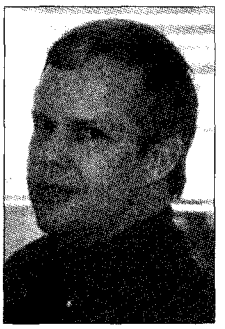
**REFEREED ARTICLE**
Evaluation Journal of Australasia, Vol. 1 (new series), No. 1, March 2001, pp. 8–17

# Measuring customer satisfaction

## Comparing traditional and latent trait approaches using the Auditor-General's client survey

**Scott Bayley**[1]



Consistent with trends across OECD member countries, Australian public sector agencies are increasingly undertaking customer satisfaction surveys in order to evaluate services and to demonstrate accountability. A review of the published literature suggests that customer satisfaction measures based upon traditional test theory suffer from a number of technical problems, including: a lack of attention to scale development; highly skewed distributions; and ordinal level measurement. In this context, the current paper then introduces latent trait measurement theory as an alternative to traditional test theory. Data from the Western Australian Auditor-General's 1998 'Client Survey' is subsequently analysed from both the traditional and latent trait perspectives. The two sets of results are then compared and the advantages offered by latent trait theory are considered. Finally, the implications of these findings for customer satisfaction surveys in evaluation practice are discussed.

## Introduction

### Towards a more customer-focussed public sector

The role of government is evolving in OECD countries in response to fundamental changes in economies and societies (OECD 1995). Members of the public are increasingly being viewed by governments as customers with needs and preferences, as opposed to citizens with rights and obligations. As a consequence, public sector agencies are moving towards adopting private sector marketing techniques such as the use of suggestion boxes, conducting focus groups to identify customer needs, and undertaking customer satisfaction surveys to evaluate services and to demonstrate accountability (Bayley, McCann & Russo 1998; OECD 1996). Similarly, evaluation practitioners in Australia appear to have become heavily reliant on performance indicators such as client satisfaction ratings. We need to remember that under the program evaluation standards it is important to ensure that evaluations use data that is both valid and reliable (The Joint Committee 1994).

*Scott Bayley is Principal Performance Analyst in the Office of the Auditor-General, Western Australia.*

### The concept of customer satisfaction

Customer satisfaction is the outcome of an evaluation process involving the perceived quality or utility provided by a good or service, and a reference point such as pre-service expectations. That is, customer satisfaction is a function of expected performance, perceived performance, and the difference between perceived performance and expectation (Yi 1991).

This particular model of customer satisfaction is known as the expectancy-disconfirmation paradigm (see Oliver 1993; Yi 1991). In this model, customer satisfaction is a measure of how well the client's expectations are being met. Clients are satisfied when the perceived standard of service equals or exceeds their expectations (Lebow 1982; Treasury Board of Canada 1992).

## Measuring customer satisfaction

### Traditional test theory

Customer satisfaction is an abstract concept that is not in itself directly measurable. The best we can do is to measure various indicators of satisfaction, such as by asking customers to respond to survey questions (Hayes 1998). From the answers of respondents, we can then construct a score that indicates each person's level of satisfaction with the product or service in question.

According to what is known as traditional test theory, observed scores consist of two elements: a true score; and measurement error (Hayes 1998; Linden & Hambleton 1997). Measurement error, in turn, consists of two aspects: systematic measurement error; and random measurement error. The basic equation of traditional measurement theory describes the relationship between observed scores, true scores, and error. The model is expressed in Equation 1:

$$X_{nj} = T_n + E_{nj} \quad \text{(Equation 1)}$$

where X is the observed score, T is the true score of person 'n', and E is measurement error. To the extent that we have a small error of measurement, the observed measure (X) is highly representative of the true score (T) (Hayes 1998). If an instrument is free of random measurement error, it is said to be highly reliable. If an instrument is free of systematic measurement error, it is said to be valid (see DeVellis 1991, for a further discussion of these points).

### Problems with current measures of customer satisfaction

According to Chakrapani (1998), the most widely used measures of customer satisfaction are indices and summated scores based upon traditional methods of test construction. Unfortunately, these types of measures have also been subjected to considerable criticism in the literature.

If customer satisfaction measures are to provide accurate and useful information they need to be properly constructed. However, it is common practice for investigators to invent their own satisfaction measures with little attention to the psychometric properties of their instrument (Larsen et al 1979; Lebow 1983; Nguyen, Attkisson & Stegner 1983; Pascoe 1983; Rickett 1992; Ryan, Buzas & Ramaswamy 1995). As a result, many researchers have proposed caution regarding the use of customer satisfaction measures because of their technical shortcomings (Godley, Fiedler & Funk 1998). The nature of these limitations is summarised on the next page.

The literature is quite clear in demonstrating that currently used measures of customer satisfaction based upon traditional test theory suffer from a number of fundamental shortcomings; see: Chakrapani (1998); Lebow (1983); and Yi (1991).

Latent trait theory is based upon two postulates: (1) the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits, or abilities; and (2) the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve ...

## Latent trait theory: A solution?

Due to the well-documented problems of traditional test theory, psychometricians have sought to develop alternative theories and models of measurement. In the last decade there has been a revolution as the application of traditional test theory has given way to 'latent trait theory' (Hambleton, Swaminathan & Rogers 1991). Latent trait theory is based upon two postulates: (1) the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits, or abilities; and (2) the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (Hambleton, Swaminathan & Rogers 1991). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases.

