Comparing Discrete Distributions: Survey Validation and Survey Experiments

Author(s): Kishore Gawande, Gina Yannitell Reinhardt, Carol L. Silva and Domonic Bearfield

Source: *Political Analysis*, Winter 2013, Vol. 21, No. 1 (Winter 2013), pp. 70-85

Published by: Cambridge University Press on behalf of the Society for Political Methodology

Stable URL: https://www.jstor.org/stable/23359693

Advance Access publication December 4, 2012

Political Analysis (2013) 21:70–85
doi:10.1093/pan/mps036

Comparing Discrete Distributions: Survey Validation and Survey Experiments

Kishore Gawande
Bush School of Government, Texas A&M University, College Station, TX 77843-4220
e-mail: kgawande@tamu.edu (corresponding author)

Gina Yannitell Reinhardt
Bush School of Government, Texas A&M University, College Station, TX 77843-4220
e-mail: greinhardt@bushschool.tamu.edu

Carol L. Silva
Center for Applied Social Research, Department of Political Science, University of Oklahoma, 455 W. Lindsey, Room 205, Norman, OK 73019-2001
e-mail: clsilva@ou.edu

Domonic Bearfield
Bush School of Government, Texas A&M University, College Station, TX 77843-4220
e-mail: dbearfield@bushschool.tamu.edu

Edited by R. Michael Alvarez

Field survey experiments often measure amorphous concepts in discretely ordered categories, with postsurvey analytics that fail to account for the discrete attributes of the data. This article demonstrates the use of discrete distribution tests, specifically the chi-square test and the discrete Kolmogorov–Smirnov (KS) test, as simple devices for comparing and analyzing ordered responses typically found in surveys. In Monte Carlo simulations, we find the discrete KS test to have more power than the chi-square test when distributions are right or left skewed, regardless of the sample size or the number of alternatives. The discrete KS test has at least as much power as the chi-square, and sometimes more so, when distributions are bi-modal or approximately uniform and samples are small. After deriving rules of usage for the two tests, we implement them in two cases typical of survey analysis. Using our own data collected after Hurricanes Katrina and Rita, we employ our rules to both validate and assess treatment effects in a natural experimental setting.

1 Introduction

Survey experiments fuse the external validity of surveys with the internal validity of experiments into one rigorous research design. As the tool proliferates across the social sciences, survey experiments are increasingly used to measure amorphous concepts, from mental health and mood in public health literature (Bourque, Siegel, and Shoaf 2002; Bourque et al. 2006) to trust in government, blame attribution, and feelings about government policy in public opinion studies (see Malhotra and Margalit 2010; Hetherington and Suhay 2011). Without naturally quantifiable scales, these concepts are typically self-assessed and measured in discretely ordered categories on 4-, 5-, or 7-point scales.

Authors' note: We appreciate Matt Henderson's valuable assistance with coding the Internet-based survey. We thank Elisabeth Gerber, Jennifer Jerit, Jason Barabas, Janet Box-Steffensmeier, Andrew Sobel, Gary Miller, Randall Calvert, Andrew Martin, and Itai Sened for helpful comments. The editor, R. Michael Alvarez, and an anonymous referee suggested revisions that improved the article significantly. For replication data, code, and instructions, see Gawande et al. (2012). Supplementary materials for this article are available on the *Political Analysis* website.

Comparisons of these concepts between groups, within groups, or in difference-in-difference (between- and within-group) analyses are crucial to postsurvey analysis. Comparing the sample to the population can validate the sample, whereas evaluating the sample pretest, with respect to the posttest distribution, assesses time and experimental effects. Critical to survey experiments fielded at one point in time, measuring the distributions in the treatment group vis-à-vis the control group, can assess whether grouping was truly randomized and can evaluate the effect of the treatment. Much of this analysis, although performed on discretely ordered responses, remains implemented with techniques designed for continuous, binomial, or normally distributed variables.

In this article, we evaluate an underutilized method, the discrete Kolmogorov–Smirnov test (discrete KS test), and compare its performance as a distribution test to the more commonly used chi-square test of independence, via Monte Carlo simulations. Each test compares entire distributions, but the discrete KS test accounts for order, whereas the chi-square test does not. We gauge the size and power of the two tests and suggest rules for their implementation. We find the discrete KS test has more power than the chi-square test when distributions are right or left skewed, regardless of the sample size or the number of alternatives. The discrete KS test has at least as much power as the chi-square, and sometimes more, when distributions are bi-modal or approximately uniform and samples are small.

We then apply the tests to a unique public opinion data set of hurricane disaster victims and observers, illustrating the utility of the discrete KS test, and our suggested rules, as tools for survey validation and experimental analysis. Our work is timely because of a recent surge in experimental survey work on public opinion, and on experimental surveys in general. Field survey experiments with small samples, and those analyzing questions with skewed distributions, should be evaluated with an exact distribution test like the discrete KS test, rather than an asymptotic test of independence like the chi-square test. Our application clarifies the advantages of the discrete KS test and emphasizes the importance of using the distribution test appropriate for one's data. Using improper tests is not only theoretically wrong, it can lead to overrejection of null hypotheses of equal distributions. A researcher might erroneously find her survey sample to be invalid when it was actually valid, or worse, might find a treatment to have an effect when in fact it did not. With proper implementation, the discrete KS allows inferences that illuminate the effects of natural experiments, exogenous shocks, and treatments in detail previously unseen.

2 Discretely Ordered Distributions

In 1932, Rensis Likert proposed measuring attitudes by allowing respondents to choose from a set of discrete, ordered options. Commonly known for a list of five alternatives (Strongly Approve, Approve, Undecided, Disapprove, and Strongly Disapprove), his example allowed for flexibility in both the number and names of alternatives (Likert 1932). It is typical to find scales with three or seven items, or without both positive and negative responses. Likert's ordinal scale revealed his intent to capture underlying continuous variables representing respondent opinions, attitudes, or feelings—variables which, if possible to be measured objectively, would be captured on an interval scale, at best (Clason and Dormody 1994, 31).

While questions with "feelings thermometers" and other sets of discretely ordered alternatives intuitively make sense to respondents, analysis of them has been problematic. Standard equality-of-means or equality-of-proportions tests lose explanatory heft when the number of alternatives expands and distributions are not normal. In 1939, the KS test (Smirnov 1939) was introduced as a method for comparing an empirical distribution with a theoretical target distribution, or for comparing two independent distributions (Panchenko 2003). Initially popular for its ability to compare entire distributions rather than only means or proportions, this exact test did not require large samples to be valid. The exactness was especially popular with scholars forced to rely on small samples. Important, too, was the power of the test: the higher the power, the less likely the researcher was to commit Type II error, and the failure to reject null hypotheses that should be rejected.

In the 1960s, Lilliefors (1967, 1969) showed the continuous KS test to have greater statistical power than the parametric chi-square test. Later, Horn (1977) showed the power superiority of the

continuous KS over the chi-square for any sample size. The same year, Pettitt and Stephens (1977) demonstrated the power advantages of the KS test when modified and used to analyze discontinuous distributions (Gleser 1985), making it a firm choice for discretely ordered Likert response variables. Yet, the chi-square retained its popularity, despite its inability to account for data order (Wood and Altavela 1978). And although the chi-square is an asymptotic test that requires a large sample size to be valid, scholars continued to use it with small sample analyses because it was computationally efficient. The KS test fell out of regular usage.

