CFI Group

The American Customer Satisfaction Index (ACSI) Technology: A Methodological Primer

By CFI Group

The American Customer Satisfaction Index (ACSI) Technology: A Methodological Primer

Introduction

Background: The *American Customer Satisfaction Index (ACSI)* has a proven relationship with customer spending,[1] shareholder value,[2, 3] cash flows,[4] business performance[5] and GDP growth.[6] The technology upon which it is based is backed by over 70 years of rigorous scientific inquiry in the fields of consumer psychology and psychometrics, coupled with advanced analytic techniques from statistics, econometrics, and chemometrics.[7] While applicability of the ACSI technology to the management of commercial product and service companies has been repeatedly demonstrated in the literature,[8, 9] this paper provides a basic description of the underlying analytic technology.

Purpose: The purpose of this paper is to provide the reader with an overview of how the ACSI technology as delivered by CFI Group meets the performance measurement requirements of the business community and improves the management and delivery of products and services to customers. This is accomplished by:

- Describing the rigorous methodology used by CFI Group to harness the power of the ACSI technology. This section focuses on highlighting the critical elements of the methodology that provides highly accurate measurement coupled with sensitive diagnostic and powerful prognostic capability.
- Summarizing the major components of the technology as implemented by CFI Group and the benefits realized by managers.
- Comparing the ACSI technology base with some of the more common alternative methodologies offered by competing firms.
- Reviewing the results achieved by CFI Group clients with a brief compendium of case studies illustrating many of the points made throughout the paper.
- Providing technical appendices with in-depth discussions of various aspects of the ACSI technology as implemented by CFI Group.
- Documenting the scientific nature of the ACSI technology by numerous references to secondary literature throughout the paper and in the bibliography.

---

[1] Claes Fornell and Roland Rust, "The effect of customer satisfaction on consumer spending growth," under review, 2005.

[2] Claes Fornell, Sunil Mithas, Forrest Morgeson, and M. S. Krishnan (2006), "Customer Satisfaction and Stock Prices: High Returns, Low Risk," *Journal of Marketing*, Vol. 70, No. 1, 3.

[3] Eugene Anderson, Claes Fornell and Sanal Maznancheryl (2004) "Customer satisfaction and Shareholder Value," *Journal of Marketing*,(October) Vol. 68, no.4, 172.

[4] Gruca, Thomas S., and Lopo L. Rego (2005) "Customer Satisfaction, Cash Flow, and Shareholder Value," *Journal of Marketing,*(July) Vol.69, 115-130.

[5] Morgan, Neil and Lopo Rego (forthcoming 2006), "The Value of Different Customer satisfaction and Loyalty Metrics in Predicting Business Performance," *Marketing Science*.

[6] Claes Fornell, Paul Damien, Marcin Kacperczyk, and Michel Wedel, "The Empirical Relationship between Buyer Satisfaction and GDP Growth under Parameter and Distributional Uncertainty," under review, 2004.

[7] The main difference between econometrics and chemometrics is that while both are focused on prediction, chemometric methods do a superior job of identifying and separating components from the underlying "noise" in a measurement system. The ACSI technology utilizes a type of chemometric analysis to extract meaning from the inter-correlations between predictors in structural equation models. See Svante Wold's article "Chemometrics: what do we mean with it, and what do we want from it?" in *Chemometrics and Intelligent Laboratory Systems*, 30 (1995) 109-115, for more details.

[8] Fornell, Claes, Michael D. Johnson, Eugene W. Anderson, Jaesung Cha and Barbara Everitt Bryant, (1996), "The American Customer Satisfaction Index: Nature, Purpose and Findings," *Journal of Marketing*, Vol. 60, October, 7-18.

[9] Anderson, Eugene W., Claes Fornell and Roland T. Rust (1997), "Customer Satisfaction, Productivity and Profitability: Differences Between Goods and Services," *Marketing Science*, Vol. 16, No. 2, 129-145, Summer.

Rigor—What is the CFI Group Advantage?

The simple essence of the CFI Group's implementation of the ACSI technology is *measurement*, *diagnosis* and *prognosis*.

Building upon the knowledge developed from 70 years of social psychology research, CFI Group *measures* the three levels of a customer's thought process resulting from an experience with a product or service:

- *Perceptions of the performance* delivered by the various facets of the product and/or service experience,
- Overall *attitudinal evaluation* of the experience, and
- *Future behavioral intentions* towards the product or service in question.

These measures are embedded in a *diagnostic model* of cause and effect linkages that helps quantify the measures while at the same time empirically connects the three measurement levels; i.e., how do perceptions affect evaluation, and how does evaluation affect future intentions. The linkages quantify the changes that are necessary at one level to effect the greatest amount of change in the subsequent measurement level.

Perceived Performance →	**Attitudinal Evaluation** →	**Future Intentions**
"Service Delivery"	***"Satisfaction"***	***"Use Service Again"***

Finally, the *diagnostic* framework is then used to provide *prognoses* about how best to invest resources in programs, practices and procedures that affect the perceived performance levels of products or services, and what can be expected (in terms of evaluation and future intentions) as a result of the investments.

For commercial enterprises this powerful set of metrics, with their cause and effect linkages, gives a company an unequaled ability to manage the economic or relationship value of its customer base by providing *marginal resource allocation* guidance for product and service quality.[10] In the following sections each of these elements (measurement, diagnosis and prognosis) is described in detail.

[10] The *marginal resource allocation* concept is sometimes called "derived" importance. It should be noted that in the cause and effect measurement networks executed by CFI Group, all experience facets are fundamentally "important" to the customer/citizen. However from a prognosis perspective the concern centers on how to achieve the greatest amount of change in a desired outcome (e.g., satisfaction), so the issue is most efficient marginal allocation of resources—not the reallocation of resources. An efficient allocation of resources is an allocation that satisfies the rule marginal benefit=marginal cost for each area of investment.

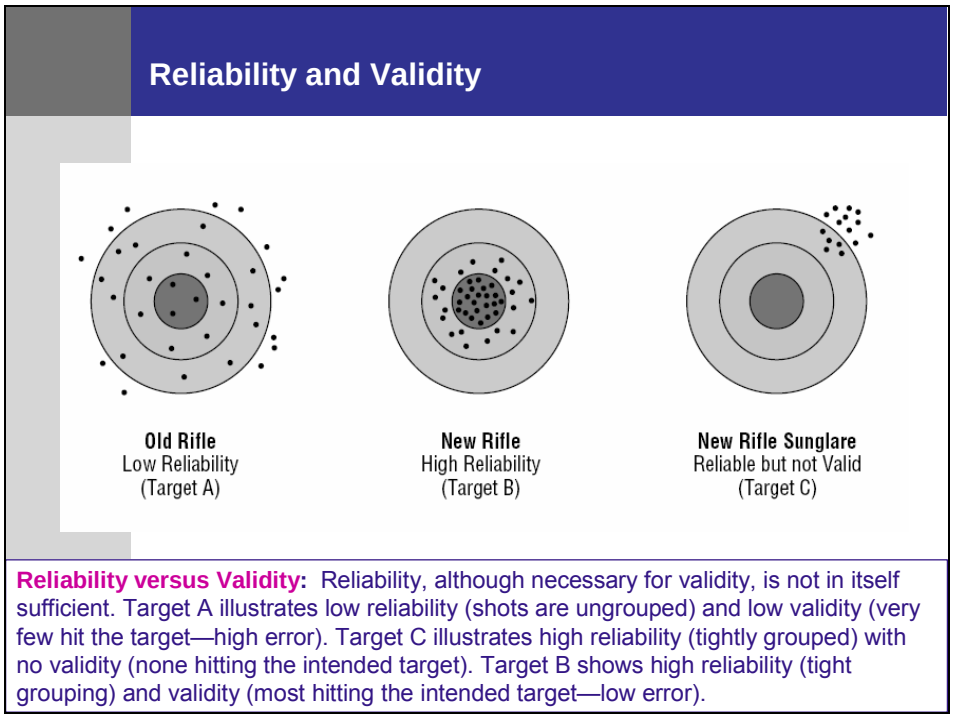
Measurement

Good measurement requires *reliability* and *validity* for statistical precision, and it also requires *sensitivity* for statistical power.

- **Reliability**: Reliability is the quality of a measurement tool that allows it to obtain similar results over time and across situations (this is also referred to as the *internal consistency* of a measure). It is the *degree to which measures are free from random error and therefore yield consistent results.*
    - Example: a rifle that is fired at a target the same way each time by the same rifleman should result in the same pattern of hits each time it is fired. If it does, then the rifle is considered to be reliable. If it doesn't, then there may be a flaw in the construction of the rifle (the sights are loose) that prevents it from being consistent.

- **Validity**: Validity is the quality of a measurement tool to *measure what we intend it to measure*. In other words, extending the rifle analogy, does the rifleman hit the bull's-eye of the target? It is the *degree to which measures are free from measurement error and reveal the truth about an object or quality of an object*.
    - For example, in measuring "intention to buy", if a question is not worded correctly there could be a systematic bias to identify brands "I wish I could afford" rather than the brand usually purchased.

- **Sensitivity**: The sensitivity of a measurement tool is important, particularly when changes in attitude, or other hypothetical constructs, are under investigation. Sensitivity refers to the ability of an instrument to identify variability in stimuli or responses over successive measurement occasions or between groups (*power to detect change*).
    - Adding additional questions or items can increase the sensitivity of a single question or single item scale.
    - In other words, because index measures allow for a greater range of possible scores, they are more sensitive than single-item scales.



Reliability versus Validity: Reliability, although necessary for validity, is not in itself sufficient. Target A illustrates low reliability (shots are ungrouped) and low validity (very few hit the target—high error). Target C illustrates high reliability (tightly grouped) with no validity (none hitting the intended target). Target B shows high reliability (tight grouping) and validity (most hitting the intended target—low error).

The ACSI technology implemented by CFI Group is based upon an advanced measurement and analysis system that combines best practices from psychometric science with an advanced causal modeling algorithm that insures potent levels of precision (validity combined with reliability) and power (sensitivity—ability to detect change).

What are the salient characteristics of the CFI Group system that make it superior to competitive measurement approaches?

- The use of "*voice of the customer*" (VOC) techniques to discover the true meaning of a customer's experience and convert the customer's "voice" into survey questions.[11] VOC techniques are far superior to alternative methods for developing questionnaires that rely upon judgment or experience of researchers.

- Reduction of measurement error through the use of *multiple measures* of important experience factors and satisfaction levels. It is a well documented scientific fact that the use of multiple item measures are far superior to single items for capturing the underlying "truth" of customer experiences and satisfaction. Multiple item measures are the best way to measure intangible psychological concepts such as performance perceptions and attitudes, since a single measure has a very high probability of "missing the target." Why this is so is addressed below.

- The derivation of *optimal measure weights* based on the cause and effect relationships between experiences, evaluations and intentions for combining the multiple measures into a single index.

How the CFI Group Measurement System Realizes Precision and Power: The ACSI technology relies upon advanced psychometric science as the basis for developing valid and reliable multiple-item measures.

In general there are three reasons for using multi-item measures.[12] The first issue relates to the *specificity* of individual items with respect to a particular trait. Single item measures usually have lower correlations with the particular phenomena being investigated and may also be correlated with other characteristics or phenomena at the same time. For example, on a spelling test, the correct spelling of the word *umpire* may reflect the spelling ability of the test taker, but it also may reflect the interest of the speller in baseball. A child who spent much time reading baseball stories might spell the word correctly even though he or she was a poor speller in general. This lack of specificity is a serious problem that can be remedied by the use of multiple measures, assuming that the measures are well designed.

A second issue in measurement is the ability of the measure to make fine *distinctions* between individuals. The greater the ability of a measure to make fine distinctions between individual respondents the more sensitive the measure for detecting changes. Dichotomous measures (e.g., yes/no, or "top-box") categorize respondents into two categories at most. A five-point or seven-point scale increases the number of distinctions to five or seven. In most measurement situations it is desirable to make as many fine distinctions among respondents as possible, and this can seldom be done with a one-item measure. An index like the ACSI made up of three ten-point items can make very fine distinctions between respondents because an individual's score is the average of the three ratings. This allows for a wide range of possible groupings from one group (if everyone in the sample answered all questions the same—an unlikely scenario) to an upper limit of *n* groups, where *n* = number of respondents.

The third issue is that individual items have considerable random *measurement error*. This is because any single item is basically unreliable in its ability to accurately measure a psychological phenomena. This can be demonstrated by asking a respondent to repeat a test procedure after a period of time. They may give a rating of 3 on one measurement occasion, and then indicate a 5 on the next repetition of the rating. This randomness in the ratings means that a single item cannot be trusted to give a reliable measurement of a psychological construct—like intelligence. This unreliability averages out when scores on numerous items are combined to obtain a total score, which is then highly reliable.

Fundamentally *the main focus in measurement should be on insuring measure validity*. While there are different types of validity the most important is *construct validity*—i.e., does the measure specifically measure what it purports to measure.