Different latent trait models utilise different mathematical functions, based on diverse sets of assumptions. A one-parameter model for dichotomous (e.g. yes/no) responses, based only on the difficulty of a set of items, was developed by Rasch (1966) and has been adopted and strongly championed by some investigators as a method capable of yielding fundamental (interval level and unidimensional) measurement (see: Andrich 1988a; Duncan 1984; Linden & Hambleton 1997). This

## Problems with current measures of customer satisfaction

### Sampling methods

Measuring customer satisfaction usually involves studying a sample of agency customers with the intention of making generalisations about all of the agency's customers. Common problems in published studies include: poor identification of agency customers; biased samples; and sample sizes that are too small (Bayley, McCann & Russo 1998; Office of the Legislative Auditor 1995).

### Reliability

A reliable scale is one that measures in a consistent manner, and hence produces consistent findings. Reliability is a function of random measurement error. Unfortunately, most researchers do not formally test the reliability of their satisfaction scales, so it is unclear how trustworthy their reported findings really are (Bayley, McCann & Russo 1998; Goldman, Rachuba & Van Tosh 1995; Rickett 1992).

### Response rates

It is quite common for individuals who do not respond to a survey to differ markedly from those who do. The survey's findings can then be biased and misleading as a result of this difference (Bayley, McCann & Russo 1998; Lebow 1983). Response rates in customer satisfaction surveys are often well under 50% which raises further concerns about representativeness and generalisability (Lebow 1982; Pascoe 1983).

### Validity

Validity concerns whether a scale is really measuring what it is intended to measure. Validity is a function of systematic measurement error. Both Chakrapani (1998) and *Fortune* (1995) claim that research has shown that satisfaction measures generally lack validity. More specifically, satisfaction measures tend to be unrelated to important variables such as repeat purchases, customer retention, and organisational profitability. Both Chakrapani (1998) and Reichheld (1996) go on to hypothesise that the source of this problem is not the concept of customer satisfaction per se, but rather with our approach to measuring this construct.

### Scale construction

The most commonly used scaling methods in the field of customer satisfaction measurement include the construction of indexes, and the Likert approach (see: Chakrapani 1998; Hayes 1998). Indexes are very popular, perhaps due to the simplicity of having a single number to gauge overall performance, while Likert summated scales are useful for promoting reliability of measurement (Hayes 1998). However, both indexes and Likert scales are at risk of suffering from certain shortcomings, including: multidimensionality; ordinal rather than interval level measurement; item-dependent person scores; person-dependent item indices; and insensitivity in relation to measuring change (Hambleton, Swaminathan & Rogers 1991; McIver & Carmines 1981; Zhu 1996).

Irrespective of which scaling approach is adopted, it is important to use multiple items (questions) when constructing a satisfaction measure. This helps to ensure that any overall score is a more reliable and valid measure. Unfortunately, the use of single-item satisfaction scales is commonplace (Hayes 1998).

### Absence of comparative benchmarks

Another common difficulty with reported measures of customer satisfaction is the absence of comparative benchmarks (Chakrapani 1998; Kessler 1996; Lebow 1983). 'Since overall satisfaction ratings have a tendency to be overstated, reporting levels of satisfaction in absolute terms and in isolation from other comparative data is often meaningless' (Treasury Board of Canada 1992, p. 6).

### Response formats

A variety of response formats are commonly used in customer satisfaction research; including 2, 4 and 5 point satisfaction formats; 4 and 5 point performance formats; scales using grades such as A, B, C, etc; and numeric scales using ratings from 1 to 10.

The choice of response format not only affects the reliability and validity of the instrument, but also influences how results are used and how easy the survey is to answer and administer. In addition, current evidence indicates that different response formats produce different ratings, and that these ratings are not directly comparable (Devlin, Dong & Brown 1993; Lebow 1983). Too frequently a scale is borrowed or made up without regard to its effectiveness as a measurement tool (Chakrapani 1998).

### Skewed score distributions

A major problem with measures of customer satisfaction is that respondents universally report high levels of satisfaction. This holds true regardless of the method used, the population sampled, or the subject of the rating (Attkisson & Greenfield 1996; Chakrapani 1998). Peterson & Wilson (1992) attempted to research this phenomenon, but were unable to identify the specific causes of this common finding. They did conclude, however, that measures of customer satisfaction are very context-dependent, and that perhaps it is normal customer behaviour to rate services as being 'above average'.

### Data analysis and reporting

Performance data need to be presented in a manner and form that enables the agency and other audiences to assess the current level of customer satisfaction and whether it is improving or deteriorating, and to what extent. A review of the literature suggests that problems with the analysis and reporting of customer satisfaction data are commonplace (Bayley, McCann & Russo 1998; Lebow 1982).

The data collected using indexes and scales based upon traditional test theory tend to be ordinal in nature, which means that only non-parametric statistical tests are appropriate (see: Hambleton, Swaminathan & Rogers 1991; Wright & Stone 1979; Zhu 1996). A review of published studies reveals that it is common for researchers to assume that such data are interval in nature, and to then apply parametric tests of significance which are inappropriate. Even if the data were interval level (which they are not), for highly skewed distributions such as most satisfaction findings, one should still report the median and interquartile range rather than the mean and standard deviation (Peterson & Wilson 1992).

model is based upon the requirement that both guessing and item differences in discrimination are negligible (Andrich 1988a). The probability that a person will answer an item correctly is the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. For a discussion of the mathematical equations used in latent trait measurement, readers are referred to Andrich (1988a) and Hambleton, Swaminathon & Rogers (1991).

In latent trait theory, the ability of persons and the difficulty of items are measured in units known as 'logits', which is short for 'log odds units'. A person's ability in logits is defined as their natural log odds for succeeding on items of the kind chosen to define the 'zero' point of the scale. And an item's difficulty in logits is its natural log odds for eliciting failure from persons with 'zero' ability (Wright & Stone 1979, p. 17). The relationship between person ability, item difficulty, and the odds and probability of a successful response is illustrated in Table 1.