By the 1990s, the KS test had "virtually disappeared" from applied research and social science statistics texts (Wilcox 1997, 16). Analysis of Likert-scaled variables continued via chi-square and difference-of-means testing, despite the results of such analysis being "more a function of sample size than respondent attitude" (Clason and Dormody 1994, 34). As computing advanced, the continuous KS test resurged as a tool for comparing distributions, but few researchers used a version modified for discretely ordered variables; most employed the continuous KS test despite its impropriety for ordinal variables. Wilcox (1997) conceded that one could try to examine various features of distributions or data, such as measures of central tendency or distributional shape, to determine under which conditions the KS test would have a power advantage over other methods, but lamented that "Currently, no such strategy can be recommended, and it seems unlikely that such a method will ever be found" (p. 16).

With the goals of both improving survey analysis and advancing methodological understanding of the KS and chi-square tests, we present Monte Carlo simulations aimed at defining precisely the strategy Wilcox seeks. We vary sample size, distribution shape, and the number of response alternatives among samples while assessing performance of the discrete KS and chi-square tests. We gauge the performance of these tests on (i) size, the probability of wrongly rejecting a null hypothesis that is true (thus committing Type I error); and (ii) power, the probability of correctly rejecting a null hypothesis that is false (thus avoiding Type II error).

3 Discrete KS and Chi-Square Tests: Statistical Size and Power

3.1 *KS Test*

The KS test (Smirnov 1939, translated by Darling 1960; also Massey 1950) determines whether a sample of independently drawn observations comes from a population with a particular distribution, or whether two samples come from the same underlying distribution. Consider the continuous KS test. Let $F(X)$ denote the true c.d.f. of the random variable X. It shows the true probability that a realization of the random variable X has a value less than or equal to x:$F(x) = P(X \leq x)$.

Suppose we have n realizations of the random variable, arranged as ordered data points x_1, $x_2, \ldots, x_n$. Let $F_n(X)$ denote the empirical c.d.f. of the data, defined as

$$F_n(x) \equiv P_n(X \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \tag{1}$$

where $I(X_i \leq x)$ is a binary indicator that takes on the value one, if the ith realization of the random variable takes a value less than or equal to x, and zero otherwise. Thus, the empirical c.d.f. measures the proportion of the sample with values less than or equal to x. By the law of large numbers, if $F(X)$ is the true c.d.f. from which the data are drawn, then the empirical c.d.f. $F_n(x)$ converges uniformly to $F(x)$. That is, as the sample size increases, the proportion of the sample taking values in the interval $(-\infty, x)$ is equal to the true probability $F(x)$ at *all* points x. One way of writing this, reflected in the KS statistic, is

$$\sup_{x \in \Re} |F_n(x) - F(x)| \to 0, \tag{2}$$

where the supremum, the smallest number equal to or greater than every number in the set, is taken over all possible real number values x. That is, the largest difference between the true probability and its measured counterpart vanishes in probability as the sample size grows infinitely large.

Suppose we wish to test the equality of the underlying distribution of the two independent samples. Let the first sample have true c.d.f. $F(x)$ and the second $G(x)$. For now, suppose both samples have size n. The simple hypothesis is that the samples are drawn from the same underlying distribution. The formal null and alternative hypotheses are, respectively,

$$H_0 : F(x) = G(x) \tag{3}$$

$$H_1 : F(x) \neq G(x). \tag{4}$$

The KS statistic is (e.g., Canner 1975; Panchenko 2003)

$$\mathrm{KS} = n^{1/2} \sup_x |F_n(x) - G_n(x)|, \tag{5}$$

where $F_n(x)$ and $G_n(x)$ are the corresponding empirical c.d.f.'s. If the sample sizes are different, m and n for the first and second samples are

$$\mathrm{KS} = \left(\frac{\mathrm{mn}}{(m+n)}\right)^{1/2} \sup_x |F_m(x) - G_n(x)|. \tag{6}$$

3.1.1 Discrete KS test

The distribution of the continuous KS statistic in equations (6) and (5) does not depend on the underlying true c.d.f.'s, and thus is "distribution free." This property is no longer true when data come from a discrete underlying c.d.f.[1] Using the continuous KS test can provide quite inaccurate inferences for discrete distributions. We thus present and implement a more suitable KS test for discrete distributions, attributable to Wood and Altavela (1978). The method requires four steps to test the hypotheses in equations (3) and (4):

- Step 1: Calculate the two empirical c.d.f.'s; and
- Step 2: Compute the KS test statistic:

$$\mathrm{KS}_d = n^{1/2} \max_x |F_n(x) - G_n(x)| \tag{7}$$

Note the slightly different coefficient from equation (5). The max is taken over the r discontinuity points of the c.d.f. (the number of discrete data points, or alternatives to a survey question).

- Step 3. Calculate the exact confidence level via simulation as follows. Generate a multivariate normal random variable Z of dimension $r - 1$, where[2]

$$E(Z_i) = 0 \tag{8}$$

and

$$E(Z_i Z_j) = \min\{\mathrm{G}(x_i), \mathrm{G}(x_j)\} - \mathrm{G}(x_i)\mathrm{G}(x_j). \tag{9}$$

This step overcomes the distribution problem of the discrete KS c.d.f. In equation (9), i and j index elements of the $(r-1) \times (r-1)$ covariance matrix of Z.

- Step 4. Perform 1,000,000 draws of Z. Calculate the empirical frequency that the draws fall in the region $\{Z_1 < KS_d, Z_2 < KS_d, \ldots, Z_{r-1} < KS_d\}$. That is, compute the proportion of

[1] As the formula below will show, the value of the discrete KS statistic is determined by calculating the maximum over the points of the distribution. Therefore, the number of discrete categories can change the value of the discrete KS statistic. For continuous functions, this is not the case.

[2] Since the c.d.f. $= 1$ at the final data point, the effective number of data points is $r - 1$.

the draws in which *each* of the $(r-1)$ Z_i's for a draw is less than the calculated KS statistic in Step 2. This is the exact (small sample) significance or p-value of the test.[3]

3.2 *Pearson's Goodness-of-Fit (Chi-Square) Test*

Pearson's chi-square goodness-of-fit statistic (see Horn 1977) is used to test if the empirical distribution function conforms to another theoretical distribution (Kempthorne 1967). Suppose we have N observations that fall into r categories, and the observed number of items in category i is n_i^{observed}, $\sum_{i=1}^{r} n_i^{\text{observed}} = N$. The chi-square goodness-of-fit test statistic is computed as

$$\chi_1^2 = \sum_{i=1}^{r} \frac{\left(n_i^{observed} - n_i^{expected}\right)^2}{n_i^{expected}}, \tag{10}$$

where n_i^{expected} is the expected value of n_i according to the theoretical distribution.

This test is approximate, not exact.[4] Its distribution is an approximation to the theoretical chi-square distribution with $(r-1)$ degrees of freedom (df). Its computational ease made it popular before the advent of cheap, fast computing, when it was often preferred to the (exact) multinomial and likelihood ratio tests.