CFI Group competitors often violate construct validity. For example, while there are a number of ways to measure satisfaction, most firms make the mistake of treating satisfaction as a simple binary concept. Simple in the sense that only one question is used; binary in the sense that customers are categorized as either satisfied or dissatisfied (a so called "Top Box" approach) – often in percentage terms (e.g., we have 80% satisfied customers) or frequency counts. This approach is flawed because it violates the three rules mentioned above and consequently does not provide

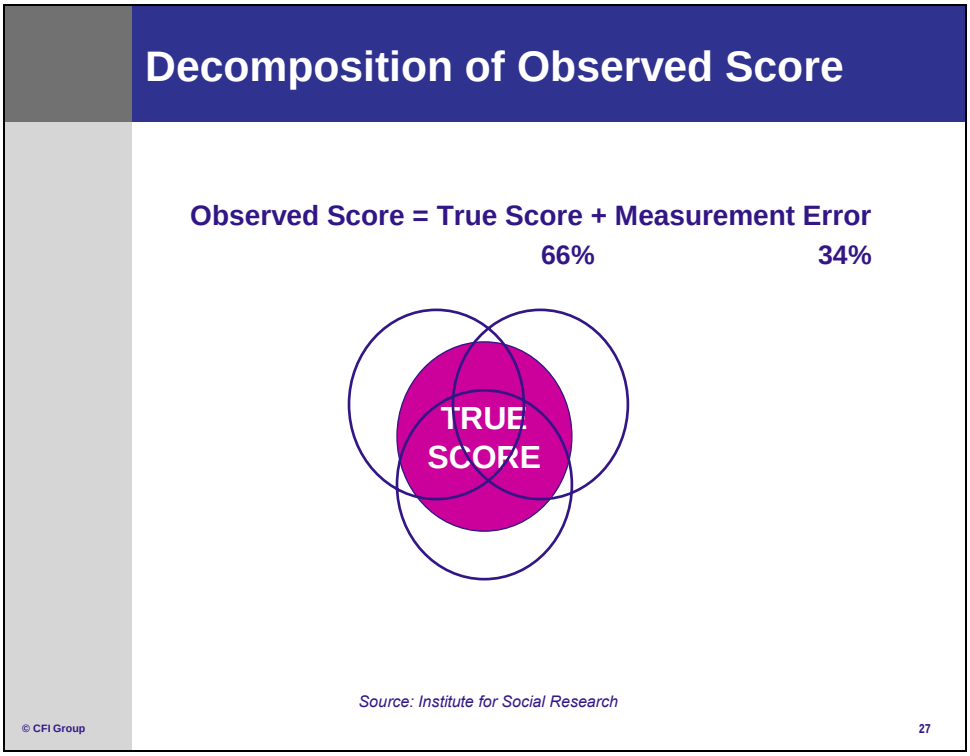
[11] Griffin, Abbie and John Hauser, "The Voice of the Customer," *Marketing Science*, Winter, 1993, 12,1,1.

[12] Nunnally, Jum C. (1978) Psychometric Theory (Second Edition), McGraw-Hill Book Company, New York, pp. 66-67.

sufficiently valid information in a reliable manner[13] (this is because there is more measurement error in "Top Box" measures and a lower likelihood of detecting a change in customer satisfaction). Given the low quality of the resulting metric it is not surprising that many firms fail to find any relationship between quality and satisfaction and between satisfaction and profit.

As an illustration, compare satisfaction, as a concept, to intelligence. Both are "multidimensional" (i.e., they possess many different aspects), and they are not directly observable (i.e., one cannot "see" intelligence or satisfaction by observing somebody). Any attempt to measure intelligence by a simple question (are you dumb or smart?) is not likely to yield useful information. It is not reasonable to think that one can assess a person's intelligence by a single question (or by a single test question). Likewise, it is not reasonable to assume that one can capture the concept of satisfaction by a single overall question (what if the target is missed? There is no "perfect" measure.).

The same logic also applies to the many different experiences that customers have with products or services. Each experience is multi-faceted. To get a "true" unbiased picture of what customers are experiencing requires a number of questions (3 to 5 is usually sufficient) to triangulate on the essence or truth of the experience. This is essential to have a valid measurement tool. As illustrated below, the more overlapped (and highly correlated) the individual measures are, the more valid (or true) the resulting combined measure is likely to be—i.e., the greater the likelihood of hitting the target.[14]



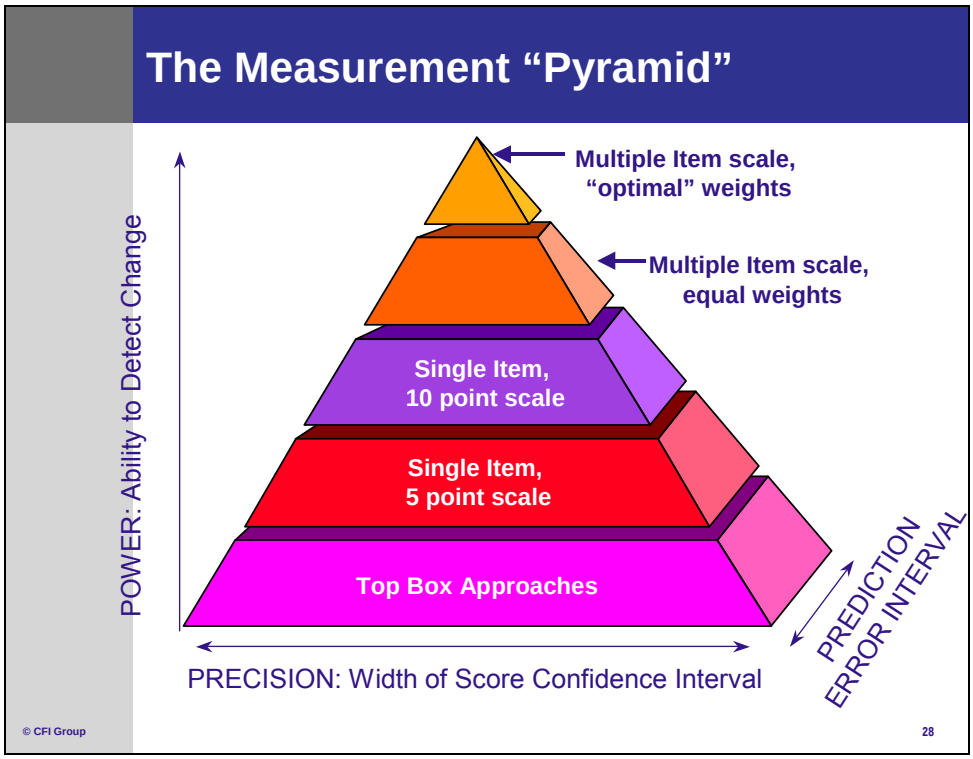
[13] Binary or dichotomous measures (also known as nominal scales) have 2 to 3 times the amount of error around the estimated population parameter (which is a proportion) than measures based on 10-point interval scaled measures (usually means) at the same confidence level.

[14] It is important to note that just because a measure uses multiple indicators does not ipso facto result in a "valid" measure. It depends on how the indicators were developed. Questionnaire items that are based on the judgment or guess work of the researcher may be completely unrelated to the concept being measured. The result will be a flawed multi-item measure that may give reliable results—but completely "miss" the target. Only by using VOC qualitative methods can one be reasonably confident that the customer measures are valid.

Clearly all measurement involves some degree of error. Ryan, Buzas and Ramaswamy (1995) found that the CFI Group measurement system leads to an increase in precision (expressed as confidence intervals) over traditional methods by 20-30%. This can lead to a direct reduction in sample size requirements on the average by 22% and still obtain the same precision as conventional methods. Also, the explanatory power with respect to the consequences of satisfaction (e.g., behavioral intentions) is 56% better than with conventional methods. This is a result of using multiple measures for overall satisfaction.[15,16] The increase in measurement precision implies that *smaller samples* can be used with the *same measurement precision* as traditional methods, which results in very high cost savings for the client (or, alternatively, in *higher precision* with the *same sample size*).

Without enough measurement precision in the satisfaction index, the achievement of a performance outcome (such as retention or repeat purchase) will suffer.[17] The reason is that lack of precision shows up as random variation in the measure. As a result, it will be much more difficult to identify how satisfaction changes as management institutes quality improvements. Overall, the importance of the gain in precision that the CFI Group system offers can hardly be understated. *In most cases, it would mean that the cost (to the client) of using CFI Group should be substantially lower than using a system by anybody else.* On the average, about 50% of the CFI Group cost is data collection and the size of the sample has a direct impact on precision.

The schematic below illustrates the relationships between precision, power and prediction error as a function of the type of measurement used. For more details about the identification of the appropriate questionnaire items see the VOC discussion in Appendix A. For an explanation about why 10-point scales are preferred in customer satisfaction measurement programs see Appendix B.



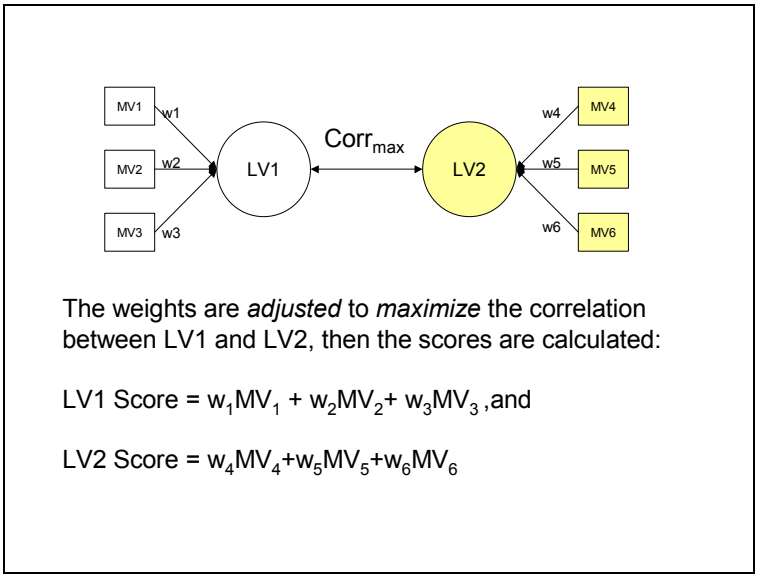
[15] Fornell, Rhee, and Yi "Direct Regression, Reverse Regression, and Covariance Structure Analysis," *Marketing Letters*, 1991, 309-320.
[16] Ryan, Michael J., Thomas Buzas and Venkatram Ramaswamy (1995), "Making Customer Satisfaction Measurement a Power Tool," Marketing Research, Vol. 7, No. 3, Summer, 11-16.
[17] Hauser, John R, Simester, Duncan I, Wernerfelt, Birger. "Internal customers and internal suppliers," *Journal of Marketing Research,* Aug 1996. Vol. 33, Iss. 3; p. 268

How Multiple Measures are Optimally Weighted: After good measures have been identified, a major issue in measurement is how best to combine the multiple measures into their respective indices—the formation of what is known as a "measurement model". The method chosen can have important effects on the analysis results, especially if the results will be used for diagnosis and prognosis.

The typical CFI Group measurement system is based upon a network of multi-dimensionally measured concepts that are linked together in a cause and effect framework. The scores of the various experience indices; the customer satisfaction index and the performance outcomes, are a function of the simultaneous optimization of the entire framework. This empirical process is superior to any other method for ensuring diagnostic and prognostic power. Competitors use methods that are piecemeal replicas by comparison.

For example, some firms in developing a satisfaction index use relative weights derived from the factor analysis[18] of a number of questions about different aspects of product or service on quality. The resulting index is simply a consequence of the shared aspects (correlation) of the questions without regard to some optimizing criterion such as a dependent variable like customer retention or other desired behavioral outcome. A particularly debilitating drawback of this approach is that if there are more questions about a particular attribute, that attribute will have a disproportionate representation in the index and can bias the resulting score. The fact that quality aspects correlate among themselves often has little to do with customers' satisfaction levels, yet some firms persist in using this confounded measure by mixing a customer's experience with their satisfaction levels—the causes are lumped together with the effects. Since the weights applied to the variables to create the satisfaction index are based on the inter-correlations among the quality measures themselves, there is little reason to expect that the resulting indices have any relationship with performance outcomes such as customer retention. Thus, this weighting scheme is based on an irrelevant criterion (inter-correlations as opposed to optimizing on an objective criterion). To be useful, a performance index or a satisfaction index must be based on a more relevant criterion (such as repurchase or willingness to pay, for example).[19]

The CFI Group system relies on a measurement model that empirically produces a system of optimally weighted indices. It is *optimal* because the weights for the product and service quality experience measures are derived based on the maximization of relationships (i.e., the correlations) between the various experience measures with customer satisfaction and future behavior. The way the system works is that the weights for all of the measures in the measurement model are "adjusted" so that the correlations between the variables along the cause and effect pathways in the measurement system are maximized. The simple two-component model shown below schematically illustrates the process.[20]



[18] The purpose of factor analysis is to discover simple patterns in the pattern of relationships among the variables. In particular, it seeks to discover if the observed variables can be explained largely or entirely in terms of a much smaller number of variables called *factors*.

[19] Other firms use even less sophisticated methods for combining individual items into a satisfaction index by relying upon summing or averaging of the ratings on the various questionnaire items.

[20] Note: The correlation is not the same as an impact. The correlation coefficient is simply used as the criteria for adjusting the weights in a manner that ensures the strongest relationships between the concepts in the model (LV1 and LV2 in the schematic) given the available information in the individual measures (MV1…MV6).

The weighting process used in the development of the measurement model is the first critical part of the CFI Group measurement system. Unlike other weighting schemes, an objective criterion of importance to managers (maximization of the relationships or correlation) is used to optimally weight the various measures in the product/service quality and customer satisfaction indices. Since the weights are determined based on the performance-satisfaction-behavior relationships in the model, this minimizes the common problem (experienced by competitors using less sophisticated weighting schemes) that an increase in a precursor index (e.g., service quality) does not lead to an increase in a successor index (e.g., customer satisfaction).

Diagnosis

Impacts versus Importance: As discussed above, the connective pathways between the experience indices, customer satisfaction and behavioral intentions play an important role in the determination of the weights used for score calculation. But these paths also provide the backbone for the second key feature of the CFI Group measurement system—impacts.

