Age Differences in Performance Awareness on a Complex Financial Decision-Making Task

Douglas A. Hershey
Oklahoma State University, Stillwater, Oklahoma, USA

Jo A. Wilson
Volant, Pennsylvania, USA

Individuals tend to be overconfident when making retrospective judgments about the quality of their decisions. However, few studies have focused on age differences in estimates of decision quality. In the present experiment performance estimates were provided by task-trained and untrained young and old individuals following completion of a series of complex financial decisions. Confidence levels were assessed by examining discrepancies between perceived and actual solution quality. Performance estimates of all 4 groups contained appreciable estimation error; however, no group showed a substantial directional bias toward underconfidence or overconfidence. Young trainees were significantly less confident in the quality of their decisions than young novices, but a comparable training effect was not found among older individuals. One's knowledge of the task, prior decision-making experience, and level of self-esteem may combine to determine the accuracy of one's retrospective performance estimates.

The ability to assess accurately the quality of one's own decision-making performance is an important metacognitive skill. As Evans (1989) pointed out, individuals display a tendency to make informal intuitive judgments about the quality of their decisions as a normative part of the decision-making process. Unfortunately, researchers have found that these intuitions are quite often inaccurate, inasmuch as individuals tend to make performance estimates that are systematically biased in their own favor. Across a wide variety of decision-making tasks, individuals' sub-

Received 20 December 1994; accepted 1 November 1996.

This work was supported by a grant awarded to David A. Walsh by the National Institute on Aging (RO1-AG08106). We thank Bert Hayslip and two anonymous reviewers for comments on an earlier draft.

Address correspondence to Douglas A. Hershey, Department of Psychology, Oklahoma State University, 215 North Murray Hall, Stillwater, OK 74078, USA.

Experimental Aging Research, 23:257–273, 1997

jective impressions of the quality of their performance typically exceed the quality of their actual performance (see Lichtenstein, Fischhoff, & Phillips, 1982, for a review). Although the direct effects of this metacognitive bias are difficult to quantify, the costs in business settings alone are potentially staggering. For instance, Russo and Schoemaker (1992) described one scenario in which managerial overconfidence cost a leading oil manufacturer, and in turn consumers, millions of dollars over a 5-year period. These authors suggested that losses of this magnitude that result from this prevalent decision bias are not at all uncommon.

From a theoretical perspective it is curious that individuals are not more accurate at evaluating the quality of their decisions, given that in most real world situations we receive some form of performance feedback. That is, one would expect that performance awareness (PA) would improve over time, as individuals realize that many of the ''good'' decisions they made are actually less than optimal. However, studies have consistently shown that perceptions of overconfidence are the norm rather than the exception; most of us suffer from a ''halo effect'' when evaluating the quality of our own cognitive efforts. From an applied perspective this is indeed a problematic state of affairs; as Devolder, Brigham, and Pressley (1990) pointed out, ''adequate performance monitoring is presumably essential for efficient and effective self-regulation of cognition and behavior'' (p. 291).

The difficulty individuals display in accurately assessing the quality of their decision-making performance has been referred to in the psychological literature as the *overconfidence effect*. Much of the knowledge about the overconfidence effect has emerged from studies of calibration accuracy, which typically require individuals to answer general knowledge questions and then assess the probability that each answer provided is correct. Although different techniques have been used to assess the accuracy of subjects' probabilistic estimates in studies of this kind (Lichtenstein et al., 1982), all are designed in some way to measure the degree of discrepancy between actual and perceived performance levels. As suggested above, in all but rare circumstances do perceptions of performance exceed actual performance levels.

A number of different aspects of the overconfidence phenomenon have been examined. For example, studies have shown that the difficulty of a decision task mediates individuals' confidence levels. Individuals tend to be quite overconfident when making difficult decisions (Lichtenstein et al., 1982; Pitz, 1974; Zakay & Glicksohn, 1992) and less confident (but overconfident nonetheless) when judging the quality of relatively easy decisions (Lichtenstein & Fischhoff, 1977). Experimental attempts have been made to improve performance awareness through practice and estimation training; however, these efforts have been met with mixed success

(Adams & Adams, 1961; Einhorn, 1980; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1980; Oskamp, 1962). Data from two published studies suggested that the nature of the cognitive operations required by the decision task can have an impact on confidence ratings (Sniezek, Paese, & Switzer, 1990; Zakay, 1985); data from at least two other studies indicated that confidence levels are not mediated by individual difference variables such as sex (Lichtenstein & Fischhoff, 1981) and intelligence (Lichtenstein & Fischhoff, 1977).

