Depression, Confidence, and Decision: Evidence Against Depressive Realism

Tiffany Fu, Wilma Koutstaal, Cynthia H. Y. Fu, Lucia Poon, and Anthony J. Cleare

Accepted August 31, 2004

This research examined how retrospective self-assessments of performance are affected by major depression. To test the validity of the depressive realism versus the selective processing hypotheses, aggregate posttest performance estimates (PTPEs) were obtained from clinically depressed patients and an age-matched comparison group across 4 decision tasks (object recognition, general knowledge, social judgment, and line-length judgment). As expected on the basis of previous findings, both groups were underconfident in their PTPEs, consistently underestimating the percentage of questions they had answered correctly. Contrary to depressive realism, and in partial support of the selective processing account, this underconfidence effect was not reduced but modestly exacerbated in the depressed patients. Further, whereas the PTPEs of the comparison group exceeded that expected on the basis of chance alone those of the depressed individuals did not. The results provide no support for the depressive realism account and suggest that negative biases contribute to metacognitive information processing in major depression.

KEY WORDS: depressive realism; overconfidence; underconfidence; bias; cognitive therapy.

According to the depressive realism hypothesis people suffering with a depressive illness are more realistic than are nondepressed individuals, who often show unrealistically positive self-evaluations, overconfidence, exaggerated perceptions of control, and unrealistic optimism (Alloy & Abramson, 1979; Colin, 1998). The depressive realism account proposes that the perceptions and inferences of depressed individuals about self-referent events are more accurate and realistic, that is, they tend not to over- or under-estimate the subjective probability of events (Abramson & Alloy, 1981; Ackermann & DeRubeis, 1991; Alloy & Abramson, 1979; Alloy & Ackerman, 1988; Clark, Beck, & Alford, 1999; Hancock, Moffoot, & O’Carroll, 1996).

The selective processing hypothesis, on the other hand, suggests that, although depressed individuals may sometimes appear to show less overconfidence, this reflects a general bias towards negativity (Beck, Rush, Shaw, & Emery, 1979). On the selective processing account, depressed individuals show a bias for sampling the negative aspects of their self-referent experiences (Beck, 1967). This bias is rooted in an automatic tendency to selectively focus on the negative features of their personal experience and to exclude or overlook the positive elements of the situation, due to their negative schemata. Based on the premise that depression is maintained by negatively biased information processing and dysfunctional beliefs (Beck, 1967), the aims of cognitive therapy for depression are to help the patient to identify and recognize these distorted evaluations and eventually to correct them.
In situations where the nondepressed are overconfident or overoptmistic, a bias towards negativity would result in more “realistic-appearing” assessments among the depressed. Thus, in situations where control participants are overconfident, the two hypotheses predict a similar outcome. However, different outcomes are predicted in cases where controls are either realistic or underconfident. Here, whereas the depressive realism hypothesis predicts that the depressed still will be more “realistic” or “accurate” in their confidence evaluations, the selective processing hypothesis suggests that the negative tendencies of the depressed will lead them to be more underconfident or negatively biased than the nondepressed.

Therefore, in order to differentially test the validity of these two accounts, it is necessary to include an experimental situation in which the accuracy of confidence or self-performance assessment in the nondepressed comparison group is either realistic or underconfident (Stone, Dodrill, & Johnson, 2001). Surprisingly, however, no studies have been conducted with clinically depressed patients in which the nondepressed participants showed underconfident performance. Furthermore, many studies of depressive realism suffer from two additional shortcomings (Ackermann & DeRubeis, 1991). First, in some studies, there were no objectively correct answers to the experimental tasks that were used (Albright & Henderson, 1995) so the degree of distortion in depressed and nondepressed participants could not be assessed. Hence, the validity of depressive realism, or the degree of distortion, demonstrated in these studies, is unknown. Second, many studies have not included persons with severe and clinically diagnosed depression (rather than self-reported dysphoria).