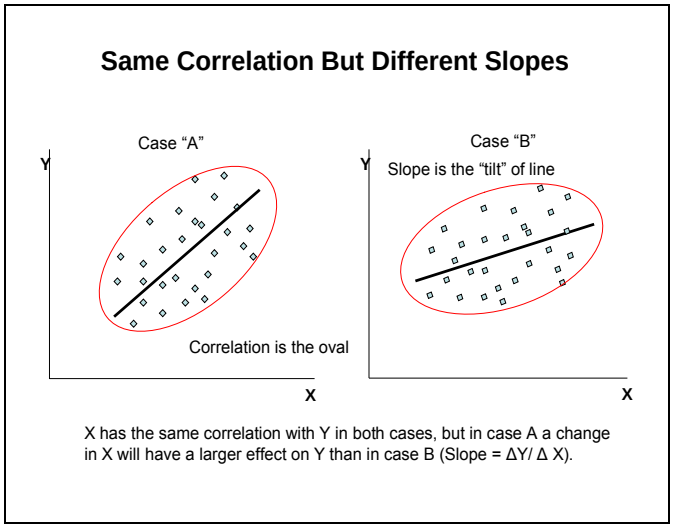
The most fundamental task of any organization (commercial or government) is the efficient allocation of scarce resources needed to accomplish desired performance outcomes. The CFI Group system quantifies the impact of experience changes on satisfaction and, in turn, the impact of satisfaction on future behavior. Managers can then use this information for efficient resource allocation. What are the properties of the CFI Group system that makes this possible?[21]

CFI Group's system is a cause-and-effect system that isolates the effects of a change in an experience on the change in customer satisfaction (and the subsequent change in desired behavioral outcomes). It is also characterized by a "simultaneous" treatment of all its components (i.e., quality, satisfaction, profit). All of these aspects make it different from other competitive approaches.

It is not well understood, but a cause-and-effect assumption is made every time a management decision is made ("if we do *x*, *y* will happen"). Unfortunately, managers often base their decisions on hunches, cross-tabs or correlation coefficients that do not support any sort of casual inferences. The CFI Group system is different. It supports causal inferences based on considerable scientific backing.

The reasons for this are several. The first is somewhat technical. The logic is the same as in path analysis and covariance structure analysis: the decomposition of correlations into causal paths. This involves a comparison of the empirical correlations in the data and the correlations imposed by the model (expected correlation matrix). If those sets of correlations are identical (within sampling error), there is evidence for the causal structure imposed by the CFI Group model (e.g., experience component *x* leads to customer satisfaction).

The second important point concerns what is meant by "effect." The CFI Group system defines this as the marginal effect of component x on y when other components are held constant — i.e., the effect of a *change* in *x* on *y*. If we graph *x* on the horizontal axis and *y* on the vertical axis, it is represented by the *slope* of the function as illustrated in the schematic below.



[21] The reader will find a more technical discussion of the CFI method for calculating impacts in Appendix A.

It is critical to understand this concept because it is different from what most other competitors provide and the results may be different from what seems intuitive to the client. Market research firms, for example, often talk about "importance" and use correlation coefficients as measures of importance. But a high correlation does not imply that a change in *x* will cause a change in *y*.

Other firms use "stated" importance measures, but these are equally flawed for the measurement of customer satisfaction. For example, Allen and Rao (2000) state that: "Few, if any, consultants advocate the stated importance framework today. Its shortcomings have been illustrated with the airline safety example in which stated and derived importance metrics lead to disparate conclusions."[22] In addition, such methods increase the length of the questionnaire by requiring shadow importance measures for every perceived performance or experience item included on the questionnaire. If ranking or constant sum scaling methods are used instead, then some kind of reduction of measures needs to be performed since respondents are psychically unable to rank or allocate points over more than 5-7 measures in a meaningful way. Plus this approach is not based on the sound psychometric principles of multiple measures and error reduction described above. Thus practitioners advocating stated importance methods are basically offering measures that have high levels or unknown levels of error in them, which is then exacerbated when the perceived performance/ importance pairs are manipulated either by multiplying or subtracting the measures to arrive at some confounded indication of "effect" or focus. Resource allocations targeted for the management of customer satisfaction and retention based on measures of this nature are akin to using a dartboard for decision-making and are ultimately doomed to failure.[23]

For management to efficiently allocate its resources, they need to know what will happen if there are changes (usually improvements) in a certain aspect of the customers' experiences – this is what CFI Group's system provides. It also means that the use of the term "important" in this context refers to what will happen as a *result of a change* in something – not what is important per se. For example, both price and quality can be highly correlated to satisfaction, but a change in one of them may produce a greater effect in terms of changing satisfaction than the other.

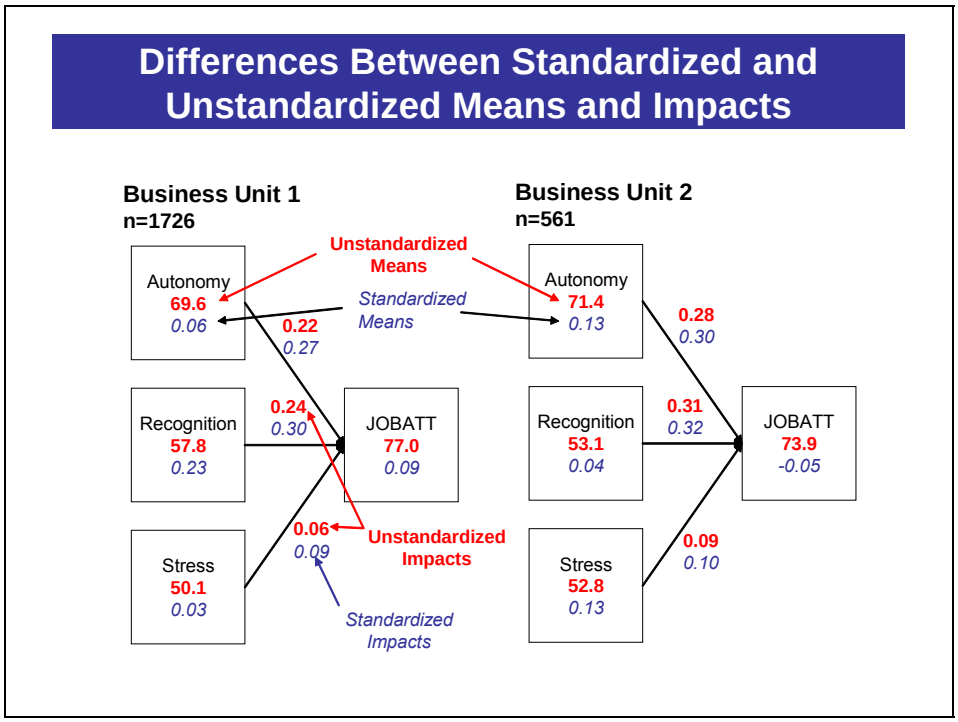
Quantifying Effects—Standardized or Unstandardized Measures?: The proper use of analysis tools is critical when quantifying effects. Other satisfaction analysts usually miss this point. For example, some firms in Europe use some of the same theoretical foundations (LV-PLS) as CFI Group, but do not understand that the core LV-PLS program is unsuitable without the CFI Group modifications. Basically, the problem is this: In order to solve the unknowns in equations with latent variables, some restrictions have to be put on the system – otherwise there would be too many unknowns. One set of restrictions, that are quite common in psychology, is to set all variances to unity and all means to zero – that is to *standardize* all variables. However, in terms of *quantifying effects*, standardization renders the results useless and destroys comparability between samples. What is then interpreted as importance is the impact of quality *x* on the spread (standard deviation) of satisfaction. This makes no sense and is, of course, very different from the CFI Group system (which does not rely on standardization). In practice, it turns out that our results are quite different from what the generic LV-PLS program provides. The modifications by CFI Group to the LV-PLS algorithm are proprietary and highly technical. They involve a solution to the multicollinearity problem and a rescaling method to insure comparability of results (see Appendix A for more detail).

The schematic below illustrates the problem with using standardized measures. The example shows two models for two different business units in the same company. The bolded (red) quantities are the unstandardized measures (component means and impacts), while the italicized (blue) quantities are the standardized measures (means and impacts). Using unstandardized measures is straightforward—for business unit 1, a 1 unit (point) change in the *Autonomy* score yields a 0.22 change in the *JobAtt* score. Using the standardized measures is less intuitive—for business unit 1, a 1 unit (standard deviation) change in *Autonomy* yields a 0.27 standard deviation change in *JobAtt*. Notice also the rather large differences in the standardized scores (*Autonomy* has a standardized score of 0.06 and *Recognition* is 0.23) of the variables both within each business unit model (reflecting the different variances for each component), as well as across business units (*Recognition* in unit 1 is 0.23, and in unit 2 it is 0.04).

[22] Allen, Derek and Tanniru Rao, *Analysis of Customer Satisfaction Data*, ASQ Quality Press 2000, p.70.

[23] One customer perceived value (CVP) practitioner advocates the misguided use of a perceived performance / stated importance measurement framework for the management of "customer loyalty" for all customers regardless of whether they are current customers or new customers. Why the concept of loyalty is germane to new customers is in itself puzzling. That aside, it is well known that retention strategies are quite different from acquisition strategies both in terms of content and costs. Consequently, the guidance dispensed from this confused measurement approach will certainly result in a mal-allocation of scarce resources for those who have unfortunately bought into this method.

This illustrates that because standardized measures are depended on the variation (or spread) in the data, which can differ from sample to sample, comparability is lost. For this reason, it is best not to compare groups using standardized means or impacts.



Multicollinearity: A very difficult problem in impact estimation is the isolation of the individual effect of each experience component from other components. This is because respondents tend to see many components as inter-related to some extent. This "halo" can contribute to high correlations between the components resulting in what is known as multicollinearity. No statistical technique is equipped to handle such multicollinearity and the result is misleading diagnosis. Normal LV-PLS and some other structural equation modeling techniques can help in reducing multicollinearity, but not enough to overcome the problem. The CFI Group system, however, is (1) able to extract the cause of multicollinearity and (2) apply a solution from the field of chemometrics to solve the problem.

Other consulting firms either ignore the problem at worst, or conduct a factor analysis of the experience components (thus grouping them together) and then regress customer satisfaction, or some other dependent variable, on the factor analysis groups. The problems with this approach are so serious that it is virtually impossible to make sense of the results.

- First, it destroys the meaning of the variables as they were originally conceived and measured; the resulting factors must be interpreted post-hoc by the analyst—raising questions of validity.
- Second, the imposed correlational structure among the factors is highly artificial and far removed from the how the respondents perceived things. The most common way is to force all the factors to be independent from each other (i.e., constrain the factors to have zero correlations with one another). This is most certainly wrong and very different from how the respondents perceived them—the "halo" effect.
- Third, usually the first factor extracted in a factor analysis solution will be totally overwhelming in terms of information (variance) content, which makes it necessary to use some sort of rotation scheme (introducing yet another artificial device) so the results can be interpreted by the analyst.
- Fourth, factor analysis plus regression represents a piecemeal two-step approach. Any errors existing in the first step are magnified by the second step—an optimal index cannot be constructed under this scenario. The post-hoc interpreted factors may not resemble those quality components that have maximal impact on satisfaction (and subsequent behaviors).

Two other approaches that are often used by firms to analyze satisfaction are stepwise regression and conjoint analysis. Stepwise regression assumes that absolutely nothing is known beforehand and everything is left to a sample of data points. In other words, the solution is an artifact of the data. As the name implies, stepwise regression is a technique for including "important" variables in a regression in a stepwise manner. The limitations of stepwise regression are:

- Notoriously unstable results;
- High likelihood of omitting a key variable;
- An inferior methodology if any theory exists;
- The results of stepwise regression cannot be evaluated by statistical significance testing; and
- The regression coefficients are biased.

Stepwise regression will almost never be used in articles published by respectable scientific journals (for the reasons given above).

Conjoint analysis is a different matter. In contrast to stepwise regression, conjoint analysis is a useful scientific method. The problem is that it is not well suited to the measurement and diagnosis of customer satisfaction. The basic problems are that it cannot handle many attributes and that there has to be a "level" of each quality attribute that the respondent is asked to evaluate. Conjoint analysis is more suitable for new product (service) development, in which respondents are asked to evaluate different prototypes (on paper) that have different levels of each attribute. For CFI Group, conjoint analysis can be used if a client is interested in finding out what customer satisfaction would be, if certain attributes were added to the product (service) and what the importance of each attribute would be. A nice benefit of conjoint analysis in this context is that it can be done on a single customer. The contrast between causal modeling methods and conjoint analysis is detailed in Appendix A.

Prognosis

The ultimate proof of a good measurement system is its ability to make accurate predictions. The models built on the principles described above provide managers with measurement-based tools for better management of intangible assets (like customers). With the patented process[24] used in the development of CFI Group measurement systems managers in commercial and public service organizations alike can be assured that they are getting valid, reliable and sensitive measures within a cause and effect framework that allows them to evaluate their decisions before they make them.

Once an initial model is built, the resultant component scores and impacts provide managers with high-powered metrics for determining the best courses of action they can take for accomplishing desired outcomes. Competing measurement systems "statically" compare self-reported importance measures against current performance measures. The CFI Group performance measurement approach provides a "dynamic" tool that tells managers what changes are important in affecting desired outcomes (e.g., increases in customer satisfaction). This distinction is a critical one for the success of resource allocation decisions that managers make daily. Without the knowledge of "what to expect" when executing a plan, decision-making devolves to a mere guessing game.

Most traditional approaches to market research either confuse comparison of levels (e.g., current performance and levels of importance as provided by customers) with marginal contributions (e.g., what should be changed), or fail to make the connections to desired performance outcomes (such as economic returns), or both. As discussed above, the CFI Group system allows for all of these features—the perceived performance comparisons, the impact of quality components on satisfaction, the impact of satisfaction on future behaviors, and the use of this information for efficient resource allocation.

The CFI Group approach provides specific and quantifiable information about the *levels* of service and quality and the *marginal contribution*, to both customer satisfaction and profits, which will result from a *change* in a process, service, aspect of quality, etc. Unlike other consulting firms, CFI Group utilizes a cause-and-effect system that isolates the effects of a change in a quality component on the change in customer satisfaction, and the subsequent change in economic returns. This is very different from focusing on what customers deem "important". It is also characterized by a "systems" treatment of all its components (i.e., quality, satisfaction, profit). All of these aspects make it different from other approaches.

[24] United States patent number 6,192,319, visit www.uspto.gov for more information.

Summary Table

The following table provides a basic summary of the key points made in the foregoing discussion about the characteristics of the ACSI technology and the resulting benefits for users.

