© *Academy of Management Journal*
1997, Vol. 40, No. 1, 189–204.

RETROSPECTIVE REPORTS IN ORGANIZATIONAL RESEARCH: A REEXAMINATION OF RECENT EVIDENCE

C. CHET MILLER
Baylor University
LAURA B. CARDINAL
Duke University
WILLIAM H. GLICK
Arizona State University

Retrospective reports of important organizational phenomena are commonly used in strategic management and organization theory research. A recent study, however, suggested that these reports tend to be inaccurate and seriously questioned their use. Our reexamination of this study suggests the situation is not so dire. Our work suggests that retrospective reporting is a viable research methodology *if* the measure used to generate the reports is adequately reliable and valid. Retrospective reports should neither be rejected nor used indiscriminately.

Retrospective reports are popular tools for learning about the past. Such reports are used in courtroom proceedings, journalistic interviews, congressional hearings, and many other investigatory endeavors. In organizational research, retrospective reports have been used extensively in studies of decision making (e.g., Bourgeois & Eisenhardt, 1988; Mintzberg, Raisinghani, & Theoret, 1976), organizational change (e.g., Huber & Glick, 1993; Kanter, 1983), and competitive strategies (e.g., Feeser & Willard, 1990; Zajac & Shortell, 1989).

Despite the popularity of retrospective reports, many researchers believe that problems associated with informant fallibility strongly influence these reports. The primary problem is that key informants may not be able to accurately recall the past. As Golden (1992), Huber and Power (1985), Wolfe and Jackson (1987), and many others have suggested, inaccurate recall in retrospective reporting can result from inappropriate rationalizations, oversimplifications, faulty post hoc attributions, and simple lapses of memory. A secondary problem is that key informants may try to present a socially de-

We thank Janice Beyer, Brian Golden, Loren Gustafsen, and the anonymous reviewers for their helpful input. We also send our thanks to the late G. Douglas Jenkins, Jr., for his help. Finally, we thank Susan Fox-Wolfgramm for the use of her data.

sirable image of themselves or their firms (Golden, 1992; Huber & Power, 1985).

In the strategic management and organization theory areas, the sources of retrospective reports are typically chief executive officers. In some studies (e.g., Bourgeois & Eisenhardt, 1988), retrospective data from chief executives are combined with data from other upper-echelon executives, but in many studies (e.g., Zajac & Shortell, 1989) only chief executives provide retrospective data. Thus, understanding the degree of inaccuracy that tends to exist in chief executive retrospective reports is critical.

In a recent empirical study, Golden (1992) examined chief executive accuracy and reported that only 42 percent of chief executives accurately selected the competitive strategies their firms were using two years prior to the retrospective reports. In Golden's words, "Nearly 60% of the retrospective reports . . . did not agree with the validated reports elicited only two years earlier" (1992: 852). Golden's study is important because it is the only study that systematically examines the issue of retrospective accuracy using a large sample of chief executives. Further, despite Golden's warning to treat his conclusions as tentative, the study has been cited a substantial number of times by authors arguing that retrospective reports are dangerous and should be avoided or treated with extreme caution. Priem and Harrison, for example, cited Golden to support the conclusion that retrospective reports suffer from "severe problems of recollection" (1994: 318). Priem and Harrison subsequently dropped retrospective reports from their review of methodologies for eliciting strategic information. In another example, Kumar, Stern, and Anderson cited Golden to support the position that informant reports and actual events may exhibit "little correspondence" (1993: 1634). We view Golden (1992) as a seminal study that has helped to define the key territory for the discussion of retrospective accuracy.

The purpose of this brief research note is to reexamine the evidence supporting Golden's conclusion that "use of retrospective accounts in management research needs to be seriously questioned" (1992: 857). Given the persistent popularity of retrospective reports and given the importance of Golden's study, we considered it important to reexamine his evidence. Our reexamination focused on three methodological issues: (1) the use of an overly pessimistic statistic to assess accuracy, (2) the use of a fairly weak questionnaire measure, and (3) the failure to separate error due to unreliability of that measure from error due to informant fallibility. After taking into account these three issues, our reanalysis of the Golden data suggests that error due to informant fallibility was not excessive, but error due to the measure used in generating the retrospective reports was excessive. Thus, our reanalysis suggests that retrospective reporting is a viable research methodology if the measure used to generate reports is adequately reliable and valid.

We make two contributions in this note. First, we provide evidence that the only major study of CEO retrospective accuracy yielded overly pessimistic results. Second, we reduce the likelihood that Golden's (1992) pessimis-

tic results will be used as a basis for indiscriminate rejection of retrospective reports. We argue, however, against the use of our more optimistic results as a basis for indiscriminate acceptance of retrospective reports. Evidence of validity and reliability should be reported routinely for any measurement approach, including retrospective reporting.