A second method of assessing PA involves asking individuals to solve a series of problems or make a number of decisions and then to provide a global rating of the overall quality of their performance (cf. Keren, 1987).[1] In studies of this kind, realism of confidence (Adams & Adams, 1961) is assessed by measuring the discrepancy (usually in the form of a difference score) between predicted performance and actual performance (Devolder et al., 1990). If an individual's predicted performance level is superior to his or her actual performance level, then the individual is said to be overconfident. If the opposite pattern is found to emerge, then the individual is classified as underconfident. In cases where there is no discrepancy between predicted and actual performance, the individual's confidence level is deemed to be appropriate given his or her level of performance on the task. It is important to point out that individuals may be found to display appropriate confidence levels regardless of the actual quality of their performance on the task. That is, confidence levels would be considered to be appropriate for those individuals who do poorly on a task and then judge the quality of their efforts to be poor. Similarly, confidence levels would be considered appropriate for those who turn in an excellent performance and then judge the quality of their work to be excellent.

In light of the substantial body of research that has been conducted on PA in decision contexts, we were surprised to find that only one study has focused on age differences in this metacognitive ability.[2] As part of a multifaceted study of PA, Devolder (1993) asked 24 young individuals (19–40 years) and 24 older individuals (59–84 years) to provide a global rating of the quality of their performance after solving seven legal and

[1] This particular method has been prevalent in the memory monitoring literature, where individuals are asked to evaluate the overall quality of their cognitive efforts following performance on a recall or recognition task.

[2] The lack of developmental research on PA in decision contexts can be contrasted with a substantial body of developmental work on PA in the memory monitoring literature (see Devolder et al., 1990, for a recent review). However, there are fundamental theoretical and methodological differences between studies of PA in memory monitoring and PA in decision making. Therefore, the memory monitoring work is not reviewed here.

seven financial problems.[3] Specifically, they were asked to estimate the number of problems they had correctly solved from among the 14 items on the test. In analyzing the estimation data, Devolder focused only on those individuals who displayed a metacognitive bias; therefore, 6 individuals (3 young, 3 old) who displayed perfect postdiction performance were excluded from the analysis. The remaining 21 members in each of the two age groups were then dichotomously classified as either underconfident or overconfident. Devolder found that the classification of individuals into these two performance categories differed as a function of age. Only 14% of younger adults overestimated the number of problems they correctly solved, whereas 86% underestimated their performance. In contrast, 71% of older adults overestimated their performance, and 29% were underconfident.

The younger adults' tendency toward underconfidence in the Devolder (1993) experiment is a curious finding in light of the robust overconfidence effect described above. Perhaps younger individuals in this study tended toward underconfidence because they had had little prior experience at solving legal and financial problems. Or, perhaps this unique finding was an artifact that was based on the dichotomous classification of individuals into one of two confidence groups (i.e., the magnitude of their estimation biases were not taken into account, but only the direction of their estimation errors was considered). In explaining the pattern of findings shown by older individuals, Devolder suggested that their tendency toward overconfidence may have been due to a general familiarity with financial and legal problems; however, data were not available to test the adequacy of this explanation. At any rate, the lack of parallel performance across age groups suggests the need for additional developmental research on PA in decision contexts. Of particular value would be studies that use sensitive (interval-level) techniques to measure the accuracy of individuals' performance estimates. Until such studies are carried out, the pattern of age differences found in the Devolder study, although interesting, should be viewed as tentative.

On the basis of what is known about normative changes in adult intellectual abilities—that is, age-related increases in knowledge of oneself and the world (Labouvie-Vief, 1992) and declines in basic processing abilities (Horn & Hofer, 1992)—it would not be unreasonable to expect to find developmental differences in this important metacognitive skill. On the one hand, it could be argued that older adults should be more accurate at estimating the quality of their decision performances than younger adults, because they witnessed a lifetime of decision outcomes across a variety of tasks and decision domains. On the other hand, it has been clearly demonstrated that numerous basic information processing abilities decline with advancing age (Craik & Salthouse, 1992; Salthouse, 1982, 1991), which in turn, one might argue, could have a deleterious effect on the processing mechanisms that underlie performance assessments. Specifically, Charness and Bieman-Copland (1992) have suggested that age-related declines in working memory may limit the ability of older adults to monitor ongoing performance results, which may lead them to have a skewed perception of the overall quality of their efforts.

