Journal of the Royal Statistical Society
SERIES C Applied Statistics
Appl. Statist. (2014)
63, *Part* 2, *pp.* 195–210

Generalizing from unrepresentative experiments: a stratified propensity score approach

Colm O'Muircheartaigh
University of Chicago, USA
and Larry V. Hedges
Northwestern University, Evanston, USA

[Received August 2012. Final revision June 2013]

Summary. The paper addresses means of generalizing from an experiment based on a non-probability sample to a population of interest and to subpopulations of interest, where information is available about relevant covariates in the whole population. Using stratification based on propensity score matching with an external populationwide data set, an estimator of the population average treatment effect is constructed. An example is presented in which the applicability of a major education intervention in a non-probability sample of schools in Texas, USA, is assessed for the state as a whole and for its constituent counties. The implications of the results are discussed for two important situations: how to use this methodology to establish where future experiments should be conducted to improve this generalization and how to construct *a priori* a strategy for experimentation which will maximize both the initial inferential power and the final inferential basis for a series of experiments.

Keywords: Generalization; Propensity score stratification; Non-probability samples

1. Introduction

The two great advances in statistical generalization in applied fields were the development of randomized experiments and of probability sampling. Randomization provides protection against confounding the effect of the treatment and the effect of potentially contaminating variables, thus giving the inference internal validity; probability sampling permits unbiased estimates of population parameters that do not depend on modelling assumptions, thus ensuring one important aspect of external validity. The purpose of experiments is often to obtain information that can inform policy choices in education, medicine and the social sciences, yet experiments with probability samples are rare in those fields. Two exceptions in education (i.e. experiments with probability samples) are the national evaluation of 'Head start' (Deming, 2009) and the evaluation of 'Upward bound' (Myers and Schirm, 1999), which are US federal government programmes providing educational assistance and social services to disadvantaged students at respectively preschool and pre-college levels.

Despite the advantages of probability sampling, there are several reasons why we believe that probability sampling is rare in social experimentation and likely to remain so. The two examples of experiments with probability samples are national evaluation studies of particular

Address for correspondence: Colm O'Muircheartaigh, Irving B. Harris School of Public Policy Studies, University of Chicago, East 60th Street, Chicago, IL 60637, USA.
E-mail: caomuirc@uchicago.edu

 0035–9254/14/63195

programmes. Such studies are important and probability sampling should be seriously considered for evaluations that address specific questions in a single (or even a few) well-defined population(s). The strict technical framework of probability sampling does not well match the realities of much social and educational policy research, for at least three reasons. Social experiments are often costly and time consuming, so a single experiment may be used to inform policy choices in a number of diverse settings. A more profound theoretical problem is that the populations of interest are not always known in advance. Third, the inference population may not be part of the population that is available for the experiment.

In this paper we describe how to generalize effectively from an assessment of interactive mathematics teaching software (SimCalc: `http://www.kaputcenter.umassd.edu/products/software/smwcomp/download/`) in Texas schools; the outcome variable is the school-average student gain scores on a test that focuses directly on the issues of proportionality, linearity and rates of change. Section 2 presents the conceptual foundation for the analysis. Causal effects in experiments are defined and causal inference from experiments is discussed. Stratification based on propensity score matching is proposed as a practical method of reducing potential bias in estimates of causal effects from experiments with non-representative sampling. An example is presented in Section 3 in which the applicability of the introduction of SimCalc in a non-probability sample of schools in Texas is assessed for the state as a whole and its constituent counties. An attempt was made to include schools that are representative of the state, but random sampling was not feasible or implemented. In our evaluation we offer one set of assumptions that could justify inferences about average treatment effects in (multiple) policy relevant inference populations. The implications of the results are discussed in Section 4.

2. Conceptual foundation

2.1. Defining causal effects in experiments

In our analysis we aim to estimate the mean effect of SimCalc in a variety of populations of Texas schools. This is the population average treatment effect (PATE) that was defined by Rubin; the estimation process is based on the framework provided by Rubin and others in the series of papers referenced below.