The two-sample counterpart to this test statistic compares two empirical c.d.f.'s (and drops the term "observed," implied in both samples),[5] where n_i and m_i are the empirical frequencies in category i, and the constants κ_1 and κ_2 are adjustments for different sample sizes, $\kappa_1 = \sqrt{M/N}$ and $\kappa_2 = \sqrt{N/M}$:

$$\chi_2^2 = \sum_{i=1}^{r} \frac{(\kappa_1 n_i - \kappa_2 m_i)^2}{n_i + m_i}. \tag{11}$$

3.3 *Monte Carlo Simulations: Size and Power*

Hilbe (2009) details the creation of synthetic ordered logit data.[6] Our simulations (Gawande et al. 2012) are based on the data generating process (DGP) in Hilbe (2010, Section 4). The continuous underlying data y_i are generated by the process

$$y_{1i} = \beta_{11} x_{1i} + \beta_{21} x_{2i} + \phi_1 e_i, i = 1, \ldots, n_1. \tag{12}$$

Each observation i is drawn independently with

- $x_{1i} = 3U[0, 1] + 1$,
- $x_{2i} = 2U[0, 1] - 1$, and
- $e_i = \Phi^{-1}(U[0, 1])$,

where $U[0, 1]$ indicates a draw from the standard uniform distribution and Φ^{-1} is the inverse of the standard normal c.d.f. Two categorical variables Y_{1i}^C are defined, one with four categories and another with ten categories:

- The four-category variable is defined as $Y_{1i}^C = 1$ if $Y_{1i} \leq 2$, $Y_{1i}^C = 2$ if $2 < Y_{1i} \leq 3$, $Y_{1i}^C = 3$ if $3 < Y_{1i} \leq 4$, $Y_{1i}^C = 4$ if $4 < Y_{1i}$.

[3] A finite sample correction increases the p-value for small samples, modifying equation (7) as in Wood and Altavela (1978, 238):

$$\text{KS}_d = n^{\frac{1}{2}} \max_x |\text{F}_m(x) - \text{G}_n(x)| - \frac{1}{\sqrt{mn/(m+n)}}. \tag{14}$$

[4] It also approximates the exact multinomial goodness-of-fit test (see, e.g., Horn 1977).

[5] The two-sample test is also known as the "test of independence."

[6] Hilbe (2007) compares discrete count data distributions by simulating synthetic Poisson and Negative Binomial random variables.

- The ten-category variable is defined as $Y_{1i}^C = 1$ if $Y_{1i} \leq 1$, $Y_{1i}^C = 2$ if $1 < Y_{1i} \leq 2$, $Y_{1i}^C = 3$ if $2 < Y_{1i} \leq 2.5$, $Y_{1i}^C = 4$ if $2.5 < Y_{1i} \leq 3$, ..., $Y_{1i}^C = 10$ if $5.5 < Y_{1i}$.

Four different types of distributions are generated as follows:

- Right skewed: $\beta_{11} = 0.80, \beta_{21} = 0.75; \phi_1 = 1$.
- Left skewed: $\beta_{11} = 1.70, \beta_{21} = 1.75; \phi_1 = 1$.
- Bimodal at the corners: $\beta_{11} = \beta_{21} = 1; \phi_1 = 4$.
- Approximately uniform across categories: $\beta_{11} = \beta_{21} = 1; \phi_1 = 1$.

The size-of-test (Type I error) simulation is conducted for a specific data distribution as follows. Two samples of size n_1 are drawn from the *same* distribution. Their chi-square and discrete KS statistics and p-values are computed. This is repeated 1000 times. The proportion of times the chi-square test and the discrete KS test each reject the null hypothesis at significance level α ($= 0.025$ and 0.050) is then reported. If the proportion exceeds α, then the test overrejects the hypothesis of equality of the distributions. If the proportion is less than α, then the test underrejects the hypothesis of equality of the distributions.

Table 1 reports the actual proportion of the 1000 simulations in which the null hypothesis that the two samples came from the same population is rejected. This is the empirical size of the test with theoretical size or significance level α. Regardless of whether a distribution has four categories or ten categories, the chi-square test rejection rate is close to α. This is consistent across all four types of distributions. We see why the chi-square test has stood the test of time. The discrete KS test rejects at approximately twice the rate of α, regardless of the type of distribution. While not reported, the same simulations using smaller samples of 500 observations yielded similar results.

To test power, data drawn from the DGP, equation (13), are compared against data drawn from the DGP:

$$y_{2i} = \beta_{12} x_{1i} + \beta_{22} x_{2i} + \phi_2 e_i, i = 1, \ldots, n_2, \tag{13}$$

whose parameters are locally different:

- Right skewed: $\beta_{11} = 0.80, \beta_{12} = 0.75; \beta_{21} = \beta_{22} = 1.25; \phi_1 = \phi_2 = 1$.
- Left skewed: $\beta_{11} = 1.70, \beta_{12} = 1.75; \beta_{21} = \beta_{22} = 1.25; \phi_1 = \phi_2 = 1$.
- Bi-modal at the corners: $\beta_{11} = \beta_{12} = 1; \beta_{21} = \beta_{22} = 1; \phi_1 = 3.5, \phi_2 = 4$.
- Approximately uniform across categories: $\beta_{11} = \beta_{12} = 1; \beta_{21} = 1.5, \beta_{22} = 1; \phi_1 = \phi_2 = 1$.

Two samples of size 500 and 5000 are drawn from the two distributions, respectively.[7] From 1000 trials, we report the proportion of times the chi-square test and the discrete KS test correctly reject the null hypothesis that their population distributions are the same, at significance level α ($= 0.025$ and 0.050). The higher the proportion of rejections, the greater the probability of rejecting the (false) null hypothesis, and the greater the test's power.

Consider the small sample cases (Table 2). The discrete KS test has greater power than the chi-square test when the frequency distribution is either right (Simulation 1) or left skewed (simulation 2). For these two types of distributions, the discrete KS test has nearly twice the power of the chi-square test when the distribution has four categories, and this factor only increases as the number of categories climbs. When the frequency distribution is either U-shaped (Simulation 3) or balanced (Simulation 4), the tests have approximately the same power.

With distributions derived from large samples (5000), both tests have much greater power than with small samples. With a right-skewed frequency distribution, the power of both tests exceeds 90% with four categories and 84% with ten categories. The discrete KS test still outperforms the chi-square test when the frequency distributions are left or right skewed. This is not surprising. The KS test statistic is based on the single category at which the (absolute) difference between the two

[7] For purposes of this article, small-sample size = 500, large-sample size = over 5000. The small-sample size conforms to a size seen in regional subsamples of the GSSs. The large-sample size conforms to a size seen in regional subsamples of the MEPS.

Table 1 Size of chi-square and discrete KS tests

Simulation	Significance α	Four categories		Ten categories	
		χ2	KS	χ2	KS
Simulation 1	0.05	0.058	0.103	0.047	0.079
(Right skewed)	0.025	0.025	0.050	0.021	0.041
Simulation 2	0.05	0.051	0.084	0.04	0.095
(Left skewed)	0.025	0.022	0.044	0.019	0.045
Simulation 3	0.05	0.059	0.096	0.042	0.098
(Bimodal at corners)	0.025	0.023	0.052	0.023	0.054
Simulation 4	0.05	0.055	0.109	0.048	0.092
(Balanced)	0.025	0.027	0.052	0.026	0.053

Probability of rejecting the hypothesis that the underlying distributions are the same for the two samples.
Notes. The underlying distributions are the *same* by design.
1. Continuous DGP (from Hilbe 2010): $y=\beta_1 x_1+\beta_2 x_2+\varphi e$, where the random variables (r.v.) x_1, x_2, and e are generated as $x_1=3\ U[0,1]+1$; $x_2=2\ U[0,1]-1$; $e=\Phi^{-1}(U[0,1])$, where $U[0,1]$ is a uniform r.v. and Φ^{-1} is the inverse of the standard normal c.d.f.
2. The categorical variable y^C is then defined as $y^C=1$, if $y\leq 2$; $y^C=2$, if $2<y\leq 3$; $y^C=3$, if $3<y\leq 4$; and $y^C=4$, if $4<y$.
For ten categories: $y^C=1$, if $y\leq 1$; $y^C=2$, if $1<y\leq 2$; and then consecutively in 0.5 intervals until $y^C=10$, if $5.5<y$.
3. Four simulations: Simulation 1: $\beta_1=0.75$; $\beta_2=1.25$; $\varphi=1$ (left skewed distribution of y and y^C);
Simulation 2: $\beta_1=1.75$; $\beta_2=1.25$; $\varphi=1$ (right skewed distribution of y and y^C);
Simulation 3: $\beta_1=1$; $\beta_2=1$; $\varphi=4$ (bimodal distribution of y and y^C with corner modes); and
Simulation 4: $\beta_1=1$; $\beta_2=1$; $\varphi=1$ (balanced distribution of y and y^C).
Each simulation has two samples of 5000 observations each. They are used to test the hypothesis that they came from the same population.
4. The table reports the empirical probability that the tests reject H_0 at the significance level α, where
H_0: $F_1=F_2$ (both distributions came from the same population) versus H_1: $F_1\neq F_2$. The null hypothesis H_0 is true by construction (both distributions come from the same DGP).
5. Simulations with a smaller sample of 500 observations each showed results similar to those reported above.