The current study addresses each of these shortcomings. Patients diagnosed with major depression (unipolar subtype) and an age-matched comparison group completed four different decision tasks and, after the completion of each task, were asked by the experimenter to provide a posttest performance estimate (PTPE), in which they stated the overall percentage of questions they thought they had answered correctly in that task. The decision tasks included tests of object recognition (episodic memory), general knowledge (semantic memory), line-length judgment (perceptual discrimination), and a modified version of Baron-Cohen, Wheelwright, Hill, Raste, and Plumb’s (2001) “Reading the Mind in the Eyes” test (social judgment). Items for each task had objectively correct answers and thus the individual’s estimated performance could be compared with their actual performance. Critically, although many studies have shown overconfidence in normal controls on cognitive tasks such as answering general knowledge questions under conditions where they are asked to give their confidence ratings on an “item-by-item” basis, immediately after answering each question (e.g., Lichtenstein, Fischhoff, & Phillips, 1982; Stankov, 1998), normal controls have been found to show either realistic or underconfident estimates of their performance when, instead, asked to retrospectively evaluate their overall level of performance on a task, in an “aggregate” manner as a PTPE (e.g., May, 1988, cited in Brenner, Koehler, Liberman, & Tversky, 1996; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Griffin & Tversky, 1992; Sniezek, Pasee, & Switzer, 1990). Thus, use of the PTPE measure of confidence should provide the opportunity to differentially test the depressive realism versus the selective processing hypotheses. Furthermore, it has been argued (Gigerenzer et al., 1991) that item-by-item confidence ratings and aggregate PTPE confidence measures are conceptually different because they require participants to refer to different “reference classes” (in the sense used by Brunswik, 1943). For example, for general knowledge questions regarding the relative time of occurrence of two inventions or discoveries, the reference class that is relevant to a particular item-by-item confidence judgment is all inventions or discoveries that were included in that specific task. By contrast, for the PTPE confidence judgment for this same task, the relevant reference class is the set of all prior experiences that the individual has had with this form of relative historical dating task. On this proposal, the PTPE confidence judgment involves greater self-reference than does the item-by-item judgment, such as attempted retrieval of relevant instances from the individual’s autobiographical past. The PTPE measure thus should provide a strong test of the selective processing versus depressive realism accounts.

If the depressive realism hypothesis is correct, the performance of depressed individuals on the PTPE measure should be more accurate or realistic than that of the nondepressed control participants. However, if the selective processing hypothesis is valid, the PTPEs of the depressed individuals should be more negatively biased (more underconfident) than those of the nondepressed comparison group.

A third group of participants also was tested. These participants previously had been depressed but, at the time of testing, no longer met clinical criteria for depression. These patients allowed examination of the realism of retrospective estimates of performance in a group who were still moderately dysphoric but who had recovered from the depressive illness.
Depression, Confidence, and Decision

METHOD

Participants

The depressed patient group included 15 individuals who fulfilled DSM-IV (American Psychiatric Association, 1994) criteria for a current episode of major depression, unipolar subtype, as based on the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID; First, Spitzer, Gibbon, & Williams, 1996). The comparison group included 22 healthy age-matched volunteers. The recovered patient group included 15 individuals who previously had been depressed but who, at the time of testing, no longer met DSM-IV criteria for unipolar depression.

Exclusion criteria for the depressed and recovered patients were history of electroconvulsive treatment in the last year; neurological disorders, for example, stroke, seizure disorder, etc.; head injury with loss of consciousness; and history of a manic/hypomanic episode as defined in DSM-IV. Most of the depressed (13/15) and recovered (13/15) patients were on medication.

Patients were recruited from, and completed the experiment at, the National Affective Disorder Unit, Alexander House 2, Bethlem Royal Hospital, Beckenham, Kent. Control participants were recruited from staff at Bethlem Royal Hospital (15 participants) and from the Aging Panel in the Department of Psychology at the University of Reading (seven participants). All participants reported normal, or corrected-to-normal, vision. Additionally, all were native speakers of English, or had begun to learn English before the age of 6 years.

