The Predictive Validity of Alternative Approaches to Evaluating Quality of a Festival

JOHN L. CROMPTON AND LISA L. LOVE

Alternative conceptualizations of the quality and satisfaction constructs are presented from both the leisure/tourism and the marketing literatures, and differences between them are noted. The predictive validity of seven alternative operationalizations of quality that were suggested by the literature were measured by evaluating them against an overall measure of quality in the context of a festival. Results showed that the most valid measures of quality were the performance-based operationalizations; the least valid were the disconfirmation-based operationalizations; and the inclusion of importance weights did not improve predictive validity of the measures. There was no significant difference between first-time and frequent visitors to the festival on any of the seven alternative operationalizations. Implications of the findings are discussed.

There is consensus that quality is central to success in the delivery of tourist and leisure services, but quality remains an elusive, indistinct, and ambiguous construct. “Tourist satisfaction with a destination area is a nebulous concept. . . . Many travel researchers and practitioners who use it daily would probably find it difficult to define” (Pizam, Neumann, and Reichel 1978, p. 315). Although this statement was made in the early days of tourism research and was related to satisfaction, it appears to be equally appropriate today when extended to the construct of quality. This elusiveness is manifested in the array of different ways in which it has been conceptualized and operationalized in the tourism, leisure, and marketing literatures, and by the continued confusion about its relationship to satisfaction. For example, in the context of measuring customer evaluation of service quality in travel agencies, LeBlanc (1992) states, “Even though quality and satisfaction are in theory concerned with the difference between expectations and perceptions, at present no theoretical distinction seems to be made between the two concepts in the literature” (p. 15). Similarly, Cronin and Taylor (1992), after reviewing the marketing literature, concluded:

The service literature has left confusion as to the relationship between consumer satisfaction and service quality. This distinction is important to managers and researchers alike because service providers need to know whether their objective should be to have consumers who are “satisfied” with their performance or to deliver the maximum level of “perceived service quality.” (p. 56)

It has been suggested that “how the service quality construct should be measured . . . is arguably among the most important concerns in services marketing” (Cronin and Taylor 1992, p. 65). Certainly, it is a key issue in tourism. This article first delineates the construct of quality and differentiates it from satisfaction. It then identifies the different ways in which quality of a destination has been operationalized in the literature and discusses the rationale underlying these operationalizations. Finally, results are reported of an empirical comparison of seven approaches that have been used in the leisure or tourism literatures to measure quality or satisfaction. They are compared in the context of measuring participants’ evaluations of the quality of a festival, and their relative validity is discussed.

DELINATION OF THE QUALITY CONSTRUCT

There appear to be two distinctive lines of thought regarding the nature of the quality construct and how it is differentiated from satisfaction. The prevailing perspective in the marketing literature was articulated by Parasuraman, Zeithaml, and Berry (1986) in their original work on service quality.

We distinguish between service quality and satisfaction in that service quality is a global judgment or attitude relating to the superiority or excellence of the service, whereas satisfaction is related to a specific transaction. The two constructs are related in that incidents of satisfaction over time result in perceptions of service quality. (p. 5)

This view is currently being debated (Cronin and Taylor 1992), but the debate focuses upon the extent to which satisfaction influences or mediates service quality (Bitner 1990; Bolton and Drew 1991a, b; Cronin and Taylor 1992) or vice versa (Teas 1993; Parasuraman, Zeithaml, and Berry 1994). The fundamental notion of some type of cumulative impact over time being the core of the differentiation between the two constructs appears to be generally accepted in the marketing field.

This is an entirely different conceptual approach to that which has evolved in the context of leisure services, where
In the services marketing literature, this notion was embraced by Langeard et al. (1981) in their “servuaction system,” which Bateson (1992) used as an organizational framework for his text. This system comprises the visible environment, the service providers, and the visitor. It recognizes that in a service encounter, production and consumption take place simultaneously, and the visitor is an integral part of the process. Thus, the outcome of an encounter is influenced by what the visitor brings to it.

There appears to be some movement in the marketing literature toward the conceptualization used by leisure scientists. Teas (1993) suggested that both service quality and customer satisfaction could be examined meaningfully from both the transaction-specific and global perspectives. Using this as a point of departure, Parasuraman, Zeithaml, and Berry (1994) revised their original notion of satisfaction as a precursor to service quality. Acknowledging the influence of recent empirical work suggesting that the reverse was more likely to be true, they offered a model that “posits a customer’s overall satisfaction with a transaction to be a function of his or her assessment of service quality, product quality, and price. This conceptualization is consistent with the ‘quality leads to satisfaction’ school of thought” (p. 121). If two additional ingredients were added to this model, it would effectively lead to a complete reconciliation with the leisure sciences’ conceptualization of satisfaction relationship: (1) that satisfaction is a psychological outcome, and (2) that intervening variables exist between level of service quality and level of satisfaction that are outside a service provider’s control so the relationship between the two constructs is unlikely to be 1.0. This latter point is implicitly recognized by Langeard et al. (1981) in their servuaction system since this embraces the notion of the encounter outcome being influenced by what the visitor brings to it. Both points appear to be accepted by Cronin and Taylor (1994):

Service quality perceptions reflect a consumer’s evaluative perceptions of a service encounter at a specific point in time. In contrast, consumer satisfaction judgments are experiential in nature, involving both an end state and a process, and reflecting both emotional and cognitive elements. (p. 127)

These shifts in thinking suggest that a consensus view of the relationship between service quality and satisfaction may emerge in the relatively near future.

The contribution that visitors bring to quality of experience is articulated by Brown (1988) in the context of outdoor recreation: “Quality of experiences can be influenced by input factors provided by managers . . . but to a considerable extent the quality of experiences depend[s] upon how recreationists use the many factors of production” (p. 413). Williams (1988) similarly notes that visitors shape and create the recreation experience, and attraction suppliers have little influence over how visitors will react to opportunities they are provided. Their experience is created “through a transaction with the physical and social setting, including what the recreationist brings to the process in terms of use, history, perceptions, companions, skills, equipment, identities, hopes and dreams” (p. 432).

Given that a recreation or tourism experience is substantially self-produced (Williams 1988), it is illogical to evaluate an organization’s performance as a tourism supplier based on visitors’ levels of satisfaction with their experience. Managers are unable to control what visitors bring to the setting, but they are able to control what is provided at an attraction's end. It is therefore necessary to focus first on providing high quality service rather than on evaluating satisfaction from end-user’s perspectives.

Attribution

Attribution theory has been used in tourism and leisure research to examine the role of relationships between service providers and their perceptions of expectations. Brown (1988) developed a model that identifies (1) actions (2) responses and (3) attributions. This framework has been expanded to include expectations (4) attributions (5) relationships and (6) attributions in the context of leisure and tourism research.

Attributes of service quality can be controlled and manipulated by recreation suppliers assuming that the necessary resources are available. In contrast, level of satisfaction is dependent not only on quality of service attributes but also on the status of a host of variables that may affect the user, such as the climate or the nature of the social group. Such variables are outside a supplier’s control and may intervene, so that a perceived high-quality service could result in a low level of satisfaction. Conversely, a high satisfaction outcome may result even when perceived service quality is low because, for example, the social group interactions are sufficiently positive to offset the low quality service. (p. 368)

Attractions to a product or event will be shaped by the perceptions visitors are prompted to experience.