Rubin (1974, 1977, 1978, 1980, 1986) has provided a framework for defining causal effects and demonstrating how they can be estimated in experiments and other research designs. Let $r_i^{(1)}$ be the response of unit i if it receives treatment 1 and $r_i^{(0)}$ be the response of unit i if it receives treatment 0. Then $\mathbf{r}_i = (r_i^{(0)}, r_i^{(1)})$ is the vector of possible responses; let W_i be a variable indicating treatment assignment such that $W = 0$ implies assignment to treatment 0 and $W = 1$ implies assignment to treatment 1. Rubin's stable unit treatment value assumption is equivalent to saying that the response of the ith unit depends on the treatment that is assigned to this unit but not that assigned to other units. In a randomized experiment where units are randomized individually, W is independent of $\mathbf{r}$, so

$$E[r^{(1)}|W=1] - E[r^{(0)}|W=0] = E[r^{(1)}] - E[r^{(0)}] = E[r^{(1)} - r^{(0)}].$$

Therefore, if units are a simple random sample from a population, and each is individually randomized to treatments, the difference between sample means is an unbiased estimator of the average causal effect of treatment 1 *versus* treatment 0 for the population—the PATE. Where the units are randomized to treatments but the sampling mechanism is non-representative, the difference between sample means is an unbiased estimator of the treatment effect for the sample units (the sample average treatment effect), but not necessarily of the PATE. In this paper we

discuss how and under what conditions data from a non-representative sample can be used to estimate the PATE.

2.2. Causal generalization from experiments

We focus on a strategy of finding a set of covariates that explain variation in treatment effects in the inference population of interest; we then estimate the treatment effect in the inference population by averaging over the conditional distribution of the treatment effects given the covariates. Because different inference populations may have different covariate distributions this leads to different estimates of average treatment effects for different inference populations. Importantly, the method leads to variance estimates that reflect the larger uncertainty of treatment effect estimates when the study sample does not match the inference population well. Note that the strategy of basing inferences about the mean of the distribution of Y on the conditional distribution of treatment effects given the covariate is not novel. This approach is used in demography where it is called standardization (i.e. standardizing a population composition so that comparisons between desired quantities in two different populations are not confounded with the difference in population composition), and in economics to form index numbers (see, for example, Kitagawa (1964)) and to isolate the effect of changes in population composition on wage differentials (see, for example, Oaxaca (1973)). The same idea is used as a tool for analysis in survey research (see, for example, Rosenberg (1962) or Kalton (1968)). It is also a basic tool in the analysis of missing data (see, for example, Rubin (1976), Little and Rubin (2002) or Groves *et al.* (2002)).

The application of the strategy above to causal generalization from experiments is straightforward in principle. Start with a finite inference population definition that includes the units that are in the experiment and a census or probability sample survey based on that population definition. Stratify the inference population and experimental sample on the basis of the covariate. (We propose to use propensity score stratification in this connection, as we discuss in more detail below.) Then construct a treatment effect estimate for each stratum: the stratum-specific treatment effect. Finally compute the estimate of the population treatment effect as the weighted average of these stratum-specific treatment effects, weighting by the population weight for each stratum.

If the stable unit treatment value assumption holds and the sampling mechanism is ignorable, causal generalization follows from an application of the same method for generalizing to population means. Let Z be an indicator of being in the sample such that $Z = 1$ for population members who are in the experimental sample and $Z = 0$ for those who are not. The sampling mechanism is ignorable given a set of observed variables $\mathbf{x}$, if two conditions are met.

Assumption 1. The first assumption is that the difference between the possible responses $r_i^{(1)} - r_i^{(0)}$ is conditionally independent of Z given $\mathbf{x}$, i.e.

$$\Pr(r_i^{(1)} - r_i^{(0)} | \mathbf{x}, Z) = \Pr(r_i^{(1)} - r_i^{(0)} | \mathbf{x}).$$

Assumption 2. The second assumption is that, for every value of $\mathbf{x}$ in the population, there is a non-zero probability that this value of $\mathbf{x}$ could be selected into the sample; thus the conditional value of the treatment effect $r^{(1)} - r^{(0)} | \mathbf{x}$ is estimable from the sample for every $\mathbf{x}$...