The average age of the depressed patients was 44 years (SD = 15). The depressed patients (13 female/2 male) obtained an average of 18.1 (SD = 3.5) on the 21-item Hamilton Depression Rating Scale (HAM-D; Hamilton, 1960), and an average of 38.5 (SD = 13.7) on the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961); they had an average of 14 years (SD = 4) of formal education. The mean age of the comparison controls was 45 years (SD = 18). The controls (14 female/8 male) obtained average scores of 0.5 (SD = 0.9) and 3.3 (SD = 4.2) for the HAM-D and BDI respectively; on average, they had 13 (SD = 4) years of formal education. The depressed patients did not differ from the controls in age or years of education (Fs < 1). The recovered patients (10 female/5 male) likewise did not differ from the controls in age (M = 52 years, SD = 11) or years in formal education (M = 13, SD = 3), Fs < 1.6; they obtained an average of 4.6 (SD = 3.9) on the Hamilton Depression scale and an average of 18.9 (SD = 15.8) on the Beck Depression Inventory. Thus, while these individuals were still moderately dysphoric, they no longer met criteria for clinical depression, and their HAM-D scores were below the standard cut off point of <10, defined as full remission from depressive illness (Nicenber & DeCecco, 2001).

Procedure

Participants were tested individually during one experimental session of about 1 h. After giving informed consent, participants first completed the BDI and thereafter were interviewed with the HAM-D. They then took part in four decision tasks, which were administered in counterbalanced order. All of the tasks were in a two-alternative forced choice format, and included object recognition (a test of episodic memory), general knowledge questions (a test of semantic memory), a social judgment task (assessing the emotions shown by pictured actors), and a line-length judgment task (a task of visual perceptual discrimination). All of the tasks were presented on a Macintosh laptop using PsyScope (Cohen, Macwhinney, Flatt, & Provost, 1993), and participants gave their answers by pressing designated keys on the computer keyboard.8

Before each task, participants were given written instructions and the experimenter also explained the instructions verbally to them. After each task participants verbally were asked to estimate the percentage of questions they had answered correctly on that task (e.g., 70%). Participants were reminded that, because each of the questions had two alternatives, they could expect to obtain an average of 50% correct simply by guessing or chance alone. After completing all four tasks, participants were paid a small honorarium and debriefed regarding the experiment.

7 The very first recovered patient tested was not administered the BDI or HAM-D. The actual distribution of HAM-D scores obtained by the recovered patient group was 0, 0, 0, 0, 2, 2, 3, 7, 7, 8, 9, 9, 9.
8 All participants also provided confidence judgments immediately after each decision (item-by-item confidence judgments). However, because these confidence ratings were made on a 1-6 scale, and were only verbally described as probability estimates when initially presented to participants, it is not possible to know whether participants consistently interpreted their confidence ratings as probability estimates (e.g., that a confidence rating of 1 corresponded to an estimated 50% probability of being correct on that trial; that a confidence rating of 2 corresponded to a 60% probability of being correct on that particular question, etc.). The FTPE responses were direct probability judgments and therefore permit valid comparisons between the participants' perceived estimates of their performance and their actual decision accuracy.
Experimental Design

The experimental design included one between-subjects variable: group (three levels—depressed patient; comparison; and recovered patient), and one within-subjects variable: task (four levels—object recognition, social judgment, general knowledge, and line-length judgment).

Stimulus Materials

Object Recognition Task

There were two phases in this episodic memory task: a study or “encoding” phase followed by the test phase. In the study phase, participants were shown 120 colored, individual objects, which appeared for 300 ms. Participants were asked to indicate if the object was larger than a 13-inch (33 cm) box by pressing one of two predesignated keys on the keyboard (stimuli and tasks as in Koutstaal, 2003; an example box was shown before the study phase). The purpose of the size-judgment task was to ensure that each participant encoded the objects in a similar manner; there was no mention of the subsequent memory test. A 5-min unrelated filler activity (letter search) was given between the study and test phase.