The data presented in this article were collected as part of a larger study of developmental differences in complex decision making conducted at the University of Southern California (USC). Our primary goal was to evaluate age differences in people's ability to assess the quality of their decisions. Toward this end, participants were required to evaluate the quality of their performance following the completion of a cognitively complex, financially oriented retirement planning task. Discrepancies between actual and perceived decision quality were evaluated using the difference score approach described above. A secondary goal of the study was to determine the extent to which knowledge of the decision domain mediated perceptions of decision quality. This was accomplished by comparing the perceived decision quality accuracy of task-trained (knowledgeable) and untrained (novice) participants in each of the two age groups.

METHOD

Participants

The young group (M = 18.6 years, SD = .92) consisted of 28 undergraduate students (14 trained, 14 untrained) attending USC. Their participation was solicited through fliers left in dormitory mailboxes and notices posted on campus bulletin boards. The old group (M = 71.1 years, SD = 6.28) consisted of 32 USC alumni or their spouses (18 trained, 14 untrained) who were recruited through an older adult subject pool maintained by the psychology department. The mean age of novices and trainees was comparable in both the young group and the old group. Mean educational levels were found to differ across age groups; those in the young group had completed an average of 13.9 years of formal education, and those in the old group had completed 16.2 years of school-

[3] All references to the Devolder (1993) study refer solely to the between-subject (young–old) postdiction-only comparison. Although she separately assessed the effects of prediction and postdiction using both between- and within-subject designs, the between-subject–postdiction comparison is most similar to the design used in the calibration accuracy literature, and it is identical to the design used in the present experiment.

ing, on average. All participants received $5 per hour for their participation, which was paid on completion of the experiment.

Procedure

Trainees attended two 3-hr group training sessions that focused on issues related to financial planning for retirement, prior to attending a third, decision-making (test) session. Participants were tested individually at the decision-making session, which involved solving a series of six retirement investment problems. For each of the six problems, participants had to decide how much money a hypothetical individual should contribute to an employer-sponsored 401k retirement savings plan. The investment decisions were made one at a time, after information related to the present and anticipated future financial situation of the hypothetical investor had been considered. No feedback on the quality of participants' performance was provided at any time during the test session. Space limitations preclude a detailed description of this complex task; however, additional information regarding the task and methodology can be found elsewhere (Hershey, 1990; Hershey, Walsh, Read, & Chulef, 1990; Walsh & Hershey, 1993). After completing the series of problems, participants rated the overall quality of their six solutions using a 7-point Likert-type scale (1 = *very poor solutions;* 7 = *very good solutions*).

To validate the overall efficacy of the training, the performance of trainees was compared to novices on a 32 item test designed to assess knowledge of three different areas of retirement and financial planning: general financial knowledge (e.g., understanding how compound interest accrues; knowing the current rate of inflation and prime interest rate); knowledge specific to financial aspects of retirement (e.g., likely sources of income in retirement; how inflation affects purchasing power over time); and knowledge associated with various types of retirement investment vehicles (e.g., how tax deferred investment vehicles operate; limits on deposits to 401k accounts and penalties for early withdrawals). Trainees' scores on this measure (M = 68%, SEM = 1.63) revealed that they were significantly more knowledgeable about financial and retirement planning than novices (M = 44%, SEM = 1.77), $t(58) = 9.72$, $p <$ .01, which provided empirical support for the effectiveness of the training program.

Computation of PA Scores

PA scores were calculated to assess the degree of discrepancy between participants' perceived decision quality and the actual quality of their decisions. Perceived decision quality scores were based on the Likert scale ratings. The actual decision quality score, in contrast, was the absolute value of the difference between participants' recommended investment amounts and the optimal investment amounts,[4] summed over the six problems. Thus, the actual decision quality marker represents an error score, based on an aggregate of unsigned deviations from the optimal solutions. Mean actual and perceived decision quality scores for each of the four groups are shown in Table 1. PA scores were then derived using the difference score method described above, by first converting both the perceived decision quality distribution and the actual decision quality distribution into standard score units (z scores) and then taking the difference between these two values for each individual.

Prior to computing the PA scores, the raw score distributions for the actual and perceived decision quality measures were plotted and inspected for the entire sample and for each of the four subgroups. Because the PA score (the critical dependent measure) is based on the difference between standardized actual and standardized perceived performance scores, it

Table 1. Actual and perceived decision quality scores as a function of age group and training status

	Decision quality			
	Actual ($)		Perceived[a]	
Group	*M*	*SD*	*M*	*SD*
Novices				
Young	29,353	29,560	4.93	1.07
Old	20,070	19,441	4.43	0.65
All novices	24,711	25,001	4.68	0.90
Trained				
Young	2,710	3,330	5.00	0.96
Old	19,812	19,113	4.72	1.18
All trained	12,330	16,712	4.84	1.08
Total *M*s	18,108	21,716	4.77	1.00

[a] Scored on a 7-point Likert-type scale.