Table 2 Power of chi-square and discrete KS tests

Simulation	Significance α	Small sample: $n_1=n_2=500$				Large sample: $n_1=n_2=5000$			
		Four categories		Ten categories		Four categories		Ten categories	
		χ2	KS	χ2	KS	χ2	KS	χ2	KS
Simulation 1	0.05	0.157	0.291	0.107	0.311	0.964	0.982	0.901	0.989
$\beta_{11}=0.80$; $\beta_{12}=0.75$	0.025	0.094	0.197	0.059	0.207	0.934	0.966	0.841	0.969
Simulation 2	0.05	0.083	0.158	0.080	0.192	0.462	0.701	0.533	0.849
$\beta_{11}=1.70$; $\beta_{12}=1.75$	0.025	0.046	0.099	0.046	0.117	0.358	0.579	0.402	0.766
Simulation 3	0.05	0.125	0.129	0.135	0.173	0.747	0.532	0.959	0.934
$\varphi_1=3.5$; $\varphi_2=4$.	0.025	0.067	0.068	0.080	0.096	0.638	0.353	0.922	0.850
Simulation 4	0.05	0.145	0.146	0.074	0.124	0.937	0.831	0.994	0.943
$\beta_{21}=1.5$; $\beta_{22}=1$	0.025	0.088	0.084	0.034	0.061	0.890	0.696	0.984	0.853

Probability of rejecting the hypothesis that the underlying distributions are the same for the two samples.
Notes. The underlying distributions are locally *different* by design.
1. Notes 1–3 to Table 1 apply here with different DGPs for generating data distributions F_1 and F_2. Distribution F_1 of y^C from the DGP: $y=\beta_{11}x_1+\beta_{21}x_2+\varphi_1 e$ and distribution F_2 of y^C from the DGP: $y=\beta_{12}x_1+\beta_{22}x_2+\varphi_2 e$.
2. The table reports the power of the test, the empirical probability of rejecting H_0 at the significance level α, where
H_0: $F_1=F_2$ (both distributions came from the same population) versus H_1: $F_1\neq F_2$. The null hypothesis H_0 is false by construction:
Simulation 1: $\beta_{11}=0.80$, $\beta_{12}=0.75$; $\beta_{21}=\beta_{22}=1.25$; $\varphi_1=\varphi_2=1$;
Simulation 2: $\beta_{11}=1.70$, $\beta_{12}=1.75$; $\beta_{21}=\beta_{22}=1.25$; $\varphi_1=\varphi_2=1$;
Simulation 3: $\beta_{11}=\beta_{12}=1$; $\beta_{21}=\beta_{22}=1$; $\varphi_1=3.5$, $\varphi_2=4$; and
Simulation 4: $\beta_{11}=\beta_{12}=1$; $\beta_{21}=1.5$; $\beta_{22}=1$; $\varphi_1=\varphi_2=4$.
Large-sample simulations generate distributions from 5000 draws, and small-sample simulations from 500 draws.

distributions is at the maximum (the ordering of categories is what allows focus on the single category to judge equality of the complete distribution). For skewed distributions generated by different DGPs, the maximum difference likely occurs at or near their peak, which is captured by the KS test.

Surprisingly, with U-shaped and balanced frequency distributions, the chi-square test has greater power. For such distributions, the KS test underrejects. A possible reason for this is that since the chi-square test aggregates the squared differences in the two distributions across all categories, it

Table 3 Distribution Check Rules: Suggested rules for use of discrete K–S versus chi-square tests

	Size (Type I error)		*Power (= 1 − Type II error)*	
	Few categories	*Many categories*	*Few categories*	*Many categories*
Right skewed frequency distribution				
Large sample (>5000)	Chi-square	Chi-square	Discrete KS	Discrete KS
Small sample (<500)	Chi-square	Chi-square	Discrete KS	Discrete KS
Left skewed frequency distribution				
Large sample (>5000)	Chi-square	Chi-square	Discrete KS	Discrete KS
Small sample (<500)	Chi-square	Chi-square	Discrete KS	Discrete KS
Bimodal (at the corners) frequency distribution				
Large sample (>5000)	Chi-square	Chi-square	Chi-square	Chi-square
Small sample (<500)	Chi-square	Chi-square	Both ok	Discrete KS
Balanced (approximately uniform) frequency distribution				
Large sample (>5000)	Chi-square	Chi-square	Chi-square	Chi-square
Small sample (<500)	Chi-square	Chi-square	Both ok	Discrete KS

captures information about significant differences at categories other than (and including) the category at which the maximum difference occurs. With bi-modal distributions with differences in both modes, the KS test may fail to reject in instances where the single-node maximum difference is borderline. But since the chi-square test aggregates the squared difference in both modes, it has a greater chance of correctly rejecting the equality of distributions.

3.3.1 Rules

These results suggest rules (Table 3) for when each test is more appropriate. We recommend the chi-square test when there is low tolerance for Type I error, and the discrete KS test when there is low tolerance for Type II error, especially when samples are small. An added advantage in small samples is that the discrete KS test is an exact test. When the power of the test matters, therefore, the discrete KS test should be preferred over the chi-square test in small samples regardless of the distribution. When comparing two samples of different sizes, the smaller sample size must dictate which rule applies.

When the frequency distribution is left or right skewed, such as might occur in surveys on a 5- or 7-point scale when most respondents report that they are either on the "satisfied" or "unsatisfied" side of a midpoint, the discrete KS test outperforms the chi-square test on power regardless of sample size. If both samples have over 5000 observations,[8] the chi-square test has better power properties than the discrete KS test, if the frequency distribution is U-shaped or balanced. There is surprisingly little that differentiates the tests based on the number of categories.

4 Application: Validation and Survey Experiment

In August and September 2005, Hurricanes Katrina and Rita caused a combined $141 billion in damage (2005 U.S. dollars), and resulted in 1952 deaths (Lott et al. 2012). One year later, we fielded a survey in *hurricane-threatened regions* of the United States: counties/parishes from Texas to North Carolina that border the coast or are separated from the coast by not more than one other county/parish.[9] The survey was administered via the Internet by Survey Sampling International, a sampling firm with a preregistered sampling frame around the world, and yielded 7024 respondents. About 22% of the sample was directly affected by either Hurricane Katrina or Rita and evacuated. Of these, 902 remained displaced from their homes a year after Katrina, when our survey was taken. This "treatment" group is the target group in our continuing research on issues such as political

[8]The MEPS, collected multiple times over the course of each year, polls more than 35,000 respondents nationwide.
[9]Further details of this and all surveys mentioned here can be found in the Supplementary Materials.

trust, government effectiveness, and willingness to return to their pre-hurricane locations. Our "control" group is composed of 4689 respondents who resided in these hurricane-threatened regions, but had not evacuated for a hurricane in 2004 or later.