In the test phase of the object recognition task, participants were shown 120 pictures of objects and were asked to make old/new judgments, in which they were to decide if they had, or had not, earlier seen the object in the size-judgment task. Of the 120 objects, 40 were exactly the same as those shown in the size-judgment task, another 40 were alternative exemplars of objects shown in the encoding phase (that is, different instances or tokens of objects that appeared in the size-judgment task), and the remaining 40 objects were new items never previously shown in the experiment. Participants were instructed to give an “old” judgment only if the object was identical to one they had seen before and to call the object “new” if they had not seen that specific object earlier. The order in which items appeared was counterbalanced in both the study and test lists.

Social Judgment Task

This task was a modified version of the task used by Baron-Cohen et al. (2001). The original version was a four-alternative forced choice task, in which participants were shown black-and-white photographs of the eye region of different actors, and were asked to judge the mental state of the actors (e.g., jealous, panicked, hateful or arrogant). However, in order to maintain the same guessing level (50%) for all of the tasks used in the current experiment, the original task was changed into a two-alternative forced choice task. In pilot work, three different versions of the social judgment task were created and administered, with the versions constructed by separately pairing the correct adjective for each question with each of the incorrect adjectives. The final version then was created by choosing item pairs so as to avoid a ceiling effect in task performance. Participants were shown each photograph with two adjectives (e.g., upset, annoyed) and asked to choose the word that best described the feeling of the actor in the photograph.

General Knowledge Task

Two types of general knowledge were tested in this semantic memory task: inventions/discoveries and flags. In the inventions task, participants were asked to indicate which one of two listed inventions/discoveries (e.g., piano, air pump) they thought took place earlier (tasks as in Koutstaal & Matthews, 2004). In the flags task, participants were asked to indicate which one of two flags shown on the computer screen matched the country designated in the question (e.g., which is the flag of Canada?). Normative data concerning levels of recognition for the inventions and countries were collected previously (Koutstaal & Matthews, 2004), and only inventions and countries that were recognized by the majority of participants in the norming study were included. Items within each type of question (i.e., inventions and flags) were randomly paired with one another. There were two randomized lists for each of the general knowledge tasks and approximately half of the participants in each group received each of the orders.

Line-Length Judgment Task

This perceptual task was a modified version of a line-length task adopted by Crawford and Stankov (1996). Participants were shown two vertical nonaligned black lines on the computer screen and were asked to choose the line that they thought was longer. Task difficulty was varied by manipulating the magnitude of the difference between the two lines.

RESULTS

Three key dependent measures are considered: proportion correct (the proportion of items or questions
Table 1. Mean Performance Scores

