Westland *Financial Innovation* (2015) 1:10
DOI 10.1186/s40854-015-0011-8

Financial Innovation
a SpringerOpen Journal

RESEARCH Open Access

The information content of financial survey response data

J Christopher Westland

Correspondence: westland@uic.edu
Information & Decision Sciences,
University of Illinois, 601 S. Morgan
Street, (UH2400), Chicago, IL
60607-7124, USA

Abstract

Market data for financial studies typically derives from either historical transactions or contemporaneous surveys of sentiment and perceptions. The research communities analyzing data from these opposing categories of source data see themselves as distinct, with advantages not shared by the other. This research investigates these latter claims in an information theoretic context, and suggests where methods and controls can be improved. The current research develops a Fisher Information metric for Likert scales, and explores the effect of particular survey design decisions or results on the information content. A Fisher Information metric outperforms earlier metrics by converging reliably to values that are intuitive in the sense that they suggest that information captured from subjects is fairly stable. The results of the analysis suggest that varying bias and response dispersion inherent in specific surveys may require increases of sample size by several orders of magnitude to compensate for information loss and in order to derive valid conclusions at a given significance and power of tests. A prioritization of quality of design, and the factors relevant to survey design are presented in the conclusions, and illustrative examples provide insight and guidance to the assessment of information content in a survey.

Keywords: Financial transactions; Information theory; Survey; Economics; Mathematical models

Survey response paradigms

Market data for financial studies typically derives from one of two broad categories of source material: (1) records and summaries of historical transactions offered from sources such as Compustat, CRSP, government statistics, and raw market transactions; and (2) survey data of sentiment and perceptions from a variety of demographics offered from sources such as Harte Hanks, Valassis, Acxiom and Experian Simmons.

The research communities analyzing data from these opposing categories of source data see themselves as distinct, with advantages not shared by the other. Survey data researchers argue that they have access to relevant data on topics that are financially important, but difficult or impossible to directly observe. Transaction data researchers argue that their information is objective, since money has changed hands (Miller 1999; Miller 2000). They may disparage survey datasets as potentially biased and lacking in accuracy due to the subjectivity of responses and faults of measurement instruments Bennett et al. (2012). Countering this, behavioral finance researchers who almost exclusively rely on survey data contend that they have evolved methodologies and controls that mitigate biases. This research investigates these latter claims in an information theoretic context, and suggests where methods and controls can be improved.

Likert scales represent a widely used approach to scaling responses in survey research, such that the term is often used interchangeably with rating scale. Likert scales require subjects to project qualitative or quantitative beliefs or opinions onto a discrete set of Likert items – fundamental observations from such experiments – typically containing between three and nine categories. The survey researchers' challenge is to record subject responses in mappings that are balanced, properly scaled, meaningful, informative, accurate and unbiased. There is considerable discussion as to the exact meaning of Likert scaling (Alphen et al. 2008; Bond and Fox 2007; Fitzpatrick et al. 2004; White and Velozo 2002; Likert 1974). Likert scales are sometimes considered an implementation of a Rasch model, though not every set of Likert scaled items can be used for Rasch measurement, but in practice data has to be thoroughly checked to fulfill the strict formal axioms of the Rasch model (Norquist et al. 2004, Bond and Fox 2007). Likert scale data can, in principle, be used as a basis for obtaining interval level estimates on a continuum by applying the polytomous Rasch model, when data can be obtained that fit this model. In addition, the polytomous Rasch model permits testing of the hypothesis that the statements reflect increasing levels of an attitude or trait, as intended.

Much of this uncertainty in application and interpretation arises from its history of development. Rather than being a direct product of statistical modeling or measurement theory, it was initially developed as an ad hoc, readily implementable method for capturing otherwise unobservable 'personal belief' information. Rensis Likert developed the 'Likert scale' during his PhD thesis work in the 1930s (Likert 1974; Jöreskog and Sörbom 1982). Since its development, the Likert scale has become popular and extensively applied in survey research – in marketing, sociology, psychology and other fields – allowing respondents to express both the direction and strength of their knowledge and opinions, albeit couched in an artificial form and non-intuitive recording structure.

Open questions on the nature and interpretation of Likert scales make it difficult to assess the adequacy of experimental or survey design, or generate reliable statistics concerning the unobservable sentiments that Likert scales supposedly measure. Even were it possible to consistently and accurately verify the truthfulness of subject responses, theory is still lacking on ways to assess whether this is accurately translated into research conclusions. Though we assume that there must be some information loss in translation from presumably continuous personal beliefs to discrete and ordered response scales, the causes and forms of loss have not been widely studied.

This paper contributes to the understanding of design trade-offs and requirements which are required to be able to assert that a particular Likert sample contains a specific amount of information concerning the specific research question that the survey is designed to answer.

Section 2 of this research reviews the prior literature, describing where the literature has provided information measures for Likert scales, and where needed guidance in survey design and Likert response statistical analysis is still wanting. Section 3 applies a standard Fisher Information measure of information content to Likert scales, and explores the consequences of particular survey design decisions or results. Section 4 provides exact solutions with Gaussian beliefs, and these are compared and contrasted with results from earlier work of (Srinivasan and Basu 1989) which also assumes Gaussian beliefs. Section 5 provides an example, comparing to Srinivasan and Basu's (1989) (Stevens and Galanter 1957) results which are often cited in support of survey statistics. Section 6 discusses the implications of a Fisher Information criterion for Likert scale interpretation, and draws conclusions germane to future survey design.

Prior research

Ordered categorical scales are widely used in marketing and the behavioral sciences to measure a variety of human responses and performance. A metric z is ordered categorical if it takes on a countable set of values $\{z_1 \dots z_m\}$ which are ordered such that $z_i > z_j$ *for all* $i > j$. Ordered categorical metrics are special cases of ordinal metrics where the numbers assigned to categories are consecutive, and are not considered to be rankings (Srinivasan and Basu 1989). Ordered categorical scales subsume a wide range of commonly encountered metrics; e.g., semantic differential scale ratings, such as the 'luxuriousness' of a car; Likert scale responses, such as level of agreement with a statement about an attitude or perception; and school assessment such as grades 'A,' B,' etc.

Ordered categorical scales are almost always used to simplify the task of measuring a completely or partially unobservable phenomenon, such as personal utilities associated with usage of a particular product. With this justification comes an assumption that an ordered categorical measurement will not be as accurate or desirable as the more preferable continuous measurement, but that the simplicity and cost-savings inherent in research implementations using ordered categorical metrics offsets the inaccuracies introduced by using a less than perfect measurement instrument. The argument is typically made that any inaccuracies introduced by the ordered categorical simplification of measurement can be counterbalanced with a larger sample size, and a regression to means of the continuous population distribution as the sample size expands to infinity (Stevens and Galanter 1957).

Unfortunately, there is little prior research that has explored the information loss – and commensurate inaccuracies injected into a research study – that comes from substituting ordered categorical metrics for continuous metrics. (Srinivasan and Basu 1989) developed what they called an "index of metric quality" for an ordered categorical metric Z that they called I_z implying that I_z somehow captured the information content of Z (thus the I notation). They assume that random variable Z is a transformation of some underlying 'true score' τ with the relationship $X = \tau + \varepsilon$ where X and ε (the error) are both Normally distributed. This is a somewhat convoluted way of describing the a situation where there are two measurement tools – an ordered categorical tool generating Z and a (hypothetical) continuous counterpart generating X where τ is unobservable and ε is a measurement error, though (Srinivasan and Basu 1989) are not clear whether error supposedly occurs in measuring Z or the (hypothetical) continuous counterpart X or both. Thus the transformation is $T: (X - \varepsilon) \rightarrow Z$. They are also mute on which of several correlation coefficients, for example ρ or r, should be used, though it is made clear from the context that they intended to measure the coefficient of determination R^2 , whose usual purpose is the prediction of future outcomes on the basis of other related information. In case of a single regressor, fitted by least squares, R^2 is the square of the Pearson product–moment correlation coefficient relating the regressor and the response variable.

