Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Economic Psychology 24 (2003) 349–376
JOURNAL OF Economic Psychology
www.elsevier.com/locate/joep

Ask and ye shall receive: The effect of the appeals scale on consumers' donation behavior

Pierre Desmet [a,1], Fred M. Feinberg [b,*]

[a] *University of Paris-IX-Dauphine and ESSEC, 20, Les Clairiéres Rouges, Cergy-Pontoise 95000, France*
[b] *University of Michigan Business School, 701 Tappan St., Ann Arbor, MI 48109-1234, USA*

Received 25 July 2001; accepted 15 July 2002

Abstract

Managers in the fundraising and public sectors face the constant challenge of soliciting donations from a population who may or may not have donated before. Rather than merely asking respondents what they wish to donate, it is standard practice to present a set of suggested amounts – the *appeals scale* – in making donation requests. We study the relationship between *what is requested* and *what is received* by incorporating prior donation history into a comprehensive, 'attraction'-based model of donation behavior.

A large-scale field trial, coupled with a unique donation database from a French charity, allows measurement of several distinct appeals scale effects while accounting for underlying heterogeneity in donation behavior. A segment-level Bayesian model for the *distribution* of donations clarifies the influence of the appeals scale on donor behavior, as well the effect of 'round' scale values, such as those appearing on common bank notes. We find that the former effect can account for as much as 12% of overall donation behavior, the latter 7%, and moreover that these effects are essentially additive. Both effects, as well as proximity of scale points to a group-wise reference level, substantially alter the distribution of donations received. The data suggest that donations can be strongly influenced by choosing appropriate quantities to ask for, suggesting avenues for improving the practice of soliciting charitable requests.

PsycINFO classification: 2240; 3900; 3920; 3940
JEL classification: C15; D12; D14; D64; D81; H31; M31; P46

[*] Corresponding author. Tel.: +734-764-4711; fax: +734-764-3240.
E-mail addresses: pierre.desmet@dauphine.fr (P. Desmet), fred.feinberg@umich.edu (F.M. Feinberg).
[1] Tel.: +33-1-44-05-44-70/34-43-30-81.

0167-4870/03/$ - see front matter
doi:10.1016/S0167-4870(02)00166-6

Keywords: Fundraising; Donation behavior; Choice theory; Reference dependence; Framing; Gibbs sampling; Bayesian methods

1. Introduction

A great deal of academic research in Marketing has examined the important problem of choosing appropriate prices or price ranges for individual goods, services, categories and product lines. In such cases, it can be argued that the consumer trades monetary resources for something concrete or tangible (in the case of goods), or at the very least something for which there is a market-driven range of values. By contrast, managers working in the fundraising and public sectors face the recurring challenge of soliciting donations without offering anything of 'objective' value in return. The same donation appeal can elicit grossly different pledges from households which appear demographically similar, no pledge at all, or different behavior from the same household at different times. As such, the relation between what one asks for – the donation appeal – and what one receives is difficult to anticipate.

Previous related research on pricing has largely focused on choice of price for a particular, isolated product or product line. In the latter case, it has been known since the work of Monroe (1973) that the range (or scale) of prices itself, as much as each individual price, serves to influence buyer behavior. Compared with the more common setting of goods and services, the charity fundraising sector possesses characteristics making it an attractive arena for studying the effects of price *scale*. For example, such effects are not strongly influenced by variations in product characteristics, as the products themselves – purely monetary donations – are intangible; nor are they strongly influenced by distribution or sales force efforts, as fundraising takes place at a distance, typically by phone or through mass mailings. Moreover, the direct marketing context facilitates the implementation of large-scale experiments in a 'natural' setting.

The influence of scaling or context effects in fundraising has been studied mainly in face-to-face interactions (e.g., Reingen, 1982), and to a lesser extent through mailed or phoned donation appeals carried out in accordance with standard experimental design methods. Such context effects have been decisively demonstrated and, generally speaking, requests for larger sums result both in a rise in average donation and in a decrease in donation frequency. However, the results obtained by various researchers are not always convergent on even these points (Schibrowsky & Peltier, 1995; Weyant, 1996), let alone more subtle ones.

The objective of the present study is to formulate a model which can aid in understanding how *what one requests* affects *what one receives*, clearly a crucial question for managers trying to plan their donation drives and, over time, to improve their efficacy. Such a model should account for the consequences of a price scale – henceforth, the *appeals scale* – taking into consideration both contextual effects due to manner of communicating the donation request (i.e., by mail, phone, etc.), as well as prior donation histories. Because appeals scales typically require choosing among a menu of specific quantities, we model it as a set of discrete 'attraction' points. As suggested by prior research, respondent groups make use of latent, internal reference values – based on their donation histories – so that proximity to appeals scale points can predispose them to accept one of the suggested donation amounts, resort to one of their own, or eschew donation entirely.

The class of models developed in this paper allows concurrent measurement of scale attraction effects, the role of common bank note denominations, and the asymmetry of attraction to higher and lower scale values (as predicted, for example, by Prospect Theory; Kahneman & Tversky, 1979). Density estimation and Geweke's (1993) heteroscedastic normal linear regression (HNLR) model allow us to disentangle a variety of scale attraction effects. We find that points appearing on the scale exert a decisive influence on donor behavior, even when such a conclusion is not apparent based on standard efficacy measures, such as the mean donation received or donation frequency. The remainder of the paper is organized as follows: the theoretical framework underlying donation response is discussed; a series of donation-group-level models are developed; the data and variable definitions are presented; followed by model estimation results and their implications for scale design.

2. Foundations

In formulating the models presented later, we incorporate effects for which there is broad prior empirical support. Extensive literature in reference pricing and scale design suggests a set of constructs which need be accounted for in a reasonable model of donation behavior. We briefly review prior research relevant to the model formulation below.

2.1. The role of the appeals scale

With regards to carrying out charitable appeals, many studies have decisively demonstrated the efficacy of manipulation techniques used in Social Psychology (Abraham & Bell, 1994; Doob & McLaughlin, 1989; Smith, 1980). When donation requests were carried out through direct mailings, Weyant (1996) found that the appeals scale served as a set of "external" (i.e., exogenous) reference points. Research has addressed the relationship between amount requested of donors and the behavior subsequently expressed, in terms of frequency of giving and the average amount of such donations (DeJong & Oopik, 1992; Schibrowsky & Peltier, 1995; Weyant & Smith, 1987). This line of research revealed other intriguing behavior: while a substantial proportion of the population tended to 'go along' with the values expressed on the scale, the central scale value was *not* the one most often chosen. Fraser, Hite, and Sauer (1988) found, by manipulating scale extremes directly (i.e., by an increase in the value requested) or indirectly, that donation frequency is not systematically, negatively related to the average amount donated.

An analogous line of inquiry has been investigated extensively by psychologists wishing to design better survey scales, relatively free of anchoring or overt contextual effects. In a major recent review of the extant literature, Schwarz (1999) emphasizes the "emergence of context effects in attitude measurement", particularly in question format and scaling. Response scales have been approached as de facto frames of reference, directly altering judgments (Schwarz, Bless, Bohner, Harlacher, & Kellenbenz, 1991), a perspective recently supported in the context of mail and telephone survey response categories (Rockwood, Sangster, & Dillman, 1997). These studies suggest the broad influence of the appeals scale on the success of donation requests. Taken together, they emphasize the effects of moderating variables and the need for an integrative model which accounts both for *scale values* and for *prior donor behavior*. The model we develop specifically addresses both issues.

2.2. The donation decision

Faced with an appeal for funds, one must choose whether to respond at all, and then possibly the conditions and amount of payment. This explicit ordering – with quantity conditional on a (binary) choice – is shorn up by the format of the fund appeal itself, which ordinarily requests an initial pledge commitment, followed by determination of the donation amount. A variety of 'alternatives' is typically offered, in the form of a discrete menu of amounts (e.g., $10, $25, $50, $100, etc.), from which the respondent chooses. Respondents are also free to select quantities not appearing on the appeals scale, either within its range or beyond its endpoints, at their own initiative. The model developed in this paper will *measure* the 'attractive' effects of the appeals scale – that is, quantify its importance across the respondent pool – while *accounting* for differences in incidence (i.e., donating or not donating) across groups with different prior donation behavior.

2.3. Reference effects and prices

Prior research has firmly established that various alternatives are evaluated not in an 'absolute' manner, but relative to a point of reference. In fact, one of three "major generalizations" in the extensive overview of consumer choice models by Meyer and Johnson (1995) dealt with the importance of reference-dependence, a view supported in the comprehensive review by Kalyanaram and Winer (1995). The relative importance of the two potential sources of information – internal (what one knows) and external/environmental (what one finds out or is told) – depends on a variety of individual-specific and environmental factors (Sherif, Taub, & Hovland, 1958; Payne, Bettman, & Johnson, 1992).

Research on reference *pricing*, in particular, has emphasized the importance of simultaneously making use of 'internal' as well as 'external' referents (Rajendran & Tellis, 1994; Zollinger, 1993). Special note should be made of the study of Mayhew and Winer (1992), who distinguish the effects of internal and external reference prices, the former a mnemonic hybrid of actual, recent or fair prices, the latter referring to externally supplied price information, typically at the time of purchase. Note, however, that for the purposes of model-building, internal reference 'prices' – a latent, unobservable construct – must be presumed or inferred. By contrast, external referents are supplied exogenously by the points of the appeals scale itself. We adopt this distinction between internal and external referents, the former operationalized through past donation behavior, the latter through which specific points are included on the appeals scale.