This model has been used to explain the tourist’s perceived benefits of their visit. More recent research by Gaynor and Van Loon (1995) examined the role of factors among which recreationists were considered as a possible domain.
attraction. Thus, it appears to be more useful and productive to focus evaluation efforts on recreation opportunities provided, rather than on the quality of recreation experience emerging from exposure to those opportunities. Hence, the recreation opportunity conceptualization was used to guide this study.

ALTERNATIVE OPERATIONALIZATIONS OF QUALITY

Although conceptualization of quality in the marketing literature has evolved differently, its operationalizations have been similar to those that have appeared in the tourism and leisure literature. In neither field is there consensus on how evaluation of quality should be operationalized. Five types of relationships have been operationalized to evaluate quality of a destination or attraction in tourism, and each was incorporated into the study reported here. They are: (1) attribute expectations; (2) attribute expectations and importance; (3) attribute performance and importance; (4) attribute performance and expectations; and (5) attribute performance, expectations, and importance. Each of these five types of relationship is reviewed in the following paragraphs. The section concludes with a review of previous research that has reported comparisons between quality evaluation measures.

Attribute Expectations

Visitor expectations are pretrial beliefs about a destination or attraction (Olson and Dover 1979). Attribute expectations have been used extensively in the measurement of destination image and in positioning studies. The conceptual underpinning for this approach in both these contexts is that a priori perceptions of a destination's attributes are used by potential visitors to make decisions about whether they will visit a particular destination. This infers that a priori perception of these attributes relates to quality of opportunity available at the vacation destinations being evaluated. And as Hunt (1975) has stated. "The perceptions held by potential visitors about an area may have significant influences upon the viability of that area as a tourist recreation region" (p. 1).

These perceptions are essentially expectations since they a priori reflect the performance that respondents anticipate will be forthcoming from a vacation opportunity. Such perceptions differ from evaluations of actual performance, which are post facto measures of an opportunity after it has been experienced.

The early tourism image studies (Mayo 1973; Hunt 1975) used generic attributes such as topography, climate, resident population life-styles, and recreational character in their efforts to differentiate regional vacation destination images. More detailed sets of destination attributes were generated by Gearing, Swart, and Var (1974) as part of their development of a touristic attractiveness index and by Scott, Schewe, and Frederick (1978) in their study of visitor preferences among northeastern states. The list of attributes that were considered increased, and the sophistication with which they were solicited and statistically treated progressed, in Pizam, Neumann, and Reichel's (1978) study of visitors' perceptions of the attributes of Cape Cod. Their approach of generating a large number of items and using a factor analysis technique to identify domains has become standard procedure with attribute expectation studies. The difference scores on domains among target markets of interest are typically used to identify the strengths and weaknesses of destinations' market positions.

Attribute Expectations and Attribute Importance

By the late 1970s, it became evident that the contribution of a priori attribute perceptions alone was limited in evaluating destination quality. It was suggested that measuring relative importance of attributes simultaneously would make this approach more useful to decision makers. The case was articulated by Crompton (1979) in his study of the image of Mexico.

The descriptive data without inclusion of the evaluative [importance] dimension is of little value for it gives no indication of the relative importance which respondents attach to particular attributes of destination image when making a vacation decision. For example, two semantic differential scales, such as "The People of Mexico are: Rich-Poor and Friendly-Unfriendly," may both receive a score of 5 by the same respondent on descriptive scales. However, a decision to go to Mexico on vacation may be greatly influenced by the latter, while the former may be of only marginal importance. The inclusion of an evaluative [importance] component identifies respondents' salient image attributes and it is these which are most likely to serve as behavior determinants. (p. 19)

Crompton's (1979) approach was relatively simplistic. He reported mean scores for each item on both the attribute perceptions and importance sections of his 30-item instrument evaluating the image of Mexico. Conclusions were drawn by simply comparing the two sets of mean scores and their rankings. This effectively identified the problem facing Mexican tourist agencies because there was "an extraordinary lack of congruency" (p. 20) between the attributes receiving the highest scores and those that respondents considered important when considering whether they should visit Mexico on a vacation.

A similar approach was used by Shih (1986) in his study of the image of Pennsylvania. He listed the mean importance and mean performance ratings for each attribute and drew conclusions. However, he introduced a refinement of this approach by noting that "the image of Pennsylvania perceived by the respondents can be further analyzed by examining the differences between mean importance and mean performance ratings on the eight primary travel selection criteria. The smaller the differences, the more significant Pennsylvania has achieved in a particular travel selection factor" (p. 10). This importance-minus-performance measure of quality was not the primary focus of the article and no underpinning rationale for including it was offered. Although Shih used the term "performance ratings" in this study, the sample included tourists who had not necessarily previously visited Pennsylvania. Hence, they were not evaluating the opportunities experienced on past trips, and so the terms a priori "perceptions" or "expectations" would have been more accurate descriptors.

An alternative relationship of the importance and perceptual dimensions of attributes was stimulated by Fishbein's (1967) work. Goodrich (1978) used the basic Fishbein model to evaluate respondents' relative a priori perceptions of nine tourism regions as vacation destinations. His basic Fishbein model stated that "an individual's attitude (Ao) toward an object (in this case, relative preference for each region as a vacation spot) is a function of (determined by, associated with) the amount of valued attributes (Ai) that the individual perceives the object (region) to have, and of the importance
scores (for example, 7 indicates that people place a high importance on the attribute). In this way, a 10-point scale can be used to give relative contributions of service quality scores.

Several studies have been conducted of using the Fishbein model (Fishbein and Ajzen, 1970; Fishbein and Ajzen, 1991; Ajzen and Fishbein, 1980). However, results of the model have been mixed, and there is a need for further research into the validity of the scores and their application.

A complete breakdown of the model is beyond the scope of this paper. However, it is clear that the model provides a useful framework for understanding consumer behavior.

Attribute Performance and Attribute Importance

The Fishbein approach was also adopted by Witter (1985) who used Goodrich’s (1978) set of attributes in her study of the comparative evaluations of a resort area by tourists and retailers. Witter’s tourist respondents, however, were actual visitors who were able to give an evaluation of the resort’s performance on the set of attributes based on their experience. Witter (1985) also implemented an additional stage of the Fishbein approach by summing the means of the perceptions-importance scores across attributes and comparing the grand mean scores of tourists and retailers. This additional step introduces the compensatory notion that trade-offs must occur between low values on one attribute and high values on another. Thus, superior performance on one attribute can compensate for “make up” for inferior performance on another attribute. This assumption is contentious and has been extensively debated in the marketing literature (Timmermans 1984).