It is useful to consider the first part of the ignorability assumption in detail. This assumption essentially requires that all the systematic variation in the treatment effect is captured by the covariate $\mathbf{x}$; the remaining variation in the treatment effect in the population is random, given $\mathbf{x}$. Obtaining the right set of $\mathbf{x}$ is the first challenge; it is possible that no set of $\mathbf{x}$ exists for which assumption 1 holds, even approximately. In this case, no modelling approach will resolve

the problem of generalizability. This issue must be addressed in every application, and the assumption justified by reference to data in the area of application. Whether this assumption is true obviously depends on the particular covariates in **x** and, for any given covariate set, it may depend on the inference population that is specified. The assumptions in this approach are analogous to the conditional independence assumptions that are used to justify estimates which control for confounding (through regression or matching, for example) in estimating treatment effects in observational studies. Our assumptions in moving from the naive sample estimate to an estimate of the population effect (the PATE) face the same challenges in finding and justifying the choice of **x**.

Even if the covariates do not explain *all* of the treatment effect variation, we argue that inferences are likely to be *less* biased than those based on study results with no matching at all. Inference using no matching at all is equivalent to assuming that the inference population matches the study sample on all relevant covariates (i.e. those that actually explain treatment heterogeneity), which seems rather far fetched in the absence of simple random sampling; it would be analogous to making no correction for a very high level of non-response in a sample survey. In fields where experiments are carried out, there tends to be more knowledge of covariates that are related to the outcome than there is in most survey situations; subject matter knowledge is crucial in justifying the selection of the covariates.

If the ignorability assumptions hold, then unbiased estimates of the treatment effect in the population (the PATE)

$$E[r_i^{(1)} - r_i^{(0)}]$$

can be obtained from a sample that includes observations for each value of **x** because

$$E_{\mathbf{x}}[E[r_i^{(1)} - r_i^{(0)} | Z = 1, \mathbf{x}]] = E_{\mathbf{x}}[E[r_i^{(1)} - r_i^{(0)} | \mathbf{x}]] = E[r_i^{(1)} - r_i^{(0)}].$$

The PATE is estimable because the inner expectation of the expression on the left-hand side is estimable from the sample given randomization. Therefore if $T(\mathbf{x})$ is the treatment effect estimate for units with the covariate value **x**, and $p(\mathbf{x})$ is the proportion of the inference population with the covariate value **x**, then for discrete **x**

$$\sum_{\mathbf{x}} p(\mathbf{x}) T(\mathbf{x})$$

is an unbiased estimate of the PATE, the treatment effect in the inference population.

2.3. Practical matching strategies

There is now a large literature on matching (see, for example, Rubin (2006)). Perhaps the most elegant and flexible class of matching strategies involves the use of propensity scores (Rosenbaum and Rubin, 1983). For purposes of this paper, the propensity score is the conditional probability of a population member being in the sample given **x**. Thus, if Z is an indicator of being in the sample such that $Z = 1$ for population members who are in the sample and $Z = 0$ for those who are not, the propensity score is

$$e(\mathbf{x}) = \text{Prob}(Z = 1 | \mathbf{x}).$$

The importance of propensity scores for matching problems lies in the fact that they are balancing scores, namely that the distribution of **Z** depends on **x** only through $e(\mathbf{x})$: $\text{Prob}(Z = 1|x) = \text{Prob}\{Z = 1|e(x)\}$ (Rosenbaum and Rubin, 1983). Thus matching on propensity scores is the equivalent to matching on all of the components of **x** simultaneously. This reduces the complicated problem of multivariate matching to a simpler univariate matching problem. The

simplicity and transparency of this method is advantageous in communicating the findings to a non-statistical audience, as is illustrated by the example presented in Section 3, below. Furthermore, the avoidance of large weights has technical benefits in terms of the variance of the population estimates.