While *intended* behavior is primarily determined by internal (i.e., latent, pre-existing) factors, manipulating the decision context through the appeals scale exerts influence on actual *choice*. This is particularly the case when a scale offering several discrete responses is presented. For example, the ability to anticipate donation response deteriorates as the appeal is placed ever further from what respondents 'intend' to donate (though this quantity can only be inferred). Indeed, Urbany, Bearden, and Weilbaker (1988) report that a donation request that is considered unrealistic (e.g., asking for far more than someone is willing to give), fails to appreciably influence donor behavior.

2.4. Internal and external referents

Compared with other salient dimensions of an alternative or product, price is typically of pivotal importance, for several reasons: the scale is continuous, expressed in readily-understood units, and is simple to use for comparison. It is important to consider the likely outcome of a donation appeal being out of line with what respondents anticipate, that is, when it is near to or far from, either above or below, their intended 'internal' donation (if any). To better understand how the appeals scale may affect overall pattern of donation behavior, it is instructive to consider a specific scenario.

An example. Although this notion is formalized later, a guiding example might involve a respondent who has some specific 'internal' donation reference value, in principle unknown to the researcher, in mind. For the purposes of illustration, we choose an amount of, approximately, \$42; [2] a cursory examination of actual donation data, however, indicates that such quantities are seldom expressed. Rather, respondents may choose to donate any of a number of other amounts, for example, nearby multiples of \$5 (\$40 or \$45), of \$10 (\$40 or \$50), of \$25 (\$25 or \$50) or the nearest points on an offered appeals scale. *The actual amount donated will depend, therefore, on how strongly respondents wish to conform to the set appeals scale, and how they compare upward and downward deviations from their internal reference donations*. In this example, respondents offered an appeals scale of {\$10, \$25, \$50, \$100} may deem \$50 'closer' than \$25 to the (latent) \$42 reference donation, and choose the former, whereas an appeals scale of \$10 multiples would result in a \$40 final donation; alternatively, respondents faced with an appeals scale with exclusively large amounts (e.g., all over \$100) may choose not to donate at all, or to choose a value

[2] We stress that it is not necessary that respondents hold such specific intended donation amounts in their heads, but merely that they can be systematically accounted for by prior donation behavior in terms of a point estimate. That is, such a value is 'internal': not explicitly expressed, a latent construct which the model must infer. Respondents may also restrict themselves to 'pre-rounded' quantities, a possibility for which the model allows.

not appearing on the scale, but close to the $42 (internal) reference donation. The main point here is not that respondents may have some 'internal' reference donation; rather, *if* they do, we wish to assess how it is influenced by the use of an explicit appeals scale, and the specific quantities which comprise it.

2.5. Framing and asymmetry in relation to reference values

Consider a situation in which what one wishes to donate, say, $75, lies midway between two points of the appeals scale, e.g., $50 and $100. If the appeals scale exerts influence to choose one of the scale values ($50 and $100) instead of the internal referent ($75), it is natural to ask whether the upward or downward pull is stronger, all else equal. Fashioning scales which encourage generosity is clearly an important question for fund-raisers.

Prior research on pricing amply demonstrates the importance of both asymmetry and framing. Studies of risky choice and mental accounting (Kahneman & Tversky, 1979; Thaler, 1985; Tversky & Kahneman, 1981) overwhelmingly show that 'gains' and 'losses' are accounted for differently; pronounced sensitivity to loss, for example, explains such behavior as aversion to extremes (Simonson & Tversky, 1992). The main consequence of this theory is an asymmetry in the effects of price variations: an increase in price (in relation to the reference point, i.e., a loss) produces a greater absolute effect than an identical decrease. Although there have been confirmatory empirical results (e.g., Hardie, Johnson, & Fader, 1993), asymmetric response may also be partially explained in aggregate by respondent heterogeneity (e.g., Gönül & Srinivasan, 1993). The model we develop explicitly allows for such asymmetric scale 'attraction' effects, relative to respondents' (inferred) reference points.

3. Conceptual model

In line with prior literature, we posit that the amount to be donated is contingent on the position of the (internal) reference donation in relation to the (external) 'request' conveyed by the appeals scale. Simply put, if the points on the scale don't mesh with what donors are prepared to give, they may take one of four actions relative to their internal reference donation: (1) donate more, (2) donate less, (3) ignore the scale entirely and donate what they had planned to, or (4) do not donate at all. The main purpose of the present study is to model how the input variables – prior donation behavior and the values constituting the appeals scale – serve to determine which course of action is taken, and to what extent.

The donation amount. We take the *internal reference* (or expected) *donation* to be determined by previous donations. The *external reference donation* is determined on the basis of the appeals scale and its mode of communication. [3] When the decision to

[3] It is very likely that prior donation behavior, and thus the external reference value, was influenced by the same variables as the current donation amount, as well as other unmeasured contextual factors. We do not account for this directly, but take the external reference, which is directly observed, as a benchmark, as in an ARMA model. We thank a reviewer for pointing this out.

donate is made, the amount of the donation will depend on the gap between the internal reference donation and the appeals scale points closest to it; when the gap is perceived as too great, the appeals scale is 'rejected' in favor of one composed of rounded values (Urbany et al., 1988). By "rounded values", we refer to the propensity for donations to be expressed in familiar denominations (as explained in the forthcoming empirical application). While it is tempting to eschew common denominations entirely, the data strongly support their use.

Scaling. In formulating a model, one must take account of the great degree of variation in donation amount across the donor pool, typically comprising three or more orders of magnitude. As is often the case with such large ranges, Weber's Law provides useful scaling guidelines. Originally put forth as a perceptual model, Weber's Law has been used as the basis of an economic theory of risk perception in stochastic choice; Kacelnik and Abreu (1998) show it can account for the Prospect-theoretic asymmetric effects typically manifested as risk-seeking for losses and risk-avoiding for gains. Zanker (1995), who employs the Weber Law in general perceptual context, argues for an explicitly logarithmic scale, a view we adopt in the present study. Given this scaling, all deviations in the model are *multiplicative*: a respondent planning to donate \$10, but asked to donate \$20, is modeled similarly to one planning to donate \$100, but asked for \$200.

3.1. Hypotheses

To sum up the discussion thus far, the main goals of the present paper are to measure, and thereby compare the relative influence of, three distinct effects.

- That of an amount's appearing on the scale presented to respondents (later termed *Scale*).
- That of an amount's being (perhaps a small multiple of) a common denomination (*Round*).
- That of attraction effects, the 'pull' of the closest scale points above and below (D_U and D_L).

To encompass the broadest set of effects attributable to the appeals scale, we will directly model (a transform of) the *overall donation cdf*, the cumulative distribution of donations. If the appeals scale does affect the overall donation distribution, it should draw the donation cdf towards the points of the appeals scale. By similar reasoning, and in line with the findings of Urbany et al. (1988), rounded values should also exert such a 'drawing' effect on the donation cdf. Because the appeals scale is communicated at the time of request, while respondents fall back on rounded values of their own volition, scale attraction effects should be stronger than those of rounded values. Hence, we hypothesize that, in terms of the overall donation distribution.

H$_1$. Appeals scale points (*Scale*) will draw donations toward them.
H$_2$. Rounded values (*Round*) will draw donations toward them.
H$_3$. The effects of *Scale* will be stronger than those of *Round*.

The final scale effects about which we anticipate distributional effects are the upper and lower deviations from adjacent scale points; we call these, respectively, D_U and D_L. Bearing in mind that we will model attraction effects on the donation cdf, we expect the lower scale point to exert a downward pull and the higher point an upward one. A prospect-theoretic interpretation would suggest that 'losses' – construed as (negative) deviations from upper scale points – will exert a stronger pull than gains of the same size. Finally, we expect neither of these 'comparative' effects to be as strong as those for *Scale* or *Round*. Formally, then, in terms of the donation distribution.

$\mathbf{H}_4$. Greater proximity to an *upper* scale point (D_U) will draw donations upward.
$\mathbf{H}_5$. Greater proximity to a *lower* scale point (D_L) will draw donations downward.
$\mathbf{H}_6$. Proximity to an upper scale point will exert a greater pull than that to a lower scale point.
$\mathbf{H}_7$. The joint effects of proximity to scale points (D_U and D_L taken together) will be weaker than those of either *Scale* or *Round*.

3.2. Empirical application: Data, models and variables

A large-scale experiment was carried out as part of a French charity's national fundraising campaign. The charity holds fund-raising drives several times per year, so it is able to maintain a database of donation requests, responses and donation amounts, as well as records of whether the donation appeal was made by mail or door-to-door canvassing. As detailed below, a typical fund drive can involve well over 100,000 donation appeals.