**TABLE 1: LOGIT-TO-ODDS/PROBABILITY CONVERSION TABLE**

| Logit difference between person ability and item difficulty | Odds of success on a dichotomous item | Probability of success on a dichotomous item (%) |
|---|---|---|
| +5.0 | 148.000 | 99 |
| +3.0 | 20.100 | 95 |
| +1.0 | 2.700 | 73 |
| 0.0 | 1.000 | 50 |
| -1.0 | 0.300 | 27 |
| -3.0 | 0.050 | 5 |
| -5.0 | 0.007 | 1 |

In applying the one-parameter Rasch model, one is not simply attempting to model the data, rather one is checking the degree to which a variable with properties of fundamental measurement has been created (Andrich 1988a). The data need to fit the model if the requirements for fundamental measurement are to be satisfied.

When the data-model fit is good, the observed results are independent of the sample of persons and of the particular items, within some broad limits (Wright & Stone 1979). Thus we have person-free item calibration and item-free person measurement. The consequence of this is that we can assess the properties of the scale in question, independently of the distribution of the person parameters. It also means that our measurements are on an interval-level scale, which makes it much easier to undertake comparisons of individuals and/or groups of people (Wright & Stone 1979).

## The Auditor-General's client survey

### Background

The role of the Western Australian Auditor-General is specified in WA's *Financial Administration and Audit Act* 1985. Basically this role involves auditing the financial statements and performance indicators of some 300 public sector agencies, and undertaking performance examinations (evaluations) of the efficiency and effectiveness of public programs. To enable the Office of the Auditor-General (OAG) to assess the attitudes of these 300 agencies towards its products and services, OAG undertakes an annual 'Client Survey'.

OAG initially developed its client satisfaction survey in 1993, and it has since been refined over time with the assistance of private sector contractors. The survey is conducted in about August of each year, with a private sector contractor undertaking the data collection, analysing the results, and then preparing a report for the OAG.

OAG's survey consists of 48 questions grouped into five content areas:

1 items 1–12 concern financial audits;
2 items 13–21 concern audits of agency performance indicators;
3 items 22–35 concern overall client comments on OAG;
4 items 36–42 concern performance audits with a focus on efficiency and effectiveness;
5 items 43–48 relate to audits with a compliance focus.

The nature of the survey's questions requires respondents to rate OAG's performance in each of the five content areas on attributes such as: being unbiased; understanding the agency's needs; timeliness of service delivery; value for money; level of expertise, etc. These ratings are done on a four-point scale (very satisfactory, satisfactory, unsatisfactory, very unsatisfactory; or strongly agree, agree, disagree, strongly disagree).

### Technical approach to data collection and analysis

For 1998, OAG advises that a stratified random sample of 55 agencies was undertaken from a population of 311. Forty-five agencies returned the mailed questionnaire, for a response rate of 82%.

OAG's contractor formed the survey questions into three index-type scales. The first index is labelled 'Quality' and consists of 11 items spread across four of the content areas. The second index is labelled 'Effectiveness' and consists of seven items across all five content areas. The third index, labelled 'Timeliness', contains four items covering four content areas.

OAG's contractor then analysed the data by providing summary statistics for each individual question, giving the percentage of respondents answering 'very satisfactory', 'satisfactory', and so

on. For each of the three indices, an overall weighted percentage of the number of respondents answering 'very satisfactory' or 'satisfactory' is calculated and presented as the 'percentage of satisfied clients'.

## Comparing traditional and latent trait approaches

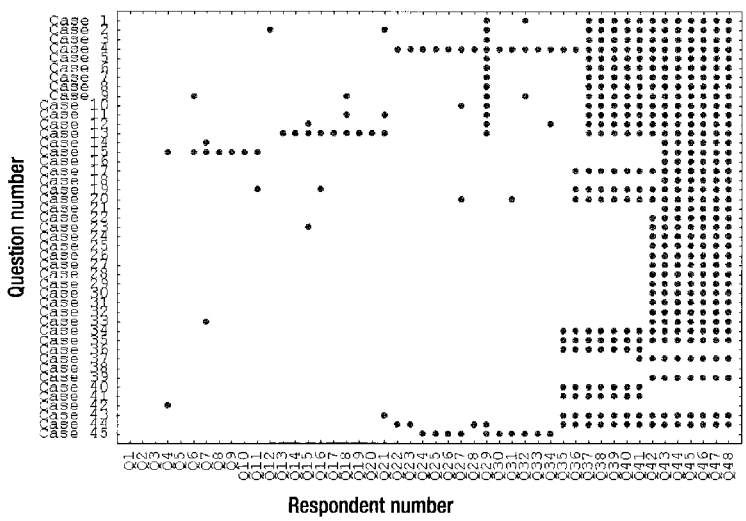
### Data analysis using traditional test theory

The psychometric properties of the three scales in OAG's survey have been analysed using the computer package Statistica[2]. This analysis was undertaken as follows:

- an examination of the raw data for missing values;
- the calculation of descriptive statistics for each question and an examination of the correlations between items;
- a reliability and item analysis was undertaken for each of the three scales;
- each of the scales was then factor analysed;
- the total summated score for every respondent on each of the three scales was calculated and then plotted.

As Figure 1 demonstrates, OAG's survey is troubled by large amounts of missing data. In addition, the pattern of correlations between items within each of the three scales is erratic. Correlations range from -.61 to .74 within both the Quality and Effectiveness scales (these two scales have several items in common); and from -.86 to .64 within the Timeliness scale. This suggests a general lack of internal consistency in the three scales.