If the study sample is a subset of the inference population, we can understand the propensity score analysis as describing the propensity of inference population members with a given covariate value $\mathbf{x}$ to be in the study sample; in a sample that is not a simple random sample, the propensities will not be the same for every value of $\mathbf{x}$. If the study sample is distinct from the inference sample, the propensity score analysis should be understood in terms of the propensity of units in the inference population to have covariate values $\mathbf{x}$ that are in the study sample. In probability samples of various kinds, the propensity scores are known. In non-probability samples, they must be estimated from the data. One way to do so is to posit a generalized linear model for the propensity scores. The most widely used such model is logistic regression, where the model is

$$\ln\left\{\frac{e(\mathbf{x})}{1-e(\mathbf{x})}\right\} = \text{logit}\{e(\mathbf{x})\}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta_0 + \mathbf{x}'.$$

Here $\mathbf{x} = (x_1, \ldots, x_p)'$ and the coefficients $\beta_0, (\beta_1, \ldots, \beta_p)' = \boldsymbol{\beta}$ are estimated from the data by pooling the inference population and the study sample (if it is not a subset of the inference population); then regress the logit of the indicator variable

$$Z = \begin{cases} 1 & \text{if the unit is in the study sample,} \\ 0 & \text{if the unit is not in the study sample} \end{cases}$$

on $\mathbf{x}$ to obtain estimates of $\boldsymbol{\beta}$. In what follows, it is often more convenient to work with the logit (estimated) propensity scores rather than the propensity scores themselves.

Once the propensity score has been estimated for each individual in the study sample, the required inferences about the mean of the inference population must be made. This can be done by stratifying on the propensity score distribution, computing the proportion of the inference population in each propensity score stratum and constructing a weighted average of the stratum-specific treatment effect estimates as the population estimate (Hedges and O'Muircheartaigh, 2010). For many distributions, Cochran (1968) showed that, for many examples of univariate $\mathbf{x}$, dividing the distribution of $\mathbf{x}$ into five strata is sufficient to control over 90% of the bias, and Rosenbaum and Rubin (1984) showed that using five propensity score strata might frequently be expected to remove at least 90% of the bias on each component of $\mathbf{x}$. Because this approach does not require weighting by the inverse of individual inverse propensity scores, it is likely to be more numerically stable than, for example, computing a weighted mean of the sample data while weighting by the reciprocal of $e(\mathbf{x})$. This approach also has advantages when applied to experimental designs that do not use simple random samples (such as cluster-randomized trials).

The propensity score strata will stratify the samples that are used in the experiment. Then the estimate of the PATE is a (population-weighted) mean of the stratum-specific treatment effects. The estimator now has a grouped version of $e(\mathbf{x})$, v, in place of $\mathbf{x}$. The estimator with five strata can be written as

$$\sum_{v=1}^{5} p(v)T(v),$$

where $p(v)$ is the proportion of the inference population in stratum v and $T(v)$ is the treatment effect estimate in stratum v.

If (as in the completely randomized design) the estimates from different strata are independent, and the uncertainty in $p(v)$ is ignored, the variance of the estimate will just be

$$\sum_{v=1}^{5} p(v)^2 V(v)$$

where $V(v)$ is the variance of the treatment effect estimate in the stratum defined by v; the covariances are 0, as the strata are independent. Thus the variance will be minimized when the strata that have large (population) weights also have small variances, i.e. when the sample size distribution in the experiment matches that in the population.

Although we have assumed that there is no uncertainty in $p(v)$, this is not exactly true. There are two sources of uncertainty in $p(v)$. One source of uncertainty arises from the estimation of v (largely from estimation of the propensity scores), which we treat as known. The other source arises from the estimation of $p(v)$, the proportion of the population in each stratum. There has been some work on the effect of using estimated propensity scores (see, for example, Abadie and Imbens (2009)). This work confirms that the variance of matching estimators by using estimated propensity scores is actually smaller than the variance that is computed by assuming that propensity scores are known. This is essentially because the estimated propensity scores more accurately reflect the particular sample at hand than do the true scores. Thus our assumption that v is known when in fact it is estimated is likely to lead to overestimation of the estimate of the variance of the PATE.