Segments and Groups. Three main donor *Segments* were made available for the present study: a segment of Irregular donors (IR), and two segments of Regular donors offering their donations through two different channels, one through Mailings (RM), the other typically through door-to-door Canvassing (RC). The regular donor segments (RM and RC) were further partitioned into *Groups* based on an index of their previous donation amounts, [4] three groups for each segment of regular donors (RM1–3 and RC1–3), whereas Irregular donors were partitioned into four groups (IR0–3), to account for smaller donations from this segment. These donation groups were determined by the charity, based on observed household-level behavior over a two-year period; neither the group divisions nor appeals scale determination were

[4] This index, and the resulting segment partitions, is standard practice for the charity, and consists of a simple average of the donations made by the household over the previous two full years. While not ideal from an experimental design perspective, it is the charity's long-trusted method of segmentation, and we had no choice but to accept it. Heterogeneity could not be directly measured, although the (HNLR) model employed below at least partly corrects for this. Ideally, some form of latent segmentation (e.g., Kamakura & Russell, 1989) might be performed, were full individual data histories available. Examination of other summary data measures (not reported) indicated that the classifications used by the charity had broad legitimacy.

Table 1
Appeals scales

Segments	Prior donation	Standard scale					
		100 FF	150 FF	250 FF	500 FF	1000 FF	Other
		Test scales					
				Center			
RM and RC	1	120 FF	180 FF	250 FF	350 FF	500 FF	Other
	2	120 FF	200 FF	350 FF	500 FF	750 FF	Other
	3	150 FF	250 FF	400 FF	600 FF	1000 FF	Other
IR	0	70 FF	120 FF	200 FF	300 FF	400 FF	Other
	1	120 FF	200 FF	350 FF	500 FF	750 FF	Other
	2	100 FF	150 FF	200 FF	350 FF	500 FF	Other
	3	150 FF	250 FF	400 FF	600 FF	1000 FF	Other

influenced by the design or goals of the present study. In a real sense, then, the data, despite its unusual richness, must be treated as something of an 'autonomous artifact'.

Data, appeals scales and subgroups. In order to better understand the role of the appeals scale on donation behavior, a *Standard* scale was employed for approximately half of donation requests, with those households chosen to receive the Standard scale randomly determined. This Standard scale was used extensively in prior donation drives, and was thus considered both reliable and useful for calibration purposes. Further, for the purpose of gauging scale effects, several alternative appeals scales were devised, henceforth referred to as *Test* scales, for the remaining households.

Each of the appeals scales used included five points, as listed in Table 1. The Standard scale, which was used for previous mail appeals, was identical for all groups (100, 150, 250, 500, 1000 FF and "other amount"). It is important to note that the small number of points on each scale allows us to literally define it by its (five) anchor points, rather than by appealing to synthetic summary indicators of its structure or distribution (such as centroid, range, density or progression; Hempel & Daniel, 1993).

Thus, there are 10 distinct groups, based on type of appeal (Mail or Canvassing), regularity of donation (IR or R) and anticipated donation. As described above, each of these 10 resulting groups was further subdivided based on a random binary draw, the first receiving the Standard scale and the second a Test scale, so that there are 20 distinct *Subgroups* to be accounted for in model estimation. [5] Summary

[5] For consistency, we henceforth use the term *Segments* to refer to the three methods of communication (*RM, RC, IR*), *Groups* to refer to *RM1–3, RC1–3, IR0–3*, and *Subgroups* to refer to the groups as split into the Standard and Test scales integral to the field experiment. Thus, there are three segments, 10 groups, and 20 subgroups.

Table 2
Average donation amounts and frequencies for each subgroup

Prior donation level		Segment IR Scales		Segment RM Scales		Segment RC Scales	
		Standard	Test	Standard	Test	Standard	Test
0	Mailings	12,390	12,439				
	Mean donation	134.3	132.1				
	Total yield (%)	5.5	5.8				
1	Mailings	5648	5666	10,810	10,812	3539	3941
	Mean donation	183.9	186.2	144.4	151.8	142.7	158.3
	Total yield (%)	11.5	11.0	42.4	40.6	13.6	13.6
2	Mailings	3623	3630	7560	7566	1836	1838
	Mean donation	272.1	268.8	267.3	274.6	285.3	285.6
	Total yield (%)	14.1	13.5	46.0	46.2	18.5	18.7
3	Mailings	2691	2702	7540	7544	1548	1551
	Mean donation	663.3	712.8	844.0	847.5	910.4	902.2
	Total yield (%)	12.0	13.7	48.1	45.6	20.9	19.5

statistics – number of mailings, response rate and mean donation amount – for each subgroup are given in Table 2.

We note three apparently anomalous features of the various Test scales. First, in each of the three donation segments (RM, RC, IR), only the Test scale for Group 3 dominates the Standard scale, and only the IR0 scale is in turn dominated by it. [6] While this in itself is not problematic for model estimation or subsequent interpretation, it does not afford the best basis for distinguishing the 10 groups' donation behavior. Second, there is what might be termed an inversion in the IR1 (120F, 200F, 350F, 500F, 750F) and IR2 (100F, 150F, 200F, 350F, 500F) Test scales; IR1's previous mean donation and yield (183.9F; 11.5%) are clearly less than the analogous quantities for IR2 (272.1F; 14.1%). So it is that in some groups, respondents are being asked to donate a greater quantity than those in other groups with historically lower donations. Again, while this presents no theoretical modeling challenges, it risks conjuring up greater within-group variation, making scaling effects harder to discern. Third, the means and medians of the various scales often vary significantly from the groupwise donation means of Table 2, again potentially inducing additional within-group variation in the subsequent donation data. While this may appear to be a serious problem, we point out that a nearly prefect match between prior donation behavior and the subsequent Test scale would amount to a confound, making it difficult to conclude anything concrete about the relative effects of prior behavior and scale attraction. As none of the scales was under experimental control, we therefore note that these three Test scale anomalies are problematic mainly in their effects on groupwise differences, biasing *against* the detection of scaling effects at the aggregate

[6] We use "domination" in the standard sense of being greater than or equal to corresponding points.

level. The HNLR regression methodology employed, as explained below, is designed to at least partially mitigate such heterogeneity problems across the 20 subgroups.

3.3. Models and variables

Because the data made available by the charity was pre-aggregated by donation subgroup, we approach the modeling of donation amount though a segment-level model. We therefore approach this as kernel density estimation, modeling the donation cdf. The cdf can take into account all twenty subgroups at once, in terms of *both* "whether" they donate (incidence) and "how much" (amount or quantity), as explained below (Section 3.5). In accordance with Weber's Law, all scale deviations are scaled logarithmically to allow the encoding of scale and attraction effects over a broad range of monetary amounts. Donations are classified in three ways: those exactly at a scale point, those at a round value, [7] and those falling between such values. By classifying donations in this manner, a picture of the cumulative donation distribution is obtained, and the resulting distribution parameters, described below, can then be estimated for each of the 10 donation groups (RM1–3, RC1–3, IR0–3).

3.4. Dependent variable: Density estimation of donation cdf by subgroup

The goal of density estimation is, as the name suggests, to recover an unknown density. This is typically accomplished by modeling the cumulative distribution function, which offers the distinct benefit of monotonicity. We follow the method set forth in Gershenfeld (1999), so that the (dependent) variable to be explained is the proportion of donations received falling into each of three types of predefined classes – a *scale* point, a *round* value, or an *intermediate* value – as follows.

Starting with gross values in French francs, donations are parceled into three types of classes: corresponding to anchor points used on one of the scales (e.g., 70, 100, 120, 180, 200, 250, 300, 350, 400, 500, 600, 750, 1000, as in Table 1), for round numbers not used on the scale (e.g., 50, 150, 1500, 2000, etc.), or in intermediate classes bounded by these values; in the last of these cases, the range is represented by geometric center (henceforth the 'log-center') of its bounds, in accordance with the logarithmic scaling employed throughout (so that, for example, endpoints of 100 and 400 would have a 'center' of 200, the geometric mean, not 250, the arithmetic one).

For example, the IR2 Test scale has adjacent scale points of 200F and 350F. A donation of exactly 200F or 350F would be classified as being on the *Scale* (thus yielding a 1 for the binary *Scale* variable), a donation of 250F would be classified

[7] While the notion of a 'round' quantity is seemingly intuitive, it is nonetheless difficult to make precise. 'Round' numbers are operationalized in the present paper as the face values of commonly used French currency notes or small integral multiples thereof. Based on the data used here, this functional definition accounts for all but a negligible proportion of off-scale donation amounts, in the sense that adding or removing additional 'rounded' values does not substantially alter any aspect of the analysis.