(Srinivasan and Basu 1989) define a metric in their study: "the descriptive measure of the ability of Z or X to predict the 'true score' "τ which has value:

$$I_Z = \frac{r^2(Z,\tau)}{r^2(X,\tau)}$$

They claim that I_z "provides an upper bound on the explanatory power (population R^2) of multiple regression models in which the ordered categorical variable is regressed against a set of predictors." Interpreting their notation of $R^2 = r^2(.,\tau)$ to mean a measure of the fit for a particular sample, this implies that they are trying to measure the information content of the (hypothetical) continuous predictors generating X with respect to the value of the ordered categorical variable Z as the "gold" standard of informativeness or predictive power against which ordered categorical scales should be judged. Yet $r^2(X,\tau)$ is dimensionless, and therefore at best can only be used to assess relative performance of particular metrics. If the value $r^2(Z,\tau)$ is 90 % of $r^2(X,\tau)$ that does not mean that Z captures 90 % of the information in either X or τ. In fact, the general interpretation of predictive value or information content associated with any particular $r^2(.,\tau)$ is not particularly meaningful. (Srinivasan and Basu 1989) also assume that $I_Z \in [0,1]$ but if τ were in fact ordinal, one could envision cases where values in excess of 1 could obtain. Nonetheless, it doesn't seem unreasonable to assume that continuous metrics will improve on prediction of ordered categorical metrics in their system. A reinterpretation of their intention would be that they want to find the information loss in using a measure $Z(\tau)$ instead of the more accurate and desirable $X(\tau)$. The authors' use of an R^2 based metric inherently assumes a linear model with a single regressor, fitted by least squares. This is a very restrictive set of assumptions, that we could reasonably assume would be tested often when measuring human response, which, in turn, is notoriously non-linear. Fundamentally $r^2(.,\tau)$ is the squared correlation – the normalized version of the covariance, obtained by dividing covariance by the standard deviations of each of the variables. Correlations are dimensionless and range from -1 *to* $+1$. Several different formulas are used to calculate correlations, but the most familiar measure is the Pearson product–moment correlation coefficient, or Pearson's correlation. Correlations are simple to interpret and to compare to each other because of their normalized range. Correlations between unobserved (latent) variables are called canonical (for continuous data like X) or polychoric (for ordinal data like Z) correlation and are distinct from the Pearson product–moment correlation coefficient that was used in (Srinivasan and Basu 1989). Correlations provide useful summarizations of large datasets into single metrics; unfortunately their parsimony comes which a significant loss of information about the sample as (Anscombe 1973) demonstrated in Fig. 1.

An example of the correlation of x and y for various distributions of (x,y) pairs was provided in Fig. 2 by (Brandstätter et al. 2002) which clearly illustrates how R^2 summarizations can mislead.

Clearly all of the scatterplots on the bottom row contain significant information about Z, X *and* τ, but in all cases $r^2(X,\tau) = 0$ and for any positive value of $r^2(Z,\tau)$ then $I_Z = \infty$. Unfortunately, Srinivasan and Basu's metric misrepresents the informativeness – i.e., their "explanatory power of multiple regression models in which the ordered categorical variable is regressed against a set of predictors" – when those predictors and categorical variables are about X and Z.

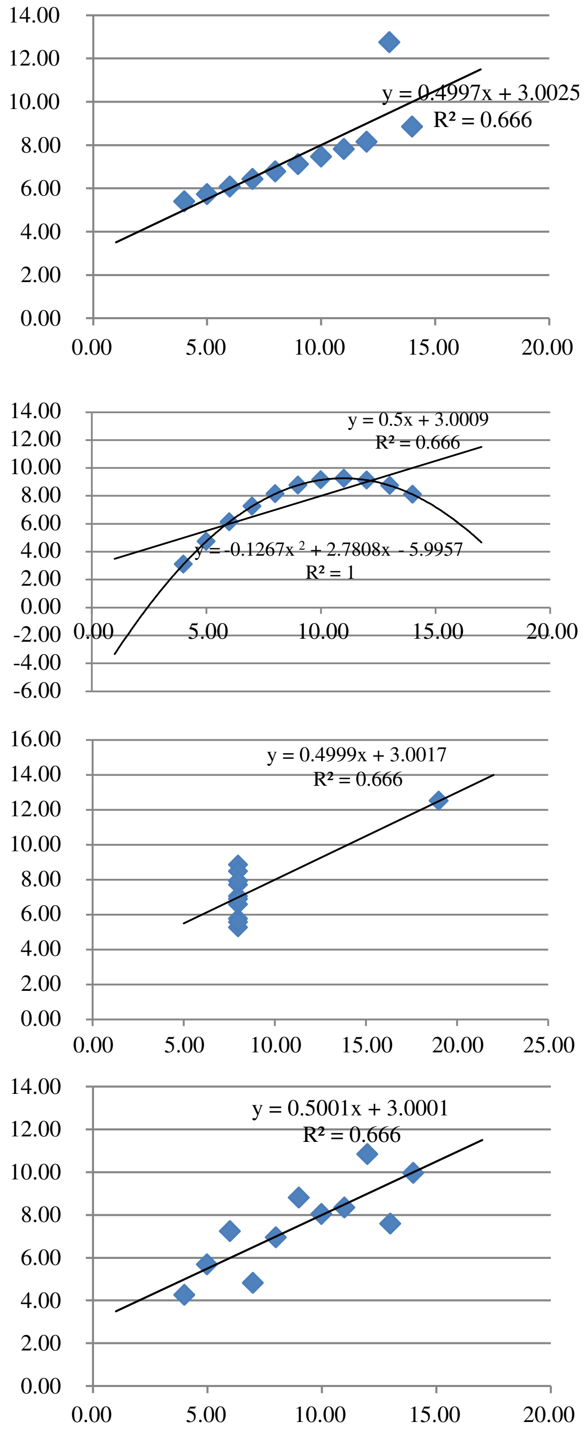

Fig. 1 Anscombe's (1973) Quartet: Four Distinct Datasets with Identical R^2

Information metrics for likert scaled data

Srinivasan and Basu (1989) couch their information metric I_z in terms of the context of a linear combination of unobserved Gaussian predictors X, and a Likert scaled observation Z. Because of the flaws inherent in using functions of R^2 as information measures, it seems productive to seek an alternative statistical measure of information that has

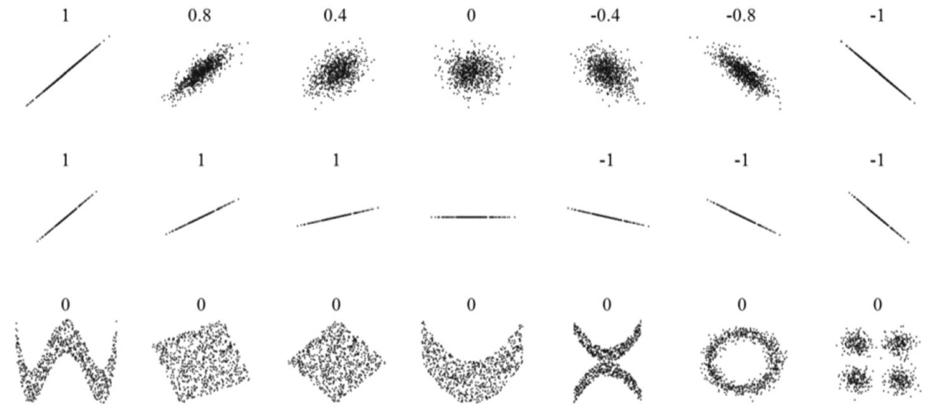

Fig. 2 Pairs of random variables with correlations of 0, −1 and +1 (Brandstätter et al. 2002)

been successfully applied elsewhere. Fortunately a widely applied measure of 'information' exists in the form of Fisher Information which is applicable to both linear-least squares contexts, and to highly non-linear, non-normal context using loss functions other than squared error.

In mathematical statistics and information theory, the Fisher information (sometimes simply called information) can be defined as the variance of the score, or as the expected value of the observed information. In Bayesian statistics, the asymptotic distribution of the posterior mode depends on the Fisher information and not on the prior. When there are N parameters, Fisher information takes the form of an N × N Fisher Information Matrix positive semidefinite symmetric matrix, defining a Riemannian metric on the N-dimensional parameter space; in that context, this metric is known as the Fisher information metric, and the topic is called information geometry. The metric is interesting in several ways; it can be derived as the Hessian of the relative entropy; it can be understood as a metric induced from the Euclidean metric, after appropriate change of variable; in its complex-valued form, it is the Fubini-Study metric. Thus Fisher Information offers a generalizable and mathematically consistent measure of the 'informativeness' of a Likert data item in measuring an underlying set of continuous parameters – the original goal of Srinivasan and Basu's metric.