<table>
<thead>
<tr>
<th>Task</th>
<th>Object-recognition</th>
<th>Social judgment</th>
<th>GK</th>
<th>LJ</th>
<th>Overall average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion correct</td>
<td>.75</td>
<td>.83</td>
<td>.68</td>
<td>.70</td>
<td>.74</td>
</tr>
<tr>
<td>SD</td>
<td>.10</td>
<td>.09</td>
<td>.05</td>
<td>.11</td>
<td>.06</td>
</tr>
<tr>
<td>PTPE</td>
<td>58.67</td>
<td>57.67</td>
<td>48.07</td>
<td>47.33</td>
<td>52.93</td>
</tr>
<tr>
<td>SD</td>
<td>13.16</td>
<td>11.93</td>
<td>16.18</td>
<td>13.48</td>
<td>10.04</td>
</tr>
<tr>
<td>Difference (Proportion correct minus PTPE)</td>
<td>-16.33</td>
<td>-25.33</td>
<td>-19.93</td>
<td>-22.67</td>
<td>-21.07</td>
</tr>
<tr>
<td>Comparison group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion correct</td>
<td>.80</td>
<td>.83</td>
<td>.74</td>
<td>.74</td>
<td>.78</td>
</tr>
<tr>
<td>SD</td>
<td>.07</td>
<td>.07</td>
<td>.08</td>
<td>.11</td>
<td>.05</td>
</tr>
<tr>
<td>PTPE</td>
<td>65.86</td>
<td>66.82</td>
<td>59.65</td>
<td>57.27</td>
<td>62.40</td>
</tr>
<tr>
<td>SD</td>
<td>12.37</td>
<td>18.26</td>
<td>16.25</td>
<td>15.72</td>
<td>10.84</td>
</tr>
<tr>
<td>Difference (Proportion correct minus PTPE)</td>
<td>-14.14</td>
<td>-16.18</td>
<td>-14.35</td>
<td>-16.73</td>
<td>-15.60</td>
</tr>
<tr>
<td>Recovered patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion correct</td>
<td>.73</td>
<td>.78</td>
<td>.71</td>
<td>.74</td>
<td>.74</td>
</tr>
<tr>
<td>SD</td>
<td>.08</td>
<td>.10</td>
<td>.07</td>
<td>.09</td>
<td>.06</td>
</tr>
<tr>
<td>PTPE</td>
<td>62.33</td>
<td>60.53</td>
<td>59.17</td>
<td>52.00</td>
<td>58.51</td>
</tr>
<tr>
<td>SD</td>
<td>12.94</td>
<td>13.36</td>
<td>16.00</td>
<td>16.99</td>
<td>10.61</td>
</tr>
<tr>
<td>Difference (proportion correct minus PTPE)</td>
<td>-10.56</td>
<td>-17.06</td>
<td>-11.28</td>
<td>-21.65</td>
<td>-15.14</td>
</tr>
<tr>
<td>Greater negative bias in depression (difference in differences for the deprived vs. comparison group)</td>
<td>2.19</td>
<td>9.15</td>
<td>5.58</td>
<td>5.94</td>
<td>5.47</td>
</tr>
</tbody>
</table>

Note. Object-recognition = object recognition; GK = general knowledge; LJ = line-length judgment.

answered correctly), PTPE (the proportion of items or questions that participants retrospectively estimated that they had answered correctly), and the relation between these (PTPE minus proportion correct). Table 1 presents the means for these three measures, separately for each task, and for all of the tasks combined. Figure 1 shows the overall PTPEs and overall decision accuracy for each of the individual depressed patients, age-matched controls, and recovered patients.

Proportion Correct

The overall average proportion correct achieved by the comparison group (.78) modestly exceeded that of the depressed (.74) and recovered patient (.74) groups. A 3 (group) × 4 (task type) mixed-factor analysis of variance (ANOVA) showed a significant main effect of group, $F(2, 49) = 3.93$, $MSE = .003$, $p = .03$. Bonferroni-corrected post hoc comparisons showed that the overall performance of the comparison group modestly but not significantly exceeded that of the depressed ($p = .08$) and the recovered patient ($p = .06$) groups (effect size $d = .79$ and .84, respectively). There was also a significant main effect of task type, $F(3, 147) = 16.73$, $MSE = .006$, $p < .001$: the social judgment task showed the highest average proportion correct ($M = 0.81$), followed by object recognition ($M = 0.77$), line-length judgment ($M = 0.73$), and general knowledge ($M = 0.71$). There was no Group × Task interaction, $F < 1.4$. For all three groups, both for each task individually and for all of the tasks overall, the proportion correct significantly exceeded chance levels (smallest $t(14) = 6.99$, all $ps < .0001$).

Posttest Performance Estimate (PTPE)

The comparison group reported the highest overall PTPE ($M = 62\%$), whereas the depressed patients provided the lowest PTPE ($M = 53\%$). The PTPE of the recovered patients ($M = 59\%$) fell between that of these two groups. A 3 (group) × 4 (task type) mixed-factor ANOVA on the PTPE measure revealed a significant effect of group, $F(2, 49) = 3.59$, $MSE = .045$, $p = .04$. Bonferroni-corrected post hoc comparisons showed that the depressed patients gave a significantly lower retrospective overall estimate of their performance than did the comparison group ($p = .03$, effect size $d = .91$). The average PTPE given by the recovered patients did not differ significantly from either of the other groups.