The uncertainty in the $p(v)$ will be influenced by the quality of the source of data from which they are derived. In our example we have a rich source of administrative data that covers the whole population, providing an accurate basis for estimating $p(v)$ if the form of the propensity score is known. Such a situation is not universal. We could nevertheless apply the same methodology in circumstances where a population database is not available as the reference set, but a survey based on a probability sample is used to provide population data. Often such a survey will be more likely to have an adequate selection of covariates available for matching.

A complication arises when the distribution shape of propensity scores in the population differs from that in the experimental sample, e.g. when the population distribution is considerably more skewed than that in the experimental sample. In this case, matching by stratification may be less effective in reducing bias in propensity scores (than when distributions to be matched have the same shape) and therefore in the underlying covariates. Tipton (2013) has studied analytically the bias that may be induced by such mismatches in distribution shape and offered approximate estimates of the relative bias reduction and variance inflation of the PATE estimate derived from the stratified estimator compared with the standard estimator. Her results provide one way to evaluate partially the bias reduction in situations where stratification provides an imperfect matching of population and experimental sample.

3. An example

Technological aids for assisting with mathematics teaching are of great interest in the USA. One such technological aid is SimCalc, which is a mathematics software program that uses a dynamic representation strategy to help middle school students to learn to solve rates and proportions problems. To evaluate the effectiveness of SimCalc, the research firm SRI International conducted two cluster-randomized trials of SimCalc in public schools in the state of Texas: a pilot study, which included 19 schools, and a full study, which included 73 schools (Roschelle *et al.*, 2010). The outcome variable is the school-average student gain scores on a test that focuses directly on the issues of proportionality, linearity and rates of change. Although every effort was made to include schools representative of the state, the sample of schools was not a probability sample of schools in Texas; the process of selection is described in Roschelle *et al.*

(2010), page 855. In particular they worked with regional education authorities to recruit the schools. The plausibility of our generalizations depends of course on the composition of the pool; this situation is analogous to that for clinical trials where only consenting patients can be assigned to the treatment.

The experiment demonstrated that the use of SimCalc led to statistically and substantively significant improvement over the controls. The mean improvement (in standard deviation units) was $T_C = 1.443$ with a standard error of 0.142. This indicates that schools using SimCalc had larger average gain scores than those carrying on as usual.

However, because the study sample was not a probability sample of the state or any other well-defined inference population, it is unclear what the policy implications of this treatment effect estimate might be. For example, would we expect that this treatment effect estimate is likely to apply to schools across the entire state of Texas? Would we expect it to apply to specific subregions of the state?

We illustrate the use of methods that were discussed earlier in this paper to explore the validity of this estimate for the two inferential populations—first, the whole state of Texas and, second, each of the 254 counties of Texas. How good is the estimate as a measure of the effect of SimCalc across all schools that have seventh-grade classrooms in the state of Texas? And, second, how good is it as a measure of impact for all or some of the state's 254 counties?

The population of schools with seventh-grade classrooms was defined by using the publicly available state academic excellence indicator system. The population was defined to be the $N = 1713$ Texas schools in the 2008–2009 academic year with seventh-grade classrooms that were not charter schools. Whereas typical public schools are completely funded and administered by the government, charter schools are privately administered under different regulations and receive lower levels of government funding. Charter schools are excluded from the data because these may differ from the typical public schools in which SimCalc was tested with respect to curriculum and student selection.

26 covariates were selected in the academic excellence indicator system for matching the schools in the experiment with those in the population. These included variables on student and teacher demographic composition, school structure and prior year school average test scores. Note that nine of these variables are teacher aggregates, 16 are student aggregates and one is a school level variable. (All these covariates, as well as their means in the sample and in the population, are listed in Table 1.)

We estimated the propensity scores $e(\mathbf{x})$ by using the logistic regression model

$$\log\left\{\frac{e(\mathbf{x})}{1-e(\mathbf{x})}\right\} = \beta_0 + \beta_1 x_1 + \ldots + \beta_{26} x_{26}.$$

Then we stratified the population of 1713 schools in Texas into $k = 5$ equal sized strata. Table 1 shows the mean value of each covariate in the experimental sample and in the population, and the bias defined as the difference between the mean in the population and the mean in the sample, before and after stratification. In Table1, the bias is rendered as the absolute standardized mean difference |SMD|: the absolute difference between the population mean and the sample mean divided by the population standard deviation. Finally, Table 1 gives the percentage reduction in bias due to stratification. Positive numbers reflect that stratification has decreased the bias, whereas negative numbers indicate that stratification has increased the bias on a covariate.