An alternative method for comparing the importance and performance dimensions of attributes is the importance-performance analysis, which was introduced by Martilla and James (1977). In this method the attribute and importance scores are plotted on a horizontal and vertical axis, respectively. The result of this procedure is that each item is categorized into one of four quadrants. This pictorial representation has the advantage of being easily interpreted by management. It was used, for example, by Guadagnolo (1985) to evaluate the quality of a 10-kilometer road race in Pittsburgh that attracted 13,000 runners, by Geva and Goldman (1991) in their evaluation of guided tours, and by Mill (1989) in his assessment of the strengths and weaknesses of Colorado as a destination for British tourists.

Attribute Performance and Attribute Expectations

There is a long history of research findings in a variety of fields suggesting that expectations create a frame of reference against which comparative judgments of service quality and satisfaction are made. Satisfaction “implies an act of judgment, a comparison of what people have to what they think they deserve, expect, or may reasonably aspire to” (Campbell, 1980, p. 22). Thus, outcomes poorer than expected (a negative disconfirmation) are rated below this reference point, whereas those better than expected (a positive disconfirmation) are evaluated above this base (Oliver 1980). This comparative framework was first used by Bultena and Kessig in 1969; “Satisfaction . . . is a function of the degree of congruency between expectations and the perceived reality of experiences” (p. 49).

This same expectancy disconfirmation paradigm was used by Parasuraman, Zeithaml, and Berry (1988) in conceptualizing their measure of service quality. Subsequently, several efforts to measure quality in the leisure and tourism field have been based on the work of Parasuraman, Zeithaml, and Berry (Crompton and MacKay 1989; Fick and Ritchie 1991; Hamilton, Crompton, and More 1991; LeBlanc 1992; MacKay and Crompton 1990; Wright, Duray, and Goodale 1992; Ostrowski, O’Brien, and Gordon 1993). This pioneering work of Parasuraman, Zeithaml, and Berry (1988) has been subjected to five main criticisms.

The first criticism relates to the form of the comparative standard against which performance judgments are made. Parasuraman, Zeithaml, and Berry (1988) indicate the standard is what an individual desires the service to deliver. This “expectations-as-ideal” standard was earlier proposed by Miller (1977), who termed it the “wished for” level of performance, and by Swan and Trawick (1980), who suggested the term “desired expectations,” which they defined as the level at which the visitor wanted the attraction or destination to perform. Thus, quality is judged against what users feel a service provider should rather than would provide. The rationale for this is that if expectation (what they expect will be delivered) rather than desire (what they believe will be delivered) is used as the comparative criterion, it might encourage managers to attain a satisfactory quality rating by lowering users’ expectations of what they could anticipate from a service. This differentiation between “would” and “should” is a key difference between how expectations are used in the satisfaction and the service quality contexts. In the service quality literature, expectations do not represent predictions about what service providers would offer (as is the case in satisfaction studies); rather, they define what providers “should” offer. In the leisure literature there is some empirical support for this desire approach, since Dorfman (1979) reported that “preferences are more highly related to overall indices of satisfaction than expectations” (p. 506). His operationalization of preference appears to be synonymous with Parasuraman, Zeithaml, and Berry’s use of desire.

The use of desire has been criticized on both theoretical and practical grounds. Cronin and Taylor (1992), from a theoretical perspective, argue that it is inconsistent with the suggestion that expectations should be based on experience norms (Woodruff, Cadotte, and Jenkins 1983) — what consumers should expect from a given service considering their experience with that specific type of service organization. From a practical perspective, the difficulty with using the desire approach is that it leads to consistently very high scores for all attributes, which people may not find meaningful or relevant. In such cases, a 10-point scale may not give real meaning to the satisfaction scores.

Second, criticism of the model is of using the same criteria on all attributes (Miller, 1970; Zeithaml, 1980). However, research has been conducted using a variety of criteria for different attributes.

A complete breakdown of the model is beyond the scope of this paper. However, it is clear that the model provides a useful framework for understanding consumer behavior.
scores (Dorfman 1979; Wall and Payne 1973). Presumably, people desire the maximum amount of quality on every attribute. If these scores are almost constant, then there is little point in including them on an instrument, since they will not give responses significantly different from using the perception scores alone.

Second, psychometricians have pointed to the problems of using discrepancy scores (Lord 1963; Cronbach and Furby 1970; Johns 1981; Brown, Churchill, and Peter 1993). However, research evidence on this point appears to be equivocal, and researchers measuring satisfaction and quality in a variety of fields continue to use discrepancy or difference scores (Williams 1988).

A third criticism is that respondents are asked to complete both the expectations and performance items at one administration. “Based on what they had experienced in the past, respondents were asked what they expected and then asked what they perceived. All respondent beliefs were entirely ex post. These expectation responses can be of little value” (Carman 1990, p. 47).

Fourth, the Parasuraman, Zeithaml, and Berry (1988) instrument has been criticized because in a retail situation it is not likely to be practical to ask a customer to complete an expectations instrument when coming in the door and then to complete the performance questions at the end of the visit (Carman 1990). However, this procedure is feasible at a tourist attraction where the length of stay is likely to be measured in hours rather than the minutes associated with a retail store visit.

A final criticism raised by Carman (1990), based on his empirical results, is that the expectation measure may not be valid in situations where visitors do not have well-formed expectations. He concluded that “expectations change with familiarity” (p. 49).

The discrepancy between expectations and performance has been measured differently in the leisure and marketing fields (Williams 1988). A common approach in consumer behavior (although it was not used by Parasuraman, Zeithaml, and Berry 1988) has been to measure perceived disconfirmation directly, by requiring consumers to respond with explicit judgments about whether a product is “better than” or “worse than” they expected (Carman 1990). In the leisure literature, discrepancy has been measured as the observed difference between ratings of expected attributes and performance on those attributes (Dorfman 1979). These two approaches are likely to produce different results (Williams 1988).

Attribute Performance, Expectations, and Importance

Carman (1990) noted that “to most service providers, the importance of a particular service attribute seems more relevant than its expected level” (p. 45). He argues: “All three variables, importance, perception [performance] and expectations, are material and play different roles in evaluating overall quality. Thus, the users of these scales will need to collect information on all three variables, not just perceptions [performance] and expectations” (p. 49).

A number of researchers have advocated that attribute importance be used to weight the performance-expectations scores, and the weighting has been operationalized in two ways. Some have used a constant-sum scale to derive importance by asking respondents to allocate 100 points among a set of attributes or dimensions of attributes (Zeithaml, Parasuraman, and Berry 1990; Whipple and Thach 1988). Others have asked respondents to complete the same set of expectations and performance Likert-scale items, for which the instructions have been adapted to measure importance (Cronin and Taylor 1992; Dorfman 1979).

RESULTS FROM PREVIOUS COMPARATIVE STUDIES

The only investigations reported in the leisure and tourism fields that have compared results from different operationalizations of quality or satisfaction were done by Dorfman (1979) and Fick and Ritchie (1991). Dorfman’s respondents were asked to rate each of 22 items related to camping in terms of their perceptions, performance, previous expectations, preference, importance, and satisfaction. These responses were then combined in various ways to form eight alternative measures of satisfaction. Each measure was compared to respondents’ overall ratings on an 11-point scale of satisfaction with the entire experience. Two findings were particularly interesting. First, for all combinations of measures, using the importance weights did not significantly change correlations from the same measures obtained without using the weights. Second, the expectations-minus-performance measure did not correlate as highly with the overall measure of satisfaction as did the straight performance measure.