There is no loss of generality in assuming that the underlying values being measured are a set of continuous phenomena. Such assumptions are widely found in consumer and behavioral research where Likert scaled survey responses typically serve as the data fed into classical statistical summarization and reporting models such as regression, ANOVA, factor analysis and summarization models. More recently, it has become popular to analyze Likert survey items with path analysis structural equation model software such as AMOS, LISREL and PLS path analysis software where there is an implicit assumption that multiple measurements need to be taken to, in effect, 'triangulate' an underlying latent or unobserved phenomenon. The classical approaches, in particular, were designed around measurements from astronomy, agriculture and physics, and were not initially formulated for highly subjective, indirectly measured constructs such as human behavioral, performance and opinion constructs. In these cases, it is common to make implicit assumptions that the underlying opinions or beliefs of subjects – the ones that are mapped into the Likert item data – are Gaussian distributed. Clearly a five or seven value discrete scale will not be Gaussian distributed, but this is often assumed to be a useful approximation. It is common to find authors finessing this assumption by invoking Central Limit Theorem convergence, for example invoking asymptotic conditions to justify this (e.g., consider the assumption that a random variable $\tilde{x} \sim \text{Poisson}\ (\lambda) \Rightarrow \tilde{x} \approx \mathbb{N}(\mu = \lambda, \sigma^2 = \lambda)$ for $\lambda > 20$). They may alternatively make the weaker and inclusive assumption that the discrete mappings of the Likert scale are approximations of the subjects' continuous belief or value systems, which either in fact or in convergence are Gaussian distributed. The latter assumption is often cited in the modeling assumption where there are social science applications of the highly popular LISREL, AMOS and PLS path analysis software packages.

On the other hand, there is a body of research that lends support to alternatives to continuous or Gaussian beliefs. For example (Kühberger 1995; Kühberger 1998; Kühberger et al. 2002; Kühberger et al. 1999; Clarke et al. 2002) conclude that people do not generally hold strong, stable and rational beliefs – that their responses are very much influenced by the way in which decisions are framed, which might serve as a caveat in the design of survey instruments.

Where Likert scale variables are modeled as a response and predictor, it is possible to use an ordered logit or probit model to handle the dependent variable; the independent variable would be categorical. If the researcher is inured to the assumption of Gaussian belief distributions, then it may be consistent to invoke the probit function or quantile function associated with the standard Gaussian distribution. But in the context of Likert scales, the logit model is probably philosophically more accurate (though in practice there is very little difference between the two models). The logit model is philosophically consistent with Likert scales for two reasons: The logit is also central to the probabilistic Rasch model for measurement, which has applications in psychological and educational assessment, among other areas. Though the modeling assumption of Gaussian distributed beliefs is widespread, justifications for the assumption are difficult to find. It may be that this has simply grown to be a research design convention, or is just convenient. In particular in LISREL, AMOS, PLS path analysis it is unlikely that Central Limit Theorem convergence is directly applicable given the structure of these models; nor is it clear what sample sizes would be necessary to assure this convergence.

Likert data summarizing gaussian beliefs

With a choice of a Fisher Information metric, we can explore the implications of the widespread assumption that survey subject beliefs or other phenomena are Gaussian distributed, but are elicited, measured and analyzed as Likert scaled data. Consider further the purpose of a Likert scale – to allow respondents to express both the direction and strength of their opinion about a topic (Likert 1974; Jöreskog and Sörbom 1982). Thus a Likert item is a statement which the respondent is asked to evaluate according to any kind of subjective or objective criteria; generally the level of agreement or disagreement is measured.

Likert scales are metrics on Likert items – i.e., mathematical distance functions which define a distance between elements of a set; generally the level of agreement or disagreement is measured. Survey researchers often impose various regularity conditions on the metrics implied in the construction of their survey instruments to eliminate biases in observations, and help assure that there is a proper matching of survey results and the analysis (Roberts et al. 2001; McArdle and Epstein 1987; Reips and Funke 2008).

A Likert item in practice is generally considered symmetric or balanced when observations contain equal amounts of positive and negative positions. The 'distance' between each successive Likert item is traditionally assumed to be equal – i.e., the psychometric distance between 1 and 2 is equidistant to 2 to 3. In terms of good research ethics, an equidistant presentation by the researcher is important; otherwise it will introduce a research bias into the analysis. A good Likert scale will present a symmetry of Likert items about a middle category that have clearly defined linguistic qualifiers for each item. In such symmetric scaling, equidistant attributes will typically be more clearly observed or, at least, inferred. It is when a Likert scale is symmetric and equidistant that it will behave like an interval-level measurement (Akaike 1974) showed that interval-level measurement better achieved by a visual analogue scale. Another perspective applies a polytomous Rasch model to infer that the Likert items are interval level estimates on a continuum, and thus that statements reflect increasing levels of an attitude or trait – e.g., as might be used in grading in educational assessment, and scoring of performances by judges.

Any approximation suffers from information loss; specifying the magnitude and nature of that loss, though, can be challenging. Fortunately, information measures of sample adequacy have a long history. These were perhaps best articulated in the 'information criterion' published in (Ludden et al. 1994) using information entropy. The Akaike information criterion (AIC) measures the information lost when a given model is used to describe population characteristics. It describes the tradeoff between bias and variance (accuracy and complexity) of a model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value (minimum information loss); it rewards goodness of fit, while penalizing an increasing number of estimated parameters. The Schwarz criterion (Pauler 1998; Dhrymes 1974) is closely related to AIC, and is sometimes called the Bayesian information criterion.

Ideally, responses to survey questions should yield discrete measurements that are dispersed and balanced – this maximizes the information contained in responses. Researchers would like respondents to make a definite choice rather than choose neutral or intermediate positions on a scale. Unfortunately, cultural, presentation and subject matter idiosyncrasies can effectively sabotage this objective (Dietz et al. 2007; Dhrymes et al. 1972; Lee et al. 2002). Cox (1980) to be more closely compressed around the central point than Western responses; superficially, this suggests that Asian surveys may actually yield less information (dispersion) than Western surveys. To improve responses, some researchers suggest that a Likert scale without a mid-point would be preferable, provided it does not affect the validity or reliability of the responses Devasagayam (1999) Friedman et al. (1981) Friedman and Amoo (1999); Matell and Jacoby (1972) Komorita and Graham (1965); Komorita (1963); Wildt and Mazis (1978) Chan (1991) have all demonstrated that as the number of scale steps is increased, respondents' use of the mid-point category decreases. Additionally, (Roberts et al. 2001; McArdle and Epstein 1987; Reips and Funke 2008) (Dawes (2012) (Dawes et al. 2002; Sparks et al. 2006) (Friedman and Amoo 1999; Allen and Seaman 2007) have found that grammatically balanced Likert scales are often unbalanced in interpretation; for instance, 'tend to disagree' is not directly opposite 'tend to agree'. Worcester and

Burns also concluded that a four point scale without a mid-point appears to push more respondents towards the positive end of the scale. The previously cited research concludes that Likert scales are subject to distortion from at least three causes. Subjects may:

1. Avoid using extreme response categories (central tendency bias);
2. Agree with statements as presented (acquiescence bias); or
3. Try to portray themselves or their organization in a more favorable light (social desirability bias).

Designing a balanced Likert scale (with an equal number of positive and negative statements) can obviate the problem of acquiescence bias, since acquiescence on positively keyed items will balance acquiescence on negatively keyed items, but there are no widely accepted solutions to central tendency and social desirability biases. Likert items are considered symmetric or 'balanced' where there are equal amounts of positive and negative positions.

The number of possible responses may matter as well. Likert used five ordered response levels, but seven and even nine levels are common as well. Allen and Seaman (2007) concluded that a five or seven point scale may produce slightly higher mean scores relative to the highest possible attainable score, compared to those produced from a ten point scale, and concluded that this difference was statistically significant. In terms of the other data characteristics, there was very little difference among the scale formats in terms of variation about the mean, skewness or kurtosis.

From another perspective, a Likert scale can be considered as a grouped form of a continuous scale. This is important in path analysis, since you implicitly treat the variable as if it were continuous for correlational analysis. Likert scales are clearly ordered category scales, as required for correlational work, but the debate among methodologists is whether they can be treated as equal interval scales.

When a Likert scale approximates an interval-level measurement, we can summarize the central tendency of responses using either the median or the mode, with 'dispersion' measured by standard deviations, quartiles or percentiles. Characteristics of the sample can be obtained from non-parametric tests such as a chi-squared test, Mann–Whitney test, Wilcoxon signed-rank test, or Kruskal–Wallis test (Jamieson 2004; Chan et al. 2000; Hill 1995).

Likert mappings may also be analyzed with respect to their resolution or granularity of measurement. Clearly a nine-point scale mapping has more resolution (or finer granularity) than a three point one. Measurement in research consists in assigning numbers to entities otherwise called concepts in compliance with a set of rules. These concepts may be 'physical,' 'psychological' and 'social'. The concept length is physical. But the question remains, 'if I report length as 6 feet in a case, what exactly does that mean? Even with physical scales, there is an implied granularity; if I say that something is 6 feet long, this implies less precision than length of 183 centimeters. In scientific pursuits, finer granularities can be pursued to almost unimaginable levels – for example, the international standard for length, adopted in 1960, is derived from the 2p10-5d5 radiation wavelength of the noble gas Krypton-86. The influence of choice of measuring stick on the results of modeling is responsible for phenomena such as Benford's Law (Mandelbrot 1982) and fractal scaling (Mandelbrot 1982, Burns and Bush 2000).