As expected, the stratified estimator reduces bias in the propensity score and the covariates; the bias is reduced by 96% for the propensity scores. For both the conventional and the stratified estimators, eight of the covariates have absolute standardized mean differences |SMD| between 0.10 and 0.20. However, for the stratified estimator, only four covariates have |SMD|s that

Table 1. Effect of stratification by propensities on bias of estimate of covariate means from the sample: bias measured as the absolute standardized mean deviation |SMD|

Covariate	*Means*			*Bias as |SMD|*		*% bias reduction*
	Population of schools with 7th-grade classrooms (Texas) (N = 1713) (A)	*Sample of schools with 7th-grade classrooms (n = 92) (B)*	*Weighted sample (stratified) (n = 92) (C)*	*In sample (= |A − B| divided by population standard deviation)*	*In sample after stratification (= |A − C| divided by standard deviation)*	
Teacher tenure (mean)	7.091	6.801	6.770	0.066	0.073	−10
Teacher experience (mean)	11.583	10.947	11.461	0.225	0.043	81
Teacher/student ratio	12.699	13.266	12.714	0.189	0.005	97
Teachers who are African American (%)	8.393	2.557	6.755	0.253	0.071	72
Teachers who are Hispanic (%)	14.721	21.571	15.168	0.322	0.021	93
Teachers in the school (total number)	39.873	42.984	40.216	0.272	0.03	89
Teachers in 1st year of teaching (%)	8.321	8.739	8.591	0.031	0.02	35
Teachers with 1–5 years of experience (%)	28.012	28.744	28.069	0.155	0.012	93
Teachers with > 20 years of experience (%)	20.251	17.695	19.869	0.221	0.033	85
Students in disciplinary alternative education programmes (%)	3.102	3.418	3.499	0.172	0.216	−26
7th-grade retention rate	1.833	1.307	1.539	0.109	0.061	44
Students who are mobile (%)	19.228	14.804	15.213	0.281	0.255	9
Students in school who are in 7th grade (%)	31.206	34.995	32.662	0.354	0.136	62
Students in school who are in 7th grade (total number)	190.399	224.247	191.230	0.326	0.008	97
Students who are African American (%)	11.792	5.114	9.595	0.31	0.102	67
Students who are Hispanic (%)	40.268	47.195	41.291	0.264	0.039	85
Students who are limited English proficient (%)	7.538	9.44	7.583	0.253	0.006	98
Students who are economically disadvantaged (%)	53.643	52.084	31.232	0.008	0.115	−1436
Students who are at risk (%)	43.467	40.607	40.543	0.135	0.138	−2
Students who are proficient in 7th-grade reading (%)	81.901	86	85.549	0.2	0.178	11
Students who are proficient in 7th-grade mathematics (%)	72.791	75.562	77.994	0.139	0.261	−88
Students who are proficient in grades 3–11 mathematics (%)	73.599	75.014	77.777	0.085	0.251	−196
Students who are proficient in grades 3–11 all subjects (%)	63.287	63.836	69.498	0.016	0.181	−1012
Students with commended performance, grades 3–11, mathematics (%)	19.612	20.315	21.005	0.053	0.105	−99
Students with commended performance, grades 3–11, reading (%)	8.71	8.877	10.770	0.003	0.037	−1201
County of school is rural	0.329	0.315	0.307	0.105	0.164	−57
Propensity scores (mean)	0.959	0.919	0.957	0.896	0.035	96

are greater than 0.20; in comparison this is the case for 11 covariates with the conventional estimator. This indicates that the reweighted experimental sample is generally more similar to the population of interest than the original sample is. However, since balance has not been achieved on all covariates, some bias remains, even after stratification.