Fick and Ritchie (1991) measured service quality in several types of tourism businesses and their findings regarding comparison of measures resembled those of Dorfman:

Mean perception of performance scores provide as good an evaluation (if not a better one) of perceived service quality than the computed quality score [of perceptions minus expectations]. It should also be noted that there is still a strong correlation between the computed quality score [i.e., the discrepancy measure] and the direct evaluation [i.e., an overall measure of quality], but much of that is probably due to the fact that a major component of the quality score is the perception of performance measure. (p. 5)

Two empirical studies in the marketing literature have similarly reported that a performance measure has explained more variation in satisfaction than discrepancy measures. Churchill and Surprenant (1982) investigated whether expectation and performance exerted independent efforts on satisfaction, in addition to their impact through disconfirmation, using a videodisc player and a chrysanthemum plant. In the case of the videodisc player, satisfaction was determined solely by its performance: “When it performed well they were satisfied with it and when it performed badly, they were dissatisfied with it, regardless of their initial expectations . . . . The direct performance-satisfaction link accounted for most of the variation in satisfaction, 88%” (p. 502). However, in the case of the plant, “Disconfirmation positively affected satisfaction as is commonly held; when subjects perceived the product performing better than expected they were satisfied with it and vice versa” (p. 502).

Cronin and Taylor (1992) gathered responses on the service quality delivered by two firms in each of four service industries. They reported that in all four industries the simple performance measure explained more of the variation in their global measure of service quality than did their other three measures: performance minus expectations, importance times (performance minus expectations), and importance times performance. Although they concluded that “the weight
of the evidence clearly supports the performance-based measures of service quality” (p. 64), they recognized that the study’s findings were derived from one sample’s experience in one city with four industries. For this reason, they recommended that future studies should incorporate multiple measures of the quality construct so that the generalizability of their findings could be assessed.

Contrary to the findings of Churchill and Surprenant (1982) and Cronin and Taylor (1992), Bolton and Drew (1991b) found that “disconfirmation explains a larger proportion of the variance quality than performance” (p. 383). However, disconfirmation was measured by better/same/worse response categories, and their study did not directly measure expectations.

Five explanations have been offered for the apparent superiority of performance measures over discrepancy measures. Carman (1990) suggests that if at the time of completing the expectations items, the expectations were not based on experience, they are likely to be poor standards against which to measure quality of performance. He explains his findings in the following terms:

Respondents at the time of completing the expectations battery had expectations, but they were not based on experience. These expectations could be described as the aspects of the service they expected should be important indicators of quality. After using the service, they were far more knowledgeable and their assessment of quality was both different and more clear. (p. 48)

A similar explanation was offered by Churchill and Surprenant (1982). In the context of tourism, Botterill (1987) reported, “analysis of the personal constructs (as defined by Kelly 1955) used by the tourists in our conversations, both before and after a vacation, suggests that it is those anticipatory constructs that are least adequate in explaining actual events that provide the basis for the tourist’s excited interest in his or her own world. . . . The unpredictability of touristic events seems to lie at the heart of touristic experience” (p. 140). Lack of familiarity with a destination may cause expectations to be tentative and uncertain, but these qualities are not captured by the expectations instrument. In such situations, considering expectations to be firm criteria against which evaluative judgments about performance are made is likely to be fallacious. After using a placement center service, respondents, many of whom were using it for the first time, were far more knowledgeable and their assessment of quality was both different and clearer than before using the service (Carman 1990). This is consistent with the view expressed by Whipple and Thach (1988) that the amount and quality of previous experience may determine the effect of expectations on satisfaction (Day 1977; Woodruff, Cadotte, and Jenkins 1983).

A related explanation for the superiority of performance measures has been articulated by Whipple and Thach (1988), who suggest that expectations may be important indicators of choice preference, which is why they are used in tourism research in studies of image, but not of quality: “There is evidence that pre-purchase choice criteria and post-purchase evaluation criteria are not the same. The results of the American Hotel and Motel Association (1985) survey of hotel choice and repeat business document the difference in the ranking of criteria before and after purchase” (p. 16). If different evaluative criteria are used after an event than before it, then the initial expectation framework is disregarded and is of little value for measuring quality.

Tse and Wilton (1988) offer a third explanation: in situations where expectation and performance items are both completed ex post facto, cognitive dissonance may be responsible for reducing the validity of discrepancy scores. If the psychological costs of accepting poor performance by an attraction are high, then a cognitive dissonance strategy would be to report lower expectations. Thus, individuals adopting a dissonance reduction strategy after a negative disconfirming experience could do this by lowering the expectations scores that they report on an ex post facto instrument.

Fourth, Woodruff, Cadotte, and Jenkins (1983) note that prior experiences may bias perceptions of performance. They observe that assimilation/contrastra theory suggests consumers may either raise or lower their performance beliefs according to how closely the perceived performance matches expected performance. Thus, they suggest that including importance weights and expectations only introduces redundancy.

A corollary of this explanation was offered by Mannell (1989). He noted that Campbell (1980) extensively used the disconfirmation approach to measure satisfaction and reported that people changed their ratings of satisfaction very little over time. Mannell (1989) suggested this constancy may exist because people adapt their level of expectation to prevailing circumstances and therefore experience little change in levels of satisfaction. If this type of shift occurs, expectations do not serve as a criterion but rather fluctuate in synchronization with performance, which makes the expectations component redundant.

A fifth explanation for the superiority of performance measures over discrepancy measures is that quality is likely to be strongly influenced by how well the offering fulfills innate needs, wants, or desires of visitors (Mannell 1989), rather than how performance compares with prepurchase predictions (Westbrook and Reilly 1983; Woodruff, Cadotte, and Jenkins 1983, 1987; Sirgy 1984).

**STUDY HYPOTHESES**

The study objective was to compare and contrast the relative predictive validity of alternative measures for evaluating the quality of a festival. The alternatives considered, derived from the literature review, were:

- **Alt 1**: Quality equals Expectations;
- **Alt 2**: Quality equals Importance minus Performance;
- **Alt 3**: Quality equals Importance times Expectations;
- **Alt 4**: Quality equals Importance times Performance;
- **Alt 5**: Quality equals Performance minus Expectations;
- **Alt 6**: Quality equals Importance times Performance minus Expectations;
- **Alt 7**: Quality equals Performance.

Procedures for determining the predictive validity of the alternative approaches for measuring quality were guided by three hypotheses that were formulated from findings in the literature review.

H1a: Performance will be a substantially better predictor of quality than the six alternative operationalizations when tested on all selected attributes of the festival.

This proposition is consistent with the findings of Dorfman (1979) and Fick and Ritchie (1991) in the leisure and tourism fields, and with the hypotheses of Butler (1985, 1988) and Pizam (1982) and is consistent with the findings of the studies in the literature review.
fields, and with those reported by Churchill and Surprenant (1982) and Cronin and Taylor (1992) in the marketing literature.