One-sample $t$-tests against the level of performance expected by chance alone (50%, given a two-alternative task), showed that the overall PTPE of the comparison group significantly exceeded the chance estimate, $t(21) = 5.37$, $p < .0001$. This also was true for the overall
PTPE of the recovered patients, $t(14) = 3.11, p < .008$. By contrast, the depressed patients' estimates of their performance were not significantly above that anticipated merely by chance, $t(14) = 1.13, p = .28$.

In addition, task type had a significant main effect on the PTPE measure, $F(3, 147) = 7.89, MSE = .02, p < .001$. Participants gave lower PTPEs for the general knowledge and line-length judgment tasks than for the object recognition and social judgment tasks and this pattern: (a) held for all three groups ($F < 1$ for the group × task interaction), and (b) generally followed the differences in the actually obtained accuracy levels of the three groups for the four tasks (see Table 1). The highest level of accuracy was obtained for the social judgment task and this was exactly equivalent in the depressed and comparison groups (83% for both groups). Under these "naturally matched" performance conditions, the depressed patients evaluated their performance as 25% lower than it actually was, compared with a 16% underestimation by the comparison group, $F(1, 35) = 2.91, MSE = .03, p < .10$, effect size $d = .58$, reflecting a moderate-to-large effect size.

Pearson correlations between overall accuracy and overall PTPE showed modest positive correlations for each of the groups (depressed patients, $r = .36$, ns; comparison group, $r = .48, p < .05$; recovered patients, $r = .45, p < .10$). A numerical count of the number of times across the four tasks that participants gave PTPEs that were below the 50% that they could expect to achieve on the basis of chance alone showed that, in the comparison group, there were a total of 12/88 (14%) such observations provided by a total of 10 different people (two persons gave below chance PTPEs for two tasks). A very similar pattern was obtained for the recovered patient group, where there were a total of 8/60 (13%) such observations provided by eight different people (one person gave below chance PTPEs for two tasks, and one person did so for all four tasks). Below chance estimates were
numerically more frequent in the depressed patient group, where there was a total of 19/60 (32%) such observations, with three persons giving below chance PTPEs for two tasks, two persons giving below chance observations for three tasks, and one person doing so for all four tasks.

The retrospective estimates of performance for each participant also were compared with their actual performance (proportion correct) by computing the difference between their estimated and actual scores for each of the tasks (i.e., PTPE minus Overall Proportion Correct). For all three groups, the overall PTPEs were significantly ($p < .0001$) lower than their overall proportion correct: the PTPE of the comparison group (62%) was 16% below their average performance (78%); that of the depressed patient group (53%) was 21% lower than their average performance (74%); and that of the recovered patients (59%) was 15% lower than their average performance. A 3 (group) × 4 (task) ANOVA on the difference scores showed a main effect of task, $F(2.55, 125.20) = 3.38, MSE = .019, p = .03$ (df reported with the Greenhouse-Geisser correction because the assumption of sphericity was not met), with no main effect of group, $F(2.49) = 1.88, p = .16$, and no group × task interaction, $F < 1$. For all three groups the magnitude of the performance underestimate was numerically larger for the object recognition and line judgment tasks than for the social judgment and general knowledge tasks but underestimation was consistently demonstrated for each of the tasks (underestimations across the various tasks for the depressed patients varied between $-16$ and $-25$%, those for the comparison group between $-14$ and $-17$%; and those for the recovered patients between $-11$ and $-22$%). Across all tasks, the magnitude of performance underestimation in the depressed patients ($-21$%) was nonsignificantly greater (not less) than that for the comparison group ($-16$%), $F(1.35) = 2.99, MSE = .009, p = .09$, effect size $d = .58$, providing no evidence for depressive realism and partial support for selective processing. A similar outcome was obtained when the overall performance accuracy of the two groups was equated by excluding the four highest scoring individuals in the comparison group ($n = 18, M = 76$%) and the four lowest scoring individuals in the depressed group ($n = 11, M = 77$%); despite these matched levels of overall performance, depressed patients still showed a (nonsignificant) tendency to more strongly understate their actual performance (mean PTPE = 54%) than did the comparison group (mean PTPE = 61%), $F(1.27) = 3.55, MSE = .009, p = .07$, effect size $d = .73$, again reflecting a moderate-to-large effect size consistent with the selective processing account. Likewise congruent with the selective processing hypothesis, a $2 \times 4$ ANOVA on the difference between predicted and actual performance in these matched performance groups (i.e., PTPE minus proportion correct) showed that the magnitude of the underestimation was significantly greater in the depressed patients (23%) than in the comparison group (15%), $F(1, 27) = 4.61, MSE = 003, p = .04$, effect size $d = .83$, with no main effect of task, $F < 1.7$, and no group × task interaction, $F < 1$. Further, whereas the posttest estimates of this performance-matched comparison group still exceeded the chance level, $t(17) = 5.06, p < .0001$, those of the depressed patients again did not, $t(10) = 1.22, p = .25$.