ASSESSMENT OF ACCURACY

The first methodological issue concerns the use of percentage of agreement (percent agreement) as an indicator of accuracy. Golden (1992) reported the percentage of chief executives who selected the same strategy in a nonretrospective report in 1984 and in a retrospective report in 1986. Unfortunately, percent agreement has two major shortcomings as an indicator of accuracy. First, it does not adjust for chance agreement (Cohen, 1960). Thus, it is sensitive to the number of coding categories used. Second, and most important, it corresponds to the sum of joint probabilities that indicate whether an organization is placed into the same category at time 1 and time 2. The real issue, however, is the probability of an organization's being placed into an appropriate category at time 2. Because of the focus on joint probabilities, percent agreement ordinarily understates the actual probability of appropriate time 2 classifications. A simple example illustrates the problem. Assume that a classification procedure for a three-category classification project has a .8 probability of placing an organization into the appropriate category. Random error prevents the probability from being 1.0. If the procedure is applied twice in a simple test-retest study, the probability of placing an organization into the appropriate category at time 1 is .8, and the probability at time 2 is .8, but the probability of placing an organization into the appropriate category at *both* time 1 *and* time 2 is .64 (.8 × .8). Similarly, the probability of placing an organization into either of the two inappropriate categories might be .1. Then, the probability of placing the organization into the first inappropriate category at *both* time 1 *and* time 2 is .01 (.1 × .1), and the probability of placing an organization into the second inappropriate category at *both* time 1 *and* time 2 is .01 (.1 × .1). Percent agreement for a sample of organizations would be on average .66 (.64 + .01 + .01), but the probability of appropriate time 2 classification is .8.

An Appropriate Statistic

Although percent agreement is inappropriate as an accuracy statistic, other indices have been proposed and are very useful when dealing with the type of categorical data used by Golden (1992). Because accuracy can be framed in terms of reliability or validity and because different indices have been proposed for these two approaches, we had to make a choice between reliability and validity (cf. Brennan & Prediger, 1981).

Intertemporal reliability concerns the extent to which a chief executive places his or her organization in the same category when nonretrospectively choosing a category at time 1 and when retrospectively choosing a time 1

category at time 2. Neither the category chosen at time 1 nor the category chosen at time 2 is assumed to be the true category: a chief executive selecting the wrong category at time 1 and then picking the same category at time 2 would contribute positively to intertemporal reliability.[1] Retrospective validity, however, concerns the extent to which a chief executive places his or her organization in the true time 1 category when assessing time 1 at time 2. To obtain high retrospective validity, the true time 1 category must be chosen at time 2; to obtain high intertemporal reliability, the same category selected at time 1 must be selected at time 2. In each case an appropriate category must be chosen at time 2, but the definition of appropriateness differs.

To assess intertemporal reliability for retrospective classifications, a researcher must know the time 1 classifications made by raters. Importantly, if the overall measurement procedure used to classify organizations into categories is not more valid at time 2 than at time 1 (and it usually is not), then intertemporal reliability sets the upper bound for retrospective validity. Thus, intertemporal reliability can typically be viewed as an indirect indicator of retrospective validity.

To assess retrospective validity, a researcher must have very good information concerning the true time 1 categories for the organizations. Otherwise, the accuracy of choices made at time 2 cannot be directly assessed. To know the classifications made by raters at time 1 is not enough, unless those classifications have been made with little or no error.

Golden (1992) initially framed his main analysis as a study of intertemporal reliability, but he subsequently argued that his time 1 data were substantially error-free and could be used to assess the validity of chief executive retrospective reports. Thus, he interpreted lack of intertemporal agreement as evidence of retrospective errors in reporting actual 1984 strategies: "The current findings indicate substantial retrospective errors, not errors in reporting strategies during the first phase of data collection" (1992: 851). He based his argument for substantially error-free time 1 data on the evidence of convergent validity reported by Shortell and Zajac (1990) for the time 1 observations. Although 24 out of 25 convergent validity coefficients were significant or approached significance ($p < .1$) in the Shortell and Zajac study that included Golden's time 1 observations, assertions about validity should be based primarily on the strength of the validity coefficients, not on statistical significance (although it is common for statistical significance to be used). Shortell and Zajac's largest convergent validity coefficient was only .36, and the average was only .17.[2] The low magnitudes of these correlations

[1] Although intertemporal reliability is closely aligned with percent agreement, the two are different. Intertemporal reliability builds on percent agreement but goes beyond it to represent the likelihood (adjusted for chance agreement) that a chief executive places his or her organization at time 2 into the category he or she would typically classify the organization into in an infinite series of classification trials.