	Elements of ACSI Technology Implementation		
	Measurement →	Diagnosis →	Prognosis
Objective	✓ Reliable (precision) ✓ Valid ✓ Sensitive (power to detect change)	✓ Impact or Key Driver Analysis ("To improve customer satisfaction what matters the most?")	✓ Change Prediction ("How do changes in experiences effect changes in satisfaction and retention?")
Characteristics	✓ "Voice of the customer" (VOC) based ✓ Multiple measures optimally weighted based on strength of relationships in measurement network ✓ Reduced measurement error ✓ Reduced confidence intervals ✓ Uses unstandardized performance scores	✓ Calculated within the context of a complex cause and effect network ✓ Based on unstandardized slopes not correlation ✓ Optimized with regard to key management objectives (i.e. CS or behaviors) ✓ Control of multi-collinearity provides more reliable impact estimation	✓ "What if" predictive tool ✓ Quantifies the effects of changes across multiple nodes (experience to evaluation to intention) ✓ Future effects are comparative across time, location or segment given planned investment levels
Benefits	✓ Accurate ✓ Meaningful—tied directly to customer experience ✓ Comprehensive—incorporates all aspects of customer experiences ✓ Understandable—simple scoring method ✓ Comparable—by using unstandardized scores	✓ Prioritizes improvement efforts ✓ Provides impacts that are additive in nature and comparable across groups ✓ Allows for more efficient allocation of resources based on the economic concept of marginality	✓ Focuses on the "dynamic" quantification of change ✓ Increased ability to envision future change in key performance outcomes

Competitive Comparisons

The following table provides a comparison of the ACSI technology as implemented by CFI Group with three classes of competitors—*Primitive*, *Naïve*, and *Pseudo-Sophisticated*.

Primitive competitors are research suppliers that compete largely on the basis of price, supplying survey information that uses "canned" questionnaires. They may be able to provide results quickly, but the results lack any diagnostic or prognostic capability. They appeal to buyers of consumer research who are unconcerned with information quality and may be looking to meet an organization requirement that customers be surveyed. However the usefulness and incorporation of the results into decision-making is rudimentary.

Comparative Criteria	**The ACSI Technology/ CFI Group**	**Types of Competitors**		
		Primitive—"Price Based"	**Naïve—"Simple Minded Solutions"**	**Pseudo-Sophisticated—"Faulty Science"**
Measurement				
>Uses VOC qualitative methods	**Yes—**all measures used by CFI are based on VOC methods	**No**—use "canned" un-customized surveys based on researcher judgment	**No**—use "canned" un-customized surveys based on researcher judgment	**Maybe**—some may use qualitative methods
>Customized measures to insure validity	**Yes**—customized measures are recommended to insure validity	**No**—repetitively use the same set of questions for all clients—validity is likely low	**No**—repetitively use the same set of questions for all clients—validity is likely low	**Some—**usually use canned surveys to save time
>Use multiple item scales to minimize measurement error	**Yes**—three to five items necessary to insure high reliability standards	**No**—use single item nominal or categorical measures—reliability very low, large confidence intervals	**No**—use single item nominal, categorical or 5 point Likert scaled measures—low reliability, large confidence intervals	**Some**—tighter confidence intervals but still 30% bigger than CFI method
>Optimal weighting for deriving scores	**Yes**—weights based on cause and effect network between components	**No**—Only report item scores—usually as percentages or proportions (i.e., "top-box")	**No**—Only report item scores as proportions or means	**No**—usually compute averages, sums or factor scores
Sample sizes required	**Small** ≤ 150-200	**Very large**—samples based on number of cells that need to be filled in a cross-tab table	**Large**—needed to get any kind of estimation precision	**Large**—needed to get any kind of estimation precision
Driver/ Impact Identification				
>Cause and Effect Network with impacts based on based on slopes	**Yes**	**NA**—no cause and effect networks are used	**No**—use correlations, difference gaps or stated importance; some may use simple regression	**Quasi**—Usually stepwise regression; or in some cases factor regression
>Control of Multicollinearity	**Yes**—proprietary PLS regression based method for allocating the "halo" in per-ceptual measures	**NA**—no estimates provided	**No**—usually ignore the existence of multi-collinearity	**Mostly No**—some may factor analyze predictors to control inter-correlations before using in a multiple regression model
>Unstandardized estimates	**Yes**	**NA**—no estimates provided	**No**	**No**
Other				
Proprietary patented analysis system	**Yes**—United States patent number 6,192,319	**No**—not available; all analyses use software packages that are purchased from third party vendors (e.g., SPSS, SAS)		

Naïve suppliers will often use similar data sources and measures as those of *Primitive* research suppliers but add some intuitively appealing analytic paraphernalia—such as performance importance matrices, difference gap analysis, etc. Unfortunately most of these so-called analytic approaches actually increase error and provide a muddled picture of reality rather than clarifying it. In addition, their ubiquitous use of a "stated performance" measurement methodology needlessly increases questionnaire length with no additional diagnostic value. They may appeal to users of research who are looking for more sophistication from their measurement suppliers than can be provided by *Primitive* suppliers. Unfortunately, such users are being hoodwinked by glib answers and simplistic solutions to complex questions of human behavior. *Naïve* suppliers are very dependent upon the inability of their customers to discriminate between what they are peddling and the kind of sound methods espoused by CFI Group.

Pseudo-Sophisticated suppliers may provide upgraded measurement and diagnostic capabilities. They do this by the "piece-meal" application of multiple measurement approaches along with limited use of multivariate statistical techniques often found in common statistical packages. These kinds of suppliers add a veneer of science to the types of measurement services provided by *Primitive* and *Naïve* suppliers.

What is often unknown to most users of the services supplied by these vendors is that the common sequential use of multivariate methods (such as using factor analysis or cluster analysis to create component scores which are then used in a multivariate regression procedure) essentially magnifies the weaknesses of both methods and creates interpretational problems that are not easily overcome. For instance, if the factor scores of a three factor solution to a data reduction problem that explains 50 % of the variance in the data used, are then regressed against a fourth variable achieving an R^2 (variance explained) value of 0.5, then what does the analysis tell the user? Never mind what the regression coefficients mean. Unfortunately many *Pseudo-Sophisticated* venders are not sufficiently cognizant of these weaknesses to adequately educate their customers about the frailties in the seemingly scientific methods they advocate. As indicated earlier in this document, users of this kind of analytic product run the danger of being seriously misled in their decision-making.

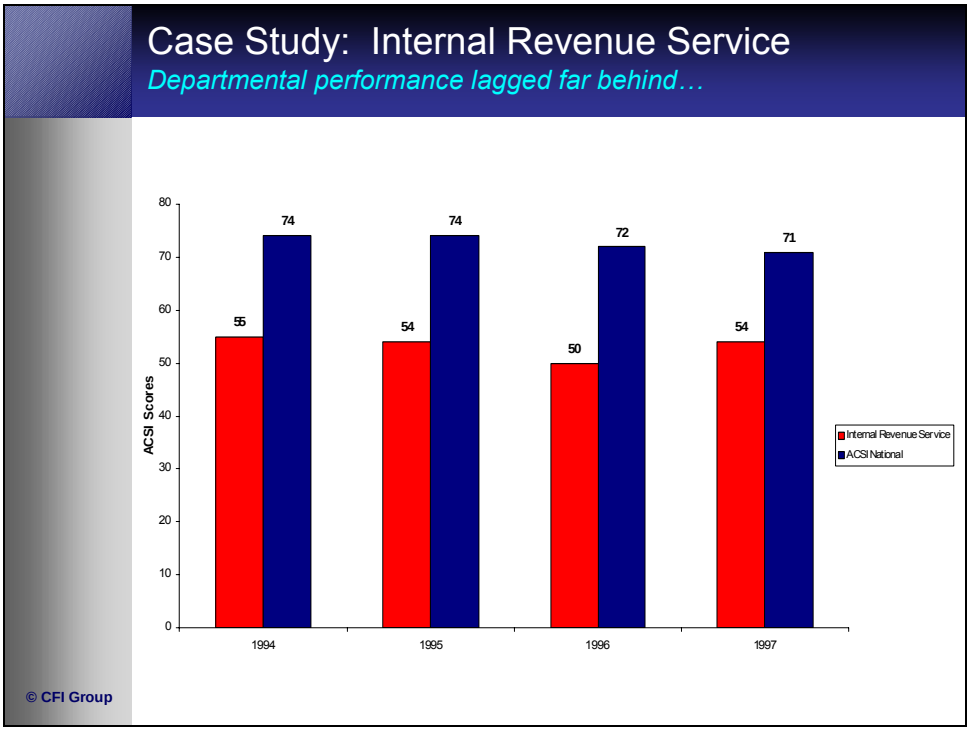
Results—How do CFI Group Clients Benefit from the ACSI Technology?

In this section examples of the how the CFI Group/ ACSI technology was applied to U.S. Government agencies are reviewed.

Internal Revenue Service (IRS):

Before working with CFI Group, the IRS suffered from:

- Disgruntled employees
- Dissatisfied taxpayers
- Declining, low ACSI Scores



This situation resulted in a 1997 Senate hearing that labeled the IRS as a "tax agency out of control". This finding was supported by witnesses and commentators making statements such as the following:

> *"As only one taxpayer representative out of thousands across the country, I have seen dozens of taxpayers severely damaged and even made homeless by the IRS collection division."* (*Anonymous Witness #1, IRS Employee*
> *Senate IRS Hearings 1997)*
>
> *"The long list of IRS horribles included arbitrary collection decisions, sale of taxpayer lien property far below value, and the cavalier mistreatment of taxpayers."* (Bob Zelnick, ABC Good Morning America, September 26, 1997)

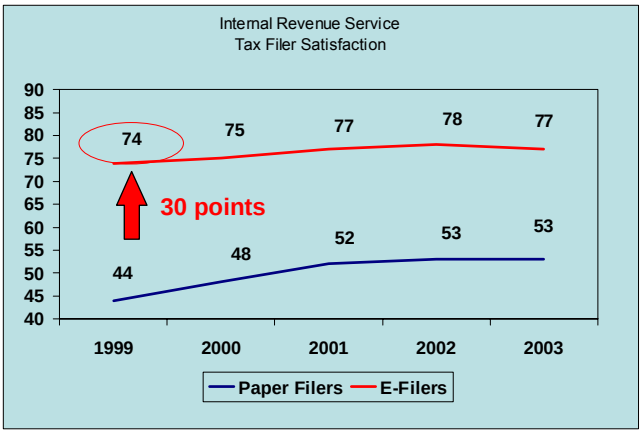
CFI Group began working with the IRS in 1999. An initial assignment discovered that the satisfaction levels for taxpayers filing on-line (eFilers) were 30 points higher than those taxpayers submitting paper returns (see first chart below). As a consequence of this finding, the IRS instituted a strategy of encouraging filers to use the on-line submission process. The result was a steady improvement in overall IRS customer satisfaction scores (see second chart below). These findings demonstrate the power of the strategic guidance provided by CFI Group to improve decision-making and subsequent customer satisfaction.

Case Study: Internal Revenue Service

A Key Discovery: ACSI Study 1999

eFilers vastly more satisfied …

- fewer errors, quick problem resolution
- earlier refunds, status tracking

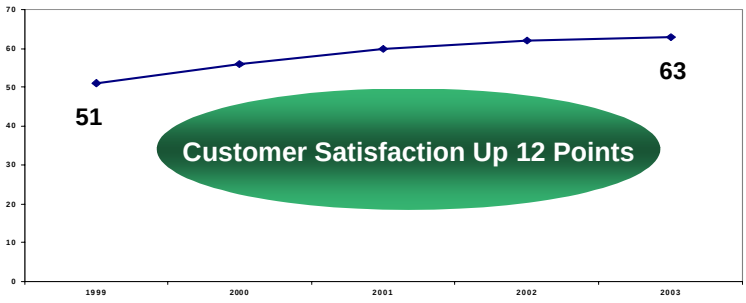


Case Study: Internal Revenue Service

A systematic improvement

IRS hears the voice of the customer …

- Commitment to customer service: kinder, gentler
- Increased awareness and usage of eFiling



Faster Trade-Up to Electronic filing…
Faster access to tax revenues?

Federal Aviation Agency (FAA):

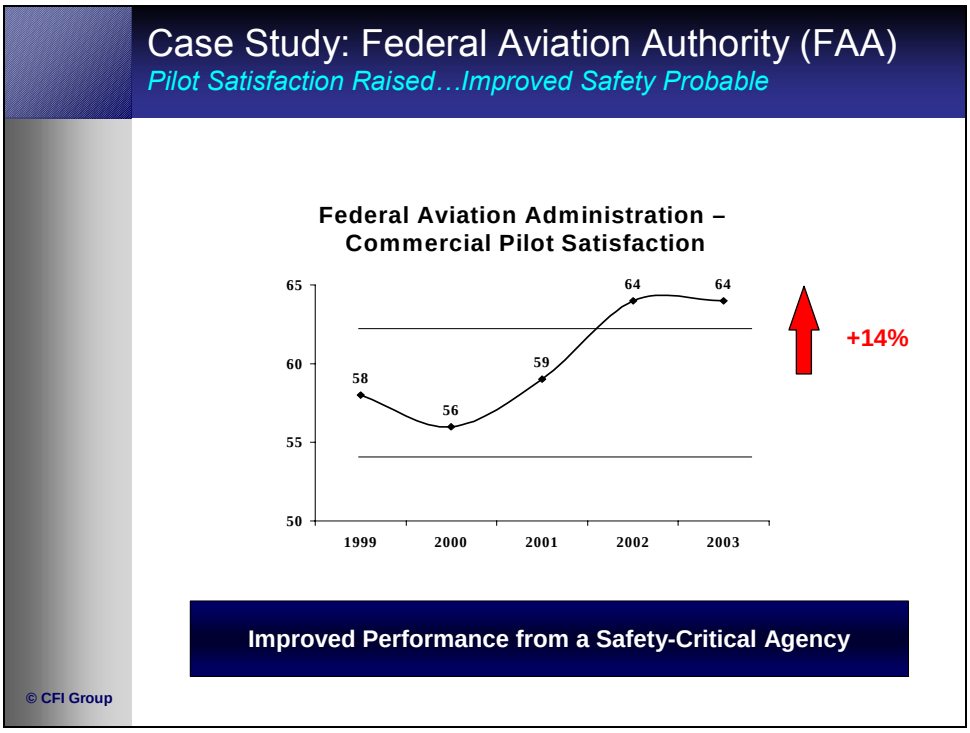
When CFI Group began working with the FAA, Satisfaction had been low, typical for an agency with largely regulatory/punitive function. CFI Group developed a model that measured three specific drivers of satisfaction:

- Quality of air traffic services
- The pilot certification process
- FAA policies, standards and regulations

The "Policies, standards and regulations" area was identified as the lowest scoring driver but with and the greatest impact on pilot satisfaction with the FAA. It was determined that pilots perceived "policies, standards and regulations" as poorly written and difficult to understand, thus failing to contribute to airline safety as well as they should.