[4] Each of the six problem scenarios was designed to have a single, optimal investment amount. As a part of the initial task analysis process, three expert financial planners were engaged as consultants to review each of the hypothetical scenarios and recommend what they believed to be the optimal investment values for each of the six problems. All three experts independently generated the same investment amount for five of the six problems, and two of the three experts generated an identical investment amount for the sixth problem. This consensus value reached by the two experts was used as the optimal investment amount for that sixth problem. See Hershey (1990) for additional details regarding this validation process.

was important to establish that the distributional properties for these two variables approximated normality prior to transformation. The skew, kurtosis, and range of scores were judged to be within reasonable limits for each of these distributions. Moreover, no unreasonably large outliers were identified in either the actual or perceived decision quality distributions, which could have distorted either the PA distribution or the PA score for any one individual.

The resulting PA scores could take on either positive or negative values. Positive scores indicate a tendency toward overconfidence, and negative scores indicate a tendency toward underconfidence. A PA score of zero indicates that an individual's perceived decision quality estimate was accurate given the actual quality of his or her solutions.

RESULTS

Absolute Value of Postdiction Errors

Initial analysis were carried out to examine the overall accuracy of participants' postdiction estimates. This was accomplished by comparing the *magnitude* of each group's mean postdiction errors to a theoretical population mean of zero. (Recall that a score of zero is indicative of perfect postdiction performance, suggesting no difference between participants' perceptions of performance and the actual quality of their decisions.) In conducting this analysis, we first took the absolute value of each participant's PA score (hereinafter referred to as *absolute PA scores*) and then calculated group means on the basis of these unsigned PA values.[5] Then, four different one-sample t tests were calculated. Two of these tests compared the means for each age group (collapsed over training conditions) against zero, and the other two compared each training condition (collapsed over age groups) against zero. Each of these four comparisons were found to be statistically significant: young participants (M = 1.06, SEM = 0.17), $t(27)$ = 6.30, $p < .01$; older participants (M = 0.97, SEM = 0.13), $t(31)$ = 7.72, $p < .01$; novices (M = 0.96, SEM = 0.13), $t(31)$ = 7.57, $p < .01$; and trainees (M = 1.07, SEM = 0.17), $t(27)$ = 6.43, $p < .01$. Taken together, this set of findings reveals that members of both age groups and both training groups made appreciable errors in estimating the quality of their decisions.

We then conducted two independent group t tests in order to determine whether the magnitude of participants' errors differed as a function of age or training. These tests contrasted (a) the absolute PA scores for the two levels of the age factor and (b) the absolute PA scores for the two levels of the training factor. The mean difference between age groups was small (M_{diff} = 0.09) and not significant, $t(58)$ = 0.42, *ns*, as was the difference between trained and untrained participants (M_{diff} = 0.12), $t(58)$ = 0.56, *ns*. These findings indicate that (a) the two age groups made comparable errors when estimating the quality of their performance, and (b) the two training groups also made comparable postdiction errors.

[5] Basing the group means on unsigned PA scores allowed us to examine the issue of how large participants' errors were (in absolute terms) without respect to whether the errors were positive (indicating overconfidence) or negative (indicating underconfidence).

Four additional one-sample t tests were calculated that compared the absolute mean PA scores for each age by training condition subgroup against zero. The mean scores for each of the subgroups were found to be significantly different from the perfect postdiction value: young novices (M = 1.34, SEM = 0.29), $t(13)$ = 4.66, $p < .01$; young trainees (M = 0.77, SEM = 0.15), $t(13)$ = 5.19, $p < .01$; older novices (M = 0.81, SEM = 0.15), $t(13)$ = 5.41, $p < .01$; and older trainees (M = 1.10, SEM = 0.19), $t(17)$ = 5.80, $p < .01$, indicating that the estimates for each of the subgroups contained appreciable error.[6]