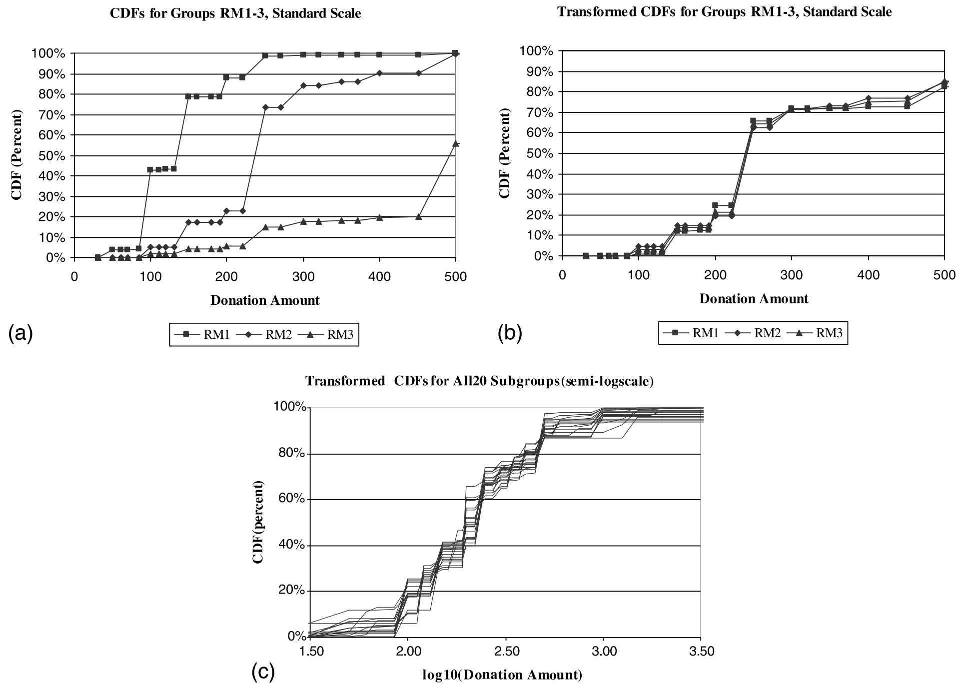

Fig. 1. (a) and (b) CDFs for three donation subgroups (RM1–3, Standard) untransformed and transformed; (c) transformed CDFs for all 20 donation subgroups.

as *Round*, and a donation of 240F would be classified as *Intermediate*. All donations can be classified in this manner, and it is from these so-called "bin counts" that the donation cdf, by subgroup, is built up (Gershenfeld, 1999; Silverman, 1986).

This process, and the scaling to be introduced to accommodate it, can be understood by considering the situation depicted in Fig. 1a and b.

The cdfs for the three different donation subgroups in the **RM** segment (RM1, RM2, RM3 for the 'Standard' scale) appear in Fig. 1a. It is quite obvious that these (sub)-groups have different patterns of donation behavior, an observation that would hold were we to reproduce a similar graph for the other two segments (*IR* and *RC*) or indeed for any three (sub)groups chosen at random. [8] In modeling appeals scale attraction effects, then, what we wish to do is to formulate a unified model which harnesses the statistical power of all 20 subgroups at once. To achieve this, then, we *use a linear transformation of the underlying scale to allow the separate subgroups' (transformed) cdfs to align as closely as possible*. So, if $y = G(x)$ is the cdf for any of the (sub)groups, we replace it by $G(\alpha + \beta G^{-1}(y))$, adopting the standard terms *location* and *scale* parameters

[8] We parenthesize "sub" in (sub)group to indicate that comparisons can be made between the Standard and Test scales *within* a group, or by aggregating the Standard and Test scales and comparing *across* groups.

for and, respectively. As shown in Fig. 1b, the resulting *transformed cdfs* retain their overall pattern of rises and falls, and conform as closely to one another as possible, up to linearity. It is possible to produce such transformed graphs for all twenty subgroups on a common axis, to offer a visual feel for the transformation process; this appears, in the semi-log scaling which typifies the model, in Fig. 1c.

3.5. Accounting for incidence rates across subgroups

Such a linear transformation does more than harness statistical power, which would, taken on its own, be decidedly ad hoc. Rather, this allows the model to *account* for the vastly different donation 'incidence' rates appearing in Table 2, which range from a low of 5.5% (IR0, Standard) to a high of 48.1% (RM3, Standard). It is important to realize that, considered from the point of view of the *cdf*, *conditioning on having made a donation is equivalent to linear transformation*. The reason for this is straightforward, if not immediately obvious. A 'mass point' consisting of all non-donations merely adds a (potentially large) discrete jump to the cdf at *0*; the *unconditional* and the *conditional* cdf – conditioning on donation – are therefore identical up to linearity, for all non-zero donations. So, similar to the use of brand-specific constants in scanner data models, which allow the brands to have different (aggregate) levels of intrinsic preference, the linear transform allows each group to have a *different intrinsic propensity to donate*. Simply put, the model allows for such between-group differences, but does not offer a theory for why they exist: different groups simply behave differently (and, given that the various groups were formed by the charity based on their prior donation behavior, this is hardly surprising). Thus, the by-group transformations help account for heterogeneity in both amount donated *and* donation incidence – both "whether" and "how much".

Because we will take errors to be normally distributed, the cumulative distribution for donations is modeled as in inverse normal transform, which maps the unit interval for the cdf into a presumably interval-scaled, unbounded variable on the whole real line. To avoid placing undue emphasis on distribution extremes, which exert disproportionate leverage, classes corresponding to cumulative frequencies under 2.5% or over 97.5% are not entered into the analysis; as usual, excluding high-leverage extremes serves to lower standard measures of model fit. In thus accounting for 95% of the observed donation frequency data, 381 data points remain, [9] across the twenty subgroups, representing the transformed cdf for donation amount.

[9] For each of the 20 subgroups, donations are parceled into one of 39 classes: 5 'scale' points, 14 'round' values related to French currency notes (i.e., $5 + 14 = 19$ discrete points) or one of the 20 'intermediate' intervals (18 between, 1 above, and 1 below). This yields $20 \times 39 = 780$ bins for density estimation. Trimming the smallest and greatest 2.5% of donations (to decrease outlier leverage, as explained) has the effect of 'removing' nearly half these bins, mainly the very large classes which received few, if any, donations (e.g., almost none over 2500F). We report that increasing the 2.5% quantity had little substantive effect on model results, while *decreasing* it caused strong distortions, so that sporadic large donations disproportionately determined the model's parameters. We stress that the model is accounting for 95% of donations, even if approximately half the potential data bins were trimmed (many of which would have been anyway, due to no donations at all falling into them).

3.6. Parameterization

There are several possibilities for specifying the structure of the model for the donation cdf, all of which can be thought of in terms of heterogeneity and parametric restrictions. All fall within the domain of fixed-effects models. For example, one approach would be to take each of the 20 subgroups as separate entities, estimating parameters for each; another would be to restrict the parameters for corresponding Test and Standard appeals scale subgroups to be identical, their difference measured by a single dummy variable. The approach taken here is to estimate several such models, searching for qualitative commonalities across them, and testing for differences between them. In this way, model implications can be separated from parametric specifications to as great an extent possible. Where there is little difference between models, the most parsimonious is presented. We first consider the manner in which explanatory variables enter the model.

Shape parameters. Donation class amounts are entered into the estimation in the form $\log[X - c_i]$, where c_i is a 'shape' parameter for each subgroup, and X is either a scale point, round value, or intermediate value (i.e., log-center of adjacent appeals scale bounds; see Footnote 9). We view these as shape-altering parameters required by the logarithmic scale, and not as strict scale 'zeros' or comparison values used by subjects. Their primary purpose is to avoid misspecifications arising from the specific shape of the *log* transform.

Presence of point on appeals scale (SCALE). Binary variable, 1 if the point is present on the scale offered, 0 otherwise. For each of the donation subgroups, there are therefore five non-zero values, corresponding to which points made up the scale for that subgroup (as in Table 1). By hypothesis (**H**$_1$), scale value attraction suggests a positive coefficient.

Round numbers (ROUND). Binary variable, 1 if the donation is a round number (whether or not it is on the scale), 0 otherwise. By hypothesis, the coefficient should be positive (**H**$_2$), but smaller than that for *Scale* (**H**$_3$).

Deviation in relation to adjacent anchor points. Consistent with logarithmic scaling, we define two variables, the upper and lower deviations (D_U and D_L), related to the logarithm of the *proportional* deviation of the donation class value or log-center (X) from the adjacent upper (A_U) and lower (A_L) anchor points (either *Scale* or *Round*):

$$D_U = \log[1 + (A_U - X)/A_U], \quad D_L = \log[1 + (X - A_L)/A_L].$$

Note that both D_U and D_L are non-negative and increasing in distance from their respective anchor points. To ensure that the associated distance is zero when X values are scale points or round values, 1 is added inside both log functions. That increased distances should result in marginally reduced effects is taken into account by the logarithmic scaling; it should be noted that, because of the logarithmic form for these variables, their coefficients have the effect of power transforms on the distance metric, so distance to anchor points is encoded multiplicatively, not linearly. Because the dependent variable is (a monotonic transform of) the cumulative donation distribution, we would anticipate the effect of being near an upper anchor as positive, and the opposite for a lower anchor. [10]