[2] These correlation coefficients representing convergent validity were calculated from *F*s

suggest that the time 1 data were not reasonably error-free (i.e., were not highly valid) and, therefore, should not have been used to assess retrospective validity. Thus, the imperfect time 1 data do not support any treatment of the issue as a question of validity. We treated our reanalysis of the Golden data as a study of reliability and treated our intertemporal coefficients as indirect rather than direct indicators of accuracy.

Having framed this as a study of reliability, we used a reliability index being used in marketing for qualitative judgments (cf. Perreault & Leigh, 1989). This index avoids the limitations that plague percent agreement as an indicator of intertemporal reliability or retrospective validity. The index incorporates an adjustment for chance intertemporal agreement, and more importantly, it focuses on underlying reliability as opposed to the joint probability of intertemporal agreement. To assess the degree to which Golden's (1992) results would have been more encouraging if an appropriate accuracy index had been used, we reanalyzed the intertemporal reliability in his data with the Perreault and Leigh (1989) index.

Reanalysis of Intertemporal Reliability in Retrospective Reports

Chief executives of 259 hospitals provided data for Golden's (1992) study. As noted above, each executive indicated his or her firm's current strategy in 1984, and two years later, each provided a retrospective report of that strategy. More specifically, each chief executive was asked to read a one-page description of Miles and Snow's (1978) four strategies and either (1) circle a number on a seven-point continuum running from defender to prospector, with analyzer in the middle, or (2) select the residual reactor category. Golden categorized their responses into a cross-tabulation table, which is reproduced as Table 1.

The intertemporal reliability of the retrospective strategy classification is .48 when all four of Miles and Snow's categories are included. However, several arguments in Miles and Snow (1978) and some empirical evidence (Doty, Glick, & Huber, 1993; Shortell & Zajac, 1990) suggest that the reactor category is a residual category that should be dropped from these analyses. When it is dropped, the intertemporal reliability is .53. Both of these estimates of intertemporal reliability are slightly less pessimistic than the 42 percent agreement reported by Golden.

---

provided in Table 2 of Shortell and Zajac (1990). The formula for converting these ratios is as follows: $corr = \sqrt{F / [F + (df, error / df, nonerror)]}$ (Rosenthal, 1991). We calculated convergent validity coefficients somewhat stronger than Shortell and Zajac's (1990) for a sample of HMOs studied by Conant, Mokwa, and Varadarajan (1990) using the same measure of strategy as Golden (1992). The average of these coefficients was .34 (the coefficients were calculated from *F*s provided in Conant et al.'s Table 3). James and Hatten (1995) also provided insights into the convergent validity of the nominal Miles and Snow measure. Using archival data, they found that less than 45 percent of organizations could be placed into the categories specified by chief executive officers. Finally, Hambrick (1981) provided convergent validity estimates on the basis of ordinal rather than nominal data: .56 was the estimate for colleges, .46 for hospitals, and .41 for insurance firms.

TABLE 1
Golden's (1992) Frequency Data[a]

Reported Time 1 Strategy	Time 1 Strategy Reported at Time 2				
	Defender	Analyzer	Prospector	Reactor	Totals
Defender	**4**	1	0	1	6
Analyzer	48	**88**	17	12	165
Prospector	7	23	**13**	4	47
Reactor	10	24	3	**4**	41
Totals	69	136	33	21	259

[a] Bold figures indicate the number of matches between retrospective and nonretrospective reports.

ATTENUATION OF RELIABILITY

The second methodological issue concerns attenuation caused by the questionnaire methodology used to assess strategy. The reliability of any methodology is not perfect. Thus, even if retrospective recall of strategic actions is perfect, a questionnaire assessment will not yield a perfect retrospective accuracy coefficient. Such coefficients are attenuated because of the simple measurement error associated with the questionnaire measure itself. If a strong retrospective accuracy coefficient is the goal, it is imperative to use a measure that has adequate reliability and validity.