**DISCUSSION**

Consistent with previous studies (e.g., Gigerenzer et al., 1991; Griffin & Tversky, 1992) the PTPE measure showed underconfidence. On this retrospective aggregate confidence measure, both clinically depressed individuals and their healthy age-matched controls systematically underestimated their actual performance accuracy on a wide range of decision tasks, including tests of episodic memory, social judgment, semantic memory, and perceptual discrimination. Contrary to the depressive realism account, there was no evidence that the clinically depressed patients offered more realistic estimations of their performance. Indeed, the underestimation was, if anything, exacerbated (and clearly not reduced) in the depressed patients relative to the comparison group. When the depressed group and comparison group were matched in level of overall performance, underestimation of the depressed patients relative to their actual performance significantly exceeded that of the comparison group. In addition, when the depressed patients were asked to evaluate or recall their performance after they had completed the tasks, unlike the overall estimate of the comparison group, and also unlike the overall estimates of the recovered patients, their average estimation of their performance did not significantly exceed the 50% level expected merely on the basis of chance. Therefore, rather than depressive realism, it is the selective processing hypothesis that offers a better interpretation of the results. Although the support for the selective processing hypothesis was only partial, the direction of the observed effects was consistently in the direction predicted by that hypothesis, and contrary to that proposed by depressive realism.

Systematic underestimation of performance also was found in persons who previously had been depressed but who, at the time of testing, no longer met clinical criteria for depression. Overall, the performance of the moderately dysphoric recovered patient group was somewhat intermediate between that of the comparison and depressed
groups, but generally was closer to that of the control group. The recovered patients’ overall accuracy on the four decision tasks was identical to that of the depressed patients yet, unlike the depressed patients but like the comparison group, the average PTPE of the recovered group significantly exceeded that expected on the basis of chance alone. In addition, the depressed patients showed at least a numerical tendency towards more frequent below-chance performance estimates for the individual tasks, where such estimates were more than twice as frequent in the depressed group (32%) than the other two groups (14 and 13% for the comparison participants and recovered patients respectively).

Although there has been considerable research aimed at investigating the validity of depressive realism, most of these studies included nonclinically diagnosed participants (e.g., Albright & Henderson, 1995; Dykman & Abramson, 1989; Johnson & DiLorenzo, 1998; Morris, 1996; Pacini, Muir, & Epstein, 1998), and they used the Beck Depression Inventory (BDI) as the only inclusion criteria. For instance, in the study of Johnson and DiLorenzo (1998), participants who had a BDI score of 9 or above were assigned to the “depressed group.” In some studies, participants meeting this criterion were called “dysphoric,” rather than depressed (Kendall, Hollon, Beck, Hammen, & Ingram, 1987). Although the BDI has been shown to be reliable at discriminating between groups of psychiatric inpatients and outpatients rated as exhibiting none, mild, moderate, or severe depression (Beck, 1967), it is a subjective inventory, compared to the HAM-D. As Haaga and Beck (1995) suggest, samples chosen from self-report depressive symptom measures alone contain many persons who do not meet diagnostic criteria for a depressive syndrome and are not severely depressed. Without a SCID interview, it is not easy to determine whether an individual is clinically depressed.