3.7. The heteroscedastic normal linear regression model

Because of the wide range of donation values and the degree of (anticipated) heterogeneity across donation (sub)groups, we cannot presume that error variances are constant, that is, that the homoscedastic assumptions underlying OLS will hold. To extract unbiased estimates of the coefficients of interest, we must estimate the relationship between the donation cdf and the various input variables without relying on such assumptions, allowing for a non-trivial degree of heteroscedasticity. To do so, we take a Bayesian approach, making use of the considerable power of the Gibbs sampling methodology. Geweke (1993) was among the first to develop such a model, which he termed *heteroscedastic normal linear regression* (HNLR). The original use for HNLR lay in explicitly modeling non-constant error variances and outliers, which is our motivation for applying it here. We briefly review the fundamentals of Geweke's (1993) formulation. Readers uninterested in Bayesian and technical modeling details can think of the model as a heteroscedastic form of OLS – coefficient interpretation is identical – and proceed to the following section. The HNLR model is given as

$$\begin{aligned} &y = X\beta + \epsilon, \\ &\epsilon \sim N[0, \sigma^2 W], \;\; W = \mathrm{diag}(w_1, w_2, \ldots, w_n), \\ &\sigma \propto 1/\sigma, \\ &r/w_i \sim ID\ \chi^2(r)/r, \\ &r \sim \Gamma(m, k), \end{aligned} \tag{1}$$

where y is an $n \times 1$ vector of dependent variable observations, X is an $n \times k$ matrix of explanatory variables, a multivariate normal prior is placed on β and a diffuse prior on σ. The parameters to be estimated are β, σ and the relative variance terms $(w_1, w_2, \ldots, w_n)$, which are assumed fixed but unknown. It may initially appear nonsensical to attempt to estimate the $n\{w_i\}$ parameters in addition to the $(k+1)$ 'ordinary' regression coefficients (β, σ), with only n observations at our disposal. However, the Bayesian approach assigns independent $\chi^2(r)/r$ prior distributions to the $\{w_i\}$ terms, so that they depend only on the hyperparameter r, and so only a single additional parameter is required to account for the n independent entries of the error covariance matrix.

The prior assigned to the w_i terms can be best understood by considering that its mean and variance are, respectively, 1 and $2/r$. Thus, large values of r reflect the

[10] Note that using the log-center of a donation class serves to increase linear distances to the upper anchor point, and therefore, relatively, lowers the coefficient for D_{U}. This scaling can only be claimed to bias in favor of the former hypothesis, so that finding for the latter constitutes fairly strong evidence in favor of asymmetry of the type expected on the basis of Prospect theory.

special limiting case $\epsilon \sim N[0, \sigma^2 I_n]$, and therefore indicate a prior belief that outliers and non-constant variances are not problematic for the data at hand. The model specification is completed by noting that values assigned to the hyperparameter r are accounted for by assigning it a $\Gamma(m, k)$ prior distribution, with mean m/k.

Geweke (1993) relates the conditional and posterior distributions for the HNLR model using a version of the Theil–Goldberger mixed estimator based on generalized least-squares. This is followed by the typical approach of Gibbs sampling, conditioning on unknown parameters as if their values were known. We refer the reader to Geweke's (1993) original paper for technical details and derivations of the three important conditional distributions, for $\beta|(\sigma, W)$, for σ, and for $W|(\beta, \sigma)$.

Given these three conditional densities, a Gibbs sampler can be formulated for the HNLR model in the usual fashion, assigning starting values for all parameters to be estimated, making successive draws, and computing posterior distributions. A large number of such 'passes' through the Gibbs sampler ensures that the true joint conditional distribution of all sampled parameters will be approached. All model computations were carried out in MATLAB, with 'burn-in' periods of 1000 passes through the chain, and 10,000 actual draws on which parameter estimates are based. A variety of chain-convergence diagnostics confirm that the model has in fact settled into its long-run state, and the asymptotically correct HNLR estimates have been achieved. [11]

3.8. Formal model statement

The model can then be stated formally as follows:

$$Y = \alpha_i + \beta_i \log[X - c_j] + \beta_{\mathrm{S}}\mathrm{Scale} + \beta_{\mathrm{R}}\mathrm{Round} + \beta_{\mathrm{U}}D_{\mathrm{U}} + \beta_{\mathrm{L}}D_{\mathrm{L}} + \epsilon,$$

where α_i and β_i are, respectively, 'location' and 'scale' parameters for each subgroup, i; c_j is a 'shape' parameter for each subgroup, j; β_{S}, β_{R}, β_{U} and β_{L} are the effects coefficients, common across subgroup; X is a scale point, round value or log-center of a subgroup; Y is $\Phi^{-1}[F]$; ϵ is as given in (1).

The final issue in specifying the model is the appropriate method to account for the 20 different subgroups, which here comprise three dimensions: solicitation method (i.e., the three *Segments*: RM, RC, IR), donation tendency (i.e., the 10 *Groups*:

[11] We appealed to five such diagnostics: (1) autocorrelation estimates; (2) Raftery–Lewis MCMC diagnostics, and Geweke's; (3) numerical standard errors (NSE); (4) relative numerical efficiency (RNE); and (5) Chi-squared means tests. The (time-series) autocorrelation estimates indicate the degree of dependence in the sequence of parameter draws; there was no evidence of significant autocorrelation at lags of 1, 5, 10 and 50. Raftery–Lewis diagnostics suggest a minimum chain length to achieve a desired degree (here 99%) accuracy for the parameters for M1; this was 937, well below the 10,000 used. NSE estimates were very small for all parameters in all models (almost all less than 0.001). RNE estimates indicate how close to IID the posterior sample was; they were indistinguishable from 1 (indicating near perfect independence). Finally, the Geweke Chi-squared tests compare parameter means for the first 20% and last 50% of the sample; in no case, for any of the models, could we reject the hypothesis of equality. In short, there is a strong indication that the MCMC routine converged to equilibrium. For additional detail, see Raftery and Lewis (1996).

RM1–3, RC1–3, IR0–3), and scale type offered (Standard and Test). There is no *a priori* reason to presume that there are not, therefore, 20 subgroups to be separately parameterized; nor is there a reason why the shape parameters $\{c_j\}$ need conform to the same subgroup breakdown as the location $\{\alpha_i\}$ and scale $\{\beta_i\}$ parameters.

A variety of models have therefore been fit, to the extent that they were sensible from an interpretive vantage point. The α_i and β_i parameters must, at the very least, be distinct in each of the 10 donation groups (RM1–3, RC1–3, IR0–3); models were therefore fit with either distinct values for the Standard and Test subgroups or constrained to a single value, requiring, respectively, either 20 or 10 parameters each for $\{\alpha_i\}$ and $\{\beta_i\}$. There is a wider range of possibilities for the shape parameters $\{c_j\}$, which can similarly take 20 or 10 values, but which may also take 2 (Test vs. Standard) or 1, the last presuming the same parameter for all subgroups. Thus, there are $2 \times 2 \times 4 = 16$ possible models to estimate, requiring anywhere from 25 through 64 parameters, depending on, respectively, location, scale and shape. All models were fit using through a constrained Newton–Raphson algorithm, the estimation proceeding in two stages, with $\{\alpha_i\}$ and $\{\beta_i\}$ fit through Gibbs sampling conditional on $\{c_j\}$. [12]

We report at the outset that we found *no* significant differences for using different values of $\{\alpha_i\}$ or $\{\beta_i\}$ for corresponding Standard and Test scale subgroups. This implies that there were no systematic differences in terms of donation *frequency* between the standard and test scale groups (given the discussion of the role of linear transformation in the model). Thus, we focus our attention solely on differences in donation *amount*, and therefore report results solely for the ‘parsimonious’ models, those which use 10 values each for $\{\alpha_i\}$ and $\{\beta_i\}$ (that is, which do not estimate different location and scale parameters for corresponding Test and Standard scale subgroups). The only remaining issue is the number of shape parameters $\{c_j\}$ used, either 1 (all subgroups identical), 2 (Standard and Test subgroups distinct), 10 (each *Group* distinct) or 20 (each *Subgroup* distinct). Formal model comparisons (available from the authors) suggested that the first option, with a single shape parameter (c_j), was preferred in terms of parsimony and overall fit; we refer to this model as **M1**.

4. Empirical results

Before referring to model implications, several findings are apparent from the raw results for the collection drive, listed in Table 2. The following relationships can be gleaned from the data without recourse to any specific model.

[12] While we do not wish to belabor purely econometric issues, it is important to understand exactly how the estimation proceeded. For each model parameterization, values for $\{c_j\}$ and r were chosen, and Gibbs sampling commenced for the remaining HNLR parameters conditional upon them. The values of $\{c_j\}$ and r were then optimized, based on the HNLR mean-squared-error estimate, through a Newton–Raphson type constrained optimization algorithm. Thus, the (Gibbs) sampling procedure yields posterior densities for $\{\alpha_i\}$ and $\{\beta_i\}$ conditional on the optimized point estimates for $\{c_j\}$ and r.

- As might be expected, the mean donation amount varies considerably by donation group, and rises with the mean level of prior donation. For the regular donor segment and the Standard scale, for example, the amount rises from 144 FF for the first level (RM1) to 267 FF, and then to 844 FF. Such mean differences are strongly significant across the three collection types (segments IR, RM, RC), and speak to the necessity of estimating distribution parameters $\{\alpha_i\}$ and $\{\beta_i\}$ for each of the 10 groups separately.
- The relationship between donation frequency and the mean level of prior donation is also positive, though not so regular as for the previous comparison. There is a systematic increase in donation frequency for the lower levels as they increase (i.e., 0–1 and 1–2), though it is unclear whether this is the case for the higher levels of prior donation (2–3).
- As also might be expected, segment-based differences suggest that the frequency of previous donation has a strong influence on present donation frequency, rising from 14.1% for irregular donors (IR2) to 46.0% for regular donors (RM2) (for the Standard scale at a high level of prior donation, in this example). By contrast, the impact of donation frequency on average donation amount is far weaker, and cannot be termed systematic.
- Scale effects are very limited at the aggregate level, and are in fact nearly negligible in relation to the effect sizes reported (later) from the model estimation. When the Test scale is displaced strictly upwards in relation to the Standard scale (group 3 of regular donors, and groups 1 and 3 of irregular donors), one expects an increase in the average donation amount and a decrease in the frequency of donations. While the data generally support these expectations, such differences between the Standard and Test scales are not always significant, and in several cases are opposite to what is expected, in particular for segment RC. *Specifically, in the absence of a particular model for these data, one may conclude that scale effects are either not present or that they actually run counter to hypothesis.* The model accomplishes this by taking account of the entire donation *distribution*, as opposed to common summary measures, such as the mean donation amount.