The measure used by Golden (1992), however, has questionable reliability and validity. With this frequently used measure, respondents are asked to read four complex paragraphs describing the four Miles and Snow strategies. Multiple, partially overlapping attributes are used in the description of each strategy. Respondents are expected to use these descriptions of the strategies to either classify an organization into a discrete category or rate the organization along a single dimension running from defender to prospector. This is a complex judgment that can introduce substantial measurement error.[3]

Empirical Evidence of Weakness in the Measure

Several available studies provide interrater agreement estimates for Golden's (1992) measure, and these estimates support our contention that the measure is problematic. These studies used the same basic measure of strategy as Golden and raters who seem to have been knowledgeable about the focal organizations, and they assessed current, not retrospective, strategy. As Table 2 shows, the estimated interrater agreement coefficients are not strong, ranging from .39 to .65, with an average of .54. These interrater agreement estimates provide a rough gauge of the amount of error produced by the strategy measure itself.

Two studies provide test-retest agreement estimates, and these estimates

[3] Doty and colleagues (1993) presented an alternative approach for assessing Miles and Snow's (1978) typology that is much more consistent conceptually with the latter.

TABLE 2
Summary of Agreement Studies

Type of Agreement	Name and Date of Study	Number of Organizations	Industry	Raters	Agreement Coefficient
Interrater	Coleman (1978)	27	Mixed	CEOs who (1) managed competing firms and (2) were located in the same metropolitan area as the firm to be rated (e.g., CEOs from firms A, B, and C rated firm D).	.39
	Meyer (1979)	19	Health care	Experts who (1) worked in the health care field and (2) were located in the same metropolitan area as the organization to be rated.	.58
	Hambrick (1981)	77	Mixed	For each of three industries, experts who (1) held jobs within the industry and (2) were located in the geographic area from which the organization had been drawn.	.65[a]
	Shortell & Zajac (1990)	8	Health care	Executives within the corporation.	.52[b]
Test-retest	Shortell & Zajac (1990)	19	Health care	CEOs of the organizations.	.71
	Conant et al. (1990)	102	Health care	Marketing directors of the organizations.	.75

[a] Hambrick (1981) reported concordance coefficients that correspond to reliability coefficients (these cannot be converted to simple agreement). In most cases, a concordance coefficient will be higher than the simple agreement coefficient. The reported value is the average concordance coefficient for three samples.

[b] Shortell and Zajac (1990) reported concentration scores (Ray & Singer, 1973), which are generally higher than simple agreement coefficients. The reported value is the average concentration score for 8 organizations.

also support the contention that the measure is an issue.[4,5] As Table 2 shows, the two test-retest analyses yielded an average disagreement rate of 27.5 percent. Because the two analyses were based on very short lags between administrations of the instrument (a few weeks at most), the test-retest agreement is probably inflated by raters simply remembering what they said at time 1 rather than independently applying the instrument at time 2.[6] Assuming that only 10 percent of the agreement found in the two analyses was due to such carryover effects, the adjusted test-retest disagreement rate is 35 percent.[7] Thus, a substantial percentage of individuals reported different strategies after approximately 14 days. Considering these results, it seems clear that the low intertemporal reliability exhibited in the Golden data was caused to a significant degree by the underlying measure of strategy.

Further Empirical Evidence: Attenuation in Retrospective and Nonretrospective Reports

Fox (1992) collected both retrospective and nonretrospective strategy data using a measure of strategy developed by Glick, Huber, Miller, Doty,

---

[4] Shortell and Zajac (1990) used the same measure as Golden (1992), but they reported test-retest results based on continuous data generated through the seven-point scale that runs from defender through analyzer to prospector (the reactor category is omitted). Shortell and Zajac reported the percentage of CEOs who described strategies that were the same or one-scale-point different from their initial responses. Conant and colleagues (1990), in the second test-retest study, used categorical data and all four categories. A third study containing test-retest data (Hambrick, 1981) is not examined here because the researcher assessed test-retest reliability with a correlation coefficient based on responses to the seven-point scale. The problem is that a correlation coefficient based directly on continuous data does not necessarily reflect the true level of underlying test-retest agreement. For example, in an extreme case, if each rater provides a response that is different from time 1 by a factor of +2, the underlying level of test-retest agreement is 0, but the correlation coefficient is 1.0. As a practical matter, the study in question probably produced agreement comparable to the two studies discussed in the text, but we cannot be sure.

[5] Test-retest agreement for nominal data equates to the sum of joint probabilities, and therefore it is subject to the same criticism we discussed earlier concerning percent agreement (interrater agreement is not subject to this criticism). Nonetheless, we did not translate the test-retest agreement coefficients to reliability coefficients because we wanted the test-retest information to be maximally comparable to Golden's (1992) percent agreement (further, in one of the two cases, we could not translate to a reliability coefficient because of the manner in which agreement was assessed).