Role of Age and Knowledge

Analyses were carried out that were designed to reveal whether age and knowledge of the task were associated with biases toward overconfidence or underconfidence. A 2 (age group) × 2 (training condition) analysis of variance (ANOVA) was conducted using participants' PA scores as the dependent measure. Unlike the absolute PA scores used in the above analyses, in this analysis unadjusted PA scores were used, representing both positive and negative values indicative of overconfidence and underconfidence, respectively. The main effect for age group was not significant, $F(1, 56) < 1$, nor was the effect of training, $F(1, 56)$ = 1.48, *ns*, MSE = 1.57. However, the age group by training interaction was found to be statistically significant, $F(1, 56)$ = 4.89, $p < .05$, MSE = 1.57. The mean PA scores for each of the four groups are shown in Figure 1. A visual inspection of these means reveals that young novices made the largest average errors (M = 0.68, SEM = 0.43); they had a tendency to overestimate the quality of their decision. In contrast, young trainees (M = −0.48, SEM = 0.23) tended to underestimate the quality of their performance, and both older trainees (M = 0.03, SEM = 0.33) and older novices (M = −0.25, SEM = 0.26) appeared to make reasonably accurate estimates. In an effort to determine

[6] To ensure that the analysiswise alpha rate did not exceed the nominal alpha level, we made Bonferroni adjustments to the p levels for each of the 10 t tests, which used absolute PA scores as the dependent measure.

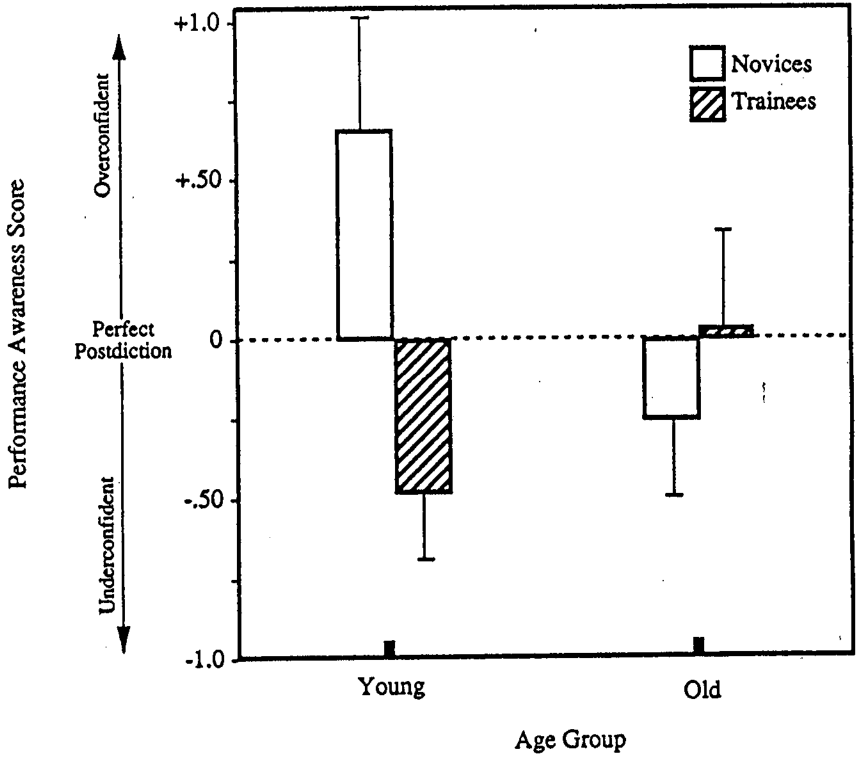

Figure 1. Mean performance awareness (PA) scores (in standard score units) and standard errors plotted as a function of age and training. These scores are based on the difference between standardized perceived solution quality scores and standardized actual solution quality values. Positive PA scores indicate overconfidence, and negative scores indicate underconfidence.

which group differences were responsible for the significant two-way interaction, post hoc pairwise comparisons were calculated using the Newman–Keuls procedure. The only statistically significant contrast was the comparison between young novices and young trainees ($p < .05$); none of the other contrasts were found to exceed the critical difference threshold.

Mean Postdiction Errors

One-sample *t* tests were used to compare the PA scores for each of the four subgroups against zero, in an effort to determine whether the groups

were significantly biased toward underconfidence or overconfidence.[7] The mean scores used in these four analyses are the same mean scores shown in Figure 1. None of these means were found to be significantly different from zero: young novices, $t(13)$ = 1.58, *ns*; young trainees, $t(13)$ = 2.12, *ns*; older novices, $t(13)$ = 0.95, *ns*; and older trainees, $t(17)$ = 0.10, *ns*.[8] These findings suggest the lack of a clear directional estimation bias for each of the four groups, despite the earlier reported finding that all four groups made appreciable errors in judging the quality of their decisions.