Gotlib (1981, 1983) conducted two studies with depressed and nondepressed psychiatric inpatients and nondepressed controls and examined their recall of evaluative feedback. Gotlib found that recall of the feedback of the depressed patients was significantly more negative than was actually the case, whereas that of the two nondepressed groups was relatively accurate. Koenig, Ragin, and Harrow (1995) examined the inference style of clinical depressives, using a proverb interpretation task in which self-generated and other-generated proverb interpretations were judged for their normative adequacy. Although both the depressed patients and the nondepressed participants were less accurate when judging themselves than when judging others, the depressed patients were significantly inaccurate in judging themselves (i.e., self-generated responses), in comparison with the nondepressed participants. This study’s outcomes are consistent with these and with other findings (e.g., Dobson & Pusch, 1995; McKendree-Smith & Scogin, 2000).

A characteristic shared by these experimental paradigms (including the present one) is the use of a “personally relevant” task or measure in the assessment of the degree of realism demonstrated by the depressed patients. The PTPE confidence measure, asking participants to recall and retrospectively evaluate their aggregate performance accuracy, may have tapped into participants’ self-schemata because the aggregate confidence judgment requires the use of abstraction, which is a process that has been shown to be disrupted in depression. It is possible that the output of abstraction is negatively distorted because of depressive schema (McKendree-Smith & Scogin, 2000). In addition, as noted in the Introduction, it has been proposed that the “reference class” for the PTPE judgment involves all of the individual’s experiences with a given type of content, including those that occurred prior to the experiment (e.g., in the case of the social judgment task, all of the person’s previous attempts to evaluate facial emotional expression in others). Overall, this study found that although (as expected) the comparison group tended to underestimate their performance on the PTPE measure, such underestimation was not eliminated in the depressed group (as would be expected on the depressive realism account) but was, instead, at least as great for each of the decision tasks used, and overall, after matching for level of actual performance, was exacerbated in the depressed.

Following Alloy and Abramson’s (1979) initial forwarding of the depressive realism hypothesis that, “depressed people are often more realistic or accurate in their perceptions and judgments than their nondepressed counterparts, who more frequently make distorted or biased judgments” (Ackermann & DeRubeis, 1991, p. 566), several studies have provided evidence supporting the hypothesis. However, for studies which included an objective measure of accuracy, although many (19) of the findings were consistent with the depressive realism hypothesis, almost as many (14) were not in line with the account (Ackermann & DeRubeis, 1991).

These mixed findings make it difficult to draw simple conclusions about depressive realism but, by and large, the evidence favors the alternative account of selective processing at least as often as the depressive realism account. The current study, in which all three important requirements for a test of depressive realism were met, including the use of a clinically diagnosed depressed sample, the use of tasks for which there was an objectively correct answer, and the inclusion of a measure on which normal controls showed either realism or underconfidence, weighs (modestly) on the side of selective processing.
Depression, Confidence, and Decision

Despite achieving an overall performance level that was well above chance (74%), the depressed patients' overall estimate of their performance was not higher than that expected on the basis of chance alone. Additionally, after matching the performance level of the comparison group with that of the depressed patients, the depressed patients' underestimations of their performance (PTPE minus proportion correct) were significantly greater than those of the comparison group. Hence, major depression did have a negative influence on self-evaluations of performance. Further studies should particularly attempt to explore the nature of the processes that contribute to such overall retrospective assessments of one's performance, and why and how they may be especially susceptible to error as a function of emotional disturbance.

REFERENCES