[6] Although carryover effects (i.e., simply remembering what was said at time 1 rather than independently applying the instrument at time 2) may be marginally acceptable in a study of intertemporal reliability, where the issue is the reliability of an informant's memory, such effects are unacceptable in a study of test-retest reliability, where the issue is the reliability of the measure itself.

[7] This figure (35 percent) reflects a 10 percent increase in disagreement as follows: .275 + [(.10 × 88 respondents exhibiting agreement)/121 total respondents]. The 35 percent could be considered acceptable for some purposes, but it is certainly not acceptable for a study of retrospective accuracy in which simple measurement error should be kept to a minimum. Using a measure with a 35 percent test-retest disagreement rate is especially inappropriate if measurement error is not measured and separated from recall error.

and Sutcliffe (1990). As described below, this measure of strategy is similar to the measure used by Golden (1992). Our analysis of Fox's results further supports the contention that the low intertemporal reliability exhibited in Golden's data was caused to a significant degree by the underlying measure of strategy. Further, our analysis suggests that CEO reliability is no lower in retrospective than in nonretrospective reports.

Fox (1992) asked multiple raters to assess firms' current strategies and, retrospectively, the strategies they had six years prior to the data collection. Thirty-one financial and banking experts in a small metropolitan community provided assessments for seven banks in the community. The informants included the CEO and 3 senior officers of each of the seven banks, and 3 finance professors at the local university. Each informant rated all seven banks. Thus, each bank was assessed by 4 inside officers, 24 outside officers from competitor banks, and 3 finance professors. The informants were asked to rate the extent to which each of the four Miles and Snow strategy descriptions characterized a bank's strategy (the same basic strategy descriptions used in the studies discussed above were used here). Thus, using four 7-point scales, the informants rated the extent to which a given bank exhibited a defender strategy, an analyzer strategy, and so on. This was done retrospectively for the time six years prior to data collection and nonretrospectively for the current time period.

In our reanalysis of Fox's (1992) data, we reduced each of the four 7-point scales to three categories (category 1 included scale points 1 and 2; category 2 included scale points 3, 4, and 5; and category 3 included scale points 6 and 7). This approach helped to make our analyses more comparable to the nominal-level analyses reflected in Golden (1992) and in the studies discussed above. To assess the rater reliability of the CEOs, we compared CEO judgments for their own banks with the judgments given by the other raters. That is, we determined whether a CEO's judgment for a given strategy (e.g., category 1, 2, or 3 for defender) matched the category that most of the other raters chose when assessing that strategy for the CEO's bank (the most chosen category for a particular bank is called the mode for that bank). After determining, for a given Miles and Snow strategy, whether each CEO had chosen the mode for his or her own bank, we calculated the percentage of CEOs who had chosen the most chosen category for their banks (e.g., 71.43 percent for the defender strategy if 5 of 7 CEOs agreed with the mode). Finally, we translated this percentage pertaining to a given Miles and Snow strategy into an estimate of CEO rater reliability using the Perreault and Leigh (1989) index.

CEO rater reliability in Fox's (1992) data was remarkably similar for both retrospective and nonretrospective reports (keep in mind that CEO retrospective data were compared to other raters' retrospective data and CEO nonretrospective data were compared to other raters' nonretrospective data when assessing CEO reliability). In both cases, CEO rater reliability was weak regardless of the comparison group (insiders, outsiders, and finance professors were used in various analyses to determine which category a CEO

should have selected). Contrary to arguments about retrospective error, CEO rater reliability was not even slightly lower for retrospective reports than for nonretrospective reports (an average .52 versus an average .51; see Table 3). Thus, CEOs were no less reliable when retrospectively rating strategy than when nonretrospectively rating strategy. This finding clearly suggests that most of the error in the Golden study was caused not by faulty retrospective thinking but by the measure itself. Although not conclusive alone, Fox's (1992) data, in conjunction with the evidence reviewed earlier, point to the measure as the major source of difficulty in the Golden study.

Plausible Adjustments for the Estimates of Intertemporal Reliability

Given a range from .39 to .75 for the prior estimates of interrater and test-retest agreement using the same basic measure of strategy as Golden (1992), given no evidence of faulty retrospective thinking in Fox's (1992) data, and given the complexity of the measure of strategy Golden used, it is plausible to assert that the low estimates of intertemporal reliability for Golden's retrospective reports (.48 and .53) are partially attributable to simple measurement error. Thus, we examined plausible adjustments of Golden's (1992) results by treating a portion of his intertemporal disagreements as being attributable to the somewhat weak measure of strategy itself. Based on the distribution of observed estimates of interrater and test-retest agreement and based on the complexity of the multifaceted stimulus paragraphs in the questionnaire measure, a conservative estimate is that more concrete, unidimensional, descriptive measures of organizational strategy would yield 15 to 55 percent fewer intertemporal disagreements. As shown in Table 4, assuming that 15, 25, 35, 45, and 55 percent of the disagreements could have been eliminated by simply using a stronger measure (not a perfect measure, just a stronger one) results in estimates of intertemporal reliability that range from marginal (.59) to very good (.83). For example, if we assume that 35 percent of the disagreements were excess disagreements caused by the measure, then eliminating them results in adjusted intertemporal reliability estimates of .71 for the full complement of four categories and .73 for the three nonreactor categories.