Ancillary Analyses

Additional analyses were conducted to determine whether participants' educational background or level of interest in the task could have contributed to the observed age differences in estimation accuracy. Toward this end, the number of years of formal education participants had completed was correlated with their PA scores. (Recall that the older participants had completed approximately two more years of education than younger ones.) The resulting correlation coefficient (r = −.10) was small and not statistically significant, suggesting that level of formal education, in and of itself, was not sufficient to account for the observed age differences in estimation accuracy. Next, a 2 (age group) × 2 (training condition) ANOVA was conducted using 7-point Likert scale responses to the following question as the dependent measure: *How interesting was it for you to work on the six problems*? This analysis failed to reveal main effects of age, training, or an interaction effect, which would suggest that participants' level of interest in the task differed. Moreover, interest levels and PA scores were not found to be correlated (r = .11), which indicated that the observed group differences could not be explained on the basis of differential amounts of task-specific motivation.

DISCUSSION

The findings from the present experiment are consistent with previous studies of PA, inasmuch as participants were shown to be poor judges of

[7] Both these analyses and the previous ANOVA and Newman–Keuls analyses are similar in that they focus on the extent to which the subgroups show directional biases in their estimates. However, the two sets of analyses are conceptually distinguishable. In the previous analysis, the mean estimation performance of each subgroup was compared to the grand mean (in the ANOVA) and to each of the other subgroups (in the Newman–Keuls analysis), whereas in the present set of analyses, the PA scores are being compared to the theoretically based "perfect prediction" mean of zero.

[8] Bonferroni adjustments were made for these four comparisons to ensure that the analysiswise alpha rate did not exceed .05.

the quality of their decisions. This conclusion is based on the set of findings that revealed that the absolute PA scores were significantly different from zero for each of the four subgroups. Conceptually, these comparisons indicate that the various groups' estimation errors were large, relative to the perfect postdiction value of zero (i.e., without taking into account directional biases). It is important to note, however, that no differences were identified in the absolute magnitude of participants' errors across the two levels of the age group factor or the two levels of the training factor. In other words, the absolute magnitude of errors made by older participants were equivalent to those made by younger ones, and the overall quality of estimates made by novices were equivalent to those made by trainees.

However, further analyses that took into consideration both the magnitude and direction of participants' estimation errors revealed a more complex pattern of performance. Specifically, a significant Age Group × Training Status interaction was found in the two-factor (Age × Training) ANOVA using the standard PA scores as the dependent measure. Post hoc comparisons indicated that young novices were significantly more confident in their decisions than young trainees but the confidence levels of older participants did not differ as a function of training status. One interpretation of these findings is that individuals who possess little knowledge about a decision domain and relatively little experience with decisions (i.e., young novices) tend to be overconfident, presumably because they do not appreciate the true complexity of the decision at hand. In contrast, individuals who possess knowledge about the decision domain and relatively little experience with decisions (i.e., young trainees) tend to be underconfident, presumably because they are overwhelmed by the complexity of the task and the intricacies of the decision domain. If, on the other hand, individuals possess a good deal of experience with decisions (i.e., older participants in both training conditions), then it appears that they are able to judge fairly accurately the quality of their decision performance irrespective of their level of domain-specific knowledge.

It is tempting to conclude that the interaction observed in the ANOVA is the result of differential life decision-making experiences. That is, older participants have had the benefit of witnessing a lifetime of decision outcomes, may of which were consistent with their perceptions of performance, and many of which were not. These countless learning experiences may have led older individuals to develop more realistic information processing strategies for assessing the quality of their decisions. Conversely, knowledge of the decision domain may play a more salient role in influencing performance estimates for individuals who lack the

benefit of extensive decision experience (i.e., younger persons). Evidence for this assertion can be found in the significant difference in the quality of estimates produced by young trainees and young novices.

A number of recent studies have suggested that experience is a necessary and sufficient condition for skill acquisition (cf. Charness, 1989; Salthouse, 1987). Bearing this in mind and recognizing that metacognitive abilities are indeed an acquired skill, one might be inclined to conclude that awareness of performance should show general age-related improvements across a variety of judgment and decision-making situations. Such an interpretation of the data must be viewed as tentative, however, in light of the fact that the strategies individuals use to make retrospective performance estimates are not well understood (see Keren, 1987, for a speculative discussion of this issue).