The assumption of Gaussian distribution of opinions or beliefs is common in the analysis of survey research, as mentioned previously. The assumption tends to be applied in the analysis stage, rather than in the design of the survey instrument. The question of whether underlying beliefs are continuous or discrete, distributed one way or another doesn’t tend to come up in the design of Likert scaled surveys, because there a few conventions that could use this information to improve the survey design. Nonetheless, the current research will argue that it matters in assessing the informativeness of Likert scaled data, which in turn can have a large impact on significance, power and other statistics reported from the research.

Information clearly is lost in the mapping of beliefs to a Likert scale; how much information is lost is probably unknowable in practice. But the loss in information from that that would exist if our modeling assumptions (e.g., Gaussian beliefs) were actually true can be assessed. At this point, let me more precisely define the concepts of (1) informativeness, (2) bias, and (3) dispersion in Likert representations of survey subject belief distributions, starting with graphical depictions of bias and dispersion in Figs. 3 and 4 respectively.

Figures 3 and 4, standardize Likert responses so that one standard deviation of the actual distribution of beliefs will shift the Likert score one point higher – this is comparable to the process of keeping the survey instrument 'on scale' measuring beliefs in similar units to the subjects normal conventions. In addition, the mode of subject beliefs (i.e., what the largest number of people believe or agree upon) is presumed to center somewhere in the range 2 through 4 of the 5-point scale, with all other values being the 'extremes' – response '1' or response '5'. This is more or less what survey researchers aspire to, where the level of agreement or disagreement is measured (i.e. is 'on scale') and the scaling is considered symmetric or 'balanced' because there are equal amounts of positive and negative positions (Kühberger 1995; Kühberger 1998; Kühberger et al. 1999; Kühberger et al. 2002; Lydtin et al. 1975; Jöreskog 1971a). Most of the weight of the Gaussian belief distribution should lie within the Likert range 2 through 4 of the 5-point scale. Survey researchers can credibly move the range around, but probably should not try to alter the subject beliefs if they are trying to conduct an objective survey.

Weaknesses in data can be effectively addressed by increasing the sample size. This works for multicolinear data, non-Gaussian data, and for Likert data as well. But since data collection is costly, it is desirable to increase sample size as little as possible. The path analysis literature is surprisingly vague on how much of an increase is needed. (Jöreskog 1971b; Joreskog 1970) suggest increases of two orders of magnitude, but without offering causes or mitigating factors. If we assume that survey costs increase commensurately with sample size, then for most projects two orders of magnitude is likely to be prohibitive.

For example, in the path analysis approaches implemented in LISREL and AMOS software, for reasonably large samples, when the number of Likert categories is 4 or higher and skew and kurtosis are within normal limits, use of maximum likelihood is justified. In other cases some researchers use weighted least squares based on polychoric correlation. (Jöreskog 1971b; Joreskog 1970; Jöreskog 1970; Jöreskog 1969; Jöreskog 1993; Westland 2010) in Monte Carlo simulation, found phi, Spearman rank correlation, and Kendall tau-b correlation performed poorly whereas tetrachoric correlation with ordinal data such as Likert scaled data was robust and yielded better fit.

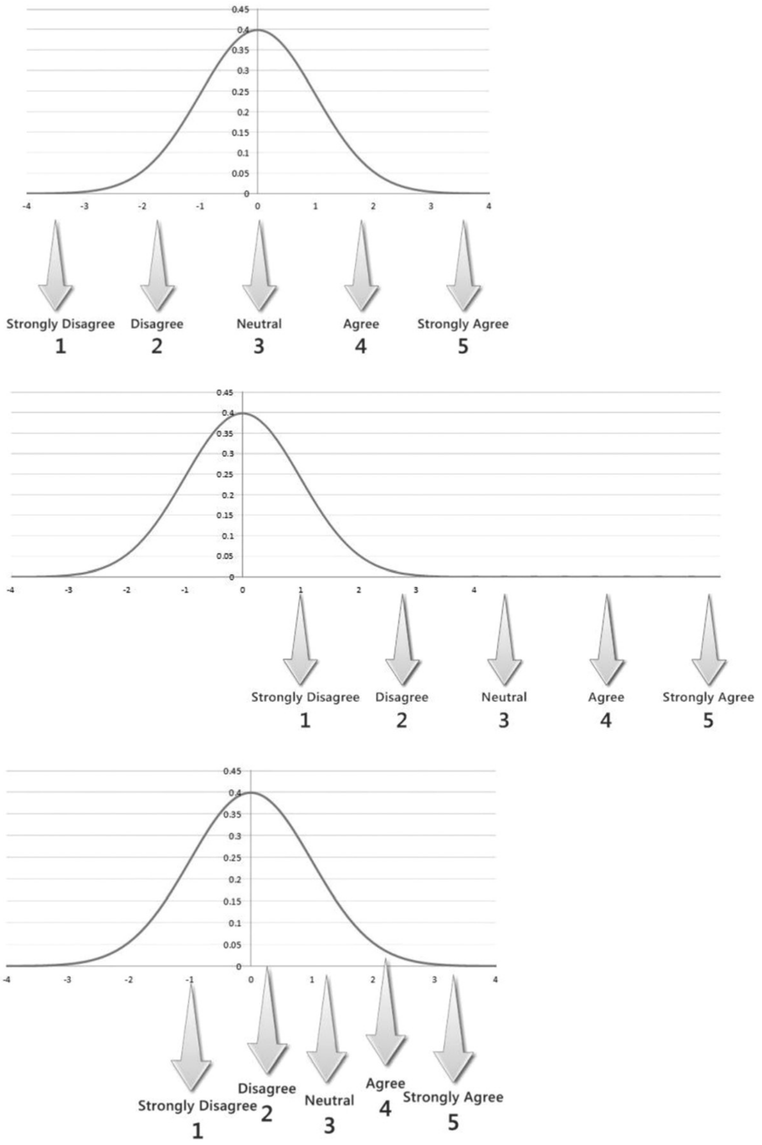

Fig. 3 Dispersion in balanced, unbalanced and mis-scaled Likert mappings

Fisher Information (denoted here as $\mathbb{I}_{sample}(parameter)$ is additive; the information yielded by two independent samples is the sum of the separate sample's information: $\mathbb{I}_{x,y}(\theta) = \mathbb{I}_y(\theta) + \mathbb{I}_x(\theta)$. Furthermore, the information in n independent sample observations is n times that in a single observation $\mathbb{I}_n(\theta) = n\mathbb{I}(\theta)$. Assume a survey collects n independent $k-point$ Likert-scale observations for each of the survey questions. Let the Likert scale represent a polytomous Rasch model, with say 5, 7 or 9 divisions (Alphen et al. 2008; Bond and Fox 2007; Fitzpatrick et al. 2004; White and Velozo 2002; Likert 1974). We can take the perspective of a polytomous Rasch model, assuming that the responses to the survey map an underlying Gaussian $\mathbb{N}(\mu, \sigma^2)$ belief distribution to a Likert item across the population of subjects surveyed for a particular question on the survey. A mapping of a Gaussian $\mathbb{N}(\mu, \sigma^2)$ belief distribution to the $k-point$ Likert-scale mapping

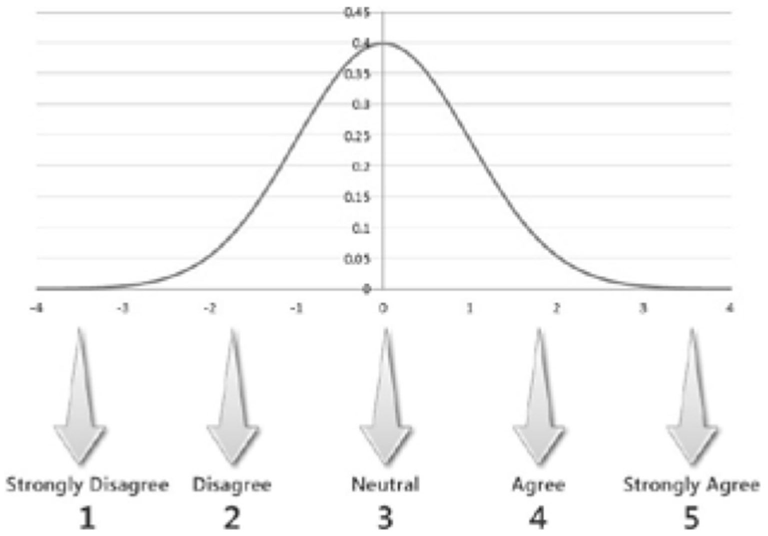

Fig. 4 Continuously varying strength of belief is approximated with 5 levels of Likert scale belief

imposed by the survey instrument might be visualized in terms of one of the four graphs in Fig. 5.