DISCUSSION

Results of our analyses suggest that a significant portion of the error reflected in Golden's (1992) data is attributable to the somewhat low reliability of the questionnaire measure of strategy. Results of interrater and test-retest analyses suggested substantial error from the measure. Our reanalysis of Fox's (1992) data failed to show lower CEO rater reliability for retrospective reports than for nonretrospective reports, reflecting a lack of evidence of CEO fallibility in recalling the past.

Although Golden (1992) did not call for complete abandonment of ret-

TABLE 3
CEO Rater Reliability[a]

Appropriate Category Defined by	Retrospective					Nonretrospective				
	Reactor	Defender	Prospector	Analyzer	Average	Reactor	Defender	Prospector	Analyzer	Average
Insiders	.60	.60	.60	.60	.60	.60	.60	.60	.39	.55
Outsiders	.39	.60	.75	.75	.62	.60	.00	.60	.75	.49
Experts	.60	.00	.75	.00	.34	.39	.39	.57	.57	.48
Averages					.52					.51

[a] Table entries are reliability coefficients.

TABLE 4
Plausible Adjustments to the Observed Intertemporal Reliability Estimates That Could Be Achieved by Using Stronger Measures

Potential Percentage Reduction in Disagreements	Three Non-Reactor Categories		Full Complement of Categories	
	Percent Agreement	Adjusted Intertemporal Reliability	Percent Agreement	Adjusted Intertemporal Reliability
0%	.52	.53	.42	.48
15	.59	.62	.51	.59
25	.64	.68	.57	.65
35	.69	.73	.63	.71
45	.74	.78	.68	.76
55	.79	.83	.74	.81

rospective reports, subsequent authors (e.g., Bergh, 1993; Boyd, Dess, & Rasheed, 1993; Kumer et al., 1993; Martell, Guzzo, & Willis, 1995; Priem & Harrison, 1994) have taken extremely cautious positions in response to Golden's results. In contrast to these extremely cautious positions, our position is that organizational researchers can continue to rely on retrospective reports provided by chief executives *if* the measures executives are asked to use are adequately reliable and valid. We emphasize here that we are not saying retrospective reports are always acceptable. Retrospective reports (or any other data) should only be used when reasonable efforts to demonstrate reliability and validity can be reported.

One infrequently used method for improving the validity of retrospective reports is to use free reports rather than forced reports. Under the free report option, an informant providing retrospective data is encouraged to say that he or she does not remember if in fact that is the case. Under the forced report option, an informant is encouraged to answer the question, and no option to skip the question is explicitly given. Although loss of data from the free report approach reduces the number of organizations available for analysis, it raises the accuracy of responses used in analyses. In a recent study designed in part to investigate why some studies in experimental and social psychology find higher levels of retrospective accuracy than others, Koriat and Goldsmith (1994) found that the free report option was associated with reasonably high accuracy (76.6 to 92.7 percent in various experiments), whereas the forced report option was associated with lower accuracy (47.6 to 67.0 percent in various experiments). Lipton (1977) and others investigating eyewitness testimony have found consistent results: eyewitnesses exhibit higher accuracy when they are asked to discuss what they saw in a free recall format (where they can say as little or as much as they wish) as opposed to being asked specific questions with an expectation for an answer to each. Similarly, Cohen and Java (1994) found that individuals using the free report option in attempting to recall personal health events were very accurate in the sense that the events they did recall had really occurred.

Other methods for improving the validity of retrospective reports are discussed at length in other sources, but are too frequently ignored. Thus, we provide a brief review here. First, researchers should utilize multiple knowledgeable informants per firm to allow the information provided by any one informant to be checked against the information provided by other informants (Bagozzi & Phillips, 1982; Phillips, 1981; Seidler, 1974; Williams, Cote, & Buckley, 1989). Second, researchers should ask about simple facts or concrete events rather than past opinions or beliefs (Glick et al., 1990; Golden, 1992; Chen, Farh, & MacMillan, 1993). A focus on facts and concrete events is likely to be less subject to cognitive biases and impression management. Questions about abstract concepts and opinions pose complex, ambiguous judgment tasks for respondents. Third, researchers should not ask informants to recall facts or events from the distant past (Huber & Power, 1985). Fourth, researchers should motivate their informants to provide accurate information. To motivate informants, confidentiality should be ensured, the duration and inconvenience of data collection should be minimized, and rich explanations of the usefulness of the project should be given (Huber & Power, 1985).