In addition to the above interpretation of the interaction, it is also possible that individuals' situationally based levels of self-esteem could have contributed to the observed effect. As one reviewer pointed out, strong confidence in one's decision processes is the mark of an individual who possesses a normal (healthy) level of self-esteem. Therefore, according to this interpretation, the relative overconfidence exhibited by young untrained individuals reflected an optimistic, psychologically adaptive view of their own cognitive capabilities. In keeping with this explanation, older participants' levels of self-esteem may have been situationally reduced during the course of the test session. It is not uncommon for older adults to experience state anxiety when given tests of memory and cognition (Kausler, 1990) and to become nervous, tense, and encounter feelings of cognitive inadequacy during an evaluative test session (Lezak, 1995). Outside the laboratory these same older adults might normally and adaptively view the quality of their decisions as better than average. Thus, the relative accuracy of the judgments exhibited by older adults in this study could have been due to a conservative self-evaluation bias that stemmed from reactivity to the task and concerns about their cognitive capabilities. Finally, young untrained participants may have tended toward underconfidence in order not to appear egotistical or presumptuous. They, like the older participants, may have felt performance pressure because they had just completed 6 hr of training moderated by the experimenter. This influence, coupled with a lack of familiarity and experience with the decision domain, may have led them to question the extent of their decision-making abilities, and ultimately, to underestimate the quality of their work. Older trainees might not have generated similarly underconfident judgments given that they presumably were more familiar with the decision domain. Of course, it is quite possible that both of these interpretations of the data are correct and that the observed pattern of findings reflect the combined influences of one's knowledge of the task, life-long decision experiences, and situationally based level of self-esteem.

Regardless of which interpretation of the interaction one prefers, the above discussion must be qualified by the set of findings that revealed that the mean PA scores for each of the four age by training subgroups were not significantly different from zero. In other words, none of the four groups showed a statistically significant directional bias in its estimates, despite the finding that the magnitude of the absolute PA error scores were large. These seemingly contradictory findings can be understood by considering the nature of the mean PA scores. Unlike the absolute PA scores, which were based on unsigned error values, the mean PA scores were derived by averaging both positive and negative error values. This in turn can create a situation where an overall mean PA score could be small (near zero), despite appreciable individual errors in both directions. Although the process of averaging positive and negative scores can lead to some conceptual confusion (Devolder et al., 1990), one might arguably conclude that calculating mean PA scores in this fashion provides the most accurate reflection of a group's directional bias.

It is difficult to make a direct comparison between the age-related estimation performance of participants in the present study and participants in the Devolder (1993) study. On a superficial level, the findings from the two studies appear to contradict one another. In the Devolder study, older adults were found to be fairly overconfident, and younger adults were substantially underconfident. In the present study, older untrained adults were somewhat underconfident, and younger untrained adults were somewhat overconfident. Possible reasons for these equivocal findings may involve differences in the methodologies used to compute estimation errors, differences in the way the data were analyzed, and differences in the nature of the experimental tasks. In the Devolder study participants were asked to estimate the number of problems they correctly answered, whereas in the present study participants rated the quality of their performance using a Likert-type scale. Perhaps the dissimilarity in the nature of these two types of ratings led to differences in self-perceptions of the quality of performance. A second distinction between the two studies involved differences in the analysis methods. In the Devolder study participants were classified as either underconfident or overconfident, and the focus of the analysis was on the proportion of individuals who were classified into either group. In the present study estimation accuracy was treated as a continuous variable. In this regard, Devolder posed the question: "Were younger and older adults overconfident, or underconfident?" In the present study we sought an answer to a slightly different question, namely, "By how much do older and younger adults differ in their ability to estimate the quality of their decisions?" Finally, in the Devolder study both legal and financial problems were used, whereas the present study focused solely on one particular type of financial problem. Given these important methodological differences, it is not surprising that there were differences in the findings of the two studies. This suggests that the nature of the age-related effects that we might expect to find in future studies of this kind may to a large extent depend on the precise nature of the task employed and the specific methods used to assess performance awareness.

It would appear that some of the issues that make it difficult to reconcile the findings between the Devolder (1993) study and the present experiment are the same issues that have plagued the study of age differences in memory monitoring (a literature fraught with equivocal findings across studies). In a review of that literature, Devolder et al. (1990) suggested a number of different factors that could account for the presence or absence of age differences in performance awareness:

> A substantive interpretation of the discrepant findings is that there are Age × Task interactions for memory monitoring; that is, age differences are more likely on some tasks than others. Alternatively, however, the discrepancies could be attributed to sample differences between studies or methodological flaws in some of the designs. General conclusions cannot be drawn with confidence from any one study, and drawing conclusions across studies is difficult because of population, procedural, and analyses variability. (pp. 291–292)

This suggests that future developmental studies of performance awareness in decision making should be carefully planned to follow logically from existing work so as to maximize the likelihood of obtaining interpretable findings.