In Fig. 5, survey responses are assumed yield an equidistant scaling of Likert items so that one standard deviation of the actual distribution of beliefs will shift the Likert score one point higher. In addition, the mean value of the mean of beliefs is presumed to center somewhere in the range 2 through 6 of the 7-point scale, with all other values being the 'extremes' – response '1' or response '7'. This is more or less what survey researchers aspire to, where the level of agreement or disagreement is measured (i.e. is 'on scale') and the scaling is considered symmetric or 'balanced' because there are equal amounts of positive and negative positions (e.g., see (Kühberger 1995; Kühberger 1998; Kühberger et al. 1999; Kühberger et al. 2002; Jöreskog 1971a)). Most of the weight of the Gaussian belief distribution should lie within the Likert support (we presumably can move the Likert support around, but we probably should not try to alter the subject beliefs if we are running an objective survey).

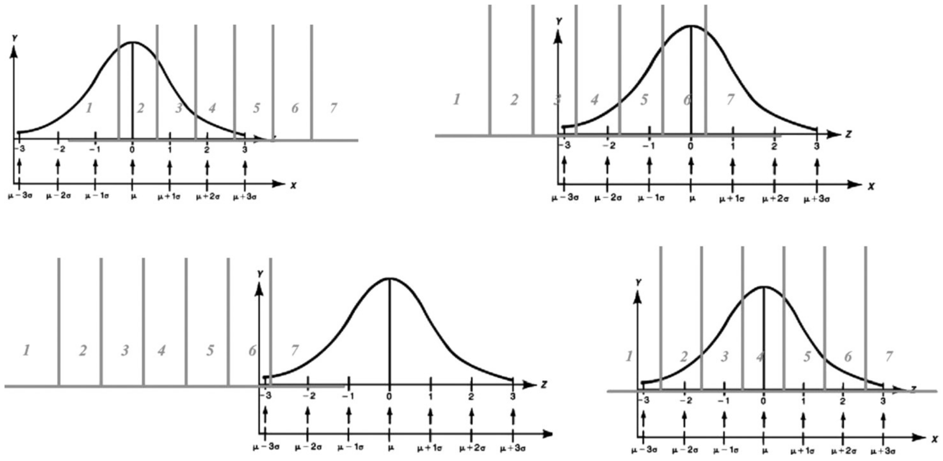

Fig. 5 Mapping Subject Beliefs to a Likert Scale

Let $F(\mu, \sigma)$ *and* $f(\mu, \sigma)$ be cdf and pdf respectively of the underlying belief distribution. Presume we use a metric scale that sets $\sigma^2 = 1$ (or alternately that the Likert 'bin' partitions are spaced σ units apart). Let the Likert 'bins' of the multinomial response distribution be the set $\left\{(x_1 \in (-\infty, 1]), \{x_i \in (i-1, i]\}_{i=2}^{k-1},\ (x_k \in (1, \infty))\right\}$ where k is the total number of bins (usually 5, 7 or 9). Then the parameters $\{p_i\}$ of the multinomial distribution of the 'bin' summing of Likert-items will be the set $\{p_1 = F(1|\mu), \{p_i = F(i|\mu) - F(i-1|\mu)\}_{i=2}^{k-1},\ p_k = 1 - F(k-1|\mu)\}$.

A particular bin i is *filled* with probability of p_i and *not filled* with probability $1 - p_i$; let n independent survey questions result in that bin being filled θ_i times, and not filled $n - \theta_i$ times. If B_i is a logical variable that indicates whether the i^{th} bin of the Likert item was chosen, then all possible outcomes for the Likert item can be represented $B_1 \vee B_2 \vee \cdots \vee B_{k-1} = \vee_{i=1}^{k-1} B_i$ since if none of the first $k - 1$ bin must have been chosen. Let the Fisher information in the i^{th} bin of a sample of n Likert items be $\mathbb{I}_{B_i}$. Since B_i is a logical variable, it can be perceived as a Bernoulli trial – a random variable with two possible outcomes, "success" with probability of p_i and "failure", with probability of $1 - p_i$. The Fisher information contained in a sample of n independent Bernoulli trials for B_i where there are m successes, and where there are $n - m$ failures is:

$$\begin{aligned}\mathbb{I}_{B_i}(p_i) &= -E_{p_i}\left[\frac{\partial^2}{\partial {p_i}^2}\ ln(f(m; p_i))\right] = -E_{p_i}\left[\frac{\partial^2}{\partial {p_i}^2}\ ln\left({p_i}^m (1-p_i)^{n-m} \frac{(m + (n-m))!}{m!(n-m)!}\right)\right] = \\ &= -E_{p_i}\left[\frac{\partial^2}{\partial {p_i}^2}\ (\mathrm{m}\ ln(p_i) + (\mathrm{n-m})\ ln(1-p_i))\right] = -E_{p_i}\left[\frac{\partial}{\partial p_i}\left(\left(\frac{m}{p_i} + \frac{(n-m)}{1-p_i}\right)\right)\right] = \\ &= -E_{p_i}\left[\left(\frac{m}{{p_i}^2} + \frac{(n-m)}{(1-p_i)^2}\right)\right] = \left(\frac{np_i}{{p_i}^2} + \frac{n(1-p_i)}{(1-p_i)^2}\right) = \frac{n}{p_i(1-p_i)}\end{aligned}$$

This is the reciprocal of the variance of the mean number of successes in n Bernoulli trials, as expected. The Fisher information contained in a sample of n independent Bernoulli trials for all possible outcomes for n Likert items $\vee_{i=1}^{k-1} B_i$ is:

$$\mathbb{I}_{\vee_{i=1}^{k-1} B_i} = \sum_{i=1}^{k-1}\left(\frac{n}{p_i(1-p_i)}\right)$$

Compare this to the Fisher Information in a sample of n observations from a Gaussian $\mathbb{N}(\mu, \sigma^2)$ belief distribution, which is estimated $\hat{\mathbb{I}}_n = \frac{n}{\sigma^2}$ (and which incidentally is independent of location parameter μ *as*: $\hat{\mathbb{I}}_n$ is the inverse of the variance). Then estimator $\hat{\omega}$ can be computed from the ratio of information content in these two different mappings from the same survey sample:

$$\omega \triangleq \frac{\frac{n}{\sigma^2}}{\sum\limits_{i=1}^{k-1}\left(\frac{n}{p_i(1-p_i)}\right)} = \frac{1}{\sigma^2 \sum_{i=1}^{k-1}\left(\frac{1}{p_i(1-p_i)}\right)}$$

Thus the lower bounds on a sample that uses a Likert mapping will need to be $\hat{\omega}$ times as large as one that assumes a full Gaussian belief distribution. Figure 6 shows how a particular Likert scale mapping of what is an inherently continuous

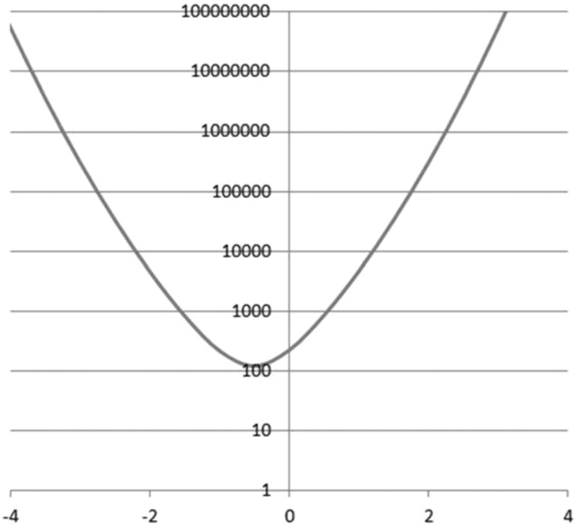

Fig. 6 Unbalanced Likert mappings cause sample size increases by the multiple shown on the y-axis when the central category on the Likert mapping is biased by the number of standard deviations shown on the x-axis

distribution of beliefs in the population results in a significant increase in the sample size needed for estimation – by a factor of at least two orders of magnitude (i.e., 100 times). Figure 7 shows how mis-scaling varies with sample size.

