001
Accepted revision received 24 January 2002

APPENDIX

TABLE A1
Sample means of the variables underlying the performance measures reported for Experiments 1a, 1b, 2a, and 2b in Table 1

		List 1				*List 2*				*All*			
		Hits		*False alarms*		*Hits*		*False alarms*		*Hits*		*False alarms*	
Experiment		*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
1a	Face recognition	.541	.017	.361	.017	.527	.022	.366	.020	.534	.014	.363	.013
	Fame judgement	.422	.024	.385	.023	.419	.019	.390	.019	.420	.015	.388	.015
1b	Object recognition	.526	.021	.413	.022	.498	.020	.412	.020	.512	.014	.412	.015
	Preference judgement	.512	.015	.494	.017	.507	.016	.483	.016	.509	.011	.488	.012
2a	Face recognition	.663	.019	.431	.016	.652	.016	.412	.015	.657	.012	.421	.011
	Fame judgement	.546	.014	.485	.015	.594	.017	.508	.017	.568	.011	.496	.011
2b	Object recognition	.604	.018	.415	.016	.603	.014	.445	.015	.604	.011	.430	.011
	Preference judgement	.510	.012	.478	.014	.508	.014	.471	.011	.509	.009	.475	.009