H1b: Importance scores will not substantially improve the predictive validity of either the performance-based operationalizations or the disconfirmation-based operationalizations of quality.

This hypothesis was developed in response to Dorfman’s (1979) findings in the context of satisfaction. He concluded that “using the importance weights does not significantly change correlations from those obtained without using the weights” (p. 497). Similar findings were reported by Cronin and Taylor (1992).

H2: Correlations of the alternative operationalizations of quality with a measure of overall quality will be higher among visitors who reported relatively frequent past visits to the festival than among those who had not previously been to the festival.

Repeat and first-time visitors could have different expectations because of varying levels of familiarity with the festival (Parasuraman, Zeithaml, and Berry 1985; Zeithaml, Berry, and Parasuraman 1993; Scott and Yalch 1980; Smith and Swinyard 1983). It was noted earlier that expectations measures may be tentative and uncertain when respondents lack familiarity with the festival, so they do not offer firm criteria against which to relate performance (Bettman 1987; Carman 1990; Churchill and Surprenant 1982; Whipple and Thach 1988). Thus, it was anticipated that the discrepancy measures reported by repeat visitors would correlate higher with a measure of overall quality than those reported by first-time visitors.

THE STUDY SETTING

The study was conducted at Dickens on the Strand, an annual Victorian Christmas celebration held in Galveston, Texas, on the first weekend in December. The Dickens theme is a logical extension of the historical waterfront’s nineteenth-century cotton-trading connections with England and its heritage of Victorian architecture. The Strand is a famous street in London and also the name of the main thoroughfare in Galveston’s historic district. The festival takes place in a 17-block section of the Strand National Historic Landmark District. During the two-day festival, the area is fenced in, and 10 gates are erected to control activities and permit the charging of admission. Typically, total attendance is between 70,000 and 80,000 for the two days.

In their pioneering work on service quality, Parasuraman, Zeithaml, and Berry (1986) reported that quality comprised five dimensions: tangibles, reliability, responsiveness, assurance, and empathy. Tangibles represent the physical facilities, equipment, and appearance of personnel. Reliability refers to the ability to perform the promised service dependably and accurately. Responsiveness is the willingness to help users and to provide prompt attention. Assurance indicates courteous and knowledgeable employees who convey trust and confidence. Empathy is offering caring, individualized attention to users.

These dimensions have guided the operationalization of quality used by others who have adapted the work of Parasuraman, Zeithaml, and Berry (1988) to the leisure and tourism fields (Crompton and MacKay 1989; Fick and Ritchie 1991; Hamilton, Crompton, and More 1991; LeBlanc 1992; MacKay and Crompton 1990; Ostrowski, O’Brien, and Gordon 1993). The Parasuraman, Zeithaml, and Berry (1988) work is based on services that involve face-to-face transactions between retailers and customers. However, leisure and tourism opportunities can be classified along a continuum according to the relative importance of people compared to facilities in their delivery (MacKay and Crompton 1988). Many tourism opportunities, such as guided high-risk activities or eating in a restaurant, require a relatively high level of interaction with personnel in service delivery. However, other opportunities, such as visits to a park or festival (e.g., Dickens) are at the opposite end of the continuum.

Fick and Ritchie (1991), after applying the Parasuraman, Zeithaml, and Berry (1988) instrument to sectors of the tourism industry, noted that the original measure did not adequately cover tangible factors. They noted that this was probably because facilities tend to be situation-specific in the tourism context and hence do not lend themselves to inclusion in a generic type of measure. In these types of situations, there is no process of delivery per se that is addressed by the dimensions of responsiveness, assurance, empathy, and to a lesser extent, reliability. Rather, the dominant dimension is tangibles. As Hamilton, Crompton, and More (1991) note in the analogous context of parks, “staff contributions tend to be indirect through maintaining or improving quality of the resource, rather than through direct involvement with park users” (p. 219).

At the Dickens festival, the tangibles dimension is dominant. Results from previous surveys of the event, discussion with the festival’s organizers, and a review of the literature led the authors to identify the key dimensions of quality at Dickens as being ambiance of the environment, sources of information on the site, comfort amenities, parking, and interaction with vendors. These dimensions were subsequently confirmed by a factor analysis of the data collected in this study. Hence, in evaluating quality of opportunity at the festival, it was these dimensions rather than the generic dimensions suggested by Parasuraman, Zeithaml, and Berry (1988) that were operationalized in this study.

METHODS

An initial list of quality attributes was developed from a literature review, previous surveys at the event, and interviews with the festival’s organizers. After pretesting, the final instrument comprised 22 items. These are listed in Table 4. Respondents were asked to complete the expectations instrument before entering the site. They rated the 22 items on a seven-point scale from very poor (1) to very good (7). The instructions on the questionnaire read: “The following set of statements relate to various features of Dickens on the Strand. Based on your previous experience at Dickens, at other festivals, and/or on information you have acquired, please tell us what your expectations are regarding each of these features.”

After completing and handing in the expectations instrument, respondents were given a questionnaire containing the importance and perceptions of performance measures and
were requested to mail it back in a prepaid envelope after they returned home. The importance measure, which was placed near the beginning of the questionnaire, listed the 22 items and used a five-point scale from not at all important (1) to extremely important (5). The instructions read: “How important was each of the following features to the quality of your overall experience at the Dickens festival?” Near the end of the questionnaire, separated from the importance items by several intervening questions, respondents were presented with the same set of items and asked to rate the festival’s performance in each area on a seven-point scale anchored with very poor (1) and very good (7). The instructions read: “We previously asked you to rate the importance of features of the Dickens event, but now we are interested in how good you thought each of them was.” In order to obtain a direct overall measure of quality, respondents also were asked to rate overall quality of the festival on a seven-point scale.

A systematic sample was selected by contacting every nth person as they entered through selected gates. During the initial heavy visitation period on Saturday morning, every 60th new visitor who passed through the same gates was interviewed. Then every 30th and 15th new visitor on Saturday and Sunday, respectively, to reflect the smaller numbers entering the site during those periods. The times at which contacts were made were randomized throughout the two days of the event, and gates were selected based on historical flow patterns recorded in previous years. A total of 586 festival visitors agreed to participate in the surveys and the refusal rate was less than 10%. Of these respondents, 80% (486) returned the mailback questionnaire encouraged by three follow-ups and an incentive. The on-site and mail surveys of each respondent were matched for analysis. Generally, all adult-child groups were consistently presented in the sample. Female respondents (62%) outnumbered males (38%). Visitors were relatively affluent, with almost 60% reporting household incomes in excess of $40,000 and more than half of those (31% of the total) reporting a household income of $60,000 or more. The festival was perceived to be an adult- rather than a child-oriented event, since only 29% of respondents groups included children.