As a result of CFI Group analysis and modelling insights, the FAA engaged in a significant overhaul of its policies, standards and regulations, doing much rewriting using plain language.

Subsequent measurements showed that pilot satisfaction with FAA increased 14%, a very large improvement in ACSI terms for a relatively short span of time.



For more case study examples visit the CFI Group website http://www.cfigroup.com/expertise/case_studies.htm .

Appendix A: A Technical Summary of the CFI Group Analytic System

CFI Group's process takes place in four stages to ensure maximum reliability, validity and inclusion of essential issues:
1) Secondary Research
2) Management Interviews
3) "Voice of the Customer" (VOC) Investigations
4) Quantitative Analysis

1) Secondary Research
Some firms might argue against the necessity of this stage, stating that vast quantities of such research had already been performed, oftentimes yielding no more information than they had had before. However, one reason firms often do not benefit from such research is that its focus tends to be scattered. One study might look at concepts of customer loyalty, while another looks at current attitudes of store personnel, and still another asks customers to focus on aspects of in-store shopping. Our purpose in performing secondary research is to build upon and *synthesize* prior research thereby gaining the maximum information available from it.

2) Management Interviews
Interviewing management personnel across relevant areas of businesses is also critical to synthesizing useful information, which might otherwise remain isolated. These interviews aid in:

- Understanding a heterogeneous customer base;
- Identifying current business issues viewed as relevant by management personnel;
- Developing a substantive knowledge of the competitive environment;
- Designing the qualitative interview guidelines for in-depth interviews with customers; and
- Determining how performance measures will be represented in the subsequent model

3) "Voice of the Customer" (VOC) Investigations
The need to talk with customers to uncover issues salient to them has become increasingly obvious over the past several years. What has not become obvious, however, are the techniques needed to uncover such issues accurately and in-depth. CFI Group's system utilizes qualitative one-on-one customer interviews specifically designed to cover both issues identified as relevant by management personnel and to allow customers to voice their opinions, concerns and desires which might otherwise be left unknown to management.

While management would likely be able to predict a large percentage of the components and issues salient to customer satisfaction, there is still a reasonable amount of information to be gained from customers, which would go unsaid if customer interview structures were too rigid.

Further, management personnel might also be unaware of the language that customers tend to use (i.e., voice of the customer) when discussing such issues or, quite importantly, all the *aspects* of a particular issue, even if correctly identified by management, relevant to the customer. CFI Group's qualitative system applies a combination of current social-psychological techniques whose power and scope exceed common research methods utilized by other firms. CFI Group's system employs the following techniques:[25]

One-on-one interviews: While focus groups can be useful in certain cases, typically what happens in such settings is that one or two strong voices emerge only to be followed by the rest of the group. The resulting information is highly biased and skewed toward the more vocal customers in the group. Although interviewers often try to avoid such biases by requiring focus group attendees to talk "in turn", they may still miss subtle (and not-so-subtle) pressures, which come from group meetings. Valuable information may be lost in such settings where the interview is highly structured.

Open ended, semi-structured interview approach: This approach allows us to ask customers about issues mentioned in secondary research and management interviews, while still leaving the opportunity for each customer to discuss "top-of-mind" issues during the course of the interview, thereby identifying salient factors which might otherwise go undetected.

Metaphors and narrative accounts: By giving customers the opportunity to tell stories and use metaphors to describe the various experiences they have had, we also encourage the identification of new and valuable information. Given innovative social-psychological research techniques, and a more conversational style interview, customers can relax and converse as they might with a friend during the interview. A skilled interviewer can keep a respondent focused on the relevant topics while still allowing them to recall experiences regarding which could be very useful to management

[25] Griffin, Abbie and John Hauser, "The Voice of the Customer," *Marketing Science*, winter, 1993, 12,1,1.

and other personnel. Similarly, simply asking someone "why" they like or dislike some aspect of a product, will not get at the real ways in which people think about things and make purchase decisions. CFI Group's qualitative system utilizes techniques which help customers to identify and discuss issues relevant to their purchasing behaviors, unlike most other consulting firms where customers are asked only to confirm or rank pre-identified and ultimately incomplete factors relevant to decision making.

Customer interviews performed by CFI Group are recorded and transcribed verbatim ensuring maximum reliability and validity in performing the analysis. Qualitative research techniques are then applied to the subsequent analysis of each transcript as well as the transcripts as a group. Unlike other firms who rely on "frequency of response" coding to identify relevant factors (thereby only increasing interviewer created bias), CFI Group's system relies on a "narrow lens approach" – a social-psychological analysis process which allows us to identify and categorize salient factors and re-group all relevant information into a subsequent model, thereby maximizing the information gained from the interviews.

CFI Group's qualitative analysis allows a specification of a preliminary model of customer satisfaction, and makes certain that attributes of each component are preserved utilizing the language of the customer. The subsequently developed questionnaire is based on the voice of the customer and helps ensure that the information gathered with it is valid.

4) Quantitative Analysis
Ultimately, the power and precision of the preliminary model is proven in the quantitative phase of CFI Group's system which is built upon three distinct points:

A. Estimating Importance, Utility, and Impact
B. Estimating Derived Importance
C. Causal Models: comparing covariance structure analysis (e.g., LISREL) and latent variable partial least squares (e.g., Wold's LV-PLS system), the two major approaches to causal models.

The objective is to identify those quality dimensions whose improvement offers the greatest returns, as measured in customer satisfaction, retention rate (and potentially related measures of individual behavior, such as spending level) and corporate financial performance. *That is, if the level of performance on a quality attribute improves by a given amount, how much will satisfaction (and, subsequently customer retention or financial performance) improve?* In evaluating a methodology, the most important criterion is whether a method can quantify the return-on-quality.

A. Estimating Importance, Utility, and Impact

Table A1: Four Approaches to Estimating Importance, Utility and Impact of Quality Improvements

Class of Methods	Quantifies Change
Explicit Self Reported Importance	No
Derived Self-Reported Importance	No
Conjoint Methods	Yes
Derived Importance Methods	Yes

In methods assessing *Explicit Self-reported Importance*, respondents directly state or rate the importance of an attribute. If respondents are asked to "Rate the importance of price on a scale from 1 to 5," attributes can only be compared in terms of their mean importance ratings. Methods of *Derived Self-Reported Importance* ask respondents to compare attributes in terms of their importance. If the question is, "Which is more important to you, price or on-time delivery?" a rank order or a derived importance scale can be calculated. Constant sum scales can also be used in this way. However, none of these approaches simultaneously calibrates the relationship from performance on a quality attribute to a consequent change in satisfaction, retention or financial performance. The best they can do is indicate on an attribute by attribute basis how important each attribute might be for satisfying the customer. But this assumes that each attribute importance measure is perfect and without error. Plus it would have to be repeated for any additional dependent variables separately extending the questionnaire length and increasing respondent fatigue. Thus, there is no way to compare the returns from quality improvements and set priorities using these methods. This is one of the main reasons that few if any consultants advocate a stated importance measurement framework.[26]

Conjoint Methods ask respondents to rate or choose between alternative profiles of products or services. The products/services are described in terms of levels of objective quality. From a pattern of preferences, we can derive *part-worths* or *utilities* of different levels of an attribute (for example, the utility of a "professional and polite employee"

[26] Allen, Derek and Tanniru Rao, *Analysis of Customer Satisfaction Data*, ASQ Quality Press 2000, p.70.

versus “warm, friendly and polite employees”). Conjoint analysis is thus able to quantify the relationship between the level of an attribute and the level of preference.

Conjoint methods, it should be noted, create a model of the individual. As a result, the ability to generalize *part-worths* to the population depends on the sampling method. A conjoint study of 30 respondents selected by a non-probability sampling method, such as convenience sampling or quota methods cannot be generalized to the population. Confidence intervals on the aggregate *part-worths* depend on sample size and method.

Table A2: Approaches to Estimating Impacts Conjoint Methods? Derived Importance Methods?

	Conjoint Methods	Derived Importance Methods
Individual Level Model	Yes	No
Population Level Model	No	Yes

The length of a conjoint questionnaire increases exponentially with the number of attributes and the number of levels to each attribute. There are some techniques for reducing the burden on the respondent, but in general the questionnaires are quite lengthy. Conjoint methods work best on product attributes with discrete concrete levels, such as colors or package designs. Conjoint is much more difficult when attributes are more subjective, such as the Employee Courtesy example, which does not identify a clear, discernible difference between “professional and polite” and “warm, friendly and polite.” Conjoint methods cannot be recommended for determining impacts.

Derived Importance Methods estimate the impact of improvement directly from the relationship between a quality factor and the level of Satisfaction. For example, the level of Satisfaction can be regressed against the levels of attributes. The regression coefficient, or impact, quantifies the relationship between Satisfaction and the level of an attribute. For example, a change of x units on an attribute results in a change of y units in Satisfaction.

The advantages and disadvantages of different methods of determining derived importance will be discussed later. For now, derived importance should be considered as a model of the population rather than the individual. That is, derived importance indicates the return from improving the level of an attribute for the *population* rather than for an individual respondent. In contrast, conjoint measures utility at the individual level and then infer population utilities using sampling statistics.

B. Estimating Derived Importance

Methods such as correlation or simple regression, examine the relationship between two variables. It is assumed that the system is not affected by any variables other than the two selected for analysis. In virtually all cases, this is an unreasonable assumption. The correlation coefficient says nothing about impact. Two variables can have the same correlation with satisfaction, but different effects because the slope of the relationship differs. This is the difference between correlation and regression.

Multiple regressions analyze the relationship between multiple variables, such as quality issues, and a single dependent variable. Basic to multiple regression is that each independent variable, each quality issue, measures a different thing. Multiple regression as well as simple regression and correlation assume that all independent variables are measured perfectly without error. Again, this is an unrealistic assumption. Error in measurement typically amounts to 30% in survey data. This error is often much greater than the sampling error. (Andrews).[27]

[27] Frank M. Andrews (1984), Construct Validity and Error Components of Survey Measures: “A Statistical Modeling Approach”, *Public Opinion Quarterly*, p.404-442.

Table A3: Single Equation Systems vs. Causal Models Bivariate Methods? Single Equation Systems? Causal Models?

	Bivariate Methods	Single Equation Systems	Causal Models
Measurement Model *(Multiple measures)*	No	No	Yes
Multiple Constructs	No	No	Yes
Multiple Objectives *(Dependent variables)*	No	Some	Yes
Complex Systems	No	No	Yes

Measurement error introduces bias and inconsistency in the estimation of importance. That is, the estimates of importance are incorrect in the sense that the expected value of the regression estimate does not equal the true importance (bias). And the regression estimates do not converge to the correct values with larger samples (inconsistency). The amount of bias and inconsistency varies in proportion to the amount of error.

Another serious problem is that many quality variables are highly related with one another (multicollinearity). This causes estimates of impacts to be imprecise with multiple regression. It should be noted that the close association between variables is a result of the nature of satisfaction. A firm's customers tend to rate the firm high or low on everything due to a strong *halo* effect. The problem of multicollinearity renders multiple regression results essentially useless. Lastly, multiple regression allows only one dependent variable, (i.e., one objective such as Satisfaction or Retention, but not both). Thus, multiple regression is inappropriate with multiple objectives and complex systems.

C. Causal Models

Causal models have all the features necessary to estimate impact. Causal models accept multiple measures to control measurement error, allow multiple objectives, and allow complex, multi-level systems of relationships. The two major approaches to causal models are covariance structure models, typified by LISREL and predictive-causal systems typified by LV-PLS. The differences between LV-PLS and LISREL are summarized below in Tables 4 and 5. The CFI Group uses a further development of the LV-PLS approach. With LV-PLS, weights and impacts are estimated to predict key variables. That is, LV-PLS will maximize our ability to predict Satisfaction or Retention. In contrast, LISREL attempts to account for covariance and maximizes the fit to the covariance matrix among all variables. Consequently, correlations between all variables are treated as equally important. The CFI Group impact is the expected (average) change on an individual score given a five-point change in a quality or experience component. Because this is the mean prediction, the prediction applies to the aggregate as well.

LISREL produces an estimate of an effect, which is meant to represent the causal effect of an unobservable variable onto another unobservable variable. However, because the unobservable variables in LISREL are *unobserved*, their scales are arbitrary. That is, a scale and origin must be assigned to each unobservable. Usually, the origin is set at zero and one of the measures of each unobservable is given a weight of one – all results are then calibrated relative to the assignments.

An alternative is to assume the unobservable variables are standardized, (i.e., have mean zero and unit variance). This is generally possible for dependent variables, but not for independent variables. The arbitrary nature of the scale assignments means that it is difficult to interpret or compare effects. That is, a unit on one unobservable may not be the same as a unit on another and hence, the effects cannot be compared directly. For example, if the dependent unobservable variables are standardized, then a change of one unit on independent variable one will produce x% change in standard deviations on the dependent variable. Similarly, a change of one unit on a different independent variable would have a different y% change in standard deviations on the dependent variable. However, comparing these x% and y% impacts is difficult since the scales of the independent variable may differ. Of course, without comparing results, it is impossible to prioritize improvements.