**FIGURE 1: MISSING DATA PLOT**




A traditional reliability and item analysis was undertaken for each of the scales. Cronbach's alpha for the 11-item Quality scale is reasonably high at .71, although the item-total correlations reveal that questions Q21, Q27, Q42 and Q48 are not measuring in a manner consistent with the other items (i.e. their item-total correlations are all less than 0.3). This lack of internal consistency confirms that the Quality scale is not unidimensional. Similar results were obtained for the Effectiveness and Timeliness scales.

Attempts to factor analyse each of the three scales were not successful due to the small sample size and large amounts of missing data for some questions. Eliminating questions with missing data and re-running the analysis produces inconsistent factor loadings with some questions having very low or negative loadings on the presumed underlying single factor. As expected, the three scales are not independent of each other, with their correlations shown in the following table:

**TABLE 2: CORRELATIONS BETWEEN SCALES**

| | Quality | Effectiveness | Timeliness |
|---|---|---|---|
| Quality | 1.00 | 0.92 | 0.63 |
| Effectiveness | 0.92 | 1.00 | 0.68 |
| Timeliness | 0.63 | 0.68 | 1.00 |

A graph of the distribution of the summated scores for each of the three scales has been undertaken with 'very satisfactory' coded as 1, and 'very unsatisfactory' is coded as 4. These graphs show the same highly skewed distributions that are common to customer satisfaction studies elsewhere. In addition, the Effectiveness scale shown in Figure 2 appears to be bimodal, although the small sample size makes visual inspection difficult.

In summary, this traditional approach to test construction and data analysis has resulted in the typical problems identified by the literature: scales with unknown validity; problems with the internal consistency of the scales; scales with debatable factor structures; and scales with highly skewed distributions.

### Data analysis using latent trait theory

Using the computer package RUMM[3], the psychometric properties of the three scales in OAG's survey have been re-analysed. RUMM software is based upon Rasch latent trait measurement theory, and is designed for use with cumulative type scales. Cumulative scales are commonly used for measures of ability or performance, while what are known as unfolding scales are thought to be more suited to attitudinal measures (see Andrich & Luo 1993; Roberts, Laughlin & Wedell 1999). Cumulative scales imply that higher levels of the latent trait, should in all probability, lead to higher item scores, which in turn should lead to higher total test scores. In contrast, the unfolding model would predict larger item scores when the item and the individual

in question are near to each other on the latent continuum. Preliminary analysis revealed the three scales to be cumulative rather than unfolding in nature, and hence the use of the RUMM software. Had the scale been found to be of the unfolding type, alternative software, RUMMFOLD, would have been used.

The RUMM analysis for the three scale was undertaken as follows:

- all items were reverse scored so that larger numbers correspond to greater levels of client satisfaction (0 = very unsatisfactory, 1 = unsatisfactory, 2 = satisfactory, and 3 = very satisfactory);
- the test-of-fit summary statistics for each scale were examined;
- the model parameters for each scale were calculated;
- the response frequencies for each scale category were examined;
- the thresholds for each item were examined;
- the fit of each individual item was assessed, with poorly fitting items being examined in further detail;
- the category probability curves for each item were examined as a check on the functioning of the response categories;
- item characteristic curves were examined to ensure that each item functioned in a similar manner across different groups of agencies;
- the fit of individual agencies was then examined, with poorly fitting agencies being studied in greater detail;
- the Guttman pattern for the scale was then assessed;
- the distribution of latent scores for both agencies and items was jointly plotted (person-item frequency distribution);
- the manifest (summated total score) and latent ability scores for each agency were plotted against each other to illustrate the relationship between the two.

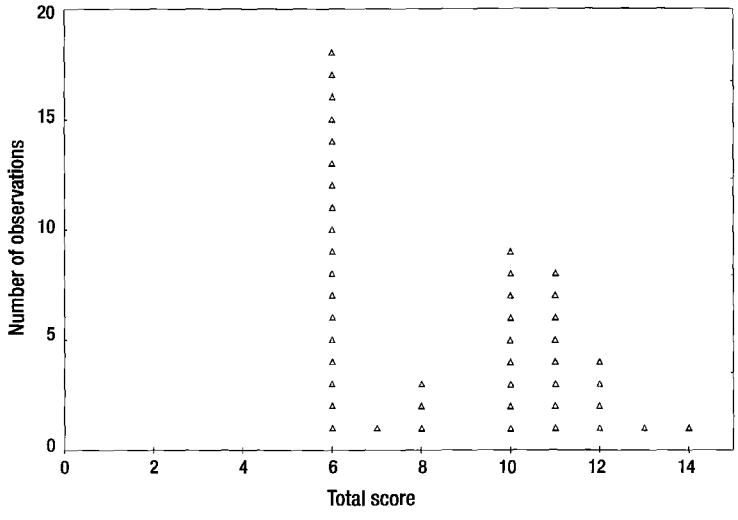
For the sake of brevity, only the analysis for the Quality scale will be presented and discussed.

### Quality scale

To enable the analysis to proceed it was necessary to delete two of the 11 items in this scale. Questions Q42 and Q48 were omitted due to problems with large amounts of missing data.

The test-of-fit summary statistics presented in Table 3 suggest that the data are a reasonable fit to the latent trait model, with the mean and standard deviation for the fit statistics of both items and agencies being acceptably close (within ± 2.0) to the desired figures of zero and one respectively. The summary statistics show that the item-trait

**FIGURE 2: DISTRIBUTION OF EFFECTIVENESS SCALE SCORES**




interaction is non-significant, which suggests that the parameter estimates are invariant (i.e. consistent) across agencies with different levels of satisfaction. The person separation index is used to measure the consistency of the ordering of agencies. This index can be interpreted in a manner similar to traditional measures of reliability based upon internal consistency. The observed result of .726 is therefore acceptable.