Beyond strengthening retrospective reports through solid measures, the free report option, and the other tactics mentioned above, we would like to see researchers strengthen retrospective reports through statistically controlling for systematic forces that cause recall errors. In an important contribution, Golden (1992) was able to explain a portion of the variance in intertemporal disagreement, and therefore was able to provide some clues as to systematic causes of CEO recall error (although our work indicates that CEO fallibility and CEO recall errors are not as pervasive as Golden's results suggest, they generally are still present to some degree). Specifically, Golden found that past strategy, extent of strategic change, and current profitability explained some variance. These results can be used in future efforts to control for systematic sources of error in retrospective informant reports of firm strategy. By measuring and controlling for these systematic sources of error, it is possible to directly improve the validity of retrospective reports.

A great deal of strategic management and organization theory research has been and continues to be based on retrospective reports. In some cases, this reliance on retrospective reports results from shortsightedness or a willingness to cut corners. In many cases, however, the reliance on retrospective reports results from researchers' inability to gain access to organizations to take multiple measures over time. In some cases, the reliance on retrospective reports is brought about by a desire to study an event whose timing could not have been anticipated (e.g., the Three Mile Island accident, the Challenger space shuttle disaster). In all cases, researchers relying upon retrospective reporting should use sound measures, should consider using the free report option, and should adhere to the other guidelines generally associated with proper retrospective data collection (cf. Huber & Power, 1985). If this were done, scholars could truly be comfortable with the idea that retrospective reports are not fiction.

REFERENCES

Bagozzi, R. P., & Phillips, L. W. 1982. Representing and testing organizational theories: A holistic construal process. ***Administrative Science Quarterly,*** 27: 458–489.

Bergh, D. D. 1993. Watch the time carefully: The use and misuse of time effects in management research. ***Journal of Management,*** 19: 683–705.

Bourgeois, L. J., & Eisenhardt, K. M. 1988. Strategic decision processes in high velocity environments: Four cases in the microcomputer industry. ***Management Science,*** 14: 816–835.

Boyd, B. K., Dess, G. G., & Rasheed, A. M. A. 1993. Divergence between archival and perceptual measures of the environment: Causes and consequences. ***Academy of Management Review,*** 18: 204–226.

Brennan, R. L., & Prediger, D. J. 1981. Coefficient kappa: Some uses, misuses, and alternatives. ***Educational and Psychological Measurement,*** 41: 687–699.

Chen, M., Farh, J., & MacMillan, I. C. 1993. An exploration of the expertness of outside informants. ***Academy of Management Journal,*** 36: 1614–1632.

Cohen, J. 1960. A coefficient of agreement for nominal scales. ***Educational and Psychological Measurement,*** 20: 37–46.

Cohen, G., & Java, R. 1994. Memory for medical history: Accuracy of recall. ***Applied Cognitive Psychology,*** 9: 273–288.

Coleman, H. J. 1978. Interindustry comparisons of strategy: Electronics and food processing. In R. E. Miles & C. C. Snow, ***Organizational strategy, structure, and process:*** 193–213. New York: McGraw-Hill.

Conant, J. S., Mokwa, M. P., & Varadarajan, P. R. 1990. Strategic types, distinctive marketing competencies and organizational performance: A multiple measures-based study. ***Strategic Management Journal,*** 11: 365–383.

Doty, D. H., Glick, W. H., & Huber, G. P. 1993. Fit, equifinality, and organizational effectiveness: A test of two configurational theories. ***Academy of Management Journal,*** 36: 1196–1250.

Feeser, H. R., & Willard, G. E. 1990. Founding strategy and performance: A comparison of high and low growth high tech firms. ***Strategic Management Journal,*** 11: 87–98.

Fox, S. 1992. ***The institutional, organizational, and issue contexts of strategic issue processing: An exploratory study.*** Unpublished doctoral dissertation, Texas Tech University, Lubbock.

Glick, W. H., Huber, G. P., Miller, C. C., Doty, H. D., & Sutcliffe, K. M. 1990. Studying changes in organizational design and effectiveness: Retrospective event histories and periodic assessments. ***Organization Science,*** 1: 293–312.

