Quantitative Marketing and Economics
https://doi.org/10.1007/s11129-019-09209-3

A study into mechanisms of attitudinal scale conversion: A randomized stochastic ordering approach

Zvi Gilula[1] · Robert E. McCulloch[2] · Yaacov Ritov[3] · Oleg Urminsky[4]

Received: 16 July 2018 / Accepted: 2 January 2019/Published online: 25 January 2019

Abstract
This paper considers the methodological challenge of how to convert categorical attitudinal scores (like satisfaction) measured on one scale to a categorical attitudinal score measured on another scale with a different range. This is becoming a growing issue in marketing consulting and the common available solutions seem too few and too superficial. A new methodology for scale conversion is proposed, and tested in a comprehensive study. This methodology is shown to be both relevant and optimal in fundamental aspects. The new methodology is based on a novel algorithm named *minimum conditional entropy*, that uses the marginal distributions of the responses on each of the two scales to produce a unique joint bivariate distribution. In this joint distribution, the conditional distributions follow a stochastic order that is monotone in the categories and has the relevant optimal property of maximizing the correlation between the two underlying marginal scales. We show how such a joint distribution can be used to build a mechanism for scale conversion. We use both a frequentist and a Bayesian approach to derive mixture models for conversion mechanisms, and discuss some inferential aspects associated with the underlying models. These models can incorporate background variables of the respondents. A unique observational experiment is conducted that empirically validates the proposed modeling approach. Strong evidence of validation is obtained.

Keywords Categorical conversion · Conditional entropy · Mixture models · Ordinal attitudinal scales · Stochastic ordering

1 Introduction

Ratings scales are widely used to measure unobserved psychological constructs, such as consumers' attitudes. These scales are defined by the (usually arbitrary)

✉ Zvi Gilula
zvi.gilula@huji.ac.il

Extended author information available on the last page of the article

number of categories, and by some combination of numeric and semantic labels defining each scale value, often arranged and communicated as ordinal categories. Scale conversion is necessary when trying to compare attitudes that had been measured on scales using different numbers of ordinal categories. This is a common problem in areas where multiple providers of the same service compete with each other.

For example, in the Israeli cell phone industry (comprising 5 competing companies), quarterly satisfaction surveys are conducted. All surveys in that industry are done within the first month of each quarter: January, March, June, and September. Each company has its own satisfaction scale and, in their surveys, they are interested in the same primary aspect of overall satisfaction, along with many other aspects. If one company manages to get the distributions of overall satisfaction for a competitor, it would be interested in a mutual scale conversion between their own categories and the competitor's categories for every quarter. Another example where scale conversion is required is in car marketing, where surveys are used to measure drivers' satisfaction with their cars. As one of us experienced as a consultant to a European marketing analytics company, it is quite common for different manufacturers to use different satisfaction scales, each with their own specific number of categories, and a need to convert may arise.The following conditions typify such situations:

(a) Both distributions result from two independent samples from the same population and are carried out in close proximity in time to each other.
(b) Both distributions relate to an attitude (e.g. satisfaction) about the same issue (satisfaction from a given service, device, etc.).
(c) During the time between the two samples, it is assumed that the relevant respondents' attitudes did not change in substance and character. Hence, the two measurements of the underlying attitude are justly comparable and conversion between the corresponding scales is valid.
(d) Conversion is between non-quantifiable ordinal categories. A category in a given scale should be converted into a category in the other scale.

In practice, researchers need to make decisions about the scale format to use. In the absence of widely adopted best practices, the same construct is often measured using different formats, by different researchers and even by the same researchers at different times. If people think about underlying constructs in continuous terms and use stable heuristics to map continuous latent constructs to ordinal scale categories (e.g, as assumed in ordinal Probit models), then the decision about number of scale items may be largely irrelevant for the research conclusions.

However, a large literature suggests that the number of scale categories does have important consequences for the effectiveness of measurement, including reliability, convergent validity, discriminability and even respondent experience (Churchill Jr and Peter 1984; Preston and Colman 2000; Taylor, West and Aiken 2006; Revilla et al. 2014). In particular, having too few scale categories limits the information that can be conveyed, but having too many scale points can introduce noise and exacerbate differential item functioning (see Cox III 1980 for a review). Furthermore, the best-

performing number of scale categories depends on the context and the participants (Chang 1994).

As a result, researchers actively consider and debate how many scale categories to use. Different marketers (and their research vendors) may use different scale categories to measure the same construct. The same consumer insights team may decide to change the scale used in ongoing longitudinal research, perhaps because of concerns about the performance of the incumbent scale. When this occurs, it is typically not clear how to integrate or compare the two measurements on different scales. Such scale changes are especially problematic in longitudinal research, such as tracking customer satisfaction. This is a particular challenge when, as is often the case, there is no empirical data on the joint distribution (i.e., when measurements for both scales were not collected for the same people at the same time). Thus, scale conversion is both a practical concern and a conceptual problem for theories of psychological measurement.

As noted by Evans, Gilula, and Guttman (2012), henceforth EGG, there is a surprisingly small rigorous statistical literature on scale conversion. This may stem from the fact that there is no universally acceptable norm for measuring attitudes that are assumed to have an underlying conceptual continuum (such as satisfaction, agreement, happiness, etc.; see for instance Miller (1956), Green and Rao (1970)). Some researchers have proposed ad hoc solutions, such as re-scaling to match endpoints (Dawes 2008), creating a conversion algorithm from prior survey data (Dolnicar and Gr un 2013) or using regression models (Colman, Morris and Preston 1997). However, these papers acknowledge the substantial theoretical and practical limitations of these approaches, which require either very strong assumptions or joint data (e.g., both scales measured at the same time among the same people).

A more theoretical initial approach is provided by EGG, who introduce a method for the special case in which assessments on a scale with fewer categories (i.e. the coarser scale) are assumed to arise by collapsing adjacent categories on a scale with more categories (i.e, the finer scale). They propose an approach for quantifying the possibility that the coarser scale is obtained deterministically from the finer scale by collapsing certain adjacent categories of the finer scale and for identifying which categories of the finer scale should be collapsed to best approximate the coarser scale. The key limitation of this approach is the strong deterministic nesting assumption that needs to be made, specifically that people's use of each category on the coarse scale represents a complete and error-free combination of one or more scale points on the finer scale, with no splitting of categories.

In this paper we propose a much broader approach to scale conversion, of which scale collapsing is a special case. Our approach is motivated by the situation in which two ordinal scales, one coarser (with R categories) and one finer (with C categories) have been separately measured, and as a result only the two distinct multinomial frequency distributions are available. The key question is how to map mutually from one scale to the other scale (from coarse to fine and from fine to coarse.) Our goal is to develop and validate scientifically sound mutual conversion mechanisms that cope well with the growingly common need of scale conversion.

The most challenging aspect is that, in practice, the only information the researcher typically has is two estimated multinomial distributions coming from

two independent samples of individuals. Under these circumstances, many intuitive approaches (e.g., linear regression or ordinal-probit models) are not applicable due to the lack of a joint distribution. It might seem that any scale conversion mechanism is doomed to be superficial and speculative in this circumstance. However, we demonstrate that, based on minimal and reasonable assumptions about how people use scales, we can generate an intuitive and accurate scale conversion mechanism

In Section 2, we introduce the notion of stochastic ordering and demonstrate its possible relevance to scale conversion. We then introduce an algorithm (named MCE) that produces a unique joint distribution, maximizing the correlation between two given independent multinomial marginal distributions. Based on this unique distribution, we build and test a mutual conversion algorithm and discuss its merits and limitations.

To test and validate our methodology, we report a unique study (in Section 3) in which four samples of individuals responded to one of four versions of the survey jointly testing two different satisfaction scales with different numbers of categories. While the benefit of our approach is that it can be used in the absence of joint data, we collect joint data to conduct a validation test of the accuracy of the MCE algorithm. In Section 4, we investigate the performance of the MCE algorithm on our data, and we find that it translates very well from one scale to the other.

In Section 5 we propose a detailed Bayesian approach for developing parsimonious mixture models for mutual scale conversion. To account for the existence of error-prone responses, the mixture is a convex combination between the scale conversion mechanism and pure error (e.g., total independence between the two underlying scales). This approach allows for an individual's background variables to help identify the mixture weight.

In Section 6, we report in detail a re-analyses of the study, using the Bayesian approach. We find very promising results, with a strong fit between predicted and actual scale conversion, leading to some recommendations regarding the conversion mechanism. Summary and recommendations for practical scale conversions are discussed in the concluding section, Section 7.

Lastly, a sound methodology for creating confidence intervals around the MCE conversion mechanism is reported in the Appendix.

2 Minimum conditional entropy conversion algorithm based on stochastic ordering

In this section we derive a conversion algorithm that has important optimal properties. This algorithm is the core of our recommended conversion mechanisms. The basic idea behind the underlying algorithm is that of stochastic ordering. The motivation for our approach is to develop a conversion mechanism that takes into account the heterogeneity of human judgment on attitude, widely recognized by experimental psychologists. This heterogeneity can presumably be observed under the following conditions (leading later to our empirical experiment).

Imagine a sample of individuals asked to express their attitude twice - first on a 3-category scale and a few days later on a 5-category scale (without being notified

in advance of the second survey). Assume that the underlying sampled individuals understand the survey and take it responsibly. Assume also that the time gap between surveys is short enough not to allow for an actual change in the underlying attitudes being measured.

Imagine that we want to convert the categorical responses on the 3-category scale (the initial scale) into categorical responses on the 5-category scale (the target scale).

In actual data, the translation from one scale is likely to involve heterogeneity. It stands to reason that not all (sensible) individuals choosing 1 on the initial 3-category scale will choose 1 on the target 5-category scale. Some will indeed choose 1 on the second scale, some will choose 2, fewer may choose 3, and even fewer (or none) will choose 4 or 5. A similar argument seems reasonable for all other categories of the 1-to-3 scale. Individuals choosing a certain category on the initial scale are likely to choose **neighboring categories** clustered around a corresponding level of the target scale.

Hence, it is expected that each category of the initial scale is associated with a distribution on the categories of the target scale. These distributions for higher-numbered responses on the initial scale are right-shifted on the target scale with respect to the distributions for the lower-numbered responses on the initial scale. The higher the category of the initial scale, the more right-shifted is its corresponding distribution on the categories of the target scale.

The pattern just described suggests the following two properties (that are shown to be strongly supported in the empirical data we present later):

(a) The two measurements provided by the individuals are likely to be strongly correlated.
(b) The joint frequency distribution of the joint scaling will induce *stochastic ordering* of the distributions of the target scale responses conditional on the categories of the initial scale, as is discussed next.

Definition Let two random variables X and Y have respective CDFs (with the same support) denoted by $F_x(\cdot)$ and $F_y(\cdot)$. X is said to be *stochastically larger* than Y if

$$F_x(t) \leq F_y(t), \text{ for all } t.$$

As argued earlier on the initial and target scales, the right-shifted distributions of target scale category responses given the initial scale categories will obey the above-given definition of stochastic order. Such stochastic order is directly monotone in the categories of the initial scale. Hence, we can expect that under the circumstances described above, the conditional distributions of the 5-category scale given the 3-category scale are stochastically ordered, from lowest to largest corresponding to the 3-categories 1, 2, and 3, respectively. In fact, if substantial violations of stochastic order were to be observed (when joint data is available), the use of any scale conversion should be reconsidered, as the data suggests actual change may have occurred.

When individuals are jointly observed on both scales, there is no need for any conversion. However, as mentioned earlier, the researcher does not commonly observe

a joint distribution of two scales pertaining to the same individuals but rather two independent multinomial samples (from the same population), each corresponding a different scale representation of the same construct under study.