The generalizability of the present findings are to some extent limited by the fact that participants in the present sample (particularly the older adults) tended to be more highly educated than the general population. However, the finding that PA scores were uncorrelated with educational level suggests that for the present decision task, prior educational experiences may not play an influential role in mediating performance awareness. Nonetheless, controlled studies that systematically explore the influence of formal education on this specific form of metacognitive performance would be a valuable contribution to the literature, inasmuch as it has been suggested that a relationship could exist between education and metacognitive abilities (Charness & Bieman-Copland, 1992).

Finally, the conclusions that can be drawn from the present study regarding developmental differences in PA are necessarily limited because only two disparate age groups were sampled. However, the findings from this experiment suggest that future studies are warranted that further examine the link between aging and PA in decision contexts. Of particular value would be studies which use multiple age groups from across the adult life span and designs that require individuals to engage in multiple, independent judgment and decision-making tasks. Comprehensive experiments such as these could help to further reveal the developmental trajectory of age differences in estimation performance and the extent to which this metacognitive skill generalizes across decision domains.

REFERENCES

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgements. *Psychological Review, 68,* 33–45.

Charness, N. (1989). Age and expertise: Responding to Talland's challenge. In L. W. Poon, D. C. Rubin, & B. A. Wilson (Eds.), *Everyday cognition in adulthood and old age* (pp. 437–456). New York: Cambridge University Press.

Charness, N., & Bieman-Copland, S. (1992). The learning perspective: Adulthood. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 301–327). New York: Cambridge University Press.

Craik, F. I. M., & Salthouse, T. A. (1992). *The handbook of aging and cognition.* Hillsdale, NJ: Erlbaum.

Devolder, P. A. (1993). Adult age differences in monitoring of practical problem-solving performance. *Experimental Aging Research, 19,* 129–146.

Devolder, P. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults. *Psychology and Aging, 5,* 291–303.

Einhorn, H. J. (1980). Overconfidence in judgement. *New Directions for Methodology of Social and Behavioral Science, 4,* 1–15.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

Hershey, D. A. (1990). The role of knowledge and experience in structuring problem solving performance (Doctoral dissertation, University of Southern California, 1990). *Dissertation Abstracts International, 51,* Section 7B.

Hershey, D. A., Walsh, D. A., Read, S. J., & Chulef, A. S. (1990). The effects of expertise on financial problem solving: Evidence for goal directed problem solving scripts. *Organizational Behavior and Human Decision Processes, 46,* 77–101.

Horn, J. L., & Hofer, S. M. (1992). Major abilities and development in the adult period. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 44–99). New York: Cambridge University Press.

Kausler, D. H. (1990). Motivation, human aging, and cognitive performance. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (3rd ed.). New York: Academic Press.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39,* 98–114.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118.

Labouvie-Vief, G. (1992). A neo-Piagetian perspective on adult cognitive development. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 197–228). New York: Cambridge University Press.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 149–171.

Lichtenstein, S., & Fischhoff, B. (1981). *The effects of gender and instructions on calibration* (Report No. PTR-1092-81-7). Eugene, OR: Decision Research.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs, 76* (28, Whole No. 547).

Pitz, G. F. (1974). Subjective probability distributions for imperfectly known quantities. In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 29–41). Potomac, MD: Erlbaum.

Russo, J. E., & Schoemaker, P. H. (1992). Managing overconfidence. *Sloan Management Review, 33,* 7–17.

Salthouse, T. A. (1982). *Adult cognition: An experimental psychology of human aging.* New York: Springer-Verlag.

Salthouse, T. A. (1987). The role of experience in aging. In K. W. Schaie (Ed.), *Annual review of gerontology and geriatrics* (Vol. 7). New York: Springer.

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging.* Hillsdale, NJ: Erlbaum.

Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46,* 264–282.

Walsh, D. A., & Hershey, D. A. (1993). Mental models and the maintenance of complex problem-solving skills in old age. In J. Cerella, J. Rybash, W. Hoyer, & M. L. Commons (Eds.), *Adult information processing: Limits on loss* (pp. 553–584). New York: Academic Press.

Zakay, D. (1985). Post-decisional confidence and conflict experienced in a choice process. *Acta Psychologica, 58,* 75–80.

Zakay, D., & Glicksohn, J. (1992). Overconfidence in a multiple-choice test and its relationship to achievement. *Psychological Record, 42,* 519–524.