Our survey is captured at one point in time after a series of exogenous shocks. Feelings of morbidity, or of mental or emotional stability, are seldom the focus in political science analysis, yet such feelings are expected side effects of catastrophic events among survivors and evacuees (see Bourque et al. 2006; Bourque, Siegel, and Shoaf 2002). Although sociologists have investigated the impact of these effects on sociological processes (Quarantelli 2002), little is known or understood about what bearing these feelings may have on one's opinions of politics or political leaders. Nationwide surveys, such as the Medical Expenditure Panel Study (MEPS) and the General Social Surveys (GSS), regularly ask questions about mental health and morbidity, enabling us to compare our survey to target populations along these dimensions.

Different from demographic questions typically used for survey validation (see Barabas and Jerit 2010, Appendix p. 3; Jerit 2009), mental and emotional assessments are more likely to require time and introspection from respondents. If attitudes and feelings are the focus of public opinion studies, attitudes and feelings should then have a place in the validation of samples designed to test those foci. Mental status, being under stress or feeling depressed, may shape important opinions about trust in government, quality of emergency response, or blame attribution. Since the MEPS and GSS surveys capture opinion before the shocks, comparing the distribution of responses on those surveys with the distribution of responses on the hurricane survey sample can help determine whether the mental or emotional makeup of the sample is different from the pre-shock population. If there is no difference between the sample and the population, evidence supports a sample of respondents representative of the population. If there is a difference between the sample and the population, further work can explore whether that difference is based on exposure to the natural treatment, and what inferences can be made. We use the discrete KS and the chi-square tests to attempt to validate our survey.

A mental health question used for validation has the additional benefit of strengthening inferences about the treatment when distributions between treatments versus control groups are compared in subsequent analyses. Survival through catastrophic events, such as hurricanes, nuclear accidents, or terrorist attacks, is a plausible reason why disaster survivors might hold different opinions from disaster observers. Loss of reason, mental instability, or the inability to make decisions are plausible reasons as well. If the distribution of responses to a mental health question in the control group is distinct from that in the treatment groups, then this difference can be the source, as GSS and MEPS surveys indicate, of different beliefs and therefore different decisions made by the affected population. How to resolve the morbidity side effects can then be an important policy question.

4.1 *Validation Tests*

4.1.1 MEPS Data

Berrens et al. (2003) find that with appropriate controls, samples from Internet surveys are sufficiently generalizable to provide robust inferences. Can a survey of displaced persons, restricted to respondents who had access to the Internet *before* Katrina hit, aspire to the same levels of validity?[10] To find out, we included in our hurricane survey a question from the MEPS, a nationally representative survey of the U.S. population administered via household computer-assisted interviews (MEPS 2007). The question asks respondents for a categorical self-assessment of their mental health (MENTALH). We use the MEPS South region sample consisting of 14,108 observations to compare with responses to the same question on the hurricane survey. Since the MEPS distribution

[10]Previous studies show reason for caution. In their 2007 Pew survey of around 2050 persons in the United States, Witt, Best, and Rainie (2008) found that 92% of those in the 18–20 age group and over 80% in the 31–62 age group had access to the Internet, but only between 46% and 57% of the over-63 age group did. Over 90% of those with (even partial) college education accessed the Internet. African Americans were somewhat less likely than Whites to use the Internet. Thus, the selection bias appears to be small for the under-63 age group, but not so for the over-63 age group.

Table 4 KS and chi-square tests

Empirical df for MENTALH w/five categories				*Empirical df for MENTALH w/four categories*			
Category	*Hurricane Sample*	*MEPS Sample*	*|Difference|*	*Category*	*Hurricane Sample*	*MEPS Sample*	*|Difference|*
Excellent	0.3910	0.4008	0.0098	Excellent	0.3910	0.4008	0.0098
Very Good	0.7438	0.6700	**0.0738**	*Good-to-Very Good*	0.9187	0.9250	**0.0063**
Good	0.9187	0.9250	0.0063	Fair	0.9825	0.9834	0.0009
Fair	0.9825	0.9834	0.0009	Poor	1	1	-
Poor	1	1	-				
Sample size (N)	6984	14,018			6984	14018	
KS statistic		5.031**				0.662	
KS p-value		0.000				0.093	
Chi-square statistic		244.3**				3.701	
Chi-square *p*-value		0.000				0.296	

Notes. H_0: MENTALH (*Mental Health Status*) distribution for hurricane-affected population and the MEPS population are equal.
1. Hurricane survey taken in September 2006.
2. MEPS taken during 2005. Sample size = 14,018 (Southern Region).
3. The more aggregate four-category comparison combines the "Very Good" and "Good" categories into a single "Good-to-Very-Good" category.
4. KS statistic from equation (7).
5. Chi-squared statistic calculated as in equation (12).
6. *p*-value is the probability that the null hypothesis of equality of the two distributions is rejected, given the null is true, that is, the (two-tailed) significance level of the test statistic.

is right skewed, the Table 3 Distribution Check Rules suggest using the discrete KS test; it has greater power than the chi-square test. Table 4 reports results comparing our 2006 Hurricane Survey with the 2005 Panel 10, Round 3 MEPS responses.

The maximal difference between the two empirical c.d.f.'s for MENTALH is 0.0738, and occurs at the category "Very Good." The KS statistic (5.031) rejects the hypothesis that the data come from the same underlying distributions. The distinction between "Good" and "Very Good" may be one of the degree and subjectivity, and it is possible that the difference between them is artificial rather than substantive. Evidence of this possibility is the difference between their empirical frequencies. The observed frequencies in the "Very Good" category are of a similar magnitude as the "Good" category, but the sign is reversed, indicating that the subjective boundary separating the two categories is fuzzy. Appendix Figure A1 depicts this difference. We therefore argue for aggregating the two categories into a single category.

The right panel of Table 4 tests the equality of distributions with "Very Good" and "Good" aggregated into a single category. Appendix Figure A2 indicates that such an aggregation makes the distributions more alike. More formally, the hypothesis that the two distributions are equal cannot be rejected at 1% by either the discrete KS test or the chi-square test. We take this as affirming the validity of the hurricane sample. We also tested the control group (the subsample that was not affected by Katrina or Rita or any other hurricane in the past 3 years) against the MEPS sample, and failed to reject equality of distribution. We take this as further evidence of the validity of the survey, since we expect the unaffected sample to respond as would a pre-shock sample.[11]

4.1.2 GSS Data

In the hurricane survey, we posed a mental health status question similar to that asked in the 2000 GSS conducted by the National Opinion Research Center at the University of Chicago (Smith et al. 2011). The GSS asked how often the respondent felt downhearted and depressed in the past 4 weeks

[11]The KS statistic is 0.175 with a *p*-value of 0.518. Full results are available from the authors.

(DOWNBLUE).[12] The GSS sample of southern states is 509, invoking the small sample Distribution Check Rules. The GSS offers respondents six answer choices, and our survey offered only five. In order to implement the discrete KS test, we must aggregate two of the GSS categories into one.[13] Table 5 reports the empirical distribution functions and results of the tests of equality of the two DOWNBLUE distributions. In the left panel of Table 5, we present results in which GSS options 2 and 3, given as "Most of the Time" and "A Good Bit of the Time," are aggregated into our category 2 that we label as "Most/Good Bit of the Time." The KS statistic is 0.773 with an exact p-value of 0.086. The chi-square test also affirms that the hypothesis that the two samples came from the same underlying population distribution cannot be rejected at the 1% or 5% levels.