**TABLE 3: QUALITY SCALE TEST-OF-FIT SUMMARY STATISTICS**

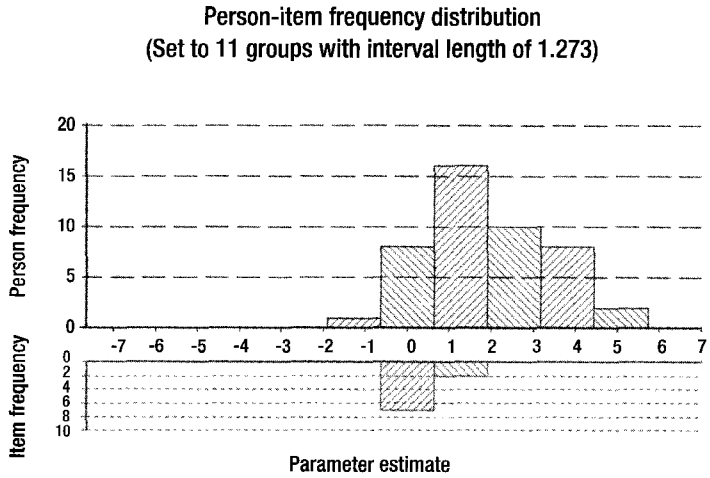
| Item-agency interaction | | | | |
|---|---|---|---|---|
| | Items | | Agencies | |
| | *Location* | *Std error* | *Location* | *Std error* |
| Mean | 0.000 | 0.045 | 1.869 | -0.296 |
| SD | 0.653 | 0.864 | 1.316 | 1.333 |

| Inter–trait interaction | | | |
|---|---|---|---|
| Total item Chi Sq | 11.128 | Person separation index | 0.726 |
| Total degree freedom | 16.000 | Cronbach | N/A |
| Total Chi Sq probability | 0.802 | | |
| Test-of-fit power | REASONABLE | | |

However, the mean and standard deviation for the locations of agencies and items reveals that the items are not well targeted to the agencies. That is, the items have a narrow range of variation in their difficulty and they are located below the satisfaction

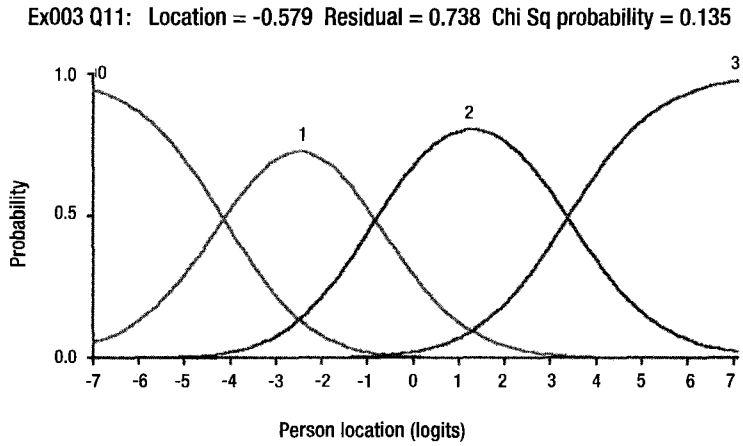
levels of most agencies. This is visually illustrated by the person-item frequency distribution graph in Figure 3 which plots the difficulty of items and satisfaction of agencies (both measured in logits) on a common scale which ranges from -7 to +7.

**FIGURE 3: TARGETING OF ITEMS TO PERSONS (AGENCIES)**




An examination of category probability curves shows that the response categories for all the items in the Quality scale are generally working as intended with the exception of question Q10. This is illustrated in Figures 4 and 5.

**FIGURE 4: EXAMPLE OF RESPONSE CATEGORIES FUNCTIONING AS INTENDED (Q11)**




This diagram illustrates that as overall Person (agency) satisfaction increases (as measured in logits), respondents move away from giving category 0 ratings (very unsatisfied) and towards category 3 ratings (very satisfied). This empirically demonstrates that the response categories for this question are working in the intended manner. In contrast, Figure 5 shows that the response categories for question Q10 are not functioning as intended.

For item Q10, as overall person (agency) satisfaction increases, respondents jump from making category 0 ratings (very unsatisfied) straight to offering category 2 ratings (satisfied) without making use of the intermediate category 1 option (unsatisfied). This indicates that the operation of the response categories for this question is disordered, and points to a need for further qualitative investigation. There could be a problem with the wording of this item, or perhaps with the manner in which agencies are attempting to respond to it.

The test-of-fit statistics presented previously showed that overall, the Quality scale items had an acceptable degree of fit to the measurement model. Figure 6 illustrates the individual fit for item Q11 across three groupings of agencies with varying levels of overall satisfaction. This graph shows that item Q11 operates in a consistent manner irrespective of whether the responding agency has a low, medium or high level of satisfaction with OAG's services.

Similarly, the test-of-fit statistics presented previously showed that taken as a whole, the responses of surveyed agencies also had an acceptable degree of fit to the measurement model for the Quality scale. However, one individual agency (not shown) was identified as being a poor fit, and this situation requires further qualitative investigation.

In summary, once items Q42 and Q48 were deleted, the Quality scale worked reasonably well. Question Q10 needs remedial attention, and the scale as a whole would benefit from some additional items at the higher end of the difficulty scale. Further qualitative work with agencies is also required in order to refine the scale's operation. In stark contrast to the findings based upon traditional test theory, the distribution of latent abilities appears reasonably normal in the RUMM analysis (as shown by Figure 3).