4.1. Model estimation results

Estimation results and relative fit statistics appear in Table 3; point estimates and standard errors for $\{\alpha_i, \beta_i\}$ appear in Table 4. First and foremost, all models fit exceptionally well, with the smallest R^2-adj value in excess of 0.975. We note that such a degree of fit is not entirely surprising, given that the model predicts a transformed value of the donation cdf as a function of a transform of the center of donation classes, an issue we will revisit and 'correct for' in a predictive context. Consequently, three points should be stressed. First, it is important to realize that, although the model can be conceptualized in purely predictive terms, its main use here is to (orthogonally) subtract systematic effects of the *monetary* scale from specific, attraction-based effects of the *appeals* scale and of rounded values. Second, while standard tests of *raw* differences in donation quantity and frequency between the

Table 3
Model comparisons ($n = 381$)

HNLR estimates and fit statistics, $n = 381$			
Model	M1: $\{\beta_S, \beta_R, \beta_L, \beta_U\}$	M2: $\beta_S = \beta_R$	M3: $\beta_L = -\beta_U$
β_S	0.292 (0.037)	0.248 (0.031)	0.299 (0.042)
β_R	0.180 (0.028)	0.248 (0.031)	0.192 (0.033)
β_L	−0.524 (0.037)	−0.547 (0.041)	±0.569 (0.071)
β_U	0.638 (0.093)	0.658 (0.102)	±0.569 (0.071)
r	3.87	2.980	4.710
R^2	0.9769	0.9736	0.9745
R^2-adj	0.9755	0.9722	0.9731
s_e	0.185	0.186	0.186
Log-likelihood	−407.47	−411.70	−410.99
-2Δ(LL) vs. M1		8.457 (1 df)	7.048 (1 df)
p vs. M1		0.004	0.008

Table 4
Estimation results for $\{\alpha_i\}$, $\{\beta_i\}$ and c, model M1

Scale, location and shape parameter estimates											
		RM			IR				RC		
		1	2	3	0	1	2	3	1	2	3
M1	α_i	−10.402	−15.782	−9.700	−8.134	−13.358	−14.270	−9.224	−10.797	−17.242	−9.805
	se(α_i)	0.556	0.427	0.175	0.393	0.559	0.372	0.213	0.627	0.459	0.214
	β_i	2.100	2.826	1.490	1.731	2.571	2.561	1.464	2.178	3.060	1.495
	se(β_i)	0.107	0.076	0.031	0.073	0.104	0.067	0.036	0.120	0.081	0.035
	c	−21.50 (All)									

Standard and Test scales fall short of significance, all such tests within the context of the various models are highly significant, as addressed below. Third, there is no sense in which the dependent variable is being used to predict itself; the shape of the donation cdf and the centers of the classes it comprises are related only in terms of predictable increase of one with the other, and strong systematic divergence from linearity is quite possible, although not observed here. [13]

4.2. Scale effects

The main goal of the research effort was stated at the outset as clarifying the effects of the appeals scale – conceptualized in terms of a set of discrete attraction points – on donor behavior. The first column of Table 3 suggests that all 'effects'

[13] We stress that the HNLR methodology accounts for heteroscedastic errors, *not* autocorrelated ones. Durbin–Watson tests indicate no significant degree of (first-order) serial error correlation for any of the models.

coefficients $\{\beta_S, \beta_R, \beta_U, \beta_L\}$ are both significant ($p < 0.0001$) and have the expected sign, offering direct support for hypotheses **H**$_1$ ($\beta_S > 0$), **H**$_2$ ($\beta_R > 0$), **H**$_4$ ($\beta_U > 0$) and **H**$_5$ ($\beta_L < 0$).

Two additional models allow us to test the hypotheses which compare parameter values, models M2 (for **H**$_3$: $\beta_S > \beta_R$) and M3 (for **H**$_6$: $\beta_U > -\beta_L$). This is accomplished by re-estimating the model under the equality constraints for the parameters in question, comparing the results through a likelihood-ratio test against (the nested model) M1. The last row of Table 3 indicates that each of these models differs significantly from the unconstrained M1, directly supporting **H**$_3$ and **H**$_6$. On the basis of these tests, we can conclude, respectively, that (1) the 'attractive' effects of a point being on the appeals scale is stronger than it being a rounded value, and (2) that proximity to an upper scale value pulls more strongly than to a lower scale value.

We point out the coefficients obtained for the scale effect are quite high ($\beta_S \approx 0.30$), far more so than that for the rounded values ($\beta_R \approx 0.18$); further, when the scale itself makes use of round values (that is, those of common denominations), effects accumulate additively, resulting in an uncommonly strong attraction. [14] However, these coefficients must not be interpreted as constant values (i.e., as having linear effects), as they are subject to an inverse normal transform to obtain a cdf value; near the scale mean, a point's appearing on the appeals scale translates into an increase of approximately 12%, while at the 10th or 90th percentiles this increase is approximately 5% (the corresponding figures for round values are and 7% and 3%). [15] We consider these to be rather pronounced effects, and the stability in relative and absolute magnitude across models is reassuring. Recall that, in the absence of any type of model, scale effects appeared to be negligible. When appropriately accounted for, it is fair to claim that as much as 12% of the cumulative donation distribution hinges on which points are chosen to appear on the appeals scale.

Both the upper and lower anchor points exert non-negligible attraction effects, as evidenced by their coefficients, $\beta_L \approx -0.50$, $\beta_U \approx 0.65$; it is important to realize that these serve as power-transforms on the absolute distance to scale anchors and, because of this (as well as the non-binary nature of the variables themselves) their values cannot be compared directly to those for the Scale and Round effects. The hypothesis of asymmetric attraction effects (**H**$_6$: $\beta_U > -\beta_L$), asserting that the upper anchor point will exert a stronger effect than the lower, is supported for these data. It is difficult to surmise whether this is an artifact of the scaling used in the problem, although we report that this effect was consistent across all models estimated, and appears robust to the specification of the other variables in the analysis. As discussed previously, this pattern of results is consistent with a Prospect-theoretic interpreta-

[14] The inclusion of a multiplicative term, that is, an interaction effect, did not alter the main effects appreciably. We note that the values of these coefficients are directly comparable due to both variables' being binary.

[15] For a unit change in parameter w, the normal transform of the cdf (F) indicates that, $\beta = [\mathrm{d}/\mathrm{d}w]\phi^{-1}[F]$, so that $\mathrm{d}F/\mathrm{d}w = \beta\varphi(\phi^{-1}[F])$. Here we have computed this for $F = 0.5, 0.1$ (and, by symmetry 0.9), where the values of $\varphi(\phi^{-1}[F])$ are $(2\pi)^{-1/2}$ and $(2\pi)^{-1/2}\exp[-(1.282)^2/2]$, or 0.399 and 0.175, respectively.

tion, and is similar in this way to that of Hardie et al. (1993); however, it would be premature to claim support for any specific theory on this basis alone. The remaining hypothesis, $\mathbf{H}_7$, will be examined in the context of prediction from a calibration to a hold-out sample.

4.3. Performance in hold-out sample

Data were provided by the charity to allow a test of the model's predictive abilities, in the form of a calibration and hold-out sample; [16] the analysis was repeated on the former using model M1. Several points should be raised in regard to a predictive context for the model. First, because the goal of the modeling effort was to measure scaling effects with aggregate data, within-donation-group variance is not an element of the prediction; in effect, heterogeneity is taken as a between-group construct, and we note that this should serve to make scaling and reference effects more difficult to discern (and was a main motivation for using the HNLR model). Second, the model should be expected to fit quite well, as location and scale parameters are estimated for each of the 10 donation groups. Calibration and hold-out fits should therefore be compared with a model which estimates *only* these 20 groupwise constants (plus the common scale parameter, c), to gauge the *marginal* explanatory power of the three effects (scale, round and deviation) the model accounts for.

Households were randomly assigned to either the calibration or hold-out group, with the 25-parameter model M1 estimated on the former group. [17] Results, in terms of a *Q–Q* plot depicting the predicted and actual cdf values, as well as their inverse-normal transforms, appear in Fig. 2.