What if the options were collapsed differently? As a sensitivity analysis, we next aggregate options 3 and 4, which in the GSS are "A Good Bit of the Time" and "Some of the Time," into our category 3 that we label as "Good Bit/Some of the Time." Since the maximal difference is the same (0.0365 in the last column) with both aggregations, the redistribution of probability mass leaves the KS test statistic unchanged. Both aggregations support the hypothesis of equality of distributions. Interestingly, the shifting of probability mass leads the chi-square test to reject the hypothesis of equality of distributions. The GSS distribution becomes tri-modal—three categories have significant probability mass. Table 3 indicates that with such distributions, the chi-square test result has greater power because of the shape of the distribution. If the distribution were skewed, the result could reverse—the discrete KS test could reject but not the chi-square test. In such a case, Table 3 would favor using the discrete KS test for its greater power.

We posed two other questions found in the GSS. They asked: (1) how much of the time (in the past 4 weeks) the respondent had been limited in work or other activities as a result of their physical health (LIMITEDP) and (2) how much of the time (in the past 4 weeks) the respondent accomplished less than they would like as a result of emotional problems, such as feeling depressed or anxious (DIDLESSE). Since we desired granularity in the responses, we offered five answer choices to these questions, ranging from "All of the Time" to "None of the Time." The 2000 GSS had posed these as binary yes–no questions. In order to compare distributions, we aggregated our five answer choices to binary yes–no outcomes: the top three choices 5, 4, and 3, starting with "All of the Time," were rolled into "yes," and the last two choices 2 and 1 (ending with "None of the Time") were rolled into "no" (Table 6). The KS test cannot reject equality of the LIMITEDP distributions or the equality of the DIDLESSE distributions from the two surveys.

Using the discrete KS test on distributions with binary outcomes may appear to be overkill, if simple difference-in-means tests achieved the same result. Whether the KS test has greater power over difference-in-means tests with binary outcomes should be investigated in future research using simulation. We proceeded under the assumption that the KS test is a better option. To investigate the sensitivity of the tests to aggregation, we rolled only the top two choices 5 and 4 into "yes" and the other three into "no." Under this arrangement, the discrete KS test rejected the equality of distributions for both variables.

4.2 *A Survey Experiment*

The hurricane survey was designed as an experiment in which the impact on the treated could be distinguished from the impact on a control group. As described above, the treated represent the population that was evacuated and remained displaced 1 year after Katrina. The control group had not evacuated for a hurricane in 2004 or later. The substance of the survey is its ability to quantify the impact of the shock on the affected population as measured by a set of trust-in-government, desire-to-return, and effectiveness-of-emergency-response questions. The survey experiment allows

[12]The 2000 GSS may be accessed at http://sda.berkeley.edu/D3/GSS06/Doc/gs06x07.htm. Clicking on the variable DOWNBLUE provides the empirical frequency table.

[13]We offer this primarily as an example of the "small sample" case of distribution checks. We acknowledge the discrepancy in question wording but do not make it our focus here. We do note that such discrepancies arise in surveys where constraints do not allow time to achieve identical wording before launching a survey, or where a survey may be validated retrospectively.

Table 5 KS and chi-square tests

	Empirical df for DOWNBLUE (mapping 1)				*Empirical for DOWNBLUE (mapping 2)*			
	Category	*Hurricane Sample*	*GSS Sample*	*|Difference|*	*Category*	*Hurricane Sample*	*GSS Sample*	*|Difference|*
1	All of the time	0.0145	0.0178	0.0033	1 All of the time	0.0145	0.0178	0.0033
2	Most/Good Bit of the time	0.0777	0.0889	0.0112	2 Most of the time	0.0777	0.0435	0.0342
3	Some of the time	0.2332	0.2688	0.0356	3 Good Bit/Some of the time	0.2332	0.2688	0.0356
4	Little bit of the time	0.6097	0.6462	**0.0365**	4 Little bit of the time	0.6097	0.6462	**0.0365**
5	None of the time	1	1	-	5 None of the time	1	1	-
	Sample size (N)	7016	509			7016	509	
	KS statistic		0.773				0.772	
	KS p-value		0.086				0.084	
	Chi-square statistics		4.3				27.72**	
	Chi-square p-value		0.365				0.000	

Notes. H_0: DOWNBLUE distribution(s) for hurricane-affected population and GSS population are equal.
1. Hurricane survey taken in September 2006. Respondents chose one answer from five ordered choices ranging from (1) all of the time to (5) none of the time. The middle categories were not explicitly described. They are labeled using GSS descriptors.
2. Responses from question posed by GSS in 2000. The sample of Southern Region was taken. Respondents chose one answer from the following six categorical choices: (1) all of the time, (2) most of the time, (3) a good bit of the time, (4) some of the time, (5) a little bit of the time, and (6) none of the time
3. Aggregation 1: GSS categories (2) and (3) were mapped into hurricane survey category (2).
Aggregation 2: GSS categories (3) and (4) were mapped into hurricane category (3).
4. KS statistic from equation (7).
5. Chi-squared statistic calculated as in equation (12).
6. p-value is the probability that the null hypothesis of equality of the two distributions is rejected, given the null is true, that is, the (two-tailed) significance level of the test statistic.
7. ** and * indicate rejection of test of equality of distribution at the 1% and 5% levels, respectively.

Table 6 KS and chi-square tests: binary outcomes

	LIMITEDP		*DIDLESSE*	
	Limited physical activity		*Accomplished less*	
Category	*Hurricane sample*	*MEPS sample*	*Hurricane sample*	*MEPS sample*
Yes	0.2213	0.2259	0.2003	0.1965
No	1	1	1	1
Sample size (N)	7016	509	7016	509
KS statistic	0.077		0.060	
KS p-value	0.427		0.440	
Chi-square statistics	0.058		0.046	
Chi-square p-value	0.810		0.830	

Notes. H_0: Distribution(s) for hurricane-affected population and GSS population are equal.
1. Responses from questions posed by GSS in 2000.
2. First two rows of the numbers indicate the empirical distributions for the four variables.
3. LIMITEDP and DIDLESSE are binary in the GSS survey. Both variables have five categories in our hurricane survey. They are aggregated down to binary, as described in the Supplementary Appendix.
4. KS statistic from equation (7).
5. Chi-squared statistic calculated as in equation (12).
6. p-value is the probability that the null hypothesis of equality of the two distributions is rejected, given the null is true, that is, the (two-tailed) significance level of the test statistic.
7. ** and * indicate rejection of test of equality of distribution at the 1% and 5% levels, respectively.

a difference-in-differences analysis of responses to these important questions. In this final section, we demonstrate a "difference-in-distributions" analysis of the same questions we have examined thus far. If the distributions on the mental health questions are different, it reveals a reason why we should expect the response of the affected group to be different from the unaffected group on substantive research questions down the line.