There are three things that should be noted concerning multipliers for sample size estimates for processing Likert data when an assumption of Gaussian data has been made in the data analysis:

> First, any difference of the actual sample standard deviation from the equidistant scale of the Likert items requires larger sample sizes; but the minimum sample size for any Likert mapped data set will be at least 100 times as large as that that would be required if you had all of the original information in the Gaussian distribution of beliefs. The information loss from using Likert scaling is at least two orders of magnitude.
>
> Second, the sample is most informative when location of the Gaussian mean coincides with the central Likert bin. This emphasizes the importance of 'balanced' designs for the Likert scaling in the survey instrument.

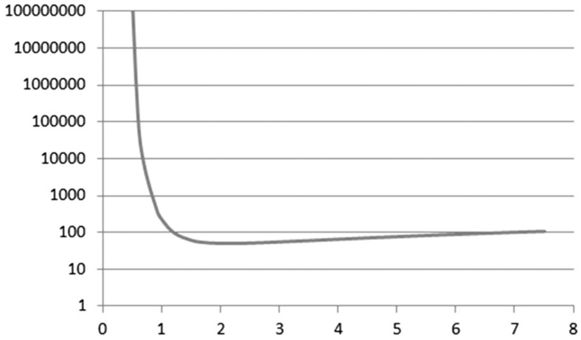

Fig. 7 Mis-scaled Likert mappings cause sample size increases by the multiple shown on the y-axis when the standard deviation of the actual belief distribution is mis-scaled by number of Likert scale intervals shown on the x-axis

Third, information in the underlying belief distribution, which has a support, does not depend on the mean of an assumed underlying Gaussian distribution of data. The Likert mapping information content does depend on the mean and is sensitive to the Likert scale being 'balanced' – this is controlled in the survey design.

Examples

In order to gain a more intuitive understanding of how the metric in this paper functions in comparison with the (Srinivasan and Basu 1989) metric, we can operationalize the Likert mapping as a survey instrument $T \otimes S$ that 'bins' Y values (i.e., the measure of unobservable underlying phenomenon X) into responses Z on a 5-point scale; $T : X \longrightarrow (Y + \tilde{\theta})$ *and* $S : Y \longrightarrow (Z + \tilde{\delta})$. Random variable $\tilde{\varepsilon}$ describes the error (information loss) in the mapping of survey instrument $T \otimes S : X \longrightarrow (Z + \tilde{\varepsilon})$. Conceptually, $\tilde{\varepsilon} = \tilde{\delta} \ + \tilde{\theta}$ where $\tilde{\theta}$ is the part resulting from misspecification of the survey instrument (bias and dispersion) and $\tilde{\delta}$ is the part resulting from approximating a continuous variable into the five bins in the Likert scale. The seven sets of responses (including a restatement of Anscombe's (1973) Quartet) in Fig. 8 and Table 1, encapsulate several challenges – skewness, kurtosis, outliers, non-informative data, and a non-linear (parabolic) data.

(Srinivasan and Basu 1989) evaluated the information content of Likert data item Z (an m-point Likert scale variable) that approximates some continuous variable $\tilde{Y}$ that in turn approximates some unobservable belief or phenomenon that the researcher wishes to measure. They assume that $\tilde{Y}$ is composed of a true 'score' $\tilde{X}$ and error $\tilde{\varepsilon}$ (which in their formulation is additive, but which we will allow to take on more complex functional forms). Then in their formulation.

$\tilde{Y} = \tilde{X} + \tilde{\varepsilon}$ where $\tilde{X} \sim N(0, 1)$ and $\tilde{\varepsilon} \sim N(0, \theta^2)$ and $\rho(\tilde{X}, \tilde{\varepsilon}) = 0$

Thus $\tilde{Y} \sim N(0, 1 + \theta^2)$ and using the fact that the Pearson correlation coefficient is invariant (up to a sign) to separate changes in location and scale in the two variables, we can recompute the (Srinivasan and Basu 1989) information metric $I_Z = \frac{\rho^2(Z,X)}{\rho^2(Y,X)}$. Since the correlation $\rho(\tilde{X}, \tilde{Y}) = \rho(\tilde{X} + \tilde{\varepsilon}, \tilde{X}) = \rho(\tilde{X}, \tilde{X}) + \rho(\tilde{\varepsilon}, \tilde{X}) = 1 + 0 = 1$ this metric is identically $I_Z \equiv \rho^2(\tilde{Z}, \tilde{X})$.

A more general formulation allows $\rho(\tilde{X}, \tilde{Y}) \in [-1, 1]$.. Figure 8 (based on data in Table 1) show five possible outcomes for $\tilde{X}, \tilde{Y}$ *and* $\tilde{Z}$. In case A, F and G $\rho(\tilde{X}, \tilde{Y}) = 1$ as assumed in Srinivasan and Basu's formulation; in cases B, C, D and E $\rho(\tilde{X}, \tilde{Y}) = 0.666$.

Table 2 shows demonstrates two weaknesses of the (Srinivasan and Basu 1989) metric I_Z

1. the value often converges outside the purported [0,1] range of the statistic (as in C,D, F and G) and
2. even small changes in survey setup, or of question interpretation by subjects can have a huge impact on reported information content.

The Fisher Information statistic does not have a value when $R^2 = 1$, but otherwise converges to values that are intuitive in the sense that they suggest that information captured from subjects is fairly stable.

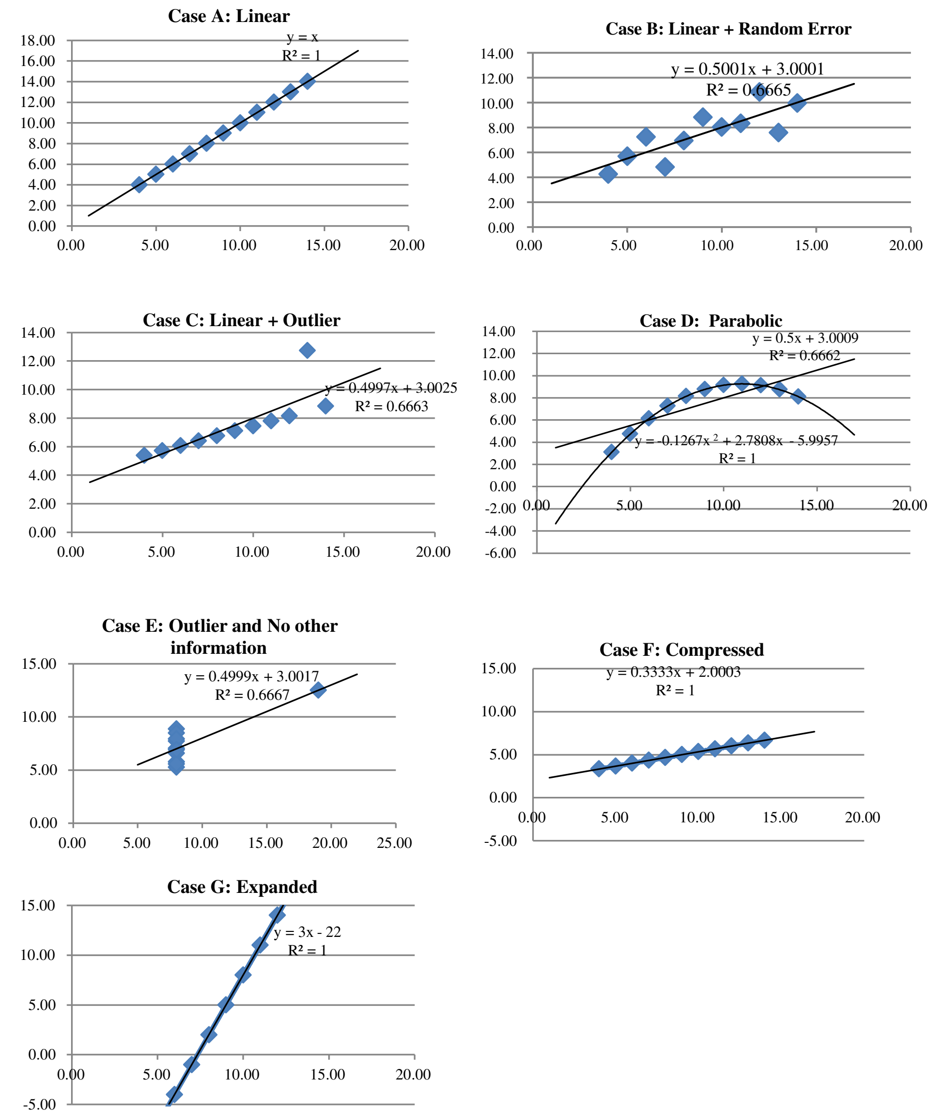

Fig. 8 $\tilde{X} \times \tilde{Y}$: Four Datasets with $R^2 = 0.666$ and Three Datasets with $\boldsymbol{R}^2 = 1$