Let Y_r and Y_c be two multinomial (categorical) variables with respective ranges consisting of the integers R and C so that $Y_r \in \{1, 2, \ldots, R\}$ and $Y_c \in \{1, 2, \ldots, C\}$. We can think of this as a standard two-way table representation of the joint distribution of two categorical variables, with Y_r as the row variable and Y_c as the column variable. We now propose a conversion algorithm applicable to the common reality of two independent samples. To do so, we must first address the following problem:

Given the marginal distributions of Y_r and Y_c, can one construct a joint distribution recovering the two multinomial marginals such that the conditional distributions of

$$< Yr \mid Y_c = j >, j = 1, \ldots, C \text{ and } < Y_c \mid Y_r = i >, i = 1, \ldots, R$$

are stochastically ordered in the corresponding order of the categories of the initial scale? A short inspection of the vast literature on the relevant concept of copulas (see for instance, Accioly & Chiyoshi [2004] and Elidan [2010]) reveals that there are infinitely many ways of constructing such joint distributions.

However, we define a unique way of constructing a joint distribution (from the two marginals) that induces stochastic ordering of conditional distributions. This unique joint distribution is shown to have the relevant optimal property of minimum conditional entropy. Specifically, in this unique joint distribution, the conditional distributions of Y_c given the categories of Y_r (and conversely, the conditional distributions of Y_r given the categories of Y_c), will not only be stochastically ordered, but will also be the most predictive. Namely, the probabilities $P(Y_c \mid Y_r)$ and $P(Y_r \mid Y_c)$ will produce the strongest (mutual) dependence of Y_c on Y_r and Y_r on Y_c, respectively. Put another way, these conditional probabilities will be the closest to 0 or 1, compared to other stochastically ordered joint distributions.

Hence, if the distributions of Y_r and Y_c represent distributions of the two underlying attitudinal scales, then the proposed joint distribution will maximize the correlation between these scales, **subject to stochastic ordering**.

The connection between minimum entropy and the strongest prediction is well known (see, for instance, the famous paper by Kullback & Liebler [1951] or Gilula and Haberman [1995]).The unique way of constructing the desired joint distribution is as follows.

Consider the following (common) notations for all i and j, $1 \le i \le R$; $1 \le j \le C$.

$$P_{i+} = P(Y_r = i); P_{+j} = P(Y_c = j)$$

The minimum conditional entropy joint distribution with probabilities

$$Q_{ij} = P(Y_r = i, Y_c = j)$$

						Total
	0	0	0	0	0	.3
	0	0	0	0	0	.53
	0	0	0	0	0	.17
Total	.2	.15	.25	.3	.1	1

Fig. 1 Initial empty table of zeros and R and C marginals

that produces stochastically ordered (conditional) distributions with the fixed marginals $< Pi_+ >$ and $< P_{+j} >$ is given by the following algorithm:

$$\sum_{l=1}^{j}\sum_{k=1}^{i} Q_{kl} = \min\left\{\sum_{k=1}^{i} P_{k+}, \sum_{l=1}^{j} P_{+l}\right\} \tag{1}$$

Algorithm (1) was first introduced by Gilula et al. (1988) in a the context of ordinal dependence and Goodmans RC family of models for ordinal categorical data.

We note that proving that $\{Q_{ij}\}$ is indeed the unique minimum conditional entropy distribution is straightforward, based on DallAglio and Bona (2011). Throughout the rest of the paper, we will refer to this algorithm as the MCE (Minimum Conditional Entropy) algorithm.

We now illustrate the step-by-step process of computing $\{Q_{ij}\}$ from the fixed marginals $< Pi_+ >$ and $< P_{+j} >$ using the MCE algorithm.Let $R = 3$, $C = 5$, and let

$$< Pi_+ >= \{.3, .53, .17\}$$

and

$$< P+_j >= \{0.2, 0.15, 0.25, 0.3, 0.1\}$$

Consider an empty 3 × 5 table with underlying fixed marginals as shown in Fig. 1.

Start with the top left cell and insert there the smaller of its two corresponding marginal, to obtain the table in Fig. 2.

In Fig. 2, the bottom row and right-most column indicate the original marginals, while the gray row and column tracks the marginal minus the quantities that have

						Total	
	.2	0	0	0	0	.1	(.3)
	0	0	0	0	0	.53	(.53)
	0	0	0	0	0	.17	(.17)
Total	.0	.15	.25	.3	.1	1	
	(.2)	(.15)	(.25)	(.3)	(.1)		

Fig. 2 Table after one iteration

						Total	
	.2	.1	0	0	0	0	(.3)
	0	0	0	0	0	.53	(.53)
	0	0	0	0	0	.17	(.17)
Total	.0	.05	.25	.3	.1	1	
	(.2)	(.15)	(.25)	(.3)	(.1)		

Fig. 3 Table after two iterations

already been put into the table. Note that in Fig. 2, the first column is already complete since the top left cell already contains the total of the first column.

Now, proceed to cell (1,2). Insert there the minimum of the two corresponding marginal probabilities shown in the gray area to obtain the table in Fig. 3. Note that the first row is now also complete.

Proceed now to cell (2,2) and re-apply the insertion method to obtain the table in Fig. 4. Repeating these steps we obtain the tables in Figs. 5, 6, 7 and 8.

In Fig. 8 we have obtained $\{Q_{ij}\}$.

A corresponding mechanism for converting row categories (1 to 3) to column categories (1 to 5) is derived by dividing all row entries in Fig. 8 by their corresponding row totals to compute the conditional distributions. This yields the following table in Fig. 9. Thus, Fig. 9 gives us the conditional distribution of Y_c given outcomes for Y_r.

To implement the randomized stochastic conversion, we apply the conditional probabilities as follows. For an individual who selects "1" on the row scale assign the category "1" on the column scale with probability .667 and the category "2" on the column scale with probability .333. Randomized assignment of categories is implemented by drawing a random number RN from the uniform distribution [0, 1] and comparing it to cutoff values to determine the category assigned. For an individual who selected "1" on the row scale, if $RN \leq .667$, assign the individual to category "1" on the column scale, and if $RN > .667$ assign the individual to category "2" on the column scale. Use this (randomized) probabilistic conversion for all individuals to convert the row scale values to probabilistically assigned column values, as prescribed by the probabilities in each row. As surprising as it may appear, this randomized conversion is indeed the unique one that maximizes the correlation between the two scales (conditional on the stochastic ordering).

On the other hand, if we would like to convert column categories to row categories, it would make sense to start the algorithm at the top left and proceed column wise rather than row-wise. However, it is easy to see that this will produce the same $\{Q_{ij}\}$. Hence, by dividing the column entries of $\{Q_{ij}\}$ by the corresponding column totals we observe

						Total	
	.2	.1	0	0	0	0	(.3)
	0	.05	0	0	0	.48	(.53)
	0	0	0	0	0	.17	(.17)
Total	0	0	.25	.3	.1	1	
	(.2)	(.15)	(.25)	(.3)	(.1)		

Fig. 4 Table after three iterations

						Total	
	.2	.1	0	0	0	0	(.3)
	0	.05	.25	0	0	.23	(.53)
	0	0	0	0	0	.17	(.17)
Total	0	0	0	.3	.1	1	
	(.2)	(.15)	(.25)	(.3)	(.1)		

Fig. 5 Table after four iterations

the analogous randomized mechanism. See Fig. 10 for the conditional distribution of Y_c given outcomes for Y_r.

In this case, it is interesting to note that for scale conversions of columns to rows (under the given marginals), randomization is only needed for column categories "2" and "4". For the other categories the conversion is deterministic: all individuals selecting "1" on the column scale are assigned "1" on the row scale, "3" on the column scale is assigned "2" on the row scale, and "5" on the column scale is assigned "3" on the row scale.

This is the optimal conversion mechanism (in the information-theoretic sense) when only two scales and their corresponding distributions are available that captures the heterogeneity of human judgment in attitude translating attitudes to scales.

To add inferential value to the MCE conversion mechanism, we discuss in the Appendix how to derive confidence intervals for the estimated joint distribution that dictates the conversion.

3 An empirical experiment

We could not find any relevant empirical study on scale conversion with publicly available data. To test the algorithm, we need a dataset where the sampled individuals' attitudes are measured jointly (i.e., during the same time period) on different ordinal scales.

We commissioned I-Panel, Israel's largest Internet panel survey company, to carry out such a study.

They recruited a sample of adult (at least 18 years old) Jewish big-city dwellers (cities with at least 100,000 residents), who were randomly assigned to one of four versions of a two-wave recontact survey, and who agreed at the end of the first survey

						Total	
	.2	.1	0	0	0	0	(.3)
	0	.05	.25	.23	0	0	(.53)
	0	0	0	0	0	.17	(.17)
Total	0	0	0	.07	.1	1	
	(.2)	(.15)	(.25)	(.3)	(.1)		

Fig. 6 Table after five iterations

						Total	
	.2	.1	0	0	0	0	(.3)
	0	.05	.25	.23	0	0	(.53)
	0	0	0	.07	0	.1	(.17)
Total	0	0	0	0	.1	1	
	(.2)	(.15)	(.25)	(.3)	(.1)		

Fig. 7 Table after six iterations

to be recontacted. Representativeness was controlled based on demographic variables listed below. All interviews were conducted on-line.All individuals were asked three questions:

1. How satisfied or dissatisfied are you with the quality of life in your city of residence?
2. How satisfied or dissatisfied are you with the quantity and diversity of places of recreation and entertainment in your city of residence?
3. How satisfied or dissatisfied are you with the cleanliness and maintenance of your city of residence?

The questions were either asked using a 5-point scale (where 1 means "very dissatisfied" and 5 means "very satisfied") or an 11-point scale (where 1 means "very dissatisfied" and 11 means "very satisfied") The two scales (5 vs. 11 categories) were chosen to be different enough from each other that we could clearly assess the accuracy of a scale conversion algorithm.

On the same date, all individuals were approached and were asked to respond to the three focal questions in the first survey. Seven days later, the same individuals were re-contacted and were asked the same three questions, using either the same scale or the other scale. The individuals were not informed that they would be approached a week later. When re-interviewed, the interviewees were not given any information about their previous ratings.

The respondents were randomly assigned to one of four conditions, as follows:

- Condition 1:

5-category scale first, followed (a week later) by the 11-category scale.

						Total	
	.2	.1	0	0	0	0	(.3)
	0	.05	.25	.23	0	0	(.53)
	0	0	0	.07	.1	0	(.17)
Total	0	0	0	0	0	1	
	(.2)	(.15)	(.25)	(.3)	(.1)		

Fig. 8 Table after seven iterations

	1	2	3	4	5
$P(Y_c \mid Y_r = 1)$	.667	.333	0	0	0
$P(Y_c \mid Y_r = 2)$	0	.094	.472	.434	0
$P(Y_c \mid Y_r = 3)$	0	0	0	.412	.588

Fig. 9 Conditional distributions of Y_c given Y_r

- Condition 2:

11-category scale first, followed by the 5-category scale.

- Condition 3:

The 5-category scale first, followed by the 5-category scale again.

- Condition 4:

The 11-category scale first, followed by the 11-category scale again.

After answering the three focal rating questions, respondents indicated their gender, age, employment status, years of residence in the city, and number of dependents:

> Gender: 1- Male, 2- Female
> Age in years
> Employment status: 1- salaried employee or army or self-employed, 2- unemployed or retired, 3- student
> Years of residence in the city: 1- less than 10 years, 2-more than 10 years
> Number of dependents

The two underlying scales were chosen to be different enough from each other (5 and 11 categories) to provide relatively strong flexibility for scale comparison. Samples 1 and 2 were aimed at 1000 individuals and samples 3 and 4 at 500 individuals. As is seen in the reported tables below, the actual numbers of validly reporting individuals were less than the target sample sizes. This reflects non-response in the second interview and data cleaning (due to some recording errors).