## Comparing traditional and latent trait findings

This section will compare and contrast the findings from the traditional and latent trait analysis, with reference to the measurement principles relevant to each approach. Consistent with the previous section, only the findings for the Quality scale will be presented for discussion.

### Quality scale

As mentioned previously RUMM is intended for use with cumulative scales, with the option of using RUMMFOLD for scales of the unfolding type. Latent trait measurement models are intended to yield fundamental measurement using either cumulative or unfolding scales, and these models are potentially falsifiable. In contrast, traditional methods assume that every scale is cumulative, the focus is on describing the data as opposed to achieving fundamental measurement, and the models are not falsifiable.

This highlights one of the key basic differences between traditional test theory and the latent trait

approach. Latent trait scaling is a theory-driven technique with the intention of achieving fundamental measurement, while traditional test theory is more atheoretical and aims to describe the data collected (see: Andrich 1996; Andrich 1988b).

An important index in traditional test theory is a scale's internal consistency reliability. In traditional test theory, the collection of data is followed by an item analysis with a view towards achieving an internally consistent scale. The traditional reliability coefficient for the Quality scale is .71, while the latent trait person separation index is .72. While the figures are nearly identical, the intention behind the two measures is not.

From the perspective of traditional test theory, reliability is quite central as it is used as a criterion for scale development, and it also leads to the calculation of standard errors. In latent trait theory, reliability can be obtained from the standard errors, but it is not central to the criterion of the model. The construction of the person separation index from the latent trait perspective makes it clear that it is a property of the estimates of person parameters for a sample provided by a test, and is not considered to be a property of the test itself, which is the traditional test theory approach (adapted from Andrich 1988a). So while both approaches express reliability as the ratio of true variance to observed variance, latent trait theory focuses on estimates of person abilities rather than on item statistics. This consistency of person ordering is more relevant to the goals of testing than the traditional test theory's emphasis on item statistics. Thus under the latent trait framework, it is more likely that insights will be gained as compared to traditional methods.

The traditional and latent trait analyses lead to different conclusions as to which of the initial 11 items should be included in the final scale. In the traditional analysis, the final Quality scale would contain the following seven questions: Q9; Q10; Q11; Q12; Q20; Q25; and Q26. In the latent trait analysis, the final scale would be comprised of a slightly different group of eight questions: Q9; Q11; Q12; Q20; Q21; Q25; Q26 and Q27 (i.e. the two scales have six items in common).

Under traditional test theory, the aim is to increase reliability, and for this the items selected should be of similar difficulty, answered correctly 50% of the time, and each individual item should accurately discriminate between those with low or high overall test scores. Under latent trait theory, the model reflects the expectation that more able people will have a higher probability of getting every item correct than less able people, and the items selected would tend to have a range of difficulty and a similar and moderate discrimination.

The latent trait analysis allows for the examination of response categories to check if they are working in the intended manner. The findings presented for the Quality scale showed that the response categories for the items were generally working as intended, with the exception of Q10. This points to the need for further qualitative analysis to try and identify the cause of this

**FIGURE 5: EXAMPLE OF RESPONSE CATEGORIES NOT FUNCTIONING AS INTENDED (Q10)**

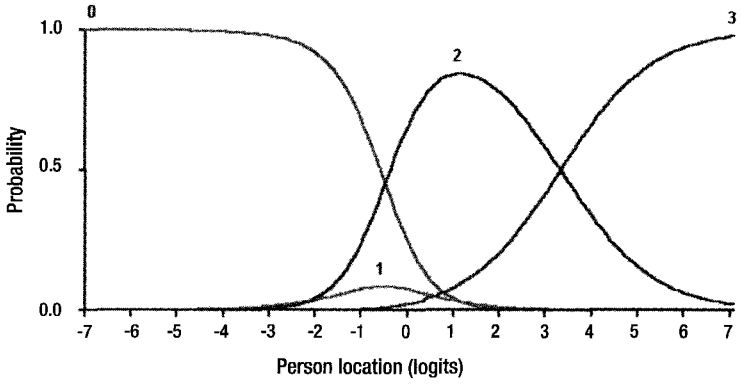
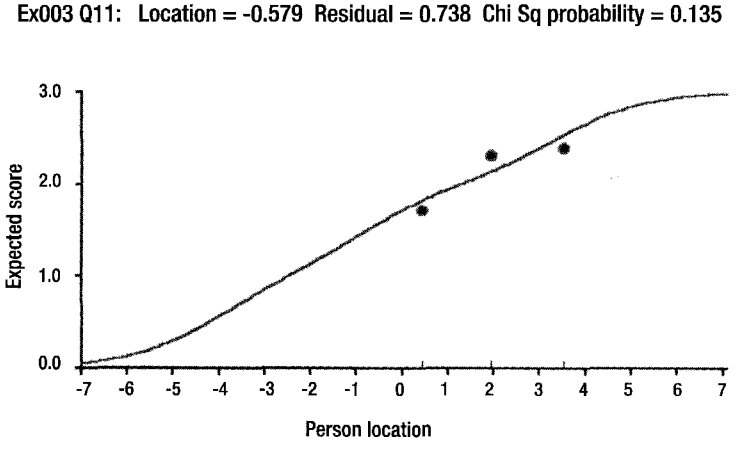



**FIGURE 6: FIT OF Q11 ACROSS THREE GROUPS OF AGENCIES**




malfunction. There are many possible causes, but the disorder may result from multidimensional rather than unidimensional responses, from different discrimination at the thresholds, from the lack of capacity to use all of the assigned number of categories, or from a genuine empirical disordering of the categories. There is never any guarantee that response categories will operate in the way intended. This ordering must be treated as a hypothesis about the data and it is important that the statistical model applied has the property that it can test this hypothesis (Andrich, de Jong & Sheridan 1997). Unfortunately, traditional test theory is unable to accommodate this requirement.