Conclusion and discussion

This paper contributes to the understanding of design trade-offs and requirements which are required to be able to assert that a particular Likert sample contains a specific amount of information concerning the specific research question that the survey is designed to answer. The research developed a Fisher Information metric and compared this to an earlier correlation based statistic. Illustrative examples using Gaussian beliefs showed that the Fisher Information metric is more informative, stable, and reliable than earlier approaches. They also accentuate the importance of balanced survey design, potentially without a midpoint, as suggested by (Devasagayam 1999) (Friedman et al. 1981) (Friedman and Amoo 1999; Matell and Jacoby 1972) (Komorita and Graham 1965; Komorita 1963; Wildt and Mazis 1978) (Chan 1991). It also suggests that where grammatically balanced Likert scales are unbalanced in interpretation, the impact on survey conclusions may be significant (Roberts et al. 2001; McArdle and Epstein 1987;

Table 1 $\tilde{X} \times \tilde{Y}$: 4 Datasets with $R^2 = 0.666$ and 4 Datasets with $R^2 = 1$ binned into Likert variable $\tilde{Z}$

	Case A									
		Case B			Case C			Case D		
Observation	Bin 1,2,3x			Bin 1y			Bin 2y			Bin 3y
		1x	1 y	1z	2x	2y	2z	3x	3y	3z
1	10.00	10.00	8.04	10.00	10.00	9.14	10.00	10.00	7.46	8.00
2	8.00	8.00	6.95	8.00	8.00	8.14	10.00	8.00	6.77	8.00
3	10.00	13.00	7.58	8.00	13.00	8.74	10.00	13.00	12.74	10.00
4	10.00	9.00	8.81	10.00	9.00	8.77	10.00	9.00	7.11	8.00
5	10.00	11.00	8.33	10.00	11.00	9.26	10.00	11.00	7.81	8.00
6	10.00	14.00	9.96	10.00	14.00	8.10	10.00	14.00	8.84	10.00
7	6.00	6.00	7.24	8.00	6.00	6.13	8.00	6.00	6.08	8.00
8	4.00	4.00	4.26	6.00	4.00	3.10	4.00	4.00	5.39	6.00
9	10.00	12.00	10.84	10.00	12.00	9.13	10.00	12.00	8.15	10.00
10	8.00	7.00	4.82	6.00	7.00	7.26	8.00	7.00	6.42	8.00
11	6.00	5.00	5.68	6.00	5.00	4.74	6.00	5.00	5.73	6.00
mean	8.36	9.00	7.50	8.36	9.00	7.50	8.73	9.00	7.50	8.18
std dev	2.15744	3.316625	2.031568	1.747726	3.316625	2.031657	2.053821	3.316625	2.030424	1.401298
skewness	-1.01393	-8.1E-17	-0.06504	-0.40869	-8.1E-17	-1.3158	-1.58382	-8.1E-17	1.855495	-0.12334
kurtosis	-0.20589	-1.2	-0.5349	-1.62132	-1.2	0.846123	1.743956	-1.2	4.384089	-0.45267
	Case E			Case F			Case G			
Observation	4x	4y	4z	5x	5y	5z	6x		6y	6z
1	8.00	6.58	8.00	10.00	8.00	8.00	10.00		5.33	6.00
2	8.00	5.76	6.00	8.00	2.00	2.00	8.00		4.67	6.00
3	8.00	7.71	8.00	13.00	17.00	10.00	13.00		6.33	8.00
4	8.00	8.84	10.00	9.00	5.00	6.00	9.00		5.00	6.00
5	8.00	8.47	10.00	11.00	11.00	10.00	11.00		5.67	6.00
6	8.00	7.04	8.00	14.00	20.00	10.00	14.00		6.67	8.00
7	8.00	5.25	6.00	6.00	-4.00	2.00	6.00		4.00	6.00
8	19.00	12.50	10.00	4.00	-10.00	2.00	4.00		3.33	4.00
9	8.00	5.56	6.00	12.00	14.00	10.00	12.00		6.00	6.00
10	8.00	7.91	8.00	7.00	-1.00	2.00	7.00		4.33	6.00
11	8.00	6.89	8.00	5.00	-7.00	2.00	5.00		3.67	4.00
mean	9.00	7.50	8.00	9.00	5.00	5.82	9.00		5.00	6.00
std dev	3.316625	2.030579	1.549193	3.316625	9.949874	3.842348	3.316625		1.105431	1.264911
skewness	3.316625	1.506818	0	-8.1E-17	-8.1E-17	0.052992	-8.1E-17		2.96E-15	0
kurtosis	11	3.151315	-1.11111	-1.2	-1.2	-2.22488	-1.2		-1.2	0.416667

Reips and Funke 2008) (Dawes (2012) (Dawes et al. 2002; Sparks et al. 2006) (Friedman and Amoo 1999; Allen and Seaman 2007). The research found that any difference of the actual sample standard deviation from the equidistant scale of the Likert items requires larger sample sizes; but the minimum sample size for any Likert mapped data set will be at least 100 times as large as that that would be required if you had all of the original information in the Gaussian distribution of beliefs. The information loss from using Likert scaling is at least two orders of magnitude. Additionally, the sample is most informative when location of the Gaussian mean coincides with the central Likert bin. This emphasizes the importance of 'balanced' designs for the Likert scaling in the survey instrument. Finally, information in the underlying belief distribution does depend on the mean and is sensitive to the Likert scale being 'balanced'.

Additionally, the research example identified a practical issue in the implementation of the (Srinivasan and Basu 1989) metric I_Z, in that its value often converges outside the purported [0,1] range of the statistic (as in examples C,D, F and G) and that even

Table 2 Information content of seven cases

	A	B	C	D	E	F	G
$I_Z = \frac{\rho^2(Z,X)}{\rho^2(Y,X)}$	0.894427	0.864447	1.014439	1.22563	0.274983	?	?
$\omega^{\hat{}} \triangleq \frac{\frac{n}{\sigma^2}}{\sum_{i=1}^{k-1}\left(\frac{n}{p_i(1-p_i)}\right)} = \frac{1}{\sigma^2\sum_{i=1}^{k-1}\left(\frac{1}{p_i(1-p_i)}\right)}$	0.144737	0.141026	0.166667	0.152778	0.141026	?	?
Info per observation $= \sum_{i=1}^{5-1}\left(\frac{1}{p_i(1-p_i)}\right)$	0.628099	0.644628	0.545455	0.595041	0.644628	0.628099	0.528926
$\mathrm{FI} = \sum_{i=1}^{5-1}\left(\frac{n}{p_i(1-p_i)}\right)$	6.909089	7.090908	6.000005	6.545451	7.090908	6.909089	5.818186

small changes in survey setup, or of question interpretation by subjects can have a huge impact on information content reported by metric I_Z.

In contrast, the Fisher Information estimator $\hat{\omega}$ developed in this research only fails to compute in the limiting case where $R^2 = 1$, but otherwise converges to values that are intuitive in the sense that they suggest that information captured from subjects is fairly stable. The contrasting information in, say, American versus Asian consumer surveys can be assumed to be comparable, even though the manner of conveyance and expression of that information may vary because of culture and other factors. This reinforces conclusions rendered by (Cox 1980) concerning survey designs across cultures.

Results from exploring the Fisher Information estimator $\hat{\omega}$ developed in this research imply that sample sizes need to increase to offset the mapping losses. The lower bounds on a sample that uses a Likert mapping will need to be many times as large as one that assumes a full Gaussian belief distribution, as in (Crook and Good 1980). There are three issues that should be considered in setting the scope of a study with Likert measurements. First, the minimum sample size for any Likert mapped data set could be several orders of magnitude larger one that would be required if you had all of the original information in the Gaussian distribution of beliefs. Second, the sample is most informative when it is balanced and centered in responses. And finally information in the underlying belief distribution is independent from the mean/mode μ, as long as this is properly controlled in the survey design.

The assumption of Gaussian belief distributions may or may not be justified in practice. Studies by (Kühberger 1995; Kühberger 1998; Kühberger et al. 2002; Kühberger et al. 1999; Clarke et al. 2002) have concluded that people do not generally hold strong, stable and rational beliefs, and that their responses are very much influenced by the way in which decisions are framed. This would tend to indicate that a class of distributions besides Gaussian distributions would be most appropriate for human beliefs. Nonetheless, the Gaussian assumption is widely used, especially in survey research using tools such as AMOS or LISREL. Widespread use of this assumption, and the amount of information loss implies that sample sizes need to increase to offset the mapping loss.