Next, we focus on analyzing the data for the first satisfaction question ("quality of life"). Data and corresponding analyses for the other satisfaction questions yielded similar results, and are available upon request. See the appendix for the data corresponding to responses to the question "How satisfied or dissatisfied are you with the quality of life in your city of residence?" on four different pairs of scales.

	$P(Y_r \mid Y_c = 1)$	$P(Y_r \mid Y_c = 2)$	$P(Y_r \mid Y_c = 3)$	$P(Y_r \mid Y_c = 4)$	$P(Y_r \mid Y_c = 5)$
1	1	.667	0	0	0
2	0	.333	1	.767	0
3	0	0	0	.233	1

Fig. 10 Conditional distributions of Y_r given Y_c

y_r	1	2	3	4	5
count	16	75	236	343	128
count/798	0.020	0.094	0.296	0.430	0.160

y_c	1	2	3	4	5	6	7	8	9	10	11
count	10	10	16	34	55	109	130	233	86	58	57
count/798	0.013	0.013	0.020	0.043	0.069	0.137	0.163	0.292	0.108	0.073	0.071

Fig. 11 Marginal observed data for Y_r (the 5-point scale) and Y_c (the 11-point scale)

3.1 A brief look at the survey data

In this section we take a brief look at responses to the question "How satisfied or dissatisfied are you with the quality of life in your city of residence?" in Condition 1, to illustrate the characteristics of the data. We have 798 observation on Y_r (the response to the first survey, on the 5-point scale) and Y_c (the response to the second survey, on the 11-point scale), and respondent characteristics. As before, the notation (Y_r, Y_c) indicates that the 5-point response is on the rows and the 11-point response is on the columns in the standard two-way table representation of the joint distribution.

Figure 11 gives the marginal counts for the observed Y_r and Y_c values. For example, we see that 43% of the respondents rated their quality of life with a 4 on the 5-point scale and 29% of the respondents rated their quality of life with an 8 on the 11-point scale. The two-way table of joint counts is given in Fig. 24 of the Appendix.

Figure 12 displays the observed marginal frequencies for Y_c and the observed conditional frequencies $Y_c \mid Y_r = y_r$ for $y_r = 1, 2, \ldots, 5$. The solid line with the solid plot symbol displays the marginal frequencies. So, for example, we can see that the observed fraction of respondents rating their quality of life as 8 on the 11-point scale is about .3 (as in Fig. 11). The conditional frequencies are given by the other lines with the numbered plot symbols indicating the value of Y_r . We see that out of the respondents that rated their quality of life as $Y_r = 4$ on the 5-point scale, about 50% rated their quality of life as 8 on the 11-point scale.Fig. 13 displays some of the respondent characteristics.

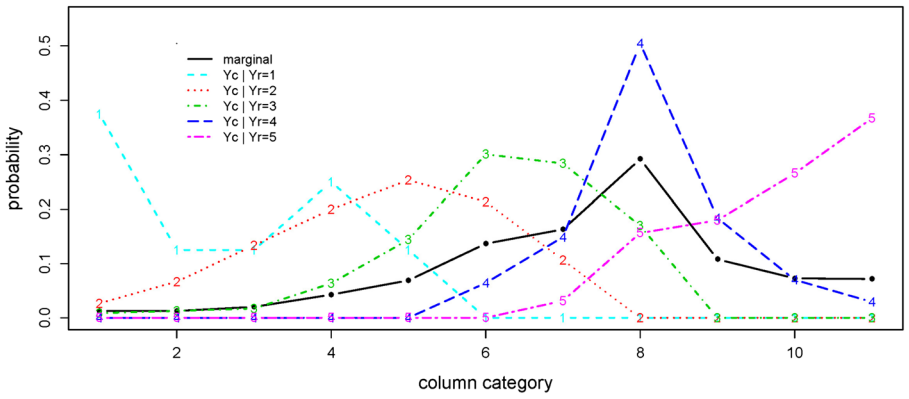

Fig. 12 Solid line gives the marginal frequencies of Y_c. Other lines give the conditional frequencies given the observed value $Y_r = y_r$, $y_r = 1, 2, 3, 4, 5$. Plot labels indicate y_r.

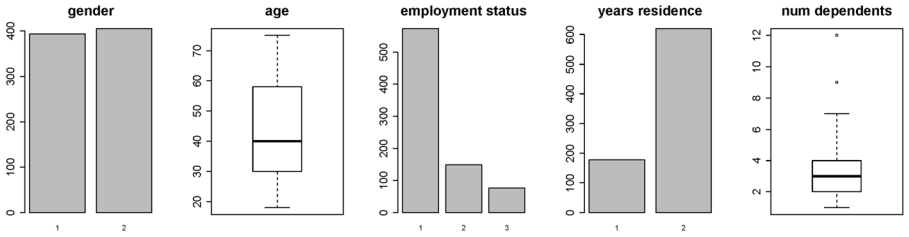

Fig. 13 The demographic variables capturing respondent characteristics: gender: 1- Male, 2- Female. age in years. Employment status: 1- salaried employee or army or self-employed, 2- unemployed or retired, 3- student. Years of residence in the city: 1- less than 10 years, 2-more than 10 years. Number of dependents

4 Applying the MCE algorithm to the observed Marginals

In this section we apply the MCE algorithm defined in Section 2 to our survey data, discussed in Section 3.1.

As an estimate of the correct joint probability table we simply use the joint frequencies (counts divided by sample size, see Fig. 24 in the appendix for the counts). To estimate the marginal probabilities we use the marginal frequencies (Fig. 11).

In our analysis, we test three approaches to the construction of a joint distribution from the marginals. First, we apply the MCE algorithm to the marginals, and we contrast this with a model that assumes independence (i.e., multiplies the marginals). Then, we use a simple mixture of the joint distribution from the MCE algorithm and independence, giving each distribution equal weight of .5.The simple mixture joint probabilities are given by

$$p(Y_r = y_r, Y_c = y_c \mid p_r, p_c) = .5\, p_M\,(Y_r = y_r, Y_c = y_c \mid p_r, p_c) + .5\, p_I\,(Y_r = y_r, Y_c = y_c \mid p_r, p_c)$$

where p_r and p_c denote the estimated row and column marginals, p_M is the joint distribution obtained by applying the MCE algorithm and p_I is the joint obtained by assuming independence.

We develop the independence model to test the possibility that the MCE algorithm by itself may be too extreme to represent real-world joint scaling. Survey responding may, at times, contain answers that are essentially random. The joint distribution under independence represents the one extreme, where all conditional distributions are the same (i.e., the target scale responses are unrelated to the initial scale responses). In contrast, the MCE represents the opposite extreme, inducing conditional distributions that are the farthest apart. The mixture approach provides a compromise between the two extremes represented by the MCE and independence approaches.

In Fig. 14 we compare the estimates obtained from the MCE (a), independence (b), and mixture (c) approaches, to the "correct" values obtained from the observed joint frequencies. To represent a joint distribution in the plots we stack the $R \times C = 5{\times}11 = 55$ joint probabilities into a single vector with 55 values. In each panel, the joint probabilities obtained from the observed joint frequencies are plotted on the horizontal axis, which is

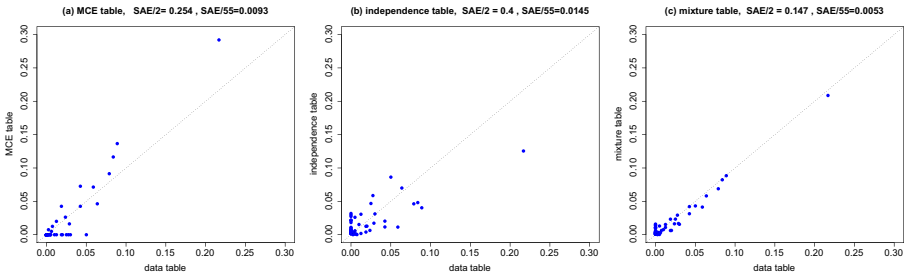

Fig. 14 Plots of the joint probabilities obtained from the joint counts (the data table) versus estimates obtained from the marginal counts. For each table we stack the columns to obtain a single vector of length $5 \times 11 = 55$. (**a**) the data table versus the MCE table, (**b**) the data table versus the independence table, (**c**) the data table versus the mixture table with mixture weight .5. Each axis of each plot is on the same scale and a line with intercept zero and slope one is drawn in each plot

labelled "data table". The estimate obtained from the marginals is plotted on the vertical axis. We obtain numerical measures of the fit obtained by looking at the sum of absolute errors $SAE = \sum_{i=1}^{5}\sum_{j=1}^{11}|p_{ij}-\hat{p}_{ij}|$.We report the average absolute error SAE/55 and a distance measure *SAE*/2. The quantity *SAE*/2 must be in the interval (0, 1) which makes it interpretable as a distance measure between the two distributions. The average absolute errors, *SAE*/55, for the MCE, independence, and mixture approaches are 0.0093, 0.0145, and 0.0053 respectively. The distance measures, *SAE*/2, are 0.2544, 0.4000, and 0.1470. Clearly, the MCE is much better than the independence approach, and the mixture is much better than the MCE.

As discussed in Section 2, we can use the tables representing the joint distributions of (Y_r, Y_c) to infer one scale given the other, simply by computing the appropriate conditional distributions. Figure 15 displays the conditionals $Y_c \mid Y_r = y_r$ for y_r equal 3, 4, and 5 (1 and 2 are omitted because they are infrequent). We display conditionals obtained from the joint data table (solid lines, "D" symbol), the MCE algorithm (dotted lines, "M"symbol) and the simple mixture table (dashed line, "S" symbol). The conditionals from the simple mixture approach tracks the data conditionals remarkably well. The MCE conditionals also do well but are too highly concentrated. Note that if we infer Y_c from Y_r by taking the most likely conditional value, then the mixture and MCE table give the same result.

Next, in Section 5, we provide a more formal treatment of these approaches as models and discuss inference for the associated parameters, including the mixture weight.