The targeting of items to individuals is another area where the traditional and latent trait approaches differ. In traditional test theory, items would be targeted to the group as a whole, and in general be of equal and moderate difficulty. In latent trait theory, items would be targeted to individuals, and tailored testing can be applied as a result of this focus. As shown in Figure 3, the latent trait analysis reveals that the items have a narrow difficulty range and that they are too easy for most agencies.

A final area of difference between the traditional and latent trait approaches is in relation to checking fit. In traditional test theory, the correlation between responses to items and the total score across persons is obtained as an index of discrimination of each item. These correlations are routinely taken to indicate the quality of the items. The latent trait approach is more sophisticated, since the fit of items is examined along with the fit of persons, and the model is potentially falsifiable.

## Conclusions and implications for customer satisfaction surveys in evaluation practice

If evaluation practitioners wish to undertake studies of customer satisfaction in order to enhance program development and accountability, it is first necessary to develop a theoretical construct of satisfaction that can be measured in a useful and meaningful way. Valid and reliable measures of theoretical constructs are required, if their distribution, causes and consequences are to be investigated and understood (McIver & Carmines 1981; Ramsay 1975).

At the present time, evaluation practitioners are using a wide range of approaches to measure customer satisfaction in both the public and private sectors. These different methods vary considerably in terms of both their technical adequacy and cost (see: Bayley, McCann & Russo 1998; Chakrapani 1998; Lebow 1983; Yi 1991). From the perspective of evaluation practitioners and the consumers of evaluation research, the adequacy of any one particular measurement approach can be assessed in terms of the following 10 criteria:

1 Sound measurement is conceptually defensible. That is to say that the operationalisation of measurement should be driven by a combination of measurement theory and substantive knowledge. For example, the choice of cumulative or unfolding response mechanisms.

2 The measurement model utilised should be potentially falsifiable. It should be possible to examine and test the fit of the data to the model.

3 Unidimensionality. The instrument should only be measuring one thing, i.e. the test items should be homogeneous (unidimensionality is a relative matter, and a test is or is not unidimensional in relation to a particular purpose).

4 The measurement model should scale both persons and items.

5 Validity. The instrument really does measure what it is intended to measure, i.e. the absence of systematic error (a test's validity is also assessed in relation to its use for a particular purpose).

6 The instrument should produce interval or ratio level measurement.

7 Person abilities and item difficulties should be estimated independently of each other. Scores should not be group or item dependent, the parameter estimates need to be invariant.

8 Reliability. The instrument should measure consistently for its intended purpose, i.e. the relative absence of random error.

9 Accuracy. The standard error of measurement should be small relative to the intended use of the scale.

10 Reproducible. The tendency for a given score to be associated with a particular pattern of responses.

For organisations and evaluation practitioners seeking to assess programs using measures of customer satisfaction, these 10 criteria suggest the following hierarchy of options:

1 A single question or a series of single questions concerning customer satisfaction are asked and reported individually.

2 Multiple questions are asked and then reported as a simple index or average.

3 Multiple questions are asked and then formed into a unidimensional scale using traditional test theory.

4 Multiple questions are asked and then formed into a unidimensional scale using Rasch latent trait theory.

The first option suffers from significant problems in the areas of reliability and validity, and as such it is not worthy of serious consideration. The second option is problematic but not completely without merit, particularly if the limitations of this approach are made clear to decision-makers. The third option is better still, but as we have seen, customer satisfaction measures based on traditional test theory suffer from a number of shortcomings. Finally, the development of customer satisfaction scales using Rasch latent trait theory is a technique almost unknown in the professional literature, but it is clearly the most technically sound approach of all.

It is worth bearing in mind that there is never any single correct type of data that must be extracted from a given set of empirical observations (Jacoby 1991). The Rasch latent trait model is advantageous when the intention is to obtain fundamental measurement, rather than to simply describe the data that might be collected (Andrich 1988a). Person-free instrument calibration and item-free person measurement are the conditions which make it possible to generalise beyond the particular scale used (Wright & Stone 1979). Conceivably, major advances in the field of customer satisfaction research will result from the fundamental measurement made possible by the Rasch latent trait model. Better measurement in turn, also provides the basis for further improving evaluation practice.

## Notes

1 The views expressed in this paper are solely those of the author, and do not reflect, in whole or in part, the position of the Auditor-General.

2 See **www.statsoft.com.**

3 Rasch Unidimensional Measurement Models, see **www.faroc.com.au/~rummlab.**

## References


Andrich, D. 1988a, *Rasch Models for Measurement*, Sage, Newbury Park.

Andrich, D. 1988b, A scientific revolution in social measurement, paper presented at the American Research Association's Conference, New Orleans.

Andrich, D. 1996, 'A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies', *British Journal of Mathematical and Statistical Psychology*, vol. 49, pp. 347–365.

Andrich, D., de Jong J. & Sheridan, B. 1997, 'Diagnostic opportunities with the Rasch model for ordered categories', in *Applications of Latent Trait and Latent Class Models in the Social Sciences*, eds J. Rost & R. Langeheine, Waxman Munster, New York.

Andrich, D. & Luo, G. 1993, 'A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses', *Applied Psychological Measurement*, vol. 17, no. 3, pp. 253–276.