Table 7 KS and chi-square tests: Hurricane Survey Experiment

	MENTALH (five category) Mental health status			*DOWNBLUE Felt down*	
Category	*Treatment Displaced*	*Control Non-evacuee*	*Category*	*Treatment Displaced*	*Control Non-evacuee*
Excellent	0.3068	0.4009	All of the time	0.0370	0.0920
Very Good	0.6439	0.7599	Most of the time	0.1445	0.0625
Good	0.8611	0.9280	Some of the time	0.3774	0.2039
Fair	0.9619	0.9865	A little of the time	0.738	0.5813
Poor	1	1	None of the time	1	1
Sample size (N)	893	4669		893	4669
KS statistic	3.157**			4.770**	
KS p-value	0.000			0.000	
Chi-square statistics	75.02**			165.3**	
Chi-square p-value	0.000			0.000	

Notes. H_0: Distribution(s) for not-evacuated population and hurricane-displaced population are equal.
1. For DOWNBLUE, hurricane survey respondents chose one answer from five ordered choices ranging from (1) all of the time to (5) none of the time. The middle categories were not explicitly described, and are labeled here descriptively rather than numerically.
2. Hurricane treatment group remained displaced as of September 2006. Hurricane control group experienced no evacuation during 2004–06, but resided in hurricane-prone regions.
3. KS statistic from equation (7).
4. Chi-squared statistic calculated as in equation (12).
5. p-value is the probability that the null hypothesis of equality of the two distributions is rejected, given the null is true, that is, the (two-tailed) significance level of the test statistic.
6. ** and * indicate rejection of test of equality of distribution at the 1% and 5% levels, respectively.

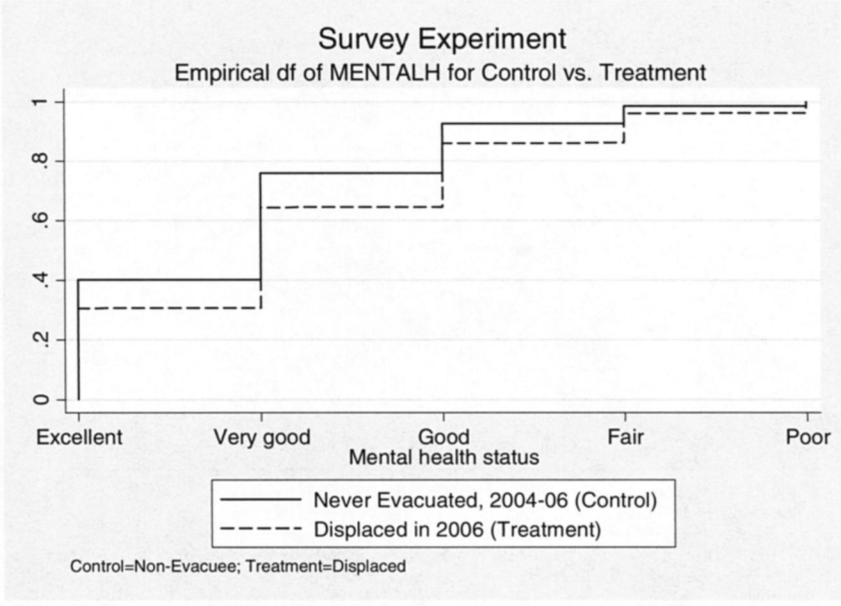

Fig. 1 Empirical distribution functions of PERCEIVED MENTAL HEALTH STATUS in hurricane sample for Treatment (Displaced as of 2006) and Control (Never experienced evacuation during 2004–06) groups.

Table 7 reports results. Figure 1 shows the distributions to indeed be different. The probability that the treatment group enjoys good mental health is lower than that for the control group. The absolute maximal difference of 0.116 at the "Very Good" category, the empirical probability that in 2006 the treated enjoyed a MENTALH outcome better than or equal to "Very Good," was 11.6% lower than for the control. This difference is too large for the two distributions to be considered alike, and the KS test strongly rejects their equality.

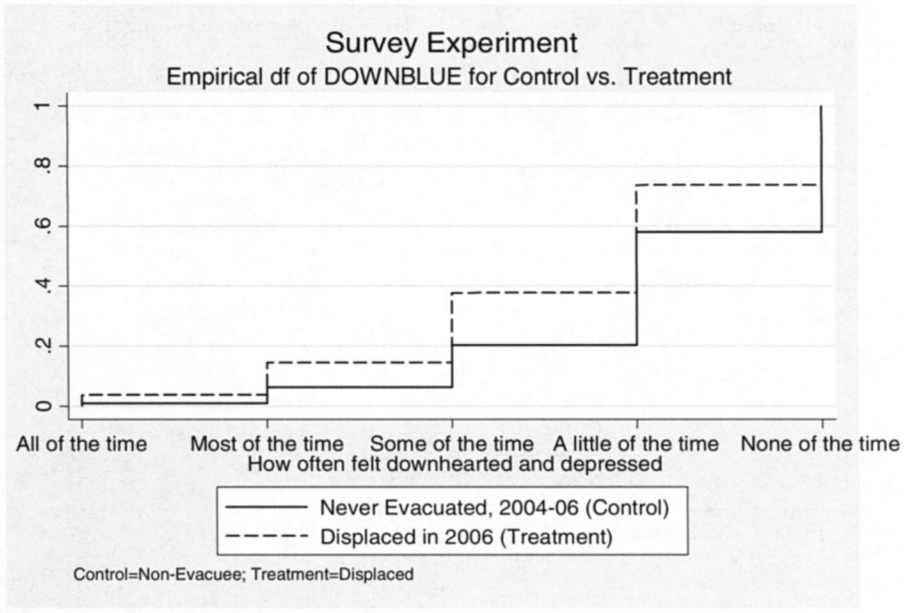

Fig. 2 Empirical distribution functions of DOWNBLUE in hurricane sample for Treatment (Displaced as of 2006) and Control (Never experienced evacuation during 2004–06) groups.

The right panel of Table 7 affirms the same about the DOWNBLUE distributions for treatment and control groups (Fig. 2 corresponds). In 2006, the empirical probability that hurricane-displaced persons felt downhearted and depressed at least "some of the time" was 17.4% higher than the same probability for the control group. These findings may help us understand responses to other substantive research questions about the impact of hurricanes on the affected population.

Finally, Table 8 shows results for the same test for the additional GSS variables LIMITEDP and DIDLESSE. These affirm the debilitating impact that the long displacement had on the treatment group. The empirical probability that in 2006 hurricane-displaced persons were limited in work or other activities as a result of their physical health at least "some of the time" was 8.1% higher than for the control group. Further, the empirical probability that as a result of emotional problems hurricane-displaced persons accomplished less than they would like at least "some of the time" was 14.1% higher than for the control group. The KS test strongly rejects the equality of these distributions for the treated versus control.

4.3 *Discussion*

Well-established questions that ask about moods, attitudes, and opinions allow researchers to validate surveys and treatment effects along dimensions that can be more revealing than standard demographic questions. These attitudes can also be the basis for decisions people make about important social science outcomes. It is therefore crucial to be able to analyze the distributions of these questions in their entirety, rather than to be constrained by focusing only on means or proportions. The nonparametric analysis in the discrete KS test enables this comparison, and reveals the intricacies of the distributions.