An alternative approach to assessing the informativeness of Likert items, with significantly reduced demands on sample size, could invoke Bayesian conjugate families of distributions. Such an approach would essentially pool prior research findings (potentially both qualitative and quantitative) in the prior distribution, with a likelihood function built from the data. Given the categorical nature of Likert mappings, a multinomial-Dirichlet conjugate family of distributions would be appropriate for Bayesian analysis of Likert survey data. Such approaches have been explored in the artificial intelligence and quality control fields (Dietz et al. 2007; Dhrymes et al. 1972; Lee et al. 2002) and statistics developed in (Crook and Good 1980 Gupta 1969). So far, the multinomial-Dirichlet conjugate family of distributions appears not to have been applied to the analysis of survey generated Likert data.

Competing interests

The author declares that he has no competing interests.

Acknowledgements

I acknowledge that work is solely my own.

Received: 21 April 2015 Accepted: 13 July 2015
Published online: 25 August 2015

References

Akaike H (1974) A new look at the statistical model identification. Automatic Control IEEE Transactions 19(6):716–723

Allen IE, Seaman CA (2007) Likert scales and data analyses. Qual Prog 40(7):64–65

Alphen A, Halfens R, Hasman A, Imbos T (2008) Likert or Rasch? Nothing is more applicable than good theory. J Adv Nurs 20(1):196–201

Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27(1):17–21

Bennett CM, Baird AA, Miller MB, Wolford GL (2012) Journal of Serendipitous and Unexpected Results. J Serendipitous Unexpected Results 1(1):1–5

Bond T, Fox C (2007) Applying the Rasch model: Fundamental measurement in the human sciences. Lawrence Erlbaum

Brandstätter E, Kühberger A, Schneider F (2002) A cognitive-emotional account of the shape of the probability weighting function. J Behav Decis Mak 15(2):79–100

Burns A, Bush RF (2000) Marketing research. Globalization 1:7

Chan JC (1991) Response-order effects in Likert-type scales. Educ Psychol Meas 51(3):531–540

Chan JPE, Tan KHC, Tay KY (2000) Nanyang Technological University. School of Accountancy and Business. In: The impact of electronic commerce on the role of the middleman. School of Accountancy and Business Nanyang Technological University, Singapore

Clarke S, Worcester J, Dunlap G, Murray M, Bradley-Klug K (2002) Using multiple measures to evaluate positive behavior support a case example. J Posit Behav Interv 4(3):131–145

Cox EP III (1980) The optimal number of response alternatives for a scale: a review. Journal of marketing research., pp 407–422

Crook J, Good I (1980) On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part II. Ann Stat 8(6):1198–1218

Dawes J (2012) Do data characteristics change according to the number of scale points used? an experiment using 5 point, 7 point and 10 point scales

Dawes J, Riebe E, Giannopoulos A (2002) The impact of different scale anchors on responses to the verbal probability scale. Canadian J Marketing Res 20(1):77–80

Devasagayam PR (1999) The effects of randomised scales on scale checking styles and reaction time. In: 1999 Marketing Management Association Conference Proceedings

Dhrymes PJ (1974) Econometrics. Springer-Verlag, New York

Dhrymes PJ, Howrey EP, Hymans SH, Kmenta J, Leamer EE, Quandt RE, Ramsey JB, Shapiro HT, Zarnowitz V (1972) Criteria for evaluation of econometric models. In: Annals of Economic and Social Measurement, Volume 1, number 3. NBER., pp 291–325

Dietz L, Bickel S, Scheffer T (2007) Unsupervised prediction of citation influences. In Proceedings of the 24th international conference on Machine learning. ACM, pp 233–240

Fitzpatrick R, Norquist J, Jenkinson C, Reeves B, Morris R, Murray D, Gregg P (2004) A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. Qual Life Res 13(2):331–338

Friedman HH, Amoo T (1999) Rating the rating scales. J Mark Manag 9(3):114–123

Friedman HH, Wilamowsky Y, Friedman LW (1981) A comparison of balanced and unbalanced rating scales. Mid-Atlantic J Business 19(2):1–7

Gupta Y (1969) Least squares variant of the Dhrymes two-step estimation procedure of the distributed lag model. Int Econ Rev 10(1):112–113

Hill TP (1995) A statistical derivation of the significant-digit law. Stat Sci 10(4):354–363

Jamieson S (2004) Likert scales: how to (ab) use them. Med Educ 38(12):1217–1218

Jöreskog KG (1969) A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34(2):183–202

Joreskog KG (1970) A general method for estimating a linear structural equation system

Jöreskog KG (1970) A general method for analysis of covariance structures. Biometrika 57(2):239–251

Jöreskog KG (1971a) Simultaneous factor analysis in several populations. Psychometrika 36(4):409–426

Jöreskog KG (1971b) Statistical analysis of sets of congeneric tests. Psychometrika 36(2):109–133

Jöreskog KG (1993) Testing structural equation models. Sage Focus Editions 154:294–294

Jöreskog KG, Sörbom D (1982) Recent developments in structural equation modeling. J Mark Res 19:404–416

Komorita SS (1963) Attitude content, intensity, and the neutral point on a Likert scale. J Soc Psychol 61(2):327–334

Komorita SS, Graham WK (1965) Number of scale points and the reliability of scales. Educ Psychol Meas 25(4):987–995

Kühberger A (1995) The framing of decisions: a new look at old problems. Organ Behav Hum Decis Process 62(2):230–240
Kühberger A (1998) The influence of framing on risky decisions: a meta-analysis. Organ Behav Hum Decis Process 75(1):23–55
Kühberger A, Schulte-Mecklenbeck M, Perner J (1999) The effects of framing, reflection, probability, and payoff on risk preference in choice tasks. Organ Behav Hum Decis Process 78(3):204–231
Kühberger A, Schulte-Mecklenbeck M, Perner J (2002) Framing decisions: hypothetical and real. Organ Behav Hum Decis Process 89(2):1162–1175
Lee JW, Jones PS, Mineyama Y, Zhang XE (2002) Cultural differences in responses to a Likert scale. Res Nursing Health 25(4):295–306
Likert R (1974) The method of constructing an attitude scale. Scaling: a sourcebook for behavioral scientists., pp 233–243
Ludden TM, Beal SL, Sheiner LB (1994) Comparison of the Akaike Information Criterion, the Schwarz criterion and the F test as guides to model selection. J Pharmacokinet Pharmacodyn 22(5):431–445
Lydtin H, Lohmöller G, Lohmöller R, Schmitz H, Walter I (1975) Hemodynamic studies on Adalat in healthy volunteers and in patients. In: International Adalat® Symposium, 2nd edn. Springer, New York, pp 112–123
Mandelbrot BB (1982) The fractal geometry of nature. Times Books
Matell MS, Jacoby J (1972) Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. J Appl Psychol 56(6):506
McArdle JJ, Epstein D (1987) Latent growth curves within developmental structural equation models. Child Dev 58(1):110–133
Miller MH (1999) The history of finance. J Portfolio Manage 25(4):95–101
Miller MH (2000) The history of finance: an eyewitness account. Journal Applied Corporate Finance 13(2):8–14
Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ 15(5):625–632
Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C (2004) Comparing alternative Rasch-based methods vs raw scores in measuring change in health. Medical care 42(1):I
Pauler DK (1998) The Schwarz criterion and related methods for normal linear models. Biometrika 85(1):13–27
Reips UD, Funke F (2008) Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. Behav Res Methods 40(3):699–704
Roberts SB, Bonnici DM, Mackinnon AJ, Worcester MC (2001) Psychometric evaluation of the Hospital Anxiety and Depression Scale (HADS) among female cardiac patients. Br J Health Psychol 6(4):373–383
Sparks R, Desai N, Thirumurthy P, Kistenberg C, Krishnamurthy S (2006) Measuring e-Commerce Satisfaction: Reward Error and the Emergence of Micro-Surveys. In: IADIS International e-Commerce Conference Proceedings
Srinivasan V, Basu AK (1989) The metric quality of ordered categorical data. Mark Sci 8(3):205–230
Stevens SS, Galanter EH (1957) Ratio scales and category scales for a dozen perceptual continua. J Exp Psychol 54(6):377
Westland JC (2010) Lower bounds on sample size in structural equation modeling. Electron Commer Res Appl 9(6):476–487
White LJ, Velozo CA (2002) The use of Rasch measurement to improve the Oswestry classification scheme. Arch Phys Med Rehabil 83(6):822–831
Wildt AR, Mazis MB (1978) Determinants of scale response: label versus position. J Mark Res 15:261–267