Attkisson, C. & Greenfield, T. 1996, 'The Client Satisfaction Questionnaire (CSQ): Scales and the Service Satisfaction Scale – 30 (SSS-30)', in *Outcomes Assessment in Clinical Practice*, eds L. Sederer & B. Dickey, Williams & Wilkins, Baltimore.

Bayley, S., McCann, P. & Russo, A. 1998, *Listen and Learn: Using Customer Surveys to Report Performance in the Western Australian Public Sector*, Office of the Auditor-General, Perth.

Chakrapani, C. 1998, *How to Measure Service Quality & Customer Satisfaction*, American Marketing Association, Chicago.

DeVellis, R. 1991, *Scale Development: Theory and Applications*, Sage, Newbury Park.

Devlin, S., Dong, H. & Brown, M. 1993, 'Selecting a scale for measuring quality', *Marketing Research*, vol. 5, no. 3, pp. 12–17.

Duncan, O. D. 1984, *Notes on Social Measurement*, Russell Sage Foundation, New York.

*Fortune*, 1995, 'Americans can't get no satisfaction', vol. 11, pp. 186–194.

Godley, S., Fiedler, E. & Funk, R. 1998, 'Consumer satisfaction of parents and their children with child/adolescent mental health services', *Evaluation and Program Planning*, vol. 21, pp. 31–45.

Goldman, H., Rachuba, L. & Van Tosh, L. 1995, 'Methods of assessing mental health consumers' preferences for housing and support services', *Psychiatric Services*, vol. 46, no. 2, pp. 169–172.

Hayes, B. 1998, *Measuring Customer Satisfaction*, ASQ Quality Press, Milwaukee.

Hambleton, R., Swaminathan, H. & Rogers, H. 1991, *Fundamentals of Item Response Theory*, Sage, Newbury Park.

Jacoby, W. 1991, *Data Theory and Dimensional Analysis*, Sage, London.

The Joint Committee on Standards for Educational Evaluation 1994, *The Program Evaluation Standards*, Sage, Thousand Oaks.

Kessler, S. 1996, *Measuring and Managing Customer Satisfaction*, ASQC Quality Press, Milwaukee.

Larsen, D., Attkisson, C., Hargreaves, W. & Nguyen, T. 1979, 'Assessment of client/patient satisfaction: Development of a general scale', *Evaluation and Program Planning*, vol. 2, pp. 197–207.

Lebow, J. 1982, 'Consumer satisfaction with mental health treatment', *Psychological Bulletin*, vol. 91, no. 2, pp. 244-259.

Lebow, J. 1983, 'Client satisfaction with mental health treatment: Methodological considerations in assessment', *Evaluation Review*, vol. 7, no. 6, pp. 729–752.

Linden, W. & Hambleton, R. 1997, *Handbook of Modern Item Response Theory*, Springer, New York.

McIver, J. & Carmines, E. 1981, *Unidimensional Scaling*, Sage, Newbury Park.

Nguyen, T., Attkisson, C. & Stegner, B. 1983, 'Assessment of patient satisfaction: development and refinement of a service evaluation questionnaire', *Evaluation and Program Planning*, vol. 6, pp. 299–314.

OECD 1995, *Governance in Transition*, OECD, Paris.

OECD 1996, *Responsive Government – Service Quality Initiatives*, OECD, Paris.

Office of the Legislative Auditor (State of Minnesota), 1995, *State Agency Use of Customer Satisfaction Surveys*, OLA, Saint Paul.

Oliver, R. 1993, 'A conceptual model of service quality and service satisfaction', in *Advances in Services Marketing and Management*, vol. 2, eds. T. A. Swartz, D. E. Bowen & S. W. Brown, Jai Press, London.

Pascoe, G. 1983, 'Patient satisfaction in primary health care: A literature review and analysis', *Evaluation and Program Planning*, vol. 6, pp. 185–210.

Peterson, R. & Wilson, W. 1992, 'Measuring customer satisfaction: Fact and artifact', *Journal of the Academy of Marketing Science*, vol. 20, no. 1, pp. 61–71.

Ramsay, J. 1975, '*Review of Foundations of Measurement* by D. H. Krantz, R. D. Luce, P. Suppes and A. Tverskey', *Psychometrika*, vol. 1, no. 40, pp. 257–262.

Rasch, G. 1966, 'An individualistic approach to item analysis', in *Readings in Mathematical Social Sciences*, eds. P. Lazarseld & N. Henry, MIT Press, Cambridge, Massachusetts.

Reichheld, R. 1996, *The Loyalty Effect*, Harvard Business School Press, Boston.

Rickett, T. 1992, 'Consumer satisfaction surveys in mental health', *British Journal of Nursing*, vol. 1, no. 10, pp. 523–527.

Roberts, J., Laughlin, J. & Wedell, D. 1999, 'Validity issues in the Likert and Thurstone approaches to attitude measurement', *Educational and Psychological Measurement*, vol. 59, no. 2, pp. 211–233.

Ryan, M., Buzas, T. & Ramaswamy, V. 1995, 'Making CSM a power tool', *Marketing Research*, vol. 7, no. 3, pp. 11–16.

Treasury Board of Canada 1992, *Measuring Client Satisfaction*, Communication and Coordination Directorate, Ottawa.

Wright, B. & Stone, M. 1979, *Best Test Design*, Mesa Press, Chicago.

Yi, Y. 1991, 'A critical Review of consumer satisfaction', in *Review of Marketing*, ed. V. Zeithmal, American Marketing Association, Chicago.

Zhu, W. 1996, 'Should total scores from a rating scale be used directly?', *Research Quarterly for Exercise and Sport*, vol. 67, no. 3, pp. 363–372.