LV-PLS relies upon Ordinary Least Squares for estimation. OLS makes no distributional assumptions. Statistical testing in LV-PLS is accomplished via jack-knifing and blindfolding. These methods are empirical and based upon case level data. In particular, these techniques do not require distributional assumptions.

Table A4: PLS vs. LISREL – Managerial Issues

Managerial Issues	PLS	LISREL
Purpose Which objective is more meaningful for managers – better prediction or best fit to covariance structure?	Minimize prediction error.	Maximize fit to covariance matrix.
Priority given to key objectives	Yes, to dependent variables	No, all variables treated equally
Component Level Scores available for benchmarking and tracking	Yes	No, component scores cannot be calculated, because scores are indeterminate.
Case level scores for further analysis, such as segmentation, descriptive, ANOVA	Yes	No, case scores are indeterminate.
Indices	Can measure one construct or form a composite, such as Overall Quality.	Measures within a component must measure one and only one construct. This is restrictive when one wants to construct a managerially useful index.
Sample Size	200 is typical but can be less (PLS fits each part of the model separately. Thereby reducing the number of cases required.)	500+ (fits entire model at one time thereby requiring more cases).

Table A5: PLS vs. LISREL: Statistical Issues

Statistical Issues	PLS	LISREL
Estimation Method	Least Squares	Typically, maximum likelihood
Assumptions	Assumes linear conditional expectation between independent and dependent variables (x is a cause of y, expected residual is zero, the residual is uncorrelated with the conditional variable. and linear measurement relationships.	Assumes linear relationships among constructs and linear measurement relationships. In addition, typically assumes multivariate normal (or related distribution) and independent observations.
Minimum specification requirements	Must specify all predictors of a dependent variable and group manifest variables into components.	Must specify all predictors of a dependent variable and group manifest variables into components. In addition, must specify all other relationships among all variables and constructs.
Feasibility of use for analysis of complex relationships	Yes	Yes
Efficiency of estimates	Predictions are consistent with minimum variance.	Yes (Parameter estimates are efficient if assumptions are met).
Consistency of estimates	Estimates of impacts are consistent. Estimated component scores are consistent at large.	Yes (If assumptions are met).
Identification (can estimate all parameters)	Not an issue	Can be problematic. To be able to estimate certain parameters, may need to make assumptions about relationships about which we have no knowledge (i.e., the covariance between residuals).

LISREL uses several methods to estimate parameters to fit (reproduce) the covariance matrix, including unweighted least squares (ULS), generalized least squares (GLS), and maximum-likelihood (ML). Statistical tests are derived under distributional assumptions and come directly from the fitting functions rather than from case level information. In particular, under the assumption that the observed variables are distributed multivariate normal, GLS and ML provide large sample estimates of the standard errors for statistical testing. Standard errors must be used with care when the assumptions of normality are not met. ULS can be justified without distributional assumptions. However, standard errors and statistical tests are unavailable for ULS.

LV-PLS makes no further assumptions about the distribution of the variables or the error terms. In particular, PLS is insensitive to non-normality of the error terms, heteroscedasticity of the error terms, and autocorrelation of the error terms. Specifically, the LV-PLS estimates are unbiased estimates. LISREL makes many more assumptions, including multivariate normality of the variables. Violations of the assumptions are generally viewed as problematic for LISREL.

In summary, the most important difference between LV-PLS and LISREL is how relationships are established. LV-PLS produces scores both for overall and for individual cases, while LISREL does not. LV-PLS makes no distributional assumptions, while LISREL requires strong distributional assumptions. For these reasons, and for others presented in the following tables, it is not possible to use a covariance structure method such as LISREL for estimating impacts. Further, scores on customer satisfaction and other components cannot be computed from a LISREL approach (they can only be estimated with the introduction of yet another source of error).

The CFI Group system is an advancement of LV-PLS. LV-PLS estimates are consistent at large.[28] That is, as the sample size increases and the number of measures increases, the scores approach their true values. Consequently, close association among related quality variables is an advantage rather than a disadvantage.[29] Moreover, convergence in measurement implies that the estimates of importance are also unbiased and consistent.[30] That is, the expected value of the impact is equal to the true importance (unbiased). And, the estimates of impact converge to the true values as sample size increases (consistent).

Thus the CFI Group system is better able to detect the true association between experience quality and satisfaction, more able to explain satisfaction, and to do so with greater accuracy than alternative methods of analysis. Whereas the basic LV-PLS is more suitable than other methods for the analysis of customer satisfaction data, it is not sufficient. Particularly, it does not handle the problems of multicollinearity and standardization well. The contribution of the CFI Group to the basic LV-PLS method is threefold:

- *It reduces* the multicollinearity by (a) using the qualitative work in model specification and (b) by extracting and isolating any remaining excess collinearity in the quantitative analysis.
- *It retains* the original scale values in analysis (the basic LV-PLS method does not do this).
- *It reduces* necessary sample size by putting the variables in the context of a comprehensive system that is estimated iteratively rather than simultaneously.

In summary, we contend that it is both cost-effective (data collection costs will be lower) and revenue effective to adopt CFI Group's system. It generates better information at lower cost than any other approach. The total cost reduction also implies a shift in the budget such that a proportionally smaller amount is spent on data collection and a larger amount on data analysis. We are eager to discuss these benefits relative to any other system.

---

[28] H. Wold (1982), "Soft Modeling: The Basic Design and Some Extensions," in K.G. Joreskog and H. Wold (Eds.), *Systems under Indirect Observation: Causality, Structure, Prediction* (Vol. 2, pp. 1-54)., Amsterdam: North Holland

[29] Claes Fornell, Byong-Duc Rhee and Youjae Yi (1991), "Direct Regression, Reverse Regression and Covariance Structure Analysis," *Marketing Letters*, Vol. 2, No. 3, p.309-320.

[30] Claes Fornell and Jaesung Cha, (1992), "Partial Least Squares," *Handbook of Marketing Research*.

Appendix B: Use of 10-point Scales

CFI Group's use of 10-point scales over commonly used 5-point scales is based on a number of statistical and managerial criteria as discussed below.

A common basis for recommending 5-point scales often rests on the assumed inability of people to reliably discriminate more than 5 levels on a scale, where offering more than 5 levels would introduce error into the measurement and offer weaker correlations and lower explanatory power. Research has clearly shown that people can handle more than 5 pieces of information at one time, particularly depending on their experience in a given area and ability. A 10-point scale is within capabilities of most people with little experience, and in areas of professional expertise people are able to and will make much finer distinctions.

Because customer satisfaction data is positively skewed (where customers less frequently use the lower ends of scales), a 5-point scale is really closer to a 3-point scale, and a 10-point scale behaves more like a 7-point scale. Since most customers don't really use the lower ends of scales (values 1 and 2 on a 5-point scale) and mostly use values 3, 4, and 5, a 5-point scale offers little opportunity to differentiate positive responses. This negative skewness introduces error into the measurement process and loss of critical, meaningful information compared with a 10-point scale.

Societal norms and the fact that customers typically "like" companies they do business with tend to limit the number of customers who use the very lower ends of response scales. In most cases, if a customer is so completely dissatisfied as to have the need to use the lower ends of the scale, they will leave and stop doing business with the company. As a result, the 5-point scale effectively turns into a 2- or 3-point scale due to limited response at values 1 and 2.

This "compression effect" also militates against the common assumption that 5-point scales offer a mid-point that can be considered as the "average response", a characteristic not present in 10-point scales. The mid-point argument is only valid if respondents use, or at least contemplate, all points of the scale, and as discussed above, they do not, and responses are consequently negatively skewed.

The use of 10-point scales significantly enhances the information that is transmitted in the surveying process. The increased information content yields:

- Greater precision of results, thereby providing opportunity to reduce sampling costs while maintaining the same precision obtained using 5-point scales – OR – Ability to reduce the number of questions on the questionnaire (which also reduces sampling costs due to reduced questionnaire length) while maintaining the same measurement reliability offered while using 5-point scales.

- Greater ability to link Satisfaction results to internal performance measures or measures of employee satisfaction due to the gains in reliability and precision.

Another critical benefit of the use of 10-point scales is in the increase explanatory (as measured by R^2) power gained.

- The gain in R-squared from using the 10-point scale is an important component of accurately identifying the drivers of Satisfaction and predicting the economic returns associated with improving Satisfaction. In addition, for business's which have inherently small populations, use of 10-point scales may make the difference between being able to discern these linkages or not.

- Further, the gain is valuable within the context of linking employee compensation to CSI. Higher correlation (R-squared) within the model ensures that targeted employee actions will be reflected in the CSI measure and will provide less error within the compensation system (i.e. reducing Type I and Type II errors, where employees are not rewarded when CSI really did change or when employees are rewarded and CSI did not really change).

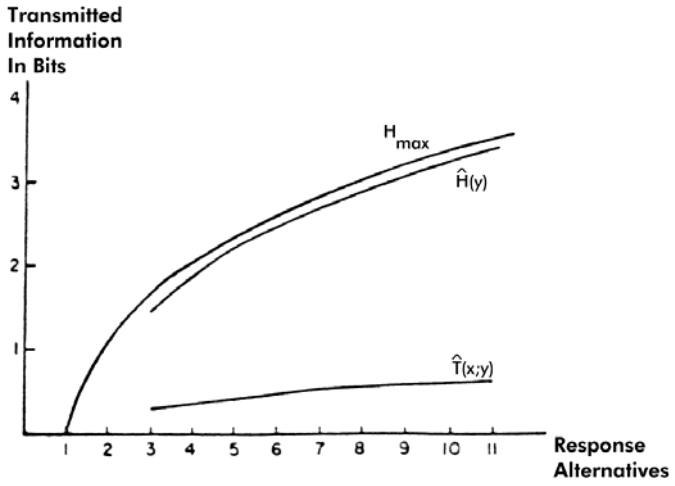
There is one area in which 10-point scales are not appropriate relative to 5-point scales – that is when there is a desire to label each response point within the scale (e.g. 1=poor, 2=not so good, 3=satisfactory, 4=good, 5=outstanding). There are several arguments for not attaching labels to response categories, most notably: 1) added error due to violation of the interval/ratio data assumption, where it can no longer be assumed that the distance between 1 and 2 is the same as the distance between 2 and 3, and so forth, and 2) respondent burden and increased questionnaire length.

Criteria for evaluating scales and supporting evidence

Cox[31] has reported the statistical benefits of 10- vs. 5-point scales.

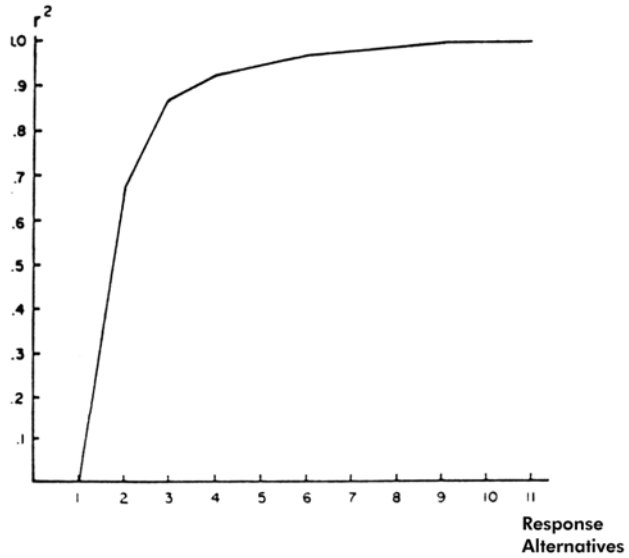
- Information content
  As Figure B1 below illustrates, more information is transmitted in 10- vs. 5-point scales – approximately 2.4 bits on a 5-point scale vs. 3.4 bits on a 10-point scale.

Figure B1: Relationship Between the Number of Response Alternatives and Transmitted Information Found by Bendig and Hughes (1953)



- Explainability and predictability (R-squared)
  The Figure B2 illustrates the significant added benefit of increasing R-squared, which we have defined as explainability (ability of the quality components to explain changes in Satisfaction) and predictability (ability of Satisfaction to explain changes in performance measures).

Figure B2: Gain in R^2 Obtained by Using a More Refined Scale



As the chart shows, the largest increased returns are achieved when employing 4- or 5-point scales, but 10-point scales continue to strengthen and tighten the relationships of the entire model.

- Mean-squared correlations
  The Table below provides strong evidence that the use of 10-point scales increases the reliability and accuracy of measures over 5-point scales. Specifically, using correlations as the benchmark level, (where higher correlations are better, indicating greater reliability) three items on a 10-point scale provide comparable reliability (0.785) to 4 items on a 5-point scale (taking the average of 0.759 and 0.813 which is 0.785).

Table B1: Mean Squared Correlations Between Observed and True Composites by the Number of Items and Response Alternatives Found by Jenkins and Taber (1977)[32]

	Categories						
Items	2	3	5	7	9	10	14
2	.551	.657	.718	.736	.744	.747	.752
3	.604	.702	.759	.776	.783	.785	.790
5	.680	.766	.813	.827	.833	.835	.839
7	.725	.804	.845	.857	.863	.865	.868
9	.756	.828	.865	.876	.880	.882	.885
10	.769	.839	.874	.885	.889	.890	.893
14	.810	.868	.899	.907	.911	.912	.915

More recently, Preston and Colman[33] in a study using ratings of service quality in restaurants and stores found:

- The rating scales that yielded the least reliable scores turned out to be those with the fewest response categories.
- According to the indices of validity and discriminating power examined, the scales with relatively few response categories performed worst.
- No corroboration with the contention that reliability and validity of scores are independent of the number of response categories and that nothing is gained by using scales with more than two or three response categories.
- Statistically, scales with small numbers of response categories yield scores that are generally less valid and less discriminating than those with six or more response categories.
- Scales with 5, 7, and 10 response categories were rated as relatively easy to use. Shorter scales with two, three, or four response categories were rated as relatively quick to use, but they were rated extremely unfavorably on the extent to which they allowed the respondents to express their feelings adequately; according to this criterion, scales with 10, 11 and 101 response categories were much preferred.
- On the whole, taking all three respondent preference ratings into account, scales with two, three, or four response categories were least preferred, and scales with 10, 9, and 7 were most preferred.
- From the multiple indices of reliability, validity, discriminating power, and respondent preferences used in the study, a remarkably consistent set of conclusions emerged.