Construct validity concerns the degree to which “a measure relates to other observed variables in a way that is consistent with theoretically derived predictions” (Bollen 1989, p. 188). The classical approach to measuring construct validity is the multitrait-multimethod (MTMM) approach proposed by Campbell and Fiske (1959). However, a central requirement of this approach is that the multiple methods used to measure the quality constructs should be truly different. Using a different method in this context in essence would mean designing a separate study to complement the one reported here. This would require a level of resources beyond those available to the authors, a problem common to those working in this area. Thus, rather than use the MTMM approach in its entirety, researchers have used the criterion of convergent validity implied in that approach (Parasuraman, Zeithaml, and Berry 1986, 1995; Cronin and Taylor 1992; Teas 1993). Parasuraman, Zeithaml, and Berry (1995) state, “a single-time overall service quality measure is appropriate for examining the convergent validity and predictive power of alternative service quality measures” (p. 27).

Correlation and regression procedures were used to measure the predictive validity of the seven alternative operationalizations with an overall measure of festival quality. Given that respondents were required to complete three sets of 22 items measuring their expectations, importance, and performance, it was anticipated there would be some missing data. It was found that only 215 (44%) of the 486 respondents completed all 66 items. Rather than lose this many cases in the analyses, the researchers identified cases in which responses to four or fewer items were missing in any of the three sets of items. Missing values for the retained responses were estimated by calculating the mean item scores for each respondent. This strategy enabled researchers to retain a substantial number of observations while minimizing the impact the missing data values may exert on the study results (Tatsuoka 1971; Wesolowsky 1976; Kirk 1982).

This process resulted in the retention of 418 cases (86%) in the analyses. When analyses were run with both the original and augmented number of cases, the results were very similar. There were differences in the magnitudes of coefficients, but not in the comparative ranking of the alternative quality operationalizations in terms of their convergent validity. The results reported here are derived from the augmented set of responses.

RESULTS

H1a: Performance will be a substantially better predictor of quality than the six alternative operationalizations when tested on all selected attributes of the festival.

H1b: Importance scores will not substantially improve the predictive validity of either the performance-based operationalizations or the disconfirmation-based operationalizations of quality.

The hypotheses were tested in two ways. First, Pearson product moment correlations were computed between each of the seven alternative quality operationalizations, using respondents' scores on the set of 22 items and their responses on a seven-point scale, anchored with very poor and very good, to the question, “I rate the overall quality of the Dickens event as . . . .” The resulting correlation coefficients are shown in Table 1. Although the correlations were moderately high (ranging from .38 to .66), there were substantial differences among them. H1a was supported since the performance measure showed the highest correlation, and it was closely followed by the other measures that linked performance with importance. The lowest correlations were those between the Expectations (Alt 1) and Importance times Expectations (Alt 3) operationalizations, and the two disconfirmation measures, Performance minus Expectations (Alt 5) and Importance times (Performance minus Expectations) (Alt 6). H1b was also supported since importance weightings assigned to the Performance and the Performance minus Expectations measures did not increase the correlations, and there was only a marginal increase when they were assigned to expectations.

The second method used to test the hypotheses was a regression analysis that regressed the alternative operationalization scores on the overall quality measure. Each of the 22 attributes was used in the regression analysis. The 22 attributes were eliminated using a backward stepwise procedure where at each step the variable with the largest probability-of-F value was removed, provided that the value was larger than 0.10. The resulting multiple-R values are shown in Table 1. Again, performance explained more of the variation in the global measure of quality than the other operationalizations, and...
TABLE 1
PEARSON CORRELATIONS AND MULTIPLE-R VALUES OF ALTERNATIVE OPERATIONALIZATIONS OF QUALITY

<table>
<thead>
<tr>
<th>Alternative Operationalizations of Quality</th>
<th>Pearson Correlation Coefficients</th>
<th>Multiple Value for Regression Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expectations</td>
<td>0.3830</td>
<td>0.3566</td>
</tr>
<tr>
<td>Importance times Expectations</td>
<td>0.4103</td>
<td>0.3818</td>
</tr>
<tr>
<td>Performance</td>
<td>0.6616</td>
<td>0.6506</td>
</tr>
<tr>
<td>Importance minus Performance</td>
<td>0.5935</td>
<td>0.5791</td>
</tr>
<tr>
<td>Importance times Performance</td>
<td>0.5859</td>
<td>0.5679</td>
</tr>
<tr>
<td>Performance minus Expectations</td>
<td>0.4212</td>
<td>0.3886</td>
</tr>
<tr>
<td>Importance times (Performance minus Expectations)</td>
<td>0.4109</td>
<td>0.3824</td>
</tr>
</tbody>
</table>

*The correlations and R values relate to an overall measure of quality.*

again it was followed by the other measures that link importance to performance. Also, importance did not increase the Performance and Performance minus Expectations multiple-R values, so H1b was again supported.

H2: Correlations of the alternative operationalizations of quality with a measure of overall quality will be higher among visitors who reported relatively frequent past visits to the festival, than among those who had never been previously to the festival.

Responses to the alternative quality operationalizations of first-time visitors and frequent past visitors were compared. Frequent past visitors were defined as those who had visited the festival at least three times including the current visit. For each group, correlations were computed in which the seven operationalizations of quality were measured against the overall quality measure. The resulting coefficients are shown in Table 2. They indicate that all of the operationalizations of first-time visitors exhibited higher convergent validity than those reported by frequent visitors. Table 3 reports the means and standard deviations of the two groups’ scores on each of the alternative operationalizations. One-way analysis of variance tests indicated that there were no statistical differences between the two groups on any of the alternative operationalizations. These results were confirmed by a t-test that revealed no significant difference in assessment of the overall quality of the festival by first-time visitors (mean = 5.535; s.d. = 1.151) and frequent visitors (mean = 5.587; s.d. = 1.454).

DISCUSSION

In the tourism field a variety of theoretically sound or intuitively appealing conceptualizations of quality have been proposed in the literature, but relatively few empirical studies have been reported that compared their relative validity, which was the primary purpose of this study. Validity was measured by testing the convergent validity of alternative operationalizations with an overall measure of quality.

Each respondent was asked to rate 22 attributes of a festival on expectations of level of quality before the experience, their importance, and the quality of their performance. In addition, respondents evaluated quality of overall performance at the festival. These values were combined in a variety of ways to operationalize seven of the most popular conceptualizations of quality that have appeared in the literature.

The major findings were unequivocal. The best predictors of quality were the performance-based operationalizations; the least accurate predictors were the disconfirmation-based operationalizations; and the inclusion of importance weights did not improve predictive validity of the measures.

The results reported here confirmed H1a, which was grounded in findings reported by others in the literature. It appears that respondents either did not form meaningful expectations or, if they were formed, did not use them as criteria against which they measured performance to determine quality.

Earlier in the article, four explanations were offered that may have accounted for these types of findings. However, two of these do not appear to be feasible explanations for the higher predictive validity of performance measures compared to discrepancy measures found in this study. First, it was suggested that cognitive dissonance may occur in situations where the expectations and performance items are both completed post facto. However, this study used a longitudinal design in which expectations were reported prior to the experience taking place, which is likely to ameliorate this problem. Second, the results shown in Tables 2 and 3 appear to refute the notion that greater familiarity with a destination may cause expectations to be less tentative so they serve as firm criteria against which evaluative judgments are made.