Golden, B. R. 1992. The past is the past—Or is it? The use of retrospective accounts as indicators of past strategy. ***Academy of Management Journal,*** 35: 848–860.

Hambrick, D. C. 1981. Strategic awareness within top management teams. ***Strategic Management Journal,*** 2: 263–279.

Huber, G. P., & Glick, W. H. 1993. ***Organizational change and redesign: Ideas and insights for improving performance.*** New York: Oxford University Press.

Huber, G. P., & Power, D. J. 1985. Retrospective reports of strategic level managers: Guidelines for increasing their accuracy. ***Strategic Management Journal,*** 6: 171–180.

James, W. L., & Hatten, K. J. 1995. Further evidence on the validity of the self typing paragraph approach: Miles and Snow strategic archetypes in banking. ***Strategic Management Journal,*** 16: 161–168.

Kanter, R. M. 1983. ***The change masters: Innovation and entrepreneurship in the American corporation.*** New York: Simon & Schuster.

Koriat, A., & Goldsmith, M. 1994. Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. ***Journal of Experimental Psychology—General,*** 123: 297–315.

Kumar, N., Stern, L. W., & Anderson, J. C. 1993. Conducting interorganizational research using key informants. ***Academy of Management Journal,*** 36: 1633–1651.

Lipton, J. P. 1977. On the psychology of eyewitness testimony. ***Journal of Applied Psychology,*** 62: 90–93.

Martell, R. F., Guzzo, R. A., & Willis, C. E. 1995. A methodological and substantive note on the performance-cue effect in ratings of work-group behavior. ***Journal of Applied Psychology,*** 80: 191–195.

Meyer, A. 1979. ***Hospital environment, strategy, and structure: The role of managerial perception and choice.*** Unpublished doctoral dissertation, University of California, Berkeley.

Miles, R., & Snow, C. 1978. ***Organizational strategy, structure, and process.*** New York: McGraw-Hill.

Mintzberg, H., Raisinghani, D., & Theoret, A. 1976. The structure of "unstructured" decision processes. ***Administrative Science Quarterly,*** 21: 256–275.

Perreault, W. D., & Leigh, L. E. 1989. Reliability of nominal data based on qualitative judgments. ***Journal of Marketing Research,*** 26: 135–148.

Phillips, L. A. 1981. Assessing measurement error in key informant reports: A methodological note on organizational analysis in marketing. ***Journal of Marketing Research,*** 28: 395–415.

Priem, R. L., & Harrison, D. A. 1994. Exploring strategic judgment: Methods for testing the assumptions of prescriptive contingency theories. ***Strategic Management Journal,*** 15: 311–324.

Ray, J. L., & Singer, J. D. 1973. Measuring the concentration of power in the international system. ***Sociological Methods and Research,*** 1: 403–437.

Rosenthal, R. 1991. ***Meta-analytic procedures for social research.*** Newbury Park: Sage.

Seidler, J. 1974. On using informants: A technique for collecting quantitative data and controlling measurement error in organization analysis. ***American Sociological Review,*** 39: 816–831.

Shortell, S. M., & Zajac, E. J. 1990. Perceptual and archival measures of Miles and Snow's strategic types: A comprehensive assessment of reliability and validity. ***Academy of Management Journal,*** 33: 817–832.

Williams, L. J., Cote, J. A., & Buckley, M. R. 1989. Lack of method variance in self-reported affect and perceptions at work: Reality or artifact. ***Journal of Applied Psychology,*** 74: 462–468.

Wolfe, J., & Jackson, C. 1987. Creating models of strategic decision making process via participant recall: A free simulation examination. ***Journal of Management,*** 13: 123–134.

Zajac, E. J., & Shortell, S. M. 1989. Changing generic strategies: Likelihood, direction, and performance implications. ***Strategic Management Journal,*** 10: 413–430.

C. Chet Miller earned his doctoral degree in organizational studies at the University of Texas at Austin. He is currently an associate professor in the Hankamer School of Business, Baylor University. His research interests center on strategic management processes, diversity in upper-echelon executive groups, and organizational design.

Laura B. Cardinal earned her doctoral degree at the University of Texas at Austin. She is currently a visiting assistant professor at the Fuqua School of Business, Duke University. Her research interests are in the areas of strategic planning and managing innovation and R&D in corporate environments.

William H. Glick is the chair of the Department of Management and a Dean's Council of 100 Distinguished Scholar at Arizona State University. He received his Ph.D. degree from the University of California, Berkeley. His research interests are in the areas of job and organizational design, business process redesign and change, and managerial cognition.