In general, it was found that scales with two, three, or four response categories yielded scores that were clearly and unambiguously the least reliable, valid, and discriminating. The most reliable scores were those from scales with between 7 and 10 response categories, the most valid and discriminating were from those with nine or more. The results regarding respondent preferences showed that scales with two, three, or four response categories once again generally performed worst and those with 10, 9, or 7 performed best. Taken together, the results reported above suggest that rating scales with 7, 9, or 10 response categories are generally to be preferred.

---

[31] Cox, Eli P. (1980), "The Optimal Number of Response Alternatives for a Scale: A Review", *Journal of Marketing Research*, XVII (November), 407-422.

[32] Jenkins, C. Douglas, Jr., and Thomas Taber, "A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability," *Journal of Applied Psychology*, (August) 1977, 62, 4, 392.

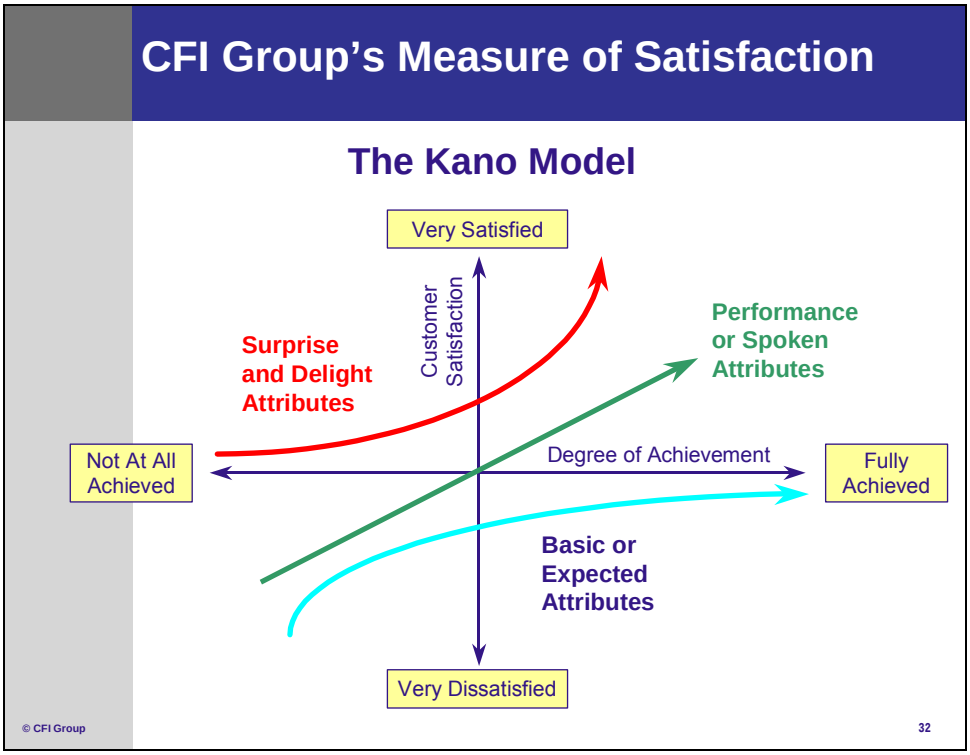
[33] Preston, Carolyn C. and Andrew M. Colman, "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica* 104 (2000) 1.

Appendix C: Why do the ACSI and CFI Group use three measures of Customer Satisfaction?

Managers must carefully evaluate the multitude of measurement options offered in the marketplace to ensure they secure the most accurate, reliable and valid measurements of customer satisfaction.

To squarely address these concerns, CFI Group has developed measures of satisfaction that blends state-of-the-art customer satisfaction research theory from leading universities with leading-edge statistical technologies. Specifically, CFI Group has used the following three concepts to measure customer satisfaction to explicitly assess the distinct dimensions of customer satisfaction. [The "Kano" Model referred to below is an oft-cited and well-accepted conceptual model of customer satisfaction.]

- *Overall Satisfaction* - This dimension assesses a customer's overall evaluation, quantifying what the Kano model characterizes as evaluation of "Performance" or "Spoken" attributes. [This dimension of satisfaction encompasses those attributes for which customers "reward" high performance and "punish" low performance in their satisfaction ratings.]
- *Meeting Expectations* – This dimension provides specific evaluation of what the Kano model characterizes as "Basic" or "Expected" attributes [i.e., those attributes that **must** be present as a condition for a person to be satisfied. A good example is "safety" on airplanes]. It addresses the disconfirmation theory of customer satisfaction, which states that an individual's satisfaction level with a product or service is strongly related to how well their experience either confirms or disconfirms what the customer thought they would experience. [The expectations dimension of satisfaction concerns those attributes where customers punish low performance with lower ratings, but do not necessarily reward performance beyond their minimum requirements for satisfaction.]
- *Being Ideal* – This question provides specific evaluation of what the Kano model characterizes as "Surprise" or "Delight" attributes [i.e., those aspects of the product or service that are unexpected and add value for the customer]. The ideal measure accounts for the fact that customers likely refer to a benchmark or standard when evaluating their experiences with a company's product or service. The ideal measure provides a more absolute evaluation of satisfaction and is based on the collection of experiences an individual has had over time and across industries. Of particular importance is that the ideal dimension complements expectations and helps explain loyalty. For example, "ideal" is why individuals don't always eat fast food. Fast food may be "satisfying" and meet "expectations," but may not always be "ideal." [This dimension of satisfaction encompasses people's attitudes toward attributes where low or absent performance is not punished, but high performance is greatly rewarded through high satisfaction ratings.]



Questions based on these three concepts are used to build a composite or multiple-item measure of customer satisfaction which, in addition to its conceptual rigor, offers superior reliability [freedom from measurement error], validity, and precision [of score estimates] over other traditional measures [especially single-item "overall" measures].[34] Three questions are necessary to achieve these benefits because, as discussed, satisfaction is made up of multiple dimensions. Asking only one question severely limits measurement coverage of customer satisfaction and subjects the measurement to bias and measurement error.

Another important point is that by consistently employing these three questions, we can validly make comparisons across different items, market segments, companies and even industries (through comparison with scores from the ACSI and soon the "EUCSI", which use the same satisfaction measure). This ability is invaluable to clients seeking a valid and relevant basis upon which to benchmark their customer satisfaction scores.

[34] Ryan, Michael, Tom Buzas, and Venkatram Ramaswamy. 1995. "Making CSM a Power Tool." *Marketing Research* 7(3):11-16.

Appendix D: Single vs. Multiple Items Measures

How should customer satisfaction, its causes and effects be measured? Is it sufficient to simply ask a customer to rate their satisfaction with a recent experience by checking a yes or no box? If the objective is to simply *screen* respondents for some further activity, then perhaps a dichotomous or other categorical response is acceptable. But if the intention is to gather data for analysis then serious problems will ensue. Single item measures, especially those with limited response categories, possess severe measurement deficiencies that limit their usefulness in advanced statistical analyses of the type usually encountered in consumer satisfaction research.

Differences Between Single and Multiple Item Measures

In addition to their ubiquitous and common use as *screeners* in many survey research designs, single item measures are often used in two additional ways:

(a) Those measuring self-reported facts that allow for the *classification* of respondents, such as years of education, age, number of previous jobs and so on; and
(b) Those purporting to measure attitudinal and behavioral *psychological constructs*, such as satisfaction, recommendation, or purchase intentions.[35]

Measuring the former with a single item is a commonly accepted practice. Errors from this usage occur largely because of response biases, i.e., a respondent may not be totally honest about their income level or age. However the use of single-item measures for psychological constructs is typically discouraged, primarily because they are presumed to have questionable validity, and low levels of reliability. This problem stems from the multifaceted and complex nature of most psychological constructs making it extremely difficult to adequately capture its meaning with a single item. There are exceptions to the norm of using only scales to measure psychological constructs. If the construct being measured is sufficiently narrow or is unambiguous to the respondent (e.g., the measurement of subjective probabilities, such as future behaviors), a single item measure may suffice. But for more complex psychological constructs (especially those based on attitudes) it is usually recommended that scales with multiple items be used.

Nunnally and Bernstein (1994), McIver and Carmines (1981), and Spector (1992) discussed the reasons for using multi-item measures instead of a single item for measuring psychological attributes. They identified the following issues:

- First, individual items have considerable *random measurement error*, i.e. are unreliable. Nunnally and Bernstein (1994) in recommending multiple item scales state, “Measurement error averages out when individual scores are summed to obtain a total score” (p. 67).
- Second, an individual item can only categorize people into a relatively small number of groups. An individual item *cannot discriminate among fine degrees of an attribute*. For example, with a dichotomously scored item one can only distinguish between two levels of the attribute, i.e. they lack precision.
- Third, *individual items lack scope*. McIver and Carmines (1981) say, “It is very unlikely that a single item can fully represent a complex theoretical concept or any specific attribute for that matter” (p. 15). They go on to say; “the most fundamental problem with single item measures is not merely that they tend to be less valid, less accurate, and less reliable than their multi-item equivalents. It is rather, that the social scientist rarely has sufficient information to estimate their measurement properties. “
- Thus their degree of validity, accuracy, and reliability is often unknowable. (p. 15). Blalock (1970) has observed, “With a single measure of each variable, one can remain blissfully unaware of the possibility of measurement [error], but in no sense will this make his inferences more valid” (p. 111).

In summary, classic measurement theory holds that single items are at a relative disadvantage to multi-item measures because more items produce replies that are more consistent and less prone to distortion from socio-psychological biases, and this enables the random error of the measure to be cancelled out. Hence they are more stable over time, more reliable, and more precise than single item measures (see Table D1 for a point by point comparison of the two types of measures).

[35] Wanous, John P., Arnon E. Reichers and Michael J. Hudy (1997) “Overall Job Satisfaction: How Good Are Single-Item Measures?” *Journal of Applied Psychology*, Vol. 82, No. 2, 247-252.

How Individuals Respond to Questions in a Survey

Many things can influence how individuals respond to survey questions (e.g., mood, events they encountered that day, etc.). They may choose *yes* to a question one day and say *no* the next day. It is also possible that people give a wrong answer or interpret the question differently over time. Using multiple item measures mitigates the tendency for individuals to be inconsistent. This is because as noted before, a multi-item measure has several questions targeting the same issue, and the final composite score is based on all questions. People are less likely to make the above mistakes to multiple items, and thus the resulting composite score is more consistent over time.

Many measured social characteristics are broad in scope and simply cannot be assessed with a single question. Multi-item measures are necessary to cover more content of the measured characteristic and to fully and completely reflect the construct domain. These issues are best illustrated with an example. To assess people's job satisfaction, a single-item measure could be as follows: I'm not satisfied with my work. (1 = *disagree*, 2 = *slightly disagree*, 3 = *uncertain*, 4 = *slightly agree*, 5 = *agree*) To this single question, people's responses can be inconsistent over time. Depending on their mood or specific things they encountered at work that day, they might respond very differently to this single question. Also, people may make mistakes when reading or responding. For example, they might not notice the word *not* and agree when they really disagree. Thus, this single-item measure about job satisfaction can be notoriously unreliable. Another problem is that people's feelings toward their jobs may not be simple. Job satisfaction is a very broad issue, and it includes many aspects (e.g., satisfaction with the supervisor, satisfaction with coworkers, satisfaction with work content, satisfaction with pay, etc.). Subjects may like certain aspects of their jobs but not others. The single-item measure will oversimplify people's feelings toward their jobs.

A multi-item measure can reduce the above problems. The results from a multi-item measure should be more consistent over time. As mentioned earlier with multiple items, random errors tend to average out. That is, with 10 items, if a respondent makes an error on 1 item, the impact on the overall score is quite minimal. More important, a multi-item measure will allow subjects to describe their feelings about different aspects of their experiences. This will greatly improve the precision and validity of the measure. Therefore, multi-item measures are one of the most important and frequently used tools in social science.

Research Evidence

There exists a lengthy stream of research findings in various fields exploring the points articulated above—for example:

- In a series of related studies, Nagy (2002), Wanous Reichers and Hudy (1997), Wanous and Hudy (2001) and Dolbier, Webster, McCalister, Mallon and Steinhardt (2005) examined the usefulness of a single-item measure of employee satisfaction. They found support of the use of a single item scale as a substitute for multi-item measures of the same construct. Loo (2002) challenged these findings by arguing for the use of single item measures as surrogates for previously validated multiple-item scales.

- Gardner, Cummings, Dunham and Pierce (1989, 1998) examined the performance of single versus multiple-item measures of "focus of attention at work". They found little difference between the two in terms of validity and common methods bias.

- Desalvo, Fan, McDonell and Fihn (2005) and Desalvo, Fisher, Tran, Bloser, Merrill and Peabody (2006) compared single- and multi-item measures of self-rated health to predict mortality and clinical events. They found that the single item measure of "general self-rated health" demonstrated good reproducibility, reliability and strong concurrent and discriminant scale performance with an established multi-item health status measure. In a similar way, Sloan**,** Aaronson, Cappelleri, Fairclough, and Varricchio (2002) described the strengths and weaknesses of single items and summated scores (from multiple items) as "quality of life" QOL measures. They concluded that no "gold standard" QOL measure can be recommended because no "one size fits all." Single items have the advantage of simplicity at the cost of detail. Multiple-item indices have the advantage of providing a complete profile of QOL component constructs at the cost of increased burden and of asking potentially irrelevant questions. The 2 types of indices are not mutually exclusive and can be used together in a single research study or in the clinical setting.

- Wirtz and Lee (2003) found that a single-item customer satisfaction measure was less reliable and explained less variance than competing six-item and four-item satisfaction measures. Gliem and Gliem (2003) reported similar findings for course evaluations made by students.