3.1. Estimation of Texas average treatment effect

To estimate the PATE, a subclassification estimator with $k = 5$ strata was used. The following regression model was estimated:

$$Y = \beta_1 * S_1 + \beta_2 * S_2 + \ldots + \beta_5 * S_5 + \beta_6 * S_1 * \text{TREAT} + \ldots + \beta_{10} * S_5 * \text{TREAT} + \varepsilon,$$

where $S_1, \ldots, S_5$ are indicators for the five strata, TREAT is an indicator for the treatment group and it is assumed that $\varepsilon \sim N(0, \sigma^2)$. Let T_S denote the estimated effect for the state of Texas (the PATE), as in Section 2.3. The stratified estimator that is used here can be written

$$T_S = T_{\text{PATE}} = \sum p(v) T(v) = \frac{1}{5} \sum T(v),$$

because the weights are all $\frac{1}{5}$ by construction.

The variance of the stratified estimator can be written

$$V(T_{\text{PATE}}) = \sum_v (\tfrac{1}{5})^2 V(T_v),$$

as the effects are estimated independently by stratum.

Our estimation assumes that the stratum-specific estimates of treatment effect variation could be pooled into a single σ^2-value; this was important here since some strata in the sample had as few as three or four experimental units, leading to imprecise stratum-specific variance estimates; an examination of the two strata where there were sufficient cases in the sample suggests that the assumption is reasonable. If the within-stratum variances are not constant, the estimated precision of the model-assisted estimates would be incorrect, though whether overestimated or underestimated would depend on the configuration of the inference population. Note that the assumption of constant residual variance is not of consequence for the point estimates of the stratum-specific treatment effects, which are mean differences (akin to those used to express the bias in each covariate for the experimental sample before and after stratification, as has been specified above).

Table 2 shows the relative weight that was given to the five strata in the experimental sample and the population, the estimated treatment effect, the standard error and the 95% confidence

Table 2. Stratum-specific estimates of treatment effects and their standard errors

Estimator	*Stratum*	*Weights*		*Estimate*	*Standard error*	*95% confidence interval*	
		Sample	*Population*				
Stratified (PATE) (T_S)	—			1.452	0.195	1.069	1.835
	1	0.530	0.200	1.597	0.191	1.222	1.971
	2	0.248	0.200	1.368	0.290	0.799	1.936
	3	0.107	0.200	0.400	0.405	−0.389	1.198
	4	0.060	0.200	1.768	0.546	0.716	2.856
	5	0.054	0.200	2.107	0.610	0.910	3.303
Conventional (sample average treatment effect) (T_C)	—	—	—	1.443	0.142	1.165	1.721

interval of the treatment effect in each stratum. Table 2 also shows that the combined estimate of the treatment effect for schools in Texas is $T_S = 1.452$ with a standard error of 0.195. This estimate differs from the conventional estimate by less than 1%, whereas the standard error is 38% larger (variance 90% larger). Thus, the conventional estimate of the treatment effect was not far from the stratification estimate (one might say that the bias was essentially nil) but the conventional analysis considerably underestimated the standard error. The proportion of schools in the population in each stratum is 20%; the proportions in the sample vary from 5.4% to 53%, indicating important differences between the sample and population compositions. The similarity of the two state estimates arises from compensating, though different, mixtures of schools from strata with high and low effects in the sample and the population. The stratification can nevertheless play a significant role in improving the estimates for subsets of the population and for other populations.

3.2. Estimation of treatment effects in Texas counties

One policy question is whether SimCalc works on average in the state. For this question the relevant inference population is the schools enrolling seventh-graders in the entire state. Another, perhaps more pertinent, policy question, however, is whether SimCalc will have benefits for schools in each county of Texas. For this question, the relevant inference populations are the schools enrolling seventh-graders in each of the counties of Texas. To carry out these analyses, we carried out essentially the same analysis using the population of schools enrolling seventh-graders in each of the 254 counties of Texas. This yields a stratified estimate of the PATE (and its standard error) in each of these 254 counties. The strata defining school types are defined for the state as a whole. There is insufficient information to construct propensity score strata for each county; we are therefore importing the strata from the state to the counties. For each county, the weights for the stratified estimate are determined by the composition of the county—the proportion of the county's schools in each of the (state) strata. (The table of results for the estimates, bias and variance for all 254 counties is available as on-line supplementary material.)