A factor that could have contributed to the inaccuracy of the expectations measure may be termed a “halo” effect. The researchers were alerted to this source of invalidity by the responses to another question on the instrument dealing with the image of the festival’s sponsors, which was unrelated to this study. Respondents were presented with six pairs of companies. Each pair comprised one company that was a festival sponsor and one that was not a sponsor but was well-known and in the same industry. One additional pair of companies was included as a control, though neither of them was a sponsor. Respondents were asked to express how they felt about these companies on a seven-point scale (anchored with very poor [1] and very good [7]), both before they entered the festival grounds and in the follow-up instrument returned by mail after their experience. For 13 of the 14 companies, the mean scores reported after the event were lower than those
reported before the event. The result was consistent across both sponsor and control companies. It appears that people completing the pre-entry questionnaire were in a more generous, positive, optimistic frame of mind regarding these companies than when they completed the mailback instrument later. It seems reasonable to conject that this phenomenon may extend to the expectations instrument and suggest that the expectations scores were affected by the emotional excitement that often accompanies anticipation of a good time. In contrast, the performance evaluation was made in the entirely different context of respondents’ normal-life routine, suggesting that a more considered, unemotional evaluative response may have been forthcoming.

The remaining two explanations for the predictive superiority of the performance-based measures over the conceptually richer discrepancy measures discussed earlier

<table>
<thead>
<tr>
<th>Alternative Operationalizations of Quality</th>
<th>First-time Visitors</th>
<th>Frequent Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>Expectations</td>
<td>5.7532</td>
<td>0.7606</td>
</tr>
<tr>
<td>Importance times Expectations</td>
<td>21.8945</td>
<td>4.8523</td>
</tr>
<tr>
<td>Performance</td>
<td>5.2244</td>
<td>0.8625</td>
</tr>
<tr>
<td>Importance minus Performance</td>
<td>-1.4433</td>
<td>0.7160</td>
</tr>
<tr>
<td>Importance times Performance</td>
<td>20.1596</td>
<td>5.3140</td>
</tr>
<tr>
<td>Performance minus Expectations</td>
<td>-0.5289</td>
<td>-0.8439</td>
</tr>
<tr>
<td>Importance times (Performance minus Expectations)</td>
<td>-1.7349</td>
<td>-3.2359</td>
</tr>
</tbody>
</table>

**TABLE 4**

**The Mean Scores and Ranks of 22 Festival Attributes When Evaluated by Three Alternative Measures of Quality**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Performance Minus Expectations</th>
<th>Importance</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Rank</td>
<td>Mean</td>
</tr>
<tr>
<td>The visual appearance of the Strand Historic District</td>
<td>-0.27</td>
<td>7</td>
<td>4.03</td>
</tr>
<tr>
<td>The quality of the entertainers</td>
<td>-0.24</td>
<td>6</td>
<td>3.87</td>
</tr>
<tr>
<td>Information booths giving site directions and performance information</td>
<td>-0.62</td>
<td>18</td>
<td>3.43</td>
</tr>
<tr>
<td>The central Christmas tree</td>
<td>-1.08</td>
<td>24</td>
<td>3.09</td>
</tr>
<tr>
<td>Feeling of safety on the site</td>
<td>0.00</td>
<td>2</td>
<td>4.33</td>
</tr>
<tr>
<td>Cleanliness of the portable restrooms</td>
<td>0.15</td>
<td>1</td>
<td>4.29</td>
</tr>
<tr>
<td>Number of places to sit down and rest</td>
<td>-0.49</td>
<td>14</td>
<td>3.73</td>
</tr>
<tr>
<td>The printed program showing event locations and performance schedules</td>
<td>-0.36</td>
<td>10</td>
<td>3.68</td>
</tr>
<tr>
<td>The Dickens Parade of Characters</td>
<td>-0.64</td>
<td>19</td>
<td>3.70</td>
</tr>
<tr>
<td>Cleanliness of the festival site</td>
<td>-0.29</td>
<td>8</td>
<td>4.12</td>
</tr>
<tr>
<td>Quality of food and beverage</td>
<td>-0.45</td>
<td>13</td>
<td>3.97</td>
</tr>
<tr>
<td>The Handbell Festival</td>
<td>-0.74</td>
<td>21</td>
<td>3.39</td>
</tr>
<tr>
<td>The number of people in Victorian costumes</td>
<td>-0.32</td>
<td>9</td>
<td>4.07</td>
</tr>
<tr>
<td>Availability of restrooms</td>
<td>-0.02</td>
<td>5</td>
<td>4.13</td>
</tr>
<tr>
<td>The decorative lighting</td>
<td>-0.58</td>
<td>17</td>
<td>3.72</td>
</tr>
<tr>
<td>Availability of English types of food</td>
<td>-0.50</td>
<td>15</td>
<td>3.29</td>
</tr>
<tr>
<td>Street maps on the site giving directions</td>
<td>-0.44</td>
<td>12</td>
<td>3.60</td>
</tr>
<tr>
<td>Friendliness of people in the Strand shops</td>
<td>-0.40</td>
<td>11</td>
<td>4.06</td>
</tr>
<tr>
<td>The indoor performances</td>
<td>-0.74</td>
<td>22</td>
<td>3.24</td>
</tr>
<tr>
<td>The characters from Dickens’ books</td>
<td>-0.57</td>
<td>16</td>
<td>3.72</td>
</tr>
<tr>
<td>The variety of gifts at the booths</td>
<td>-0.74</td>
<td>23</td>
<td>6.36</td>
</tr>
<tr>
<td>The Clydesdale horses</td>
<td>-0.64</td>
<td>19</td>
<td>3.51</td>
</tr>
</tbody>
</table>
in the article may apply in this study. Both of them are intuitively appealing explanations, but they lack meaningful empirical verification in the literature. The first is the suggestion that pre-purchase choice criteria and post-purchase evaluation criteria are not the same. Thus, the a priori importance attached to attributes at the expectations stage may change after the experience, leading to attribute performance being evaluated from an entirely different perspective.

Second, assimilation/contrast theory suggests there may have been a convergence of expectation and performance scores. This theory states that visitors are likely to perceptually distort discrepant stimuli up to some threshold, so that the stimuli will coincide with prior expectations; differences that exist beyond this threshold will be amplified. Thus, when the perceptual distance between performance and expectation is outside the threshold, amplification that heightens disconfirmation is likely because the perceived discrepancy between performance and expectation fell within the threshold point.

An additional contributing explanation, which Mannell (1991) suggests, is that evaluation of quality was guided by the extent to which visitors’ drives, motives, needs, or wants were met, rather than by their prior expectations about the festival.

Parasuraman, Zeithaml, and Berry (1994) have suggested that a partial explanation for the superior predictive validity of performance scores that have been reported may be an artifact of the “shared method variance” between the dependent and independent variable. In this context, Fiske (1982) states that the term “method” refers to the form of measurement and may include the content of specific items, scale type, response format, and the general context. Each of these potential method effects, which Fiske uses as examples, applies to the measures used in this study. Bagozzi and Yi (1991) note the use of similar or highly correlated methods tends to inflate the correlation between constructs. When the same method is used to measure different constructs, shared method variance always inflates the observed between-measure correlation because the correlation between the methods is 1.0 (which is higher than any possible correlation between the measures). (p. 426)

In this study, the dependent variable itself was a performance-based measure and, as such, was more similar to the performance than to the disconfirmation measures. To address this shared variance problem, future studies may adopt a triangulation approach and use multiple measures to assess the convergent validity of performance and disconfirmation measures (Campbell and Fiske 1959). This would enable distinctions to be made between the substantive variance of the two constructs and the unwanted method variance. However, it was noted earlier in the article that in this context multiple measures will probably involve designing a series of complementary studies because of the difficulty of incorporating them into a single study.