- Drolet and Morrison (2001) advocate trading off the higher reliability of fewer multi-item scales against the greater information content of many single-item measures in survey research with customers. While Shamir and Kark (2004) suggest the use of single-item measures as a way to control "common methods bias".

Overall the above examples from the literature provide a taste for the research examining the use of single item measures. Table D.1 provides a summary of the key research findings regarding the characteristics, advantages and disadvantages and best uses for each type of scale.

Table D1: Comparison of Single and Multiple Item measures

Points of Comparison	**Single Item measures**	**Multiple Item Measures**
Validity—ability to capture the true value of construct	Varies—can be acceptable if correlated with another validated measure of the construct. Without evidence of such convergent validity it is impossible to assess.	Moderate to high potential for a valid measure. Has a greater likelihood of capturing multiple facets of psychological constructs.
Reliability—ability to be free of random variation; consistency of measurement	Usually low—internal consistency cannot be evaluated, is best assessed by repetitive measures with same respondent.	Moderate to high potential for measures to be reliable. Coefficient alpha (the basic reliability metric) can be easily computed.
Information Content	Relatively low—because of limited number of scale points typically used (e.g., 1-3, 1-5, etc.).	Relatively high because of multi-faceted nature. Greater specificity is possible due to multiplier effect.
Statistical power (sensitivity)—ability to accurately detect changes in its value over time	Low if scale is dichotomized (e.g., "top-box" or "NPS"), acceptable if 7-10 point intervals are used (e.g., behavioral intention type measures).	Highest levels of sensitivity possible because the number of distinctions between individuals is higher.
Simplicity of administration, analysis, and managerial use	High in all areas.	Low because of the need for more questionnaire items and multivariate analytic techniques. Managerial understanding is often stretched.
Summary of Strengths	• Easy to administer • Can be collected quickly • Suitable for very large samples or census studies • Useful for screening respondents • Good for collecting factual information (e.g., age, income, etc.) • Useful for low-level descriptive and comparative analyses	• Greater sensitivity to variations between respondents allowing finer distinctions among them • Allows for greater coverage of the different aspects of an unobservable construct (e.g., beliefs, attitudes and intentions) • Measure reliability can be readily assessed • Higher levels of potential construct validity • Best for advanced statistical analyses
Summary of Weaknesses	• Unsuitable for measuring multifaceted attitudinal constructs • Require calibration with multi-item scales to establish validity • Reliability can only be established with repeated measures • Low sensitivity to variation between respondents	• Require longer questionnaires and more time to collect • May require larger sample sizes to meet "degrees of freedom" requirements and to adequately assess validity • Potential for "common methods bias". • Meaning is often difficult for practitioners and managerial users to understand

Table D.1 (continue)

Points of Comparison	Single Item measures	Multiple Item Measures
Best Uses	• Screening survey respondents for further treatment or survey actions • Classifying survey respondents into research relevant groups (usually demographic or behaviorally related) • Simplifying previously validated multiple measure constructs by using a single measure shown to be highly correlated with base construct (convergent validity)	• Providing accurate measurements of psychological constructs and other unobservable concepts • Detecting changes in the qualities of psychological constructs • Controlling variability arising from random measurement errors

Some Statistical Examples

The following examples illustrate the statistical properties of a top box measure (a typical approach to measurement used by many research companies) compared with a single 10-point measure, and multiple-item scale measure of the same concept of "perceived product benefits":

Example of the differences in precision for three different types of measures for the concept of "Product Benefits".

		Measures		
Points of Comparison		**Top Box**	**10 Point**	**7 Item**
N	Valid	698	698	708
	Missing	14	14	4
Mean		**0.285**	**5.837**	**66.291**
Std. Deviation		0.452	3.331	24.069
Minimum		0	1	0
Maximum		1	10	100
Std error		0.017	0.126	0.905
Std Error as % range		1.71%	1.40%	0.91%
95% CI +/-		**0.034**	**0.247**	**1.773**
Low end of CI		0.252	5.590	64.518
High end of CI		0.319	6.084	68.064
Ability to detect a 5% increase in mean value				
Test mean value		**0.299**	**6.129**	**69.605**
Critical value		0.319	6.084	68.064
Z score for alt mean		1.126	-0.355	-1.704
Beta		0.8700	0.3613	0.0442
Power (1-beta)		**0.1300**	**0.6387**	**0.9558**
		Note: Benchmark for Power is 80%		
Data Source: Orlando Sentinel				

In the following example a comparison is made between the statistical properties of the ACSI with those of single item measures of future behaviors.

Example of the differences in precision for Recommend, NPS and CSI				
		Measures		
Points of Comparison		**Recommend**	**NPS***	**CSI**
N	Valid	697	697	711
	Missing	15	15	1
Mean		**7.527**	**23.5%**	**74.243**
Std. Deviation		3.085	0.88	22.227
Minimum		1	-1	0
Maximum		10	1	100
Std error		0.117	0.033	0.834
Std Error as % range		1.30%	1.67%	0.83%
95% CI +/-		**0.229**	**6.5%**	**1.634**
Low end of CI		7.297	17.0%	72.610
High end of CI		7.756	30.1%	75.877
Ability to detect a 5% increase in mean value				
Test mean value		**7.903**	**25%**	**77.956**
Critical value		7.756	30%	75.877
Z score for alt mean		-1.260	1.607	-2.493
Beta		0.1038	0.9460	0.0063
Power (1-beta)		**0.8962**	**0.0540**	**0.9937**
		Note: Benchmark is 80%		
Data Source: Orlando Sentinel				

* Fred Reichheld, “The Ultimate Question,” Harvard Business School Press, 2006

Bibliography

Anderson, Eugene, Claes Fornell and Sanal Maznancheryl (2004) ”Customer Satisfaction and Shareholder Value,” *Journal of Marketing*,(October), Vol. 68, no.4, 172.

Anderson, Eugene W., Claes Fornell and Roland T. Rust (1997), "Customer Satisfaction, Productivity and Profitability: Differences Between Goods and Services," *Marketing Science*, Vol. 16, No. 2, 129-145, Summer.

Andrews, Frank M. (1984), “Construct Validity and Error Components of Survey Measures: A Statistical Modeling Approach”, *Public Opinion Quarterly*, p.404-442.

Allen, Derek and Tanniru Rao, *Analysis of Customer Satisfaction Data*, ASQ Quality Press 2000.

Blalock, H.M. Jr. (1970), “Estimating measurement error using multiple indicators and several points in time,” *American Sociological Review*, *35* (1), 101-111.

Carmines, E.G., and R.A. Zeller (1979), *Reliability and Validity Assessment*, Thousand Oaks, CA: Sage.

Cox, E. P. (1980) “The optimal number of response alternatives for a scale: a review,” *Journal of Marketing Research*, 17, 407.

Desalvo, Karen B., Vincent S. Fan, Mary B. McDonell and Stephan D. Fihn (2005), “Predicting Mortality and Healthcare Utilization with a Single Question,” *Health Research and Educational Trust*, 40, 4 (August) 1234-1246.

DeSalvo, K.B., W.P. Fisher, K. Tran, N. Bloser, W. Merrill, and J. Peabody (2006), “Assessing measurement properties of two single-item general health measures,” *Quality of Life Research*, 15 (2), March, 191-201.

Dolbier, C.L., J.A. Webster, K.T. McCalister, M.W. Mallon, and M.A. Steinhardt (2005), “Reliability and validity of a single-item measure of job satisfaction,” *American Journal of Health Promotion*, 19(3), (Jan-Feb), 194-198.

Drolet, Aimee L. and Donald G. Morrison (2001), “Do We Really Need Multiple-Item Measures in Service Research?,” *Journal of Service Research*, 3,3, 196-204.

Falk, R. Frank and Nancy B. Miller, *A Primer for Soft Modeling*, The University of Akron Press, 1992.

Fornell, Claes and Roland Rust (2005) “The effect of customer satisfaction on consumer spending growth,”

Fornell, Claes, Sunil Mithas, Forrest Morgeson, and M. S. Krishnan (2006) ”Customer Satisfaction and Stock Prices: High Returns, Low Risk” Journal of Marketing, Vol. 70, No. 1; p. 3.

Fornell, Claes, Paul Damien, Marcin Kacperczyk, and Michel Wedel (2004) “The Empirical Relationship between Buyer Satisfaction and GDP Growth under Parameter and Distributional Uncertainty,”

Fornell, Claes, Michael D. Johnson, Eugene W. Anderson, Jaesung Cha and Barbara Everitt Bryant, (1996), "The American Customer Satisfaction Index: Nature, Purpose and Findings," *Journal of Marketing*, Vol. 60, October, 7-18.

Fornell, Claes and Jaesung Cha, (1994) "Partial Least Squares," Richard Bagozzi (ed)., *Advanced Methods of Marketing*, 52-78.

Fornell, Claes and David F. Larcker (1981) ”Evaluating Structural Equation Models with Unobserved Variables and Measurement Error,” *Journal of Marketing Research*, (February) , 18, 1, 39.

Fornell, Claes and David F. Larcker (1981) ”Structural Equation Models with Unobserved Variables and Measurement Error: Algebra and Statistics,” *Journal of Marketing Research*, (August), 18, 3, 382.

Fornell, Claes, B.D. Rhee, and Y. Yi (1991) “Direct Regression, Reverse Regression, and Covariance Structure Analysis,” *Marketing Letters*, 20 (3), 309.

Gardner, Donald G., Randall B. Dunham, L.L. Cummings, and Jon L. Pierce (1989), “Focus of Attention at Work: Construct Definition and Empirical Validation,” *Journal of Occupational Psychology*, 62, 61-77.

Gardner, Donald G., L.L. Cummings, Randall B. Dunham and Jon L. Pierce (1998), “Single-Item Versus Multiple-Item Measurement Scales: An Empirical Comparison,” *Educational and Psychological Measurement*, 58,6, 898-915.

Gliem, Joseph A. and Rosemary R. Gliem (2003), “Calculating, Interpreting, and Reporting Cronbach’s Alpha Reliability Coefficient for Likert-Type Scales,” *2003 Midwest Research to Practice Conference in Adult, Continuing and Community Education*, 82-88.

Gruca, Thomas S., and Lopo L. Rego (2005) “Customer Satisfaction, Cash Flow, and Shareholder Value,” *Journal of Marketing,*(July) Vol.69, 115-130.

Haladyna, T. M. (1994). *Developing and Validating Multiple-Choice Test Items.* Hillsdale, NJ: Lawrence Erlbaum.

Hauser, John R, Simester, Duncan I, Wernerfelt, Birger (1996) “Internal customers and internal suppliers,” *Journal of Marketing Research,* (August), Vol. 33, Iss. 3; p. 268

Lohmoeller, Jan-Bernd (1989) *Latent variable Path Modeling with Partial Least Squares*, New York: Springer-Verlag.

Loo, Robert (2002), “A caveat on using single-item versus multiple-item scales,” *Journal of Managerial Psychology*, Vol. 17, No. 1. 68-75.

Louviere, Jordan J. and Towhidul Islam (2004) “A Comparison of Importance Weights/Measures Derived from Choice-Based Conjoint, Constant Sum Scales and Best-Worst Scaling,” *Centre for the Study of Choice (CenSoC)* University of Technology, Sydney, Working Paper No. 04-003.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional Scaling*. Thousand Oaks, CA: Sage.

Morgan, Neil and Lopo Rego (2006) “The Value of Different Customer satisfaction and Loyalty Metrics in Predicting Business Performance,” *Marketing Science*.

Nagy, Mark S. (2002), “Using a single-item approach to measure facet job satisfaction,” *Journal of Occupational and Organizational Psychology*, 75, 77-86

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Peterson, Robert A. and William R. Wilson (1992), “Measuring Customer Satisfaction: Fact and Artifact,” *Journal of Academy of Marketing Science*, 20 (Winter), 61-71

Preston, Carolyn C. and Andrew M. Colman (2000) “Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences,” *Acta Psychologica* 104, 1.

Ryan, Michael J., Thomas Buzas and Venkatram Ramaswamy (1995), "Making Customer Satisfaction Measurement a Power Tool," *Marketing Research*, Vol. 7, No. 3, Summer, 11-16.

Schott, G. R. and W. Bellin (2001), ”An Examination of the Validity of Positive and Negative Items on a Single-Scale Instrument,” *Evaluation and Research in Education*, Vol. 15, No. 2, 84-94.

Shamir, Boas and Ronit Kark (2004), “A single-item graphic scale for the measurement of organizational identity,” *Journal of Occupational and Organizational Psychology*, 77, 115-123.

Sloan, Jeff A., Neil Aaronson, Joseph Cappelleri, Dioane Fairclough, and Claudette Varricchio (2002), “Assessing the Clinical Significance of Single Items Relative to Summated Scores,” *Symposium on Quality of Life in Cancer Patients, Mayo Clinic Proceedings*, 77, 479-487.

Spector, P. (1992), *Summated Rating Scale Construction,* Thousand Oaks, CA: Sage.

Verlegh, Peeter W.J., Hendrik N.J. Schifferstein, Dick R. Wittink (2002), “Range and Number-of-Levels Effects in Derived and Stated Measures of Attribute Importance,” *Marketing Letters;* (February) 13.

Wanous, John P., Arnon E. Reichers and Michael J. Hudy (1997) “Overall Job Satisfaction: How Good Are Single-Item Measures?” *Journal of Applied Psychology*, Vol. 82, No. 2, 247-252.

Wanous, John P., and Michael J. Hudy (2001), “Single-Item Reliability: A Replication and Extension,” *Organizational Research Methods*, Vol. 4,No. 4, (October), 361-375

Wirtz, Jochen and Meng Chung Lee (2003), “An Examination of the Quality and Context-Specific Applicability of Commonly Used Customer satisfaction Measures,” *Journal of Service Research*, 5, 4, 345-355.

Wold, H. (1975). *Soft Modelling by Latent Variables: The Nonlinear Iterative Partial Least Squares Approach*. In: *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, ed. J. Gani, Academic Press, London.