5 Models based on the MCE algorithm with Bayesian inference

In this section we present a Bayesian analysis of models based on the MCE algorithm. Let $p_r =< P_{i+} >$ be the vector of marginal probabilities for Y_r . That is, the i^{th} element of p_r is $P\,(Y_r = i)$. Similarly, let p_c be the vector of marginal probabilities for Y_c. Given (p_r, p_c), the MCE algorithm gives us a joint distribution for (Y_r, Y_c). We can view the MCE algorithm as a model for the joint distribution with parameters (p_r, p_c). In Section 3, we examined a mixture of the MCE table and the independence table, with each table getting weight .5. We let the mixture weight be the parameter w giving us a model for

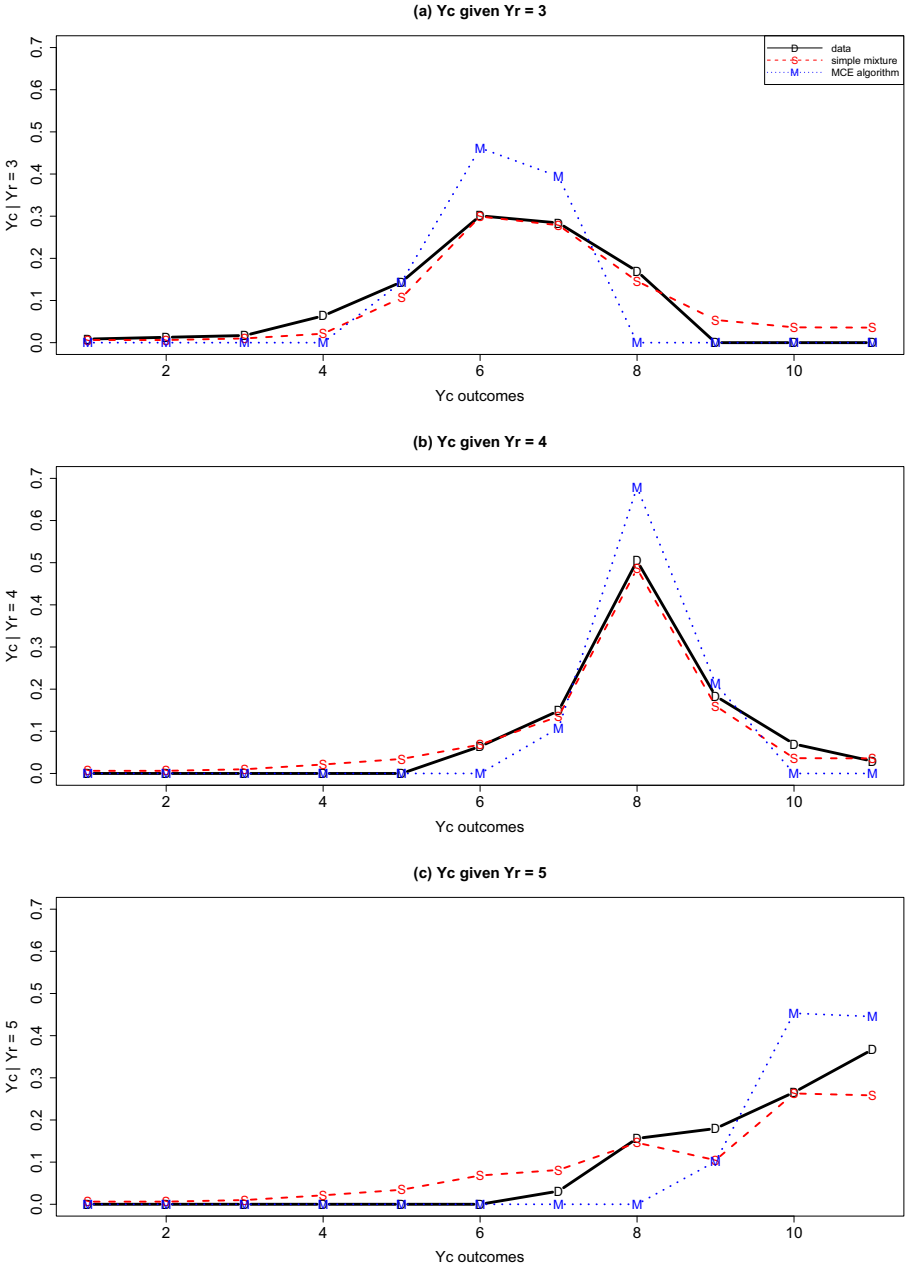

Fig. 15 Plots of the conditional probabilities obtained from the the data table, the MCE table and the simple mixture table. The conditionals are for $Y_c \mid Y_r = y_r$, $y_r = 3, 4, 5$. Panels (**a**), (**b**), and (**c**) correspond to y_r equal 3,4, and 5 respectively

the joint distribution indexed by (p_r, p_c, w). In addition, we consider models where w depends on respondent characteristics.

For these models, the Bayesian approach is appealing because it it gives exact small sample inference with relatively simple prior choices. The margin preserving property of the MCE algorithm enables a Markov Chain Monte Carlo (MCMC) algorithm for obtaining posterior draws. We use a Gibbs sampler with a Metropolis within-Gibbs step, where the Metropolis step uses an independence proposal.

We discuss models based on the MCE algorithm next (in Section 5.1), followed by the MCMC algorithms (in Section 5.2).

5.1 Models based on the MCE algorithm

Suppose we observe $Y_r = y_r$ for a sampled individual. What do we expect for Y_c, the response for the same individual on a different scale? If we impose maximal dependence, we can let p_M (p_r, p_c) be the MCE joint distribution and then compute the conditional $Y_c|Y_r = y_r$. However, the respondent might have made errors in reporting y_r and y_c. When asked for Y_c the respondent may produce a result more in line with the average population result which is captured by p_c. The errors could be due to inattention in responding or, more fundamentally, the individual's attempt to reconcile the two scales. We can capture this possibility by shrinking the MCE-based result to the result obtained from the model in which Y_r and Y_c are independent.

Let p_I (p_r, p_c) be the joint distribution obtained by assuming Y_r and Y_c are independent with marginals p_r and p_c. Our mixture model for the joint disrtibution of (Y_r, Y_c) given the marginals p_r and p_c and mixture weight w is:

$$p(Y_r = y_r, Y_c = y_c|p_r, p_c, w) = \\ w\, p_M(Y_r = y_r, Y_c = y_c|p_r, p_c) + (1-w)\, p_I(Y_r = y_r, Y_c = y_c|p_r, p_c).$$

Shrinkage ideas pervade current approaches to "data science" where they are often called "regularization." This approach can also be seen as the categorical analogue to continuous "true plus error" psychometric models. Our mixture model allows us to shrink individual responses to population responses.

Given measured characteristics x of individuals in our population we can further elaborate our model by letting the mixture weight depend on x. Such dependence allows the model to capture systematic differences across people in test-retest reliability, among other things. A basic logit type specification gives

$$p(Y_r = y_r, Y_c = y_c|p_r, p_c, \theta, x) = p\left(Y_r = y_r, Y_c = y_c|p_r, p_c, w = F\left(x'\theta\right)\right)$$

with

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Here, as usual, we have assumed x expresses the information in terms of numeric features and θ is a vector of coefficients.

We now have three candidate models: the feature-dependent mixture model with parameters (p_r, p_c, θ), the simple mixture model with parameters (p_r, p_c, w), and the MCE model with parameters (p_r, p_c). We will refer to these models as the "mixture

model", the "simple mixture model" and the "MCE model". The simple mixture model is a restricted version of the mixture model and the MCE model is a restricted version of the simple mixture model.

5.2 Markov chain Monte Carlo posterior computation

Given data $\{(y_{ir}, y_{ic}, x_i)\}_{i=1}^{n}$ we construct a Gibbs sampler for the mixture model using the obvious blocking scheme

$$p_r \,|p_c,\ \theta,\ p_c|p_r,\ \theta,\ \theta|p_r,\ p_c.$$

That is, we draw p_r conditional on the other parameters and the data, p_c conditional on the other parameters and the data, and θ conditional on the other parameters and the data.Similary, for the simple mixture model we have

$$p_r|p_c, w,\ \ p_c|p_r, w,\ \ \ w|p_r, p_c,$$

and

$$p_r|p_c \quad\ p_c|p_r$$

for the MCE model.

To draw $p_r\,|\cdots$(and $p_c|\cdots$) we used an independence proposal Metropolis Hastings (MH) step, based on the margin preserving property of the mixture model. Since both the MCE joint p_M and the independence joint p_I have marginal probabilities p_r and p_c, so does the mixture. We propose p_r values from the posterior, given the marginal data $\{y_{ir}\}$, and then accept these proposals using the full conditional likelihood from the mixture model.

The likelihood based on the full mixture model is nonstandard due to the MCE component. However, the marginal likelihoods have the simple multinomial form, so we can use the standard conjugate Dirichlet prior for p_r. We draw proposals from the Dirichlet posterior and then the priors cancel out in the MH acceptance ratio.

Let $p_r \sim$ Dirichlet(α) be the prior. As usual $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_R)$ and $p(p_r) \propto \prod p_{jr}^{\alpha_{j-1}}$. Let $q(p_r)$ be the Dirichlet posterior given the Dirichlet prior and counts from the marginal data $\{y_{ir}\}$. That is, let c_j be the count of the number of times $y_{ir} = j$. Then our (conditionally) conjugate updating, based on the marginal data, gives us a Dirchlet $\tilde{\alpha}$ posterior where $\tilde{\alpha}_j = \alpha_j + c_j$. Let $L_m(p_r)$ be the multinomial likelihood based on these marginal counts. Finally, let $L(p_r)$ be the mixture model likelihood from the full data conditional on (p_c, θ).

To compute the mixture model likelihood contribution of an observation ($Y_r = i$, $Y_c = j$, x), we first compute $w = F(x'\theta)$ and then the probability is w times the (i, j) entry of the MCE table obtained from the marginals p_r and p_c, plus $(1 - w)$ times the i^{th} element of p_r times the j^{th} element of p_c. The obvious simplication applies for the simple mixture model.

The density of our proposal is then $q(p_r) = kp(p_r)\ L_m(p_r)$ and the density of the distribution we wish to draw from is $cp(p_r)\ L(p_r)$ where k and c are constants.

Given a current value of p_r (and other parameters), our independence proposal MH step for updating p_r is:

- draw a proposed value $p_r^* \sim q(p_r)$.
- compute the ratio:

$$r = \frac{cp(p_r^*)L(p_r^*)kp(p_r)L_m(p_r)}{cp(p_r)L(p_r)\ kp(p_r^*)L_m(p_r^*)} = \frac{L(p_r^*)\ L_m(p_r)}{L(p_r)\ L_m(p_r^*)}$$

- Let $p_{accept} = \min\{1, r\}$.
- With probability p_{accept} our next draw is p_r^*, otherwise we repeat p_r.

This is just the standard independence proposal MH algorithm in the case where our proposal is a posterior obtained from the desired prior but an approximate likelihood.

The major advantages of this approach are that (i) it is simple, (ii) it is reasonably fast, (iii) it does not require tuning, and (iv) it facilitates use of the well known Dirichlet prior. In all our examples we used $\alpha_j = 2$ in the Dirichlet prior.

Figure 16 displays the draws from the posterior of the fourth element of p_r using the data analyzed in Section 6. In the top row we have 500 draws of the fourth element of p_r obtained by taking every 10^{th} draw from the full mixture model MCMC. Thus, the 500 draws displayed were obtained by thinning 5,000 draws. At left (panel a) we plot the draws and at right (panel b) we plot the ACF (autocorrelation function) of the draws. We see the repeating we expect from our MH procedure, but the ACF is quite reasonable. In the bottom row (panels c and d), we show 5,000 draws obtained from taking every 80^{th} from 400,000. This is the run our actual results presented in Section 6

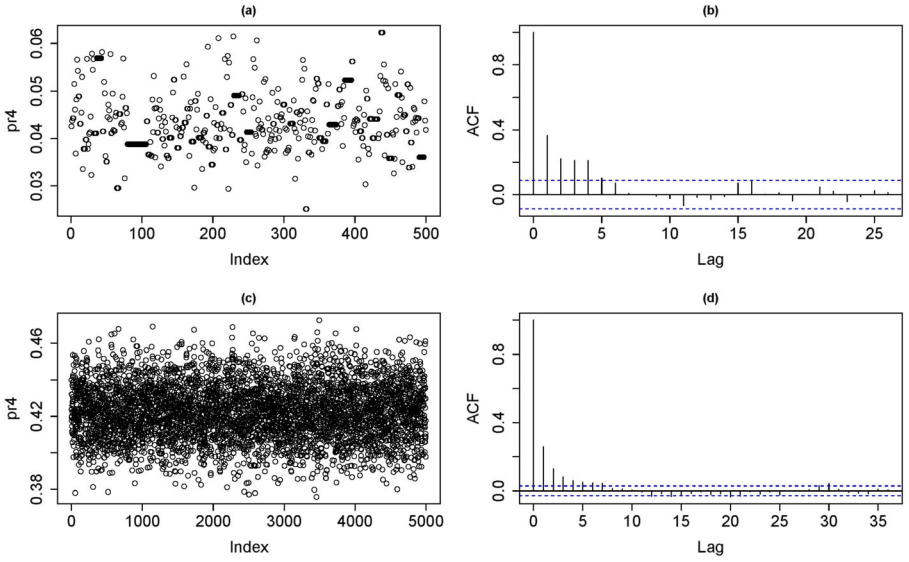

Fig. 16 Thinned draw of the fourth component of p_r with ACFs. Top: thinned to every $10th$ draw. Bottom: thinned to every $80th$ draw. In our examples we use 5000 draws thinned using every $80th$ from 400,000

are based on. As currently coded, it takes about 1 second to get 1,000 draws of p_r conditional on the other parameters. We have coded the mixture model likelihood evaluation in C++ but the other computations, including the basic MCMC loop, are in R. The ACF's of all model parameters exhibit low levels of dependence along the line of that displayed in Fig. 16.