Perhaps the most surprising findings were those reported in Tables 2 and 3. It is reasonable to expect that greater experience will lead to firmer expectations and, hence, superior discrepancy scores. There is empirical evidence to support this supposition (Scott and Yalch 1980; Smith and Swinyard 1983). In the context of satisfaction, Whipple and Thach (1988) noted that “the amount and quality of previous experience may determine the effect of expectations on satisfaction” (Day 1977; Woodruff, Cadotte, and Jenkins 1983). Therefore, first-time purchasers’ satisfaction should depend on performance” (p. 16). This study found that among frequent visitors, discrepancy measures were not superior to performance measures, nor were the discrepancy measures reported by frequent visitors superior to those reported by first-timers. Further, no differences were found on any of the alternative operationalizations between first-time and repeat visitors.

Three suggestions are offered that may contribute to explaining this unexpected finding. First, previous experience is only one source that may be used to establish expectation norms. Cadotte, Woodruff, and Jenkins (1987) suggested that these norms may also be derived from experience at other festivals. Such influence in forming expectations may come from the most liked, the last visited, the most visited, or the average performance an individual believes to be typical of festivals. In addition, Parasuraman, Zeithaml, and Berry (1985) acknowledged the importance of external communications to visitors in shaping expectations. Because Dickens on the Strand is a particularly photogenic attraction, it receives extensive media coverage, especially from television and the press. Hence, first-time visitors may have elected to come because media exposure created relatively clear expectations and suggested the festival incorporated attributes that were important to them. Thus, the relative homogeneity of first-time and frequent visitors’ expectation and importance scores may be attributable to their similar experience levels, derived either directly from past visitation or indirectly from media images.

Second, the unexpected finding may be an artifact of operationalization. “Relatively frequent” past visits was defined as two or more previous visits. Since this was the 18th year of the festival, some respondents’ past visits may have been rather infrequent and have taken place many years previously, possibly resulting in a degree of fuzziness or tenuousness in their expectations. Therefore, these expectations may not have served as firm criteria against which evaluative judgments were made. Third, festivals are dynamic attractions. Repeat visitation is the key to their longevity and in order to attract people back they have to change. If the nature of expected attributes is altered so they are perceived to be different from previous visits, the utility of expectations derived from former experience as measuring criteria is reduced and may be no greater than that of first-time visitors.

**IMPLICATIONS**

Paraphrasing Dorfman’s (1979) observation of his evaluation of different measures of satisfaction, the research reported here illustrated that the quality of a festival (and presumably other elements of tourism) can be conceptualized and measured in many ways. It is unreasonable to assume that two systems, each designed to measure festival quality, are measuring the same aspect or are even highly correlated. Research employing different measures of quality as dependent variables may therefore obtain different results due to the particular system used. It seems reasonable to suggest that there may be no single best way to measure festival quality, but rather that different systems of measurement may be useful for different purposes.

Ostensibly, the superiority of the performance-based measure appears to be an important finding, because this is the most popular conceptualization and measurement of quality used by practitioners. The performance-based measure’s superiority from a predictive-validity standpoint has been
consistently supported in recent literature (Carman 1990; Cronin and Taylor 1992; Boulding et al. 1993; Babakus and Boller 1992; Babakus and Mangold 1992). Parasuraman, Zeithaml, and Berry (1994) now recognize “the somewhat lower predictive power of the P-E measure relative to the P-O only measure” (p. 120). Given the consistency of the results in this study with this emerging consensus, the appropriate conclusion would appear to be that the more cumbersome disconfirmation-based approaches and the use of importance weights should be discarded. However, the authors believe this conclusion is inappropriate for both pragmatic and conceptual reasons.

From a pragmatic perspective, considering only performance ratings may mean that key insights about a festival’s quality are lost and suboptimal investment of resources may be made in efforts to improve quality. As Parasuraman, Zeithaml, and Berry (1994) note, “Executives in companies that have switched to a disconfirmation-based measurement approach tell us that the information generated by this approach has much more diagnostic value” (p. 116). This is illustrated in Table 4, which reports the Performance minus Expectations, Importance, and Performance mean scores and rankings for the 22 festival features. The items ranked 20 and 21 on the performance scales were ranked 5 and 1, respectively, on the Performance minus Expectations measure. These items related to availability and cleanliness of restrooms. In absolute performance terms they were perceived as being lower quality than the other festival features. However, “cleanliness of portable restrooms” was the only feature for which expectations were exceeded and the other restroom feature almost met visitor expectations, suggesting that quality of these two features may not constitute a major problem. From a managerial perspective, it would seem important to track trends of the extent to which expectations are met over time as well as trends in performance. However, if people adapt their level of expectation to prevailing circumstances as Mannell (1989) suggests, then little change in levels of quality over time would occur.

Similar pragmatic insights are offered by incorporation of the importance scale. The study’s results were consistent with those of Dorfman (1979) in showing that weighing attributes by importance did not substantially improve either the performance or the discrepancy measures. Again, however, disregarding importance may mean losing useful insights. If respondents perceive performance on some attributes to be relatively low, then the first reaction of festival organizers may be to invest in improving those features. However, such investment may only be wise if those attributes are relatively important to visitors. For example, the items “information booths which gave site directions and performance information” and “cleanliness of the festival site” were ranked approximately equally (13 and 14) on importance (Table 4). However, the payoff to the festival is likely to be much greater if effort is focused on improving cleanliness of the festival site than on the information booth, because the former was ranked 4 on the importance scales, while the latter was ranked 18.

Parasuraman, Zeithaml, and Berry (1994) observed that it is important in assessing the practical value of alternatives to performance-based measures to ask “Are managers that use service-quality measures more interested in accurately identifying service shortfalls, or in explaining variance in an overall measure of perceived service? We would submit that managers are much more interested in accurate diagnosis of service quality problems” (p. 116). They argue that the superior diagnostic power of other measures more than offsets the reduction in their predictive power. In a more general vein, they observe that “a need and an opportunity exist for explicitly incorporating practical criteria (e.g., potential diagnostic value) into the traditional scale-assessment paradigm that is dominated by psychometric criteria” (p. 120).

The conceptual reason for continuing to use discrepancy measures stems from their conceptual richness. The leisure and marketing literatures have long accepted discrepancy as central to descriptions of the satisfaction and quality constructs. Hence, the authors recognize that operationalizations used in this study may not be accurate enough to capture the richness of the concept. That is, the findings may reflect inadequacy of current operationalizations of discrepancy and expectations, rather than inadequacy of the concept.

REFERENCES