5 Conclusion

We set out to resurrect the discrete KS test as a method for comparing discretely ordered distributions. We began describing the nonparametric (discrete) KS test and the parametric chi-square test, discussing when each is appropriate to compare discretely ordered responses. Monte Carlo simulations evaluated the size and the power of the discrete KS test when compared to the

Table 8 KS and Chi-square tests: Hurricane survey experiment

	LIMITEDP *Limited physical activity*		*DIDLESSE* *Accomplished less*	
Category	*Treatment: Displaced*	*Control: Non-evacuee*	*Treatment: Displaced*	*Control: Non-evacuee*
All of the time	0.0594	0.0403	0.0426	0.0195
Most of the time	0.1411	0.0975	0.131	0.0668
Some of the time	0.2867	0.2054	0.3191	0.1778
A little of the time	0.4580	0.3877	0.5577	0.3821
None of the time	1	1	1	1
Sample size (N)	893	4669	893	4669
KS statistic	2.207**		4.790**	
KS p-value	0.000		0.000	
Chi-square statistics	29.68**		120.3**	
Chi-square p-value	0.000		0.000	

Notes. H_0: Distribution(s) for non-evacuee population and hurricane-displaced population are equal.
1. See notes to Table 7.

chi-square test, and we used them to derive a set of Distribution Check Rules. In situations with a low tolerance for Type I error, we found the chi-square test has better size properties in all sample sizes and with all distribution types. When Type II error should be minimized, the discrete KS test is superior on power for right- and left- skewed distributions (regardless of sample size or number of categories), and as good as or superior on power for small samples (regardless of distribution). For large samples (over 5000), the chi-square test has better power properties than the discrete KS test when the frequency distribution is U-shaped or balanced.

We then applied these rules in two contexts common to survey experiment researchers: survey sample validation and treatment versus control group tests. As an example, we employed the discrete KS and chi-square tests on our unique survey administered in hurricane-threatened regions of the United States, 1 year after Hurricanes Katrina and Rita. Our survey included replica questions from the MEPS and GSS and given to those directly affected by, as well as to those who did not directly experience, either hurricane. We compared distributions on the questions between groups in our survey, and between our sample and MEPS/GSS populations, using both the discrete KS test and the chi-square test, and illuminated the distinctions between them.

We draw two main lessons from our work. First, the discrete KS test is a useful test for examining distributions, now that its computational intensity is no longer a problem. It is especially relevant where the power of the test is paramount, as in the validation of costly primary surveys. When implementing our Distribution Check Rules appropriately, the power of the discrete KS test can properly validate small-sample surveys. Second, the discrete KS test is useful in survey experiments where a difference-in-differences analysis can be conducted for distributions. Such difference-in-distributions of control versus treatment groups may highlight important aspects of the data that simple difference-in-means analyses ignore. The effects of field survey treatments on attitudes and opinions will be illuminated in relief and gauged with fitting methods.

Funding

This work was supported by the National Science Foundation [#0554875].

References

Barabas, Jason, and Jennifer Jerit. 2010. Are survey experiments externally valid? *American Political Science Review* 104(2):226–42.

Berrens, Robert P., Alok K. Bohara, Hank Jenkins-Smith, Carol Silva, and David L. Weimer. 2003. The advent of Internet surveys for political research: A comparison of telephone and Internet samples. *Political Analysis* 11(1):1–22.

Bourque, Linda B., Judith M. Siegel, and Kimberley I. Shoaf. 2002. Psychological distress following urban earthquakes in California. *Prehospital and Disaster Medicine* 17(2):81–90.

Bourque, Linda B., Judith M. Siegel, Megumi Kano, and Michele M. Wood 2006. Morbidity and mortality associated with disasters. In *Handbook of disaster research*, eds. Havidan Rodriguez, Enrico L. Quarantelli, and Russell R. Dynes, 97–112. New York: Springer.

Canner, Paul L. 1975. A simulation study of one- and two-sample Kolmogorov–Smirnov statistics with a particular weight function. *Journal of the American Statistical Association* 70:209–11.

Clason, Dennis L., and Thomas J. Dormody. 1994. Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education* 35(4):31–35.

Darling, Donald A. 1960. On the theorems of Kolmogorov–Smirnov. *Theory of Probability and Its Applications* 5(4):356–61.

Gawande, Kishore, Gina Yannitell Reinhardt, Carol L. Silva, and Domonic A. Bearfield. 2012. Replication data for: Comparing discrete distributions: Surveys and survey experiments. IQSS Dataverse Network V1 [Version]. http://hdl.handle.net/1902.1/18921 (Accessed October 8, 2012).

Gleser, Leon Jay. 1985. Exact power of goodness-of-fit tests of Kolmogorov type for discontinuous distributions. *Journal of the American Statistical Association* 80:954–58.

Hetherington, Marc, and Elizabeth Suhay. 2011. Authoritarianism, threat, and Americans' support for the war on terror. *American Journal of Political Science* 55(3):546–60.

Hilbe, Joseph M. 2007. *Negative Binomial Regression*. New York: Cambridge University Press.

———. 2009. *Logistic Regression Models*. Boca Raton, FL: Chapman & Hall/CRC.

Hilbe, Joseph. 2010. Creation of synthetic discrete response regression models. *Stata Journal* 10(1):104–24.

Horn, Susan D. 1977. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics* 33:237–47.

Jerit, Jennifer. 2009. How predictive appeals affect policy opinions. *American Journal of Political Science* 53(2):411–26.

Kempthorne, Oscar. 1967. The classical problem of inference-goodness-of fit. In: *Proceedings of Fifth Berkeley Symposium On Mathematical Statistics and Probability*, Vol. 1, 235–49. Berkeley: University of California Press.

Likert, Rensis. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 140:1–55.

Lilliefors, Hubert W. 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62:399–402.

———. 1969. On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association* 64:387–89.

Lott, Neal, Adam Smith, Tamara Houston, Karsten Shein, and Jake Crouch. 2012. *Billion dollar U.S. weather/climate disasters, 1980 November 2011*. National Climate Data Center.

Malhotra, Neil, and Yotam Margalit. 2010. Short-term communication effects or longstanding dispositions? The public's response to the financial crisis of 2008. *Journal of Politics* 72(3):852–67.

Massey, Frank J. 1950. The Kolmogorov–Smirnov test for goodness-of-fit. *Journal of the American Statistical Association* 46:68–77.

MEPS (Medical Expenditures Panel Survey). 2007 *MEPS-HC sample design and collection process. Agency for healthcare research and quality*. Rockville, MD. http://www.meps.ahrq.gov/survey_comp/hc_data_collection.jsp (Accessed December 2008).

Panchenko, Dmitry. 2003. *Lecture notes on MIT OpenCourseWare site, Lectures 11 and 14*. http://ocw.mit.edu/OcwWeb/Mathematics/18-443Fall-2006/LectureNotes/index.htm (Accessed August 2008).

Pettitt, A. N., and M.A. Stephens. 1977. The Kolmogorov–Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics* 19:207–10.

Quarantelli, Enrico L. 2002. The Disaster Research Center field studies of organized behavior in the crisis time period of disasters. In *Methods of disaster research*, ed. Robert A. Stallings, 94–126. Philadelphia: Xlibris.

Smirnov, Nikolai V. 1939. An estimate of divergence between empirical curves of a distribution in two independent samples. *Bulletin Math. de l'Univ. de Moscou* 2:3–14 (in Russian).

Smith, Tom W., Peter Marsden, Michael Hout, and Jibum Kim. 2011. *General social surveys, 1972–2010: cumulative codebook*. Chicago: National Opinion Research Center.

Wilcox, Rand R. 1997. Some practical reasons for reconsidering the Kolmogorov–Smirnov test. *British Journal of Mathematical and Statistical Psychology* 50:9–20.

Witt, Evans, Jonathan Best, and Lee Rainie. 2008. Internet access and use: Does cell phone interviewing make a difference? *Paper for the 2008 Conference of the American Association for Public Opinion Research*. http://pewresearch.org/assets/pdf/cellphone-\linebreak pewinternet.pdf (Accessed July 2009).

Wood, Constance L., and Michele M. Altavela. 1978. Large-sample results for Kolmogorov–Smirnov statistics for discrete distributions. *Biometrika* 65:235–39.