To draw the mixing parameter w in the simple mixture model, we also use an independence proposal. We use a uniform prior on (0,1) and propose from a distribution proportional to the likelihood with p_r and p_c fixed at their marginal observed frequencies.

To draw θ in the mixture model, we use a griddy Gibbs sampler. We standardize all x coordinates and then use a grid of 100 values from -3 to 3 for each θ coordinate. Our prior is uniform on the grid. This is a little slow but required no tuning and was certainly fast enough to obtain the results we wanted. While the speed could be improved using a random walk Metropolis proposal, the griddy Gibbs provides a simple approach that yields good performance without tuning.

We will also consider inference for the MCE model based on marginal data alone. In this case, rather than having the data $\{y_{ir}, y_{ic}\}$ we would only have marginal data on $\{y_{ir}\}$ and on $\{y_{ic}\}$. Draws of p_r and p_c would be obtained directly from the Dirichlet posterior and these draws can then be plugged into the MCE algorithm to obtain draws of p_M (pr, pc).

6 Analysis of the experimental data

Next, we present Bayesian inference for the models discussed in Section 5, using the survey data discussed in Section 3. In Section 3.1 we looked at the data from the sample in which respondents first used a 5 point scale and then, a week later, an 11 point scale (see Fig. 24 in the Appendix for the observed joint frequency distribution). In Sections 6.2 and 6.3 we present detailed results for this data. All posterior inference will be based on 5,000 draws thinned from a longer run of the MCMC chain, as discussed in Section 5. In Section 6.4 we briefly present results for some of the other samples.

6.1 Data and prior configuration for modeling

Since our logit specification for the mixture weights uses a linear function of the conditioning information in x, we convert the categorical variables to dummy variables in the usual manner. We use one dummy for gender=2 (female), one dummy for years of residency=2 (more than 10 years), one dummy for employment status=2 (unemployed or retired), and one dummy for employment status=3 (student). This gives 6 numeric variables after we include age and number of dependents. Counting the intercept, our logit specification will have seven coefficients. Each of the 6 numeric variables is scaled to be between 0 and 1 and then the mean is subtracted. The scaling helps in specifying priors for coefficients and subtracting out the mean makes the intercept more interpretable and easier to estimate.In all examples the prior used is the same as in Section 5. Each θ is

uniform on a grid from −3 to 3, w is uniform on (0,1), and $p_r \sim$ Dirichlet($2 \times 1_R$) and $p_c \sim$ Dirichlet($2 \times 1_C$) where 1_N is a vector of N ones.

6.2 Mixture weight inference

In this section we report inference for the mixture model weight parameters. As discussed in Section 5, our mixture models puts weight w on the MCE table and weight $(1 - w)$ on the independence table. The mixture model has parameters (p_r, p_c, θ) with mixture weights given by $w(x, \theta) = \exp(x'\theta)/(1 + \exp(x'\theta))$. We also report on the parameter w from the simple mixture model having parameters (p_r, p_c, w).

Figure 17 displays the marginal posteriors of the θ_i, i = 1, 2, … , 7. In the top panel, boxplots are used to display the draws from the seven marginals. In the bottom panel, density smoothings are used.

We see that the posterior for the intercept is concentrated on positive values. The posterior mean is .455. If we let all the explanatory variables equal their mean (so that the demeaned values are zero) and plug in $\theta_1 = .455$ we get a weight of exp(.455)/(1 +

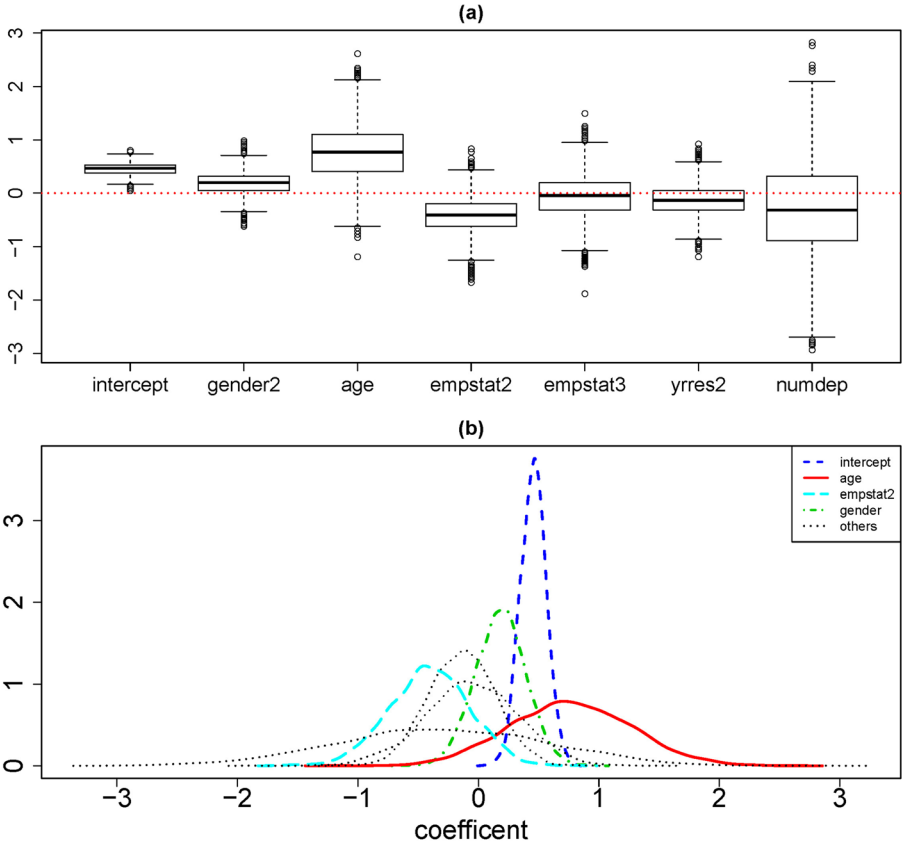

Fig. 17 Marginal distributions of the elements of θ. Top panel (**a**) reports boxplots of the draws for each θ_i and bottom panel (**b**) reports density smooths. 94% of the draws of the age coefficient are greater than zero

exp(.455)) = .61 so that 60% weight is given to the MCE joint table and 40% weight is given to the independence joint table.

There is a 94% posterior probability that the coefficient for age is positive, Recall that empstat2 indicates unemployed or retired and gender indicates a female. There is a 89% probability that the coefficient of empstat2 is negative, and a 81% probability that the coefficient for gender2 is positive. The remaining three coefficients could be close to zero.

The evidence for heterogeneous (i.e., predictor-dependent) weights is not definitive. However, the suggestion that older females and respondents who are not retired or unemployed have more highly dependent responses (i.e., less random responding) is informative.

Figure 18 reports inference for the mixture weights. The top two panels report results using the mixture model with parameters (p_r, p_c, θ). The bottom panel reports results for the simple mixture model with parameters (p_r, p_c, w). For the top two panels, we compute the values $w_{ij} = w(x_i, \theta_j)$ where θ_j is the j^{th} draw from the posterior and x_i is the i^{th} observation of the explanatory variables.

In panel (a), we display the density smooth of the 798 posterior means of the weights for each of the x in our sample. The posterior means are estimated by averaging the w_{ij} values over draws θ_j for each fixed x_i. The average posterior mean is .61 and 95% of the values are in the interval (.5, .7). In panel (b), we display the density smooth of draws from the posterior distribution of the weight averaged over our sample. The posterior draws are obtained by averaging the w_{ij} values over observations x_i for each fixed θ_j. The posterior mean is .61 and 95% of the draws are in the interval (.56, .66). Panel (c) displays a density smooth of draws of the parameter w in the simple mixture model. The posterior mean is .62 and 95% of the draws are in the interval (.57, .66).

Clearly, these analyses show that the data strongly supports mixture weights of about .6. There is relatively modest variation in the weight across the sample (top panel (a) of Figure 18).

6.3 Estimates of the joint (Y_r, Y_c) distribution

In this section, we estimate the joint distribution of (Y_r, Y_c). As discussed in Section 5, there are three different models:

- the mixture model with parameter (p_r, p_c, θ).
- the simple mixture models with parameter (p_r, p_c, w).
- the MCE model, with parameter (p_r, p_c).

For each of these models we have MCMC draws from the posterior distribution of the parameter given the $n = 798$ observations in our data. From these draws we can compute the predictive distribution of (Y_r, Y_c) simply by averaging the joint tables computed at each draw. Note that in the case of the mixture model, we condition on x so that we are computing $(Y_r, Y_c) \mid X = x$. For the MCE model we obtain draws of the parameter (p_r, p_c) from both the marginal version of the data in which we separately observe data sets $\{y_{ir}\}_{i=1}^{n}$ and $\{y_{ir}\}_{i=1}^{n}$ and the joint data $\{y_{ir}, y_{ic}\}_{i=1}^{n}$.

We also compute two "plug-in" joint table estimates where we plug in parameter estimates. We consider the table obtained by plugging in the observed marginal

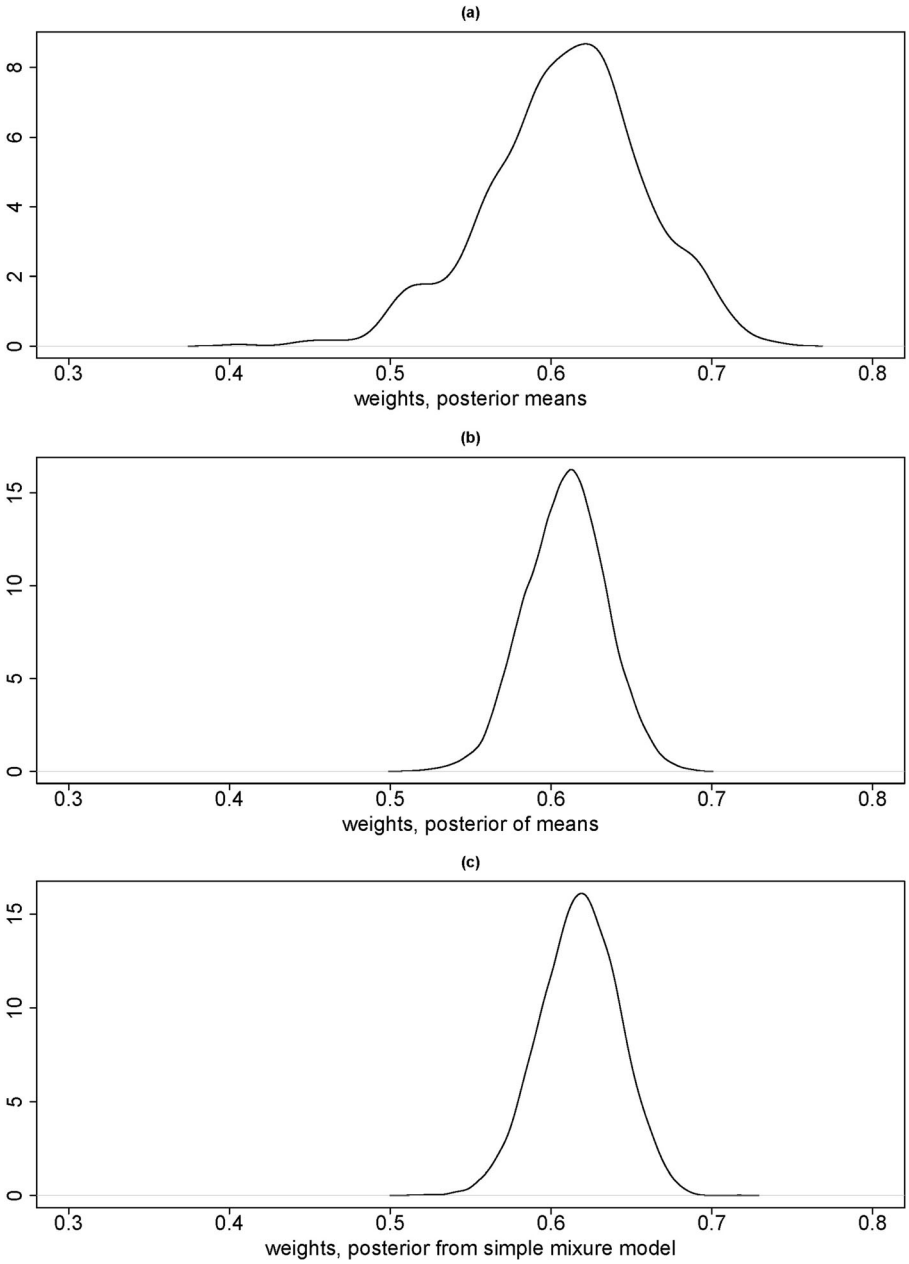

Fig. 18 Inference for weights. Let $w_{ij} = w(x_i, \theta_j)$ where θ_j is the j^{th} draw from the posterior and x_i is from the i^{th} observation. Panel (**a**) is the w_{ij} values averaged over j: variation in the expected weights over x values. Panel (**b**) is the w_{ij} values averaged over i: variation in the average weight over θ draws. Panel (**c**) is the posterior distribution of w in the simple mixture model

frequencies (see Fig. 11) to the MCE algorithm. We also consider the table obtained from the simple mixture model with the observed marginal frequencies and $\hat{w}$ = .6

plugged in. The plug-in version of the MCE table has the fundamentally appealing property that it only depends on the marginal frequencies which are easily obtained from the marginal data. The plug-in version of the simple mixture table depends on our choice of $\hat{w} = .6$ which we clearly learned from the joint data (see Fig. 18). However, it is a straightforward procedure, and we will see in Section 6.4 that it performs well in our other data sets.

This gives us six different estimates of the table representing the joint distribution. We have the three predictives from our three models and the joint data, the predictive from the MCE model and the marginal data, and the two plug-in approaches. Note that in the case of the mixture model, we average the conditional predictives over the observed x_i to obtain a single (Y_r, Y_c) joint. We also have the joint frequencies obtained from the joint data (see Fig. 24 in the Appendix). We refer to the table obtained directly from the joint data frequencies as the "data table".

To compare the joint distribution obtained from the different models and estimation methods, we simply stack the columns of the 5×11 table giving a vector of length 55 for each approach.

Panel (a) of Fig. 19 plots the data table vs. MCE predictive based on the joint data. The Pearson correlation in this plot is .953 and we can see that the MCE tracks the data

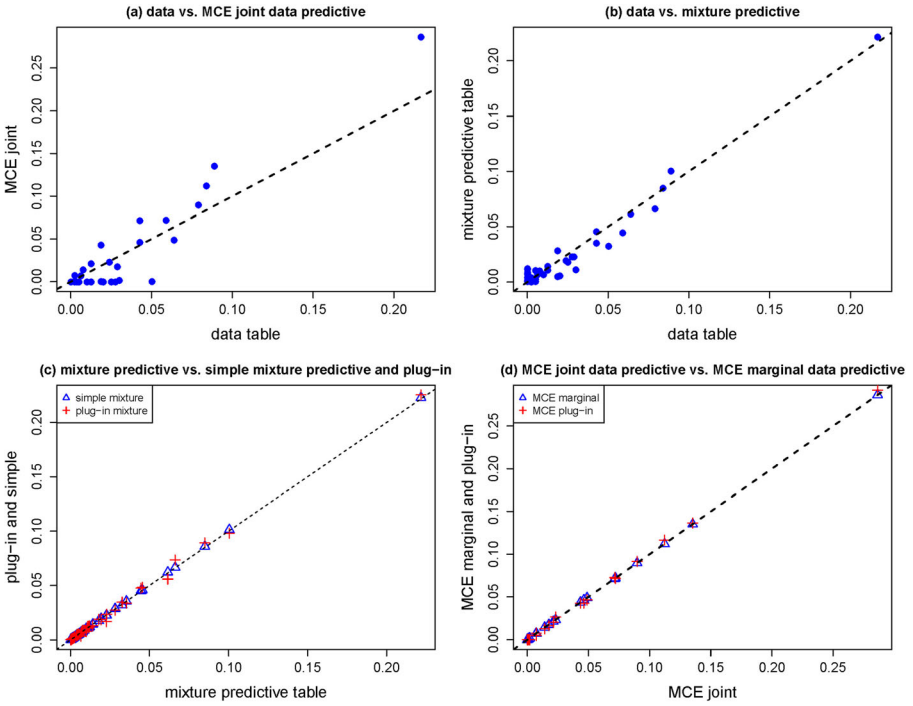

Fig. 19 Inference for tables. We plot the various estimates of the joint distribution $P(Y_r = y_r, Y_c = y_c)$. Each 5×11 table representation of the joint distribution is column stacked to give a vector of 55 numbers. The "data table" is the table estimated by the observed frequencies. (**a**) data table vs. predictive obtained from the MCE model and the joint data. (**b**) data table vs. predictive obtained from the mixture model, averaged over x. (**c**) mixture predictive vs. both the simple mixture predictive and the simple mixture plug-in. (**d**) MCE predictive (joint data) vs. the MCE predictive (marginal data) and the plug-in MCE

very well. Panel (b) plots data table vs. the mixture predictive. The Pearson correlation is .98 and we can see that the fit to the data is very good and markedly improved over the MCE fit. Panel (c) plots the three tables from the mixture models against each other. We see that they are very similar. Panel (d) plots the three tables obtained from the MCE algorithm against each other and they also are nearly identical.

This illustrates that simply applying the MCE algorithm, using only the observed marginal frequencies, estimates the joint distribution well. Mixing the MCE table and the independence table with a 60%-40% weight, both constructed using only the observed marginal frequencies, results in even better recovery of the joint distribution.

In Fig. 20 we provide a more intuitive look at the table estimates by computing the conditional probabilities $P\ (Y_c = y_c|Y_r = y_r)$. Since Fig. 19 indicates that the three mixture approaches give very similar results, we only show results for the predictive distribution for the simple mixture model. Similarly, we only show results for the predictive distribution for the MCE model using the marginal data.The five panels (labeled (a)-(e)) show the conditional distribution $Y_c|Y_r = y_r$ for $y_r = 1, 2, 3, 4, 5$. In each panel the conditionals obtained from the data table are labeled D, the conditionals obtained from the MCE table are labeled M, and the conditionals obtained from the simple mixture model are labeled S. While the MCE conditionals track the data quite well, the fit from the simple mixture is noticeably better. Note that we don't have many observations at the low level of Y_r so that the data values are most reliable for $y_r = 3, 4, 5$. If we focus on the bottom three panels of Fig. 20, the fit of the simple mixture model is very good, but the MCE model is consistently over-concentrated at the most likely values.

Note that (per Fig. 19) using the plug-in versions of the MCE table and the simple mixture table would yield an identical-looking figure. In Section 6.4 we will examine the fit of the simple plug-in estimation approaches (with $\hat{w} = .6$ for the mixture model) to our other samples (i.e., the other conditions in the study) and we find that they perform very well. It is important to note that these constitute out-of-sample experiments, in that they involve different responses than those used to calibrate the model.

While the Bayesian predictive distribution, where we average out the parameter, is the correct distribution to base decisions on, informally it can be of interest to look at posterior distributions of quantities of interest to gauge the underlying uncertainty. In Fig. 21 we look at the marginal posterior of $P(Y_c = 8|Y_r = 4, \gamma)$ where $\gamma = (p_r, p_c)$ for the MCE model, $\gamma = (p_r, p_c, w)$ for the simple mixture model, and $\gamma = (p_r, p_c, \theta)$ for the mixture model. In the case of the mixture model, we have to compute the average of $P(Y_r, Y_c|p_r, p_c, \theta, x)$ over x values and then compute the conditional.

Given our MCMC draws, we compute $P\ (Y_c = 8|Y_r = 4, \gamma)$ at each draw of γ. The first two boxplots in Fig. 21 display the draws of $P(Y_c = 8|Y_r = 4)$ from the MCE model using the marginal and joint data. The distributions are nearly identical. This indicates that there is no additional information from the joint data when using the MCE approach. The third and fourth boxplots display draws from the simple mixture model and the mixture model, respectively. This pair of boxplots is also very similar. If we average out x, the two models give the same inference. The horizontal line is drawn at the value obtained from the joint data. As we saw in Fig. 20, the MCE approaches give estimates that are too large. Here we see that the data value is not even in the support of the marginal distribution.

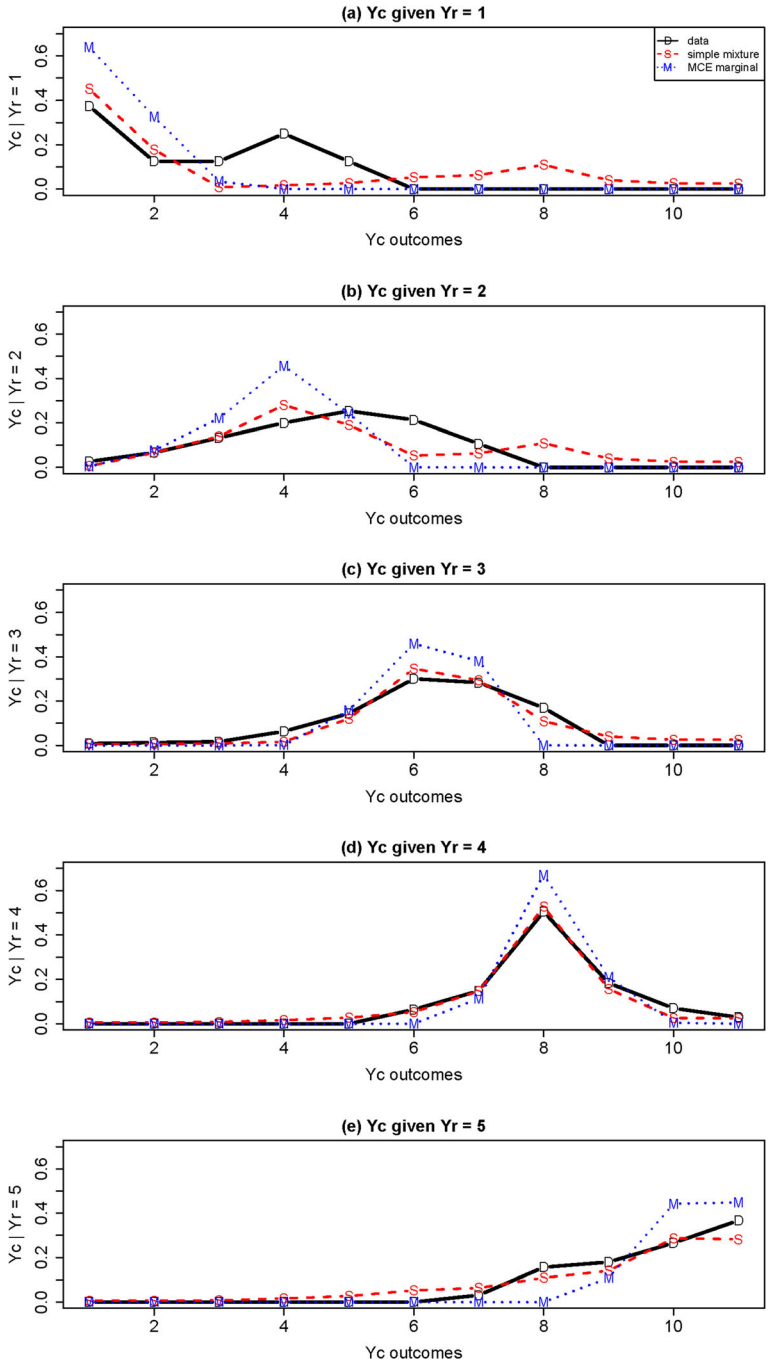

Fig. 20 Conditional distributions of $Y_c \mid Y_r$ from various fitted tables representing the joint distribution of (Y_r, Y_c). Panels (**a**)-(**e**) (top to bottom) present results for conditioning on Y_r equal to 1,2,3,4, and 5 respectively. Within each panel conditionals are computed from the observed data table, plotted with "D"; the table obtained from the MCE predictive and marginal data, plotted with "M"; the predictive obtained from the simple mixture model and joint data, plotted with "S"

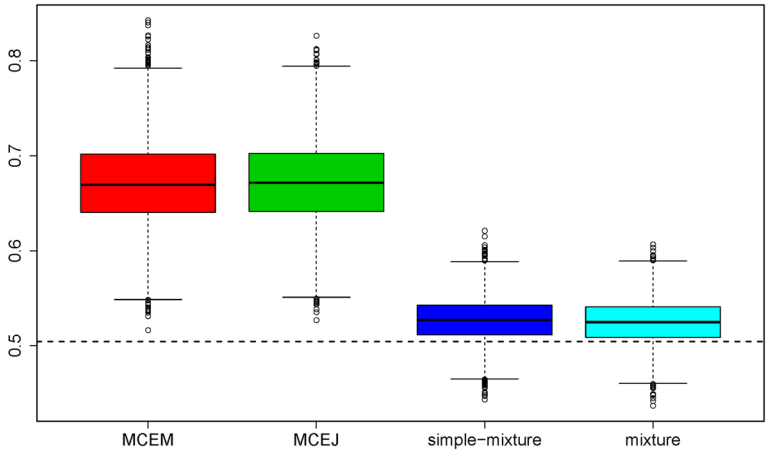

Fig. 21 Posterior distributions of $P(Y_c = 8|Y_r = 4)$ from various models. MCEM: draws of $P(Y_c = 8|Y_r = 4, p_r, p_c)$ where (p_r, p_c) are drawn conditional on the marginal data. MCEJ: draws of $P(Y_c = 8|Y_r = 4, p_r, p_c)$ where (p_r, p_c) are drawn conditional on the joint data. Simple-mixture: draws of $P(Y_c = 8|Y_r = 4, p_r, p_c, w)$. Mixture: draws of $P(Y_c = 8|Y_r = 4, p_r, p_c, \theta)$. For each draw of (p_r, p_c, θ) the joint distributions Y_r, $Y_c|x$ is averaged over x in the sample. The conditional is then computed from this average joint. The dashed horizontal line is drawn at the value obtained from the joint data

6.4 Results for the other three survey conditions

In this section we report on the fit of the plug-in versions of the simple mixture model and MCE model when applied to the other samples we collected, without using the joint distribution. Again, we use the results obtained in response to the question: "How satisfied or dissatisfied are you with the quality of life in your city of residence?". In the data set we have looked at in detail above, respondents were asked to use a 5-point scale and then an 11-point scale. In the three additional survey conditions considered in this section, respondents either used 5-point then 5-point, 11-point then-5 point, or 11-point then 5-point In all cases, Y_r refers to the first scale and Y_c to the second. See Section 3 for discussion of the surveys. The two-way tables of joint counts are given in the Appendix.

In each case we we fit the simple mixture model by plugging in the observed marginal frequencies as estimates of the marginal probabilities and the weight $\hat{w}$ = .6 on the MCE table. Note that with this approach, we are *not* using any information from the joint observations of (Y_r, Y_c) to estimate the model. Our estimation is therefore truly out of sample in the sense that we use an estimate of the key weight parameter w that had been calibrated on different responses. This is in contrast to the usual approach of evaluating out-of-sample performance by using a sub-sample for estimation.

Figure 22 presents the results. The three rows of plots correspond to the three data sets (one per survey condition). In each row, the first plot is the data table vs. the MCE table, the second plot is the data table vs. the independence table, and the third plot is the data table vs. the simple mixture plug-in table. In the first two rows (5-point to 5-point and 11-point to 5-point data sets), the mixture table fits the joint frequencies very well. In the third row (11-point to 11-point

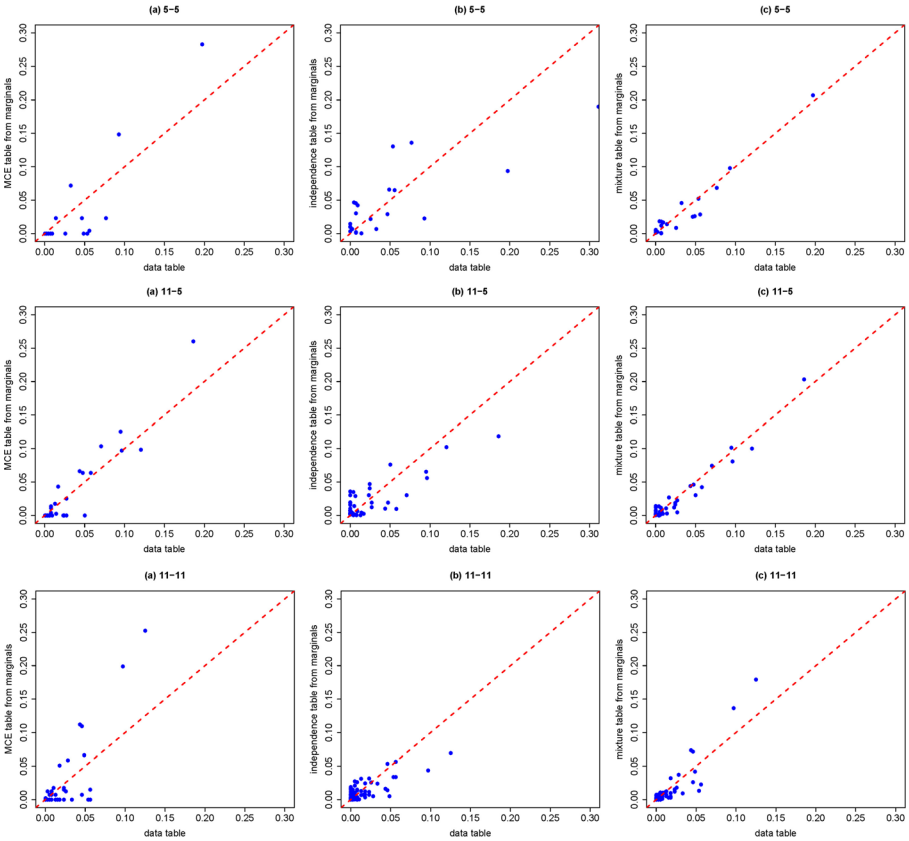

Fig. 22 The three rows of plots correspond to the 5 point scale and then 5 point scale data set, the 11 point scale and then 5 point scale data set, and the 11 point scale and then 11 point scale data set, respectively. In the first column we plot the data table vs. the plug-in MCE table. In the second column we plot the data table vs. the independence table. In the third column we plot the data table vs. the plug-in simple mixture table with $\hat{w} = .6$

data set), the mixture does not fit as well, but it is still much better than either component on its own.

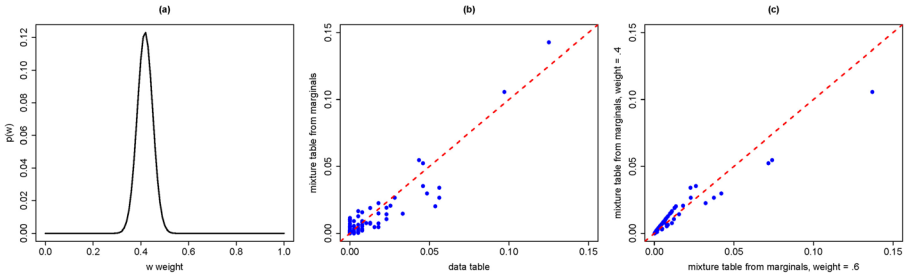

Fig. 23 Additional results for the 11 point scale to 11 point scale data set. Left panel (**a**): posterior distribution of w in the simple mixture model. Middle panel (**b**): data table versus mixture table with $\hat{w} = .4$. right panel (**c**): mixture table with weight $\hat{w} = .6$ versus mixture table with weight $\hat{w} = .4$

To improve the performance in the 11-point to 11-point data set we considered re-estimating w rather than just plugging in .6. The left panel (a) of Fig. 23 shows the posterior of w (conditional on the sample frequencies for $\underline{p_r}$ and p_c). The posterior is concentrated close to $w = .4$, indicating a higher weight on independence, consistent with lower test-retest reliability for the 11-point scale. In the middle panel (b) of Figure 23 we plot the data table versus the mixture table using the weight $\hat{w} = .4$. Again, the fit looks quite good. In the right panel of Fig. 22 we plot the two mixture tables obtained using weights .6 and .4 against each other. This analysis illustrates that external data on test-retest reliability could be useful to inform the plug-in parameters used in an implementation of the mixture model.

7 Conclusion

In this paper, we introduce a novel approach to ordinal-categorical scale conversion. The advantage of the proposed approach over the (very few) existing techniques for scale conversion is that it tries to model how individuals would convert their own categories if offered two categorical attitudinal scales with different numbers of categories. Our approach is based on stochastic ordering, which is presumed to reflect (to some degree of accuracy) the actual process of human scale conversion. Conversion by stochastic ordering (the MCE) is argued to be optimal in the sense that it maximizes the correlation between the two underlying scales (subject to stochastic ordering). No other technique, to our knowledge, satisfies this criterion.

With all due respect to MCE, it is likely that people do not exactly use this optimal conversion in real life. To account for deviations from the MCE, we have introduced simple mixture models. We have considered a convex combination of the joint distribution of the response on two different scales, where we mix a joint obtained from applying the MCE algorithm to the marginals with a joint obtained by assuming independence. The mixture balances a strong form of dependence between scales captured by the MCE algorithm with a complete lack of dependence between scales.

In this paper, we first implemented a frequentist approach, using just the MCE algorithm with no mixture (i.e., zero weight for independence). We then study mixtures, discussing different mixture parameters. Next, we develop a Bayesian approach to mixture models where we show how mixture weights can be estimated from the data. All conversion mechanisms are accompanied by inferential procedures and can accommodate background variables.