Predictive accuracy is quite good; R^2-adj for the hold-out sample is 0.960. As evidenced by Fig. 2a, there is a relatively even dispersal of the actual values about the regression line; a χ^2-test for residual normality fails to reject ($p > 0.1$). However, a *Q–Q* plot for the predicted vs. the actual cdf, F, shows that, not very surprisingly, prediction is nearly perfect at the extremes, but less so for intermediate values. Because the model was formulated in terms of the (inverse normal) transformed cdf, Fig. 2a is arguably a better gauge of model accuracy than the *Q–Q* plot of Fig. 2b. Overall, it is reasonable to conclude that M1 achieves a high degree of predictive accuracy, and that the Scale and Round value attraction effects are apparently not artifacts of overfitting in the calibration sample.

[16] The results presented thus far are for the entire data set, the calibration and hold-out sample combined. Because we did not have access to the individual-level data, combining the two smaller sets offered greater stability in estimation, particularly for the donation bins (classes) in which few, if any, donations were received.

[17] As a check, this analysis was repeated for a less parsimonious model (34 parameters; one c value per segment), which provided marginally superior predictive accuracy. Qualitative results were not appreciably affected. Summary statistics of the type in Table 2 (not reported, but available from the authors) show the calibration and hold-out group to not differ appreciably.

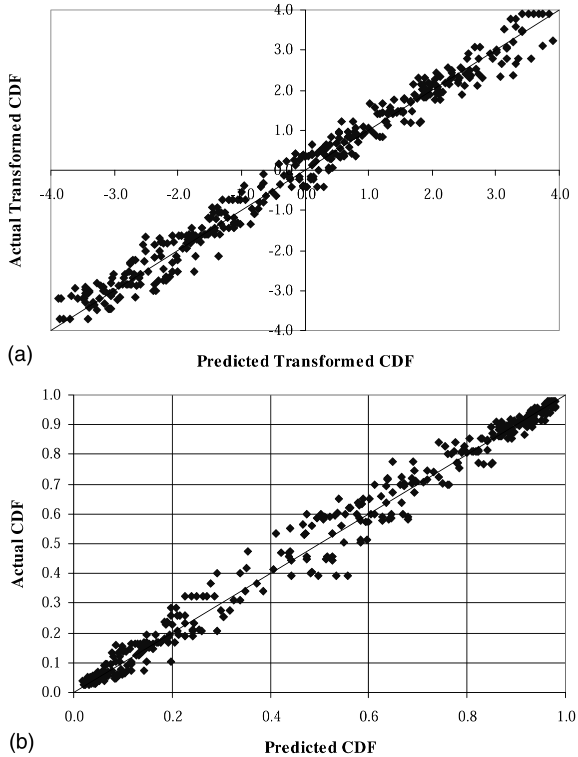

Fig. 2. (a) and (b) Out-of-sample prediction for CDF (F), and for $M^{-1}[F]$.

4.4. Model performance correcting for group differences

Because this is an aggregate model which estimates 2 per-donation-group parameters, it is reasonable to question whether the degree of within-sample or hold-out fit is in any way surprising. To gauge the extent to which the three constructs under study-scale, round values and anchor point deviation – increase fit and forecast accuracy, we re-estimate the model for the calibration data under all permutations of restricting each of these three effects to zero; this yields seven sub-models, as in Table 5 (note that the upper and lower attraction variables are only sensible if included or excluded together).

First, we note that the values of the Scale and Round coefficients are relatively insensitive to the presence of the other effects parameters. This suggests both that these effects are relatively robust to the model specification and, moreover, that they operate more or less independently of one another, so that an additive interpretation of their joint effects is reasonable. However, the values of the upper and lower anchor point attraction coefficients, β_L and β_U, appear to be strongly affected by the presence of the Scale variable; this is not surprising, given the role that Scale plays rela-

Table 5
Model comparisons for calibration and hold-out samples

Estimation results for various effects combinations (Calibration sample)											Hold-out
Effects	β_S	β_R	β_L	β_U	c	LL	p vs. 'Null'	p vs. 'Full'	R^2	R^2-adj	R^2-adj
$\{0, 0, 0, 0\}$					−3.1	−425.79	–	0.0000	0.9351	0.9310	0.9079
$\{\beta_S, 0, 0, 0\}$	0.372				−5.6	−421.88	**0.0052**	0.0000	0.9492	0.9460	0.9255
$\{0, \beta_R, 0, 0\}$		0.257			2.3	−423.15	**0.0216**	0.0000	0.9453	0.9415	0.9194
$\{0, 0, \beta_L, \beta_U\}$			−0.669	0.018	−41.5	−410.12	**0.0000**	0.0003	0.9699	0.9681	0.9516
$\{\beta_S, \beta_R, 0, 0\}$	0.338	0.168			−1.8	−421.11	0.0093	**0.0000**	0.9513	0.9482	0.9282
$\{\beta_S, 0, \beta_L, \beta_U\}$	0.305		−0.545	0.564	−25.5	−405.39	0.0000	**0.0077**	0.9743	0.9728	0.9571
$\{0, \beta_R, \beta_L, \beta_U\}$		0.203	−0.655	0.111	−31.9	−406.93	0.0000	**0.0014**	0.9730	0.9714	0.9555
$\{\beta_S, \beta_R, \beta_L, \beta_U\}$	0.288	0.187	−0.537	0.651	−17.9	−401.84	0.0000	–	0.9769	0.9755	0.9604

tive to D_L and D_U, which take a zero value at a Scale point. While all the various models, construed as combinations of the three effects, fit and predict rather well, it is clear that they do so significantly better in tandem, as evidenced by log-likelihood differences and corresponding χ^2 tests. In short, the model with all three effects is arguably superior to restricted submodels on both the calibration and hold-out samples. However, an important question remains unanswered: To what extent is model fit and predictive ability driven merely by between-donation-group differences?

To answer this question, it is instructive to consider the estimates in Table 5 as follows. The first line restricts all three effects (i.e., β_S, β_R and $\{\beta_L, \beta_U\}$) to be zero, so it is possible to consider this as a baseline model against which all the others can be tested: that is, once the between-group differences have been 'regressed out' in the form of the location, scale and shape parameters, do the three effects explain a significant portion of the remaining variation? Viewed in this way, the additional explanatory power offered by the three effects under study here can be measured and compared. We further note that while the 'no effects' model is nested within each of the others, each of them is in turn nested within the 'all effects' model, M1, the last line of Table 5.

It is immediately clear that *all* parametric combinations fit significantly better than the 'no effects' model, and that each fits *less* well than the 'all effects' model. The bolded p-values in Table 5 are particularly relevant: the first three compare the model with a single 'effect' – either Scale, Round or Deviation – to the 'no effects' model, while the remaining three compare the 'all effects' model to one which *does not* include one of the three effects. In this way, it becomes clear what the additional explanatory power is of each effect in the model, over and above the others (a similar comparison can be made of the models' R^2 values, although this does not indicate significance levels).

In each case, the likelihood-ratio test indicates that the scaling and attraction effects add very significantly to the model's explanatory power. Surprisingly, we find that hypothesis $\mathbf{H}_7$ is not supported: the model including only deviations (D_L and D_U) appears to be superior to the models which include only *Scale* or only *Round*, and the model *excluding* only D_L and D_U is inferior to those which exclude only *Scale* or only *Round*. We must interpret this carefully. Note that the coefficients for D_L and D_U in the model which contains those two variables only ($\beta_L = -0.669$, $\beta_U = 0.018$) differ markedly from their values in models which contain *Scale*, such as M1 ($\beta_L = -0.537$, $\beta_U = 0.651$). Thus, while D_L and D_U appear to be most important purely in terms of explanatory power, models which include them alone risk a serious potential misspecification.

As such, it must be stressed that the notion that the aggregate model (M1) fits well merely because it accounts for between-donation-group variation is misguided; in a marginal sense, each of Scale, Round and Deviation effects adds significantly to the model. This is clarified by the final two columns of Table 5, which list R^2-adj values for both the calibration and hold-out samples. Taking the 'no effects' model as the base – that is, regressing out between-group differences – the model with all three effects explains approximately 60% of the remaining variance in both the calibration and hold-out samples. [18]

Overall, there is clear evidence that the model both fits and predicts well, that each of the three constructs under study adds significantly to the interpretation of the aggregate donation pattern, that the strength of the effects are relatively robust to the model specification and that their actions can be considered distinct from one another. Further, the Scale coefficient's (β_S) stable value of approximately 0.3 translates, as before, into a 12% upward shift near the center of the cdf for values on the scale, while the analogous figure ($\beta_R \approx 0.19$) for Round is 7%, and moreover that these figures are essentially additive. These must be taken as very large proportional values by any standards, and constitute strong evidence of scaling and attraction effects. We stress once again that, in the absence of a model for such effects, responses – both in terms of donation frequency and quantity – on the Standard and Test scales appeared nearly identical, as in Table 2.

5. Conclusion

The objective of this study is to help fundraisers by more fully understanding the effects of scale anchor points on alternative evaluation and subsequent choice behavior. While the results are broadly convergent with those of several prior researchers, we further conclude that the framing of donation decisions is in fact influenced by external communications, as represented by the so-called appeals scale. Both points appearing on the scale and 'round' values displayed clear, strong attraction effects, effects obscured in simple comparisons of raw mean quantities received from groups given a Standard and a Test scale. These effects were not only pronounced, but asymmetric; deviations from upper scale values figured in as 'steeper' than equal (proportional) deviations from a lower scale value, relative to a reference point.