Of particular interest is how close the state and county estimates of treatment effects may be. For example, policies regarding adoption of innovations are often made centrally to apply to all counties in a state. Whether this is wise may depend partly on how well the effects of the innovation are known on average in a state, but also on the heterogeneity of those effects across subunits of the state such as counties. We might characterize the situation as asking how biased the state average estimate of a treatment effect is as an estimate of the subunit-specific (e.g. county-specific) estimate.

We define the relative bias in state estimates as estimates of county effects as the difference between the county estimate and the whole-state estimate, expressed as a percentage of the whole-state estimate. To focus on the magnitude of this bias, we use just the absolute relative bias.

For each county, the bias relative to the PATE (the whole-state estimate) depends on how well the county matches the state on the covariates. If the county is similar to the state as a whole—the distribution of schools in the county mirrors that in the state—then there may be little or no bias in the county estimate. Otherwise, the bias may be substantial. If the distribution of the schools in a county were to match the state exactly, a fifth of the schools in the county would fall in each stratum. In fact, the schools in each county are far from being evenly distributed in the strata, and this unevenness occurs in different ways in different counties. This suggests that (unless these differences compensate in unexpected ways) there may be biases in the estimated treatment effects across counties and these biases might be different from one another.

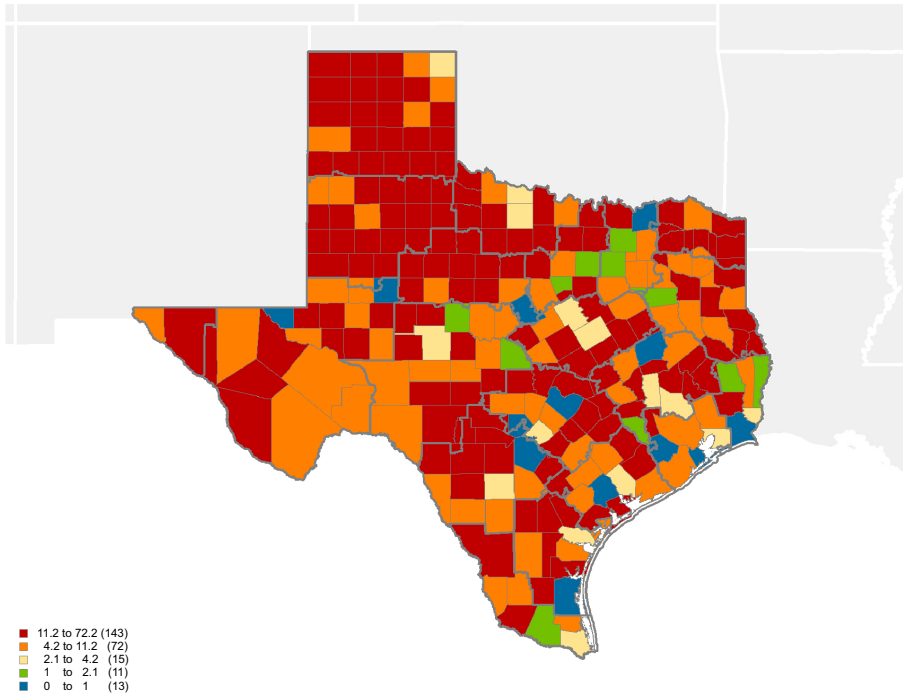
We found that the maximum (absolute) bias for any county was −72%, corresponding to a treatment effect in estimate in that county of 0.404, compared with the state estimate of 1.443.

This is only weak evidence that the county estimates are substantially different from the state estimate. The study is underpowered to identify potentially important differences in effect sizes. Only a substantially larger sample of schools could reduce the standard errors sufficiently to establish the existence of real variation between counties. The graphical presentation and discussion below suggest that such effects are confined to a relatively small proportion of counties. The configuration of schools in the counties where substantial effects may be present are different in terms of sociodemographic characteristics that might reasonably be expected to be related to the effectiveness of the programme.