In order to empirically test the proposed approach, we have conducted an observational study that is detailed in the paper.

The analysis of the empirical study provides validation for the proposed modeling approach, and for the models tested. Using just the MCE algorithm (with no mixture, no weightings and no background demographics) yields a highly accurate reconstruction of the observed joint scaling (a Pearson correlation of 0.93 between the 55 entries of observed data, Figure 22 in the Appendix, and their corresponding MCE estimates, Fig. 23). We don't claim that the simple MCE mechanism is exactly the way by which individuals actually convert their own scaling. However, the closeness between the observed joint scaling and the estimated joint scaling obtained from the MCE algorithm demonstrates the usefulness of the MCE approximation. Adding total

independence, as part of the mixture model, yields a highly noticeable improvement in the re-construction of observed joint scaling (Figs. 19, 20, and 22). We find some evidence that the weighting depends on respondent characteristics (see Fig. 17).

We also considered approximate inference for a model in which the marginal probability vectors p_r and p_c depend on x using the standard multinomial model. In our data, we found very little evidence for such dependence and have omitted this model investigation from further discussion. We note, however, that this might be of interest in future studies and that a full MCMC analysis using independence proposals could be attempted for this more general model.

The scale-conversion need that survey institutes are most likely to encounter is probably the case where the researcher only has data from two distributions, pertaining to two different scales independently measuring the same construct without any additional information. In such cases, based on the results in this paper, the researcher may want to adopt a simple and straightforward frequentist MCE mechanism with or without an arbitrary mixture. If the frequentist convex mixture with weights are to be considered, then the results of our study suggest using a 0.4-0.6 weight for the independence and MCE tables. Clearly, the Bayesian mixture model is the best option, but it may require some information on joint scaling that might not be available.

We believe that our approach may be practically useful to researchers who are conducting scale conversion. Our code for all conversion mechanisms presented in this paper, with or without background variables, is available upon request.

These approaches may also motivate further research on scale conversion. Our proposed mixture is only between two extreme components (MCE and total independence). Similar to the methodology of data fusion offered by Gilula and McCulloch (2013), in which a mixture between 21 components was proposed, many stochastic ordering structures of lesser extremity could be used for scale conversion. We intend to investigate this avenue, as well as other extensions, in the future.

Appendix 1

Data tables

Two way data tables for the question "How satisfied or dissatisfied are you with the quality of life in your city of residence?"

	yc										
yr	1	2	3	4	5	6	7	8	9	10	11
1	6	2	2	4	2	0	0	0	0	0	0
2	2	5	10	15	19	16	8	0	0	0	0
3	2	3	4	15	34	71	67	40	0	0	0
4	0	0	0	0	0	22	51	173	63	24	10
5	0	0	0	0	0	0	4	20	23	34	47

Fig. 24 Observed counts, 5 point scale to 11 point scale

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]
[1,]	10	6	0	0	0	0	0	0	0	0	0
[2,]	0	4	16	34	21	0	0	0	0	0	0
[3,]	0	0	0	0	34	109	93	0	0	0	0
[4,]	0	0	0	0	0	0	37	233	73	0	0
[5,]	0	0	0	0	0	0	0	0	13	58	57

Fig. 25 5 point scale to 11 point scale, counts obtained from the MCE algorithm.

	yc				
yr	1	2	3	4	5
1	6	3	1	0	0
2	3	14	20	4	0
3	1	11	85	33	2
4	0	3	23	134	24
5	0	0	3	21	40

Fig. 26 Observed counts, 5 point scale to 5 point scale

	yc				
yr	1	2	3	4	5
1	10	3	1	0	0
2	6	6	0	0	0
3	3	6	2	0	0
4	0	13	21	0	0
5	4	11	37	0	0
6	0	7	55	19	0
7	0	4	74	94	3
8	0	0	39	145	19
9	0	0	0	75	21
10	0	0	0	18	34
11	0	0	0	5	45

Fig. 27 Observed counts, 11 point scale to 5 point scale

	yc									
yr	1	3	4	5	6	7	8	9	10	11
1	3	0	1	1	0	0	0	0	0	0
2	0	3	0	1	0	0	0	0	0	0
3	0	1	2	1	2	2	0	0	0	0
4	0	1	4	3	2	2	0	0	0	0
5	2	2	0	7	5	7	0	0	0	0
6	1	1	2	7	18	10	9	0	0	0
7	0	0	1	5	13	38	18	3	0	3
8	0	0	0	0	5	22	49	22	7	0
9	0	0	0	0	1	2	21	17	9	1
10	0	0	0	0	0	0	3	6	11	3
11	0	0	0	0	0	0	2	2	9	19

Fig. 28 Observed counts, 11 point scale to 11 point scale

Appendix 2

Inferential aspects of the MCE algorithm

The MCE algorithm can be considered as the function $Q = \Psi(P, P')$, where $P = (P_{1+}, \ldots, P_{R+})$, $P' = (P'_{+1}, \ldots, P'_{+C})$ are probability vectors and Q_{ij}, $i = 1, \ldots, R, j = 1, \ldots, C$ is a contingency vector with marginals P and P'. In practice it is used with sampled value, $\hat{Q} = \Psi\left(\hat{P}, \hat{P}^{'}\right)$, where $\hat{P}$ and $\hat{P}^{'}$ are estimators based on a sample.

We want to construct confidence intervals for the estimators $\hat{Q}$.

Let $\mathbb{P}_i = \sum_{k=1}^{i} P_{k+}$ the cumulative distribution function corresponding to P. Define similarly other cdf's (e.g., $\mathbb{P}'$ and $\mathbb{S}$), and $\mathbb{Q}_{ij} = \sum_{k=1}^{i}\sum_{l=1}^{j} Q_{kl}$. The CME algorithm is simply

$$\mathbb{Q}_{ij} = \min\left\{\mathbb{P}_i, \mathbb{P}'_j\right\}. \tag{2}$$

We consider a standard fix point asymptotics. We assume.

A1. $\hat{P}$, $\hat{P}^{'}$ are independent and based on a multinomial samples with sizes n,n' respectively.

A2. For all $i = 1, \ldots, R$ and $j = 1, \ldots, C$, if $\mathbb{P}_i = \mathbb{P}_j$ then $i = R$ and $j = C$.

In that case the asymptotics is simple, since if $\mathbb{P}_i < \mathbb{P}_j$ then $\mathbb{Q}_{ij} = \mathbb{P}_i$, and the CI of $\mathbb{P}_i$ is the CI of $\mathbb{Q}_{ij}$ as well.

As for Q_{ij} itself,

$$Q_{ij} = \mathbb{Q}_{ij} + \mathbb{Q}_{i-1,j-1} - \mathbb{Q}_{i-1,j} - \mathbb{Q}_{i,j-1}, i = 1, \ldots, R, j = 1, \ldots, C, \tag{3}$$

where $\mathbb{Q}_{0,j} \equiv \mathbb{Q}_{i,0} \equiv 0$. Hence

$$Q_{ij} = \min\left\{\mathbb{P}_i, \mathbb{P}'_j\right\} + \min\left\{\mathbb{P}_{i-1}, \mathbb{P}'_{j-1}\right\} - \min\left\{\mathbb{P}_{i-1}, \mathbb{P}'_j\right\} - \min\left\{\mathbb{P}_i, \mathbb{P}'_{j-1}\right\},$$

where $\mathbb{P}_0 = \mathbb{P}'_0 = 0$. Without loss of generality, there are three possibilities:

i. Suppose $\mathbb{P}_i < \mathbb{P}_{j-1}$. Then clearly $\mathbb{P}_{i-1} < \mathbb{P}_i < \mathbb{P}_{j-1} < \mathbb{P}_j$ and hence $Q_{ij} = 0$. Thus $P_r\left(\hat{Q}_{ij} = Q_{ij}\right) \to^{\mathrm{p}} 1$.
ii. Suppose $\mathbb{P}_{i-1} < \mathbb{P}_{j-1} < \mathbb{P}_j < \mathbb{P}_i$. In this case $Q_{ij} = \mathbb{P}_j - \mathbb{P}_{j-1} = \mathbb{P}_{+j}$, and an asymptotic $1 - \alpha$confidence interval for Q_{ij} is given by $\hat{Q}_{ij} \pm z_{\alpha/2} n'^{-1/2} \hat{P}'_{+j}\left(1 - \hat{P}'_{+j}\right)$.
iii. Suppose $\mathbb{P}_{i-1} < \mathbb{P}_{j-1} < \mathbb{P}_i < \mathbb{P}_j$. In this case $Q_{ij} = \mathbb{P}_i - \mathbb{P}_{j-1}$, and an asymptotic $1 - \alpha$ confidence interval for Q_{ij} is given by

$$\hat{Q}_{ij} \pm z_{\alpha/2} \left(\frac{\hat{P}_{i+}\left(1 - \hat{P}_{i+}\right)}{n} + \frac{\hat{P}'_{j-1,+}\left(1 - \hat{P}'_{j-1,+}\right)}{n'} \right)^{1/2}.$$

References

Accioly, R., & Chiyoshi, F. (2004). Modeling dependence with copulas: A useful tool for field development decision process. *Journal of Petroleum Science and Engineering, 44*, 8391.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205–215.

Churchill, G. A., Jr., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, 360–375.

Colman, A. M., Morris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports, 80*(2), 355–362.

Cox, E. P., III. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 407–422.

DallAglio, G., & Bona, E. (2011). The minimum of the entropy of a two-dimensional distribution with given marginals. Electronic Journal of Statistics. April, 1–14.

Dawes, J. G. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5 point, 7 point and 10 point scales. *International Journal of Market Research, 51*(1).

Dolnicar, S., & Gr un, B. (2013). Translating between survey answer formats. *Journal of Business Research, 66*(9), 1298–1306.

Elidan, G. (2010). Copula bayesian networks. In J. Lafferty, C. K. I. Williams, J. ShaweTaylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems 23 (pp. 559567). Red hook*. New York: Curran Associates.

Evans, M., Gilula, Z., & Guttman, I. (2012). Conversion of ordinal attitudinal scales: An inferential Bayesian approach. *Quantitative Marketing and Economics, 10*(3), 283–304.

Gilula, Z., Kriger, A. M., & Ritov, Y. (1988). Ordinal Association in Contingency Tables: Some interpretive aspects. *Journal of the American Statistical Association., 83*(402), 540–545.

Gilula, Z., & Haberman, S. J. (1995). Dispersion of categorical variables and penalty functions: Derivation, estimation, and comparability. *Journal of the American Statistical Association., 80*, 1438–1446.

Gilula, Z., & McCulloch, R. (2013). Multilevel categorical data fusion using partially fused data. *Quantitative Marketing and Economics, 11*(3), 353–377.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery-how many scales and responses categories to use? *Journal of Marketing, 34*, 33–39.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency, annals of. *Mathematical Statistics, 2*, 7986.

Miller, G. A. (1956). The magical number of seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1–15.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agreedisagree scales. *Sociological Methods & Research, 0049124113509605*.

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement, 66*(2), 228–239.

Affiliations

Zvi Gilula[1] · Robert E. McCulloch[2] · Yaacov Ritov[3] · Oleg Urminsky[4]

Robert E. McCulloch
Robert.Mcculloch@asu.edu

Yaacov Ritov
yritov@umich.edu

Oleg Urminsky
oleg.urminsky@chicagobooth.edu

[1] Department of Statistics, Hebrew University, Jerusalem, Israel

[2] School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA

[3] Department of Statistics, University of Michigan, Ann Arbor, MI, USA

[4] University of Chicago Booth School of Business, Chicago, IL, USA