[18] This can be seen by computing the ratio of variation left unexplained, using either R^2 or R^2-adj. Taking values for the former, a comparison of the 'no effects' and 'all effects' (M1) models in the hold-out sample yields a ratio of $[1 - (1 - 0.9604)/(1 - 0.9079)]$, or 57%.

However, is must be admitted that the observed effects of scale variation appear to assume far less importance than those due to individual donor characteristics. As is found so frequently in Marketing studies, prior behavior is typically the best indication of future behavior, so that understanding the manifold effects of scale manipulation can only take place if donor heterogeneity, vis-à-vis prior behavior, has been appropriately accounted for. That is, though individual donor characteristics are not controllable, scale variation *can* be manipulated, so one must alter the latter while controlling for the former. In terms of our ability to generalize the model, the lack of range of the upper and lower anchors is problematic; we have not addressed how these should be optimally set, and it is an open question whether 'distant' values have an overall positive or negative effect on donation behavior. Recent work by Krishna, Wagner, and Yoon (2001), for example, suggests that the presence of even one product at an absurdly high price point can affect perceptions of normally-priced items. [19]

Averaging across the entire donor pool, scale manipulation appears to have at most modest effects on aggregate-level results, whether for overall yield or for the average donation. However, at the more detailed level of donation *groups*, effects induced by scale point displacement stand in bold relief. It may well be the case that the near absence of aggregate effects is the result of compensating individual-level effects at each of the scale points. Such hypotheses can be addressed by future studies in which *household-level* appeals scales are under direct experimental control (and thus facilitating the incorporation of heterogeneity in an endogenous manner).

The model presented here is intended to explain the displacement of donations induced by varying appeals scale anchor points. Estimated on donation data in a manner similar to the present study, the model should make it possible to select those scales which are potentially best-adapted to specific situations, including those outside an overt charitable context. For example, the model makes it possible, through interpolation, to estimate the effects of untested anchor points, and therefore suggests a path toward optimization. Modeling scale effects through a transform of the donation cdf offers a striking degree of fit and predictive accuracy for the collected donation data, and can be construed as a first step in formulating a model allowing better comprehension of price scaling effects.

The resulting model suggests the existence of three reference systems: one involving reference to points on an appeals scale, a second to the values of common denominations, and a third to an intended or anticipated donation amount. When donation appeals forego taking all three effects into account, a substantial proportion of potential donors may resist the 'default' options represented by the appeals scale and merely engage in habitual behavior, if they choose to donate at all.

[19] For example, the test scale's extreme (lower or higher) anchors never exceeded those of the standard scales. Because the data were silent on this issue, we cannot speculate on the effects of using scales whose extreme points extend far beyond standard practice. We thank a reviewer for noting this important data limitation.

A closer look at the donation behavior adjustments themselves offers several caveats regarding model limitations. First, adjustment is not in perfect conformity with the model for certain scale points, including several of the most frequent (such as 50, 100 or 120 FF). Second, because of anchor point attraction effects, the model can be made, under certain circumstances, to predict negative frequencies for certain scale values, notably those observed with low frequency. Lastly, the logarithmic transformation of donation values induces a risk of bias: a frequency error in the 5000 FF donor class does not have the same consequences as an error of the same size in the 50 FF class. To a large extent, these are all scaling artifacts, and are likely part and parcel of any attempt to produce a uniform explanation of donation behavior across a monetary scale of two or more orders of magnitude.

Our study suggests that fundraisers should not think of scales as simply a way to facilitate donations, but as an active tool in optimizing them. For example, donors could be contacted with a scale consisting solely of scale points and round values at least as high as what they've previously donated. The data suggest that, so long as the request is not greatly out of line with expectations, there will be a pronounced upward effect on donation quantity, without a commensurate decrease in frequency. Of course, such prognostications can only be verified in the context of studies designed specifically to do so.

Acknowledgements

The authors wish to thank Profs. Albert Bemmaor, Aradhna Krishna, Peter Lenk and Carolyn Yoon for their suggestions on an earlier version of this manuscript, as well as three anonymous reviewers. Authors are listed alphabetically. The usual disclaimer applies.

References

Abraham, M., & Bell, R. (1994). Encouraging charitable contributions – an examination of three models of door-in-the-face compliance. *Communication Research, 21*(2), 131–153.

DeJong, W., & Oopik, A. (1992). Effect of legitimizing small contributions and labeling potential donors as 'helpers' on response to a direct mail solicitation for charity. *Psychological Reports, 71*, 923–928.

Doob, A., & McLaughlin, D. (1989). Request size and donations to a good cause. *Journal of Applied Social Psychology, 19*(12), 1049–1056.

Fraser, C., Hite, R., & Sauer, P. (1988). Increasing contributions in solicitation campaigns: The use of large and small anchor points. *Journal of Consumer Research, 15*(September), 284–287.

Gershenfeld, N. A. (1999). *The nature of mathematical modeling* (pp. 169–174). Cambridge: Cambridge University Press.

Geweke, J. (1993). Bayesian Treatment of the Independent Student-t Linear Model. *Journal of Applied Econometrics, 8*, 19–40.

Gönül, F., & Srinivasan, K. (1993). Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues. *Marketing Science, 12*(3), 213–229.

Hardie, B. G., Johnson, E. J., & Fader, P. S. (1993). Modeling loss aversion and reference dependence effects on brand choice. *Marketing Science, 12*(4), 378–394.

Hempel, D., & Daniel, H. (1993). Framing dynamics: Measurement issues and perspectives. *Advances in Consumer Research, 20*, 273–279.

Kacelnik, A., & Abreu, F. B. E. (1998). Risky choice and Weber's Law. *Journal of Theoretical Biology, 194*(2), 289–298.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

Kalyanaram, G., & Winer, R. S. (1995). Empirical generalizations from reference price research. *Marketing Science, 14*(3), G161–G169.

Kamakura, W.A., & Russell, G.J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, *26*(November), 379–390.

Krishna, A., Wagner, M., & Yoon, C. (2001). The effects of extreme-priced products on consumer reservations prices. *Working paper*, University of Michigan Business School.

Mayhew, G. E., & Winer, R. S. (1992). An empirical analysis of internal and external reference prices using scanner data. *Journal of Consumer Research, 19*(1), 62–70.

Meyer, R., & Johnson, E. J. (1995). Empirical generalizations in the modeling of consumer choice. *Marketing Science, 14*(3), G180–G189.

Monroe, K. (1973). Buyers' subjective perceptions of price. *Journal of Marketing Research, 10*(February), 70–80.

Payne, J., Bettman, J., & Johnson, E. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology, 43*, 87–131.

Raftery, A. E., & Lewis, S. M. (1996). The number of iterations, convergence diagnostics and generic metropolis algorithms. In Gilks, Richardson, & Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Rajendran, K. N., & Tellis, G. J. (1994). Contextual and temporal components of reference price. *Journal of Marketing, 58*(1), 22–34.

Reingen, P. (1982). Test of a list procedure for inducing compliance with a request to donate money. *Journal of Applied Psychology, 67*(1), 110–118.

Rockwood, T. H., Sangster, R. L., & Dillman, D. A. (1997). The effect of response categories on questionnaire answers – context and mode effects. *Sociological Methods and Research, 26*(1), 118–140.

Schibrowsky, J., & Peltier, J. (1995). Decision frames and direct marketing offers: A field study in a fundraising Context. *Journal of Direct Marketing, 9*(1), 8–16.

Schwarz, N. (1999). Self-reports – how the questions shape the answers. *American Psychologist, 54*(2), 93–105.

Schwarz, N., Bless, H., Bohner, G., Harlacher, U., & Kellenbenz, M. (1991). Response scales as frames of reference – the impact of frequency-range on diagnostic judgments. *Applied Cognitive Psychology, 5*(1), 37–49.

Sherif, M., Taub, D., & Hovland, C. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology, 55*(2), 150–156.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. CRC Press.

Simonson, I., & Tversky, A. (1992). Choice in context: Trade-off contrast and extremeness aversion. *Journal of Marketing Research, 29*(August), 291–295.

Smith, S. (1980). Giving to charitable organizations: A behavioral review and framework for increasing commitment. In: J. C. Olson (Ed.), *Advances in consumer research* (*Vol. 7*, pp. 753–756). Ann Arbor, MI: The Association for Consumer Research.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science, 4*(Summer), 199–214.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Urbany, J., Bearden, W., & Weilbaker, D. (1988). The effect of plausible and exaggerated reference prices on consumer perceptions and price search. *Journal of Consumer Research, 15*(June), 95–110.

Weyant, J. (1996). Application of compliance techniques to direct-mail requests for charitable donations. *Psychology and Marketing, 13*(2), 157–170.

Weyant, J., & Smith, S. (1987). Getting more by asking for less: The effects of request size on donations of charity. *Journal of Applied Social Psychology, 17*(4), 392–400.

Zanker, J. M. (1995). Does motion perception follow Weber Law. *Perception, 24*(4), 363–372.
Zollinger, M. (1993). Le concept du prix de référence dans le comportement du consommateur: d'une revue de la littérature à l'élaboration d'un modèle prix de référence-acceptabilité. *Recherche et Applications en Marketing, 8*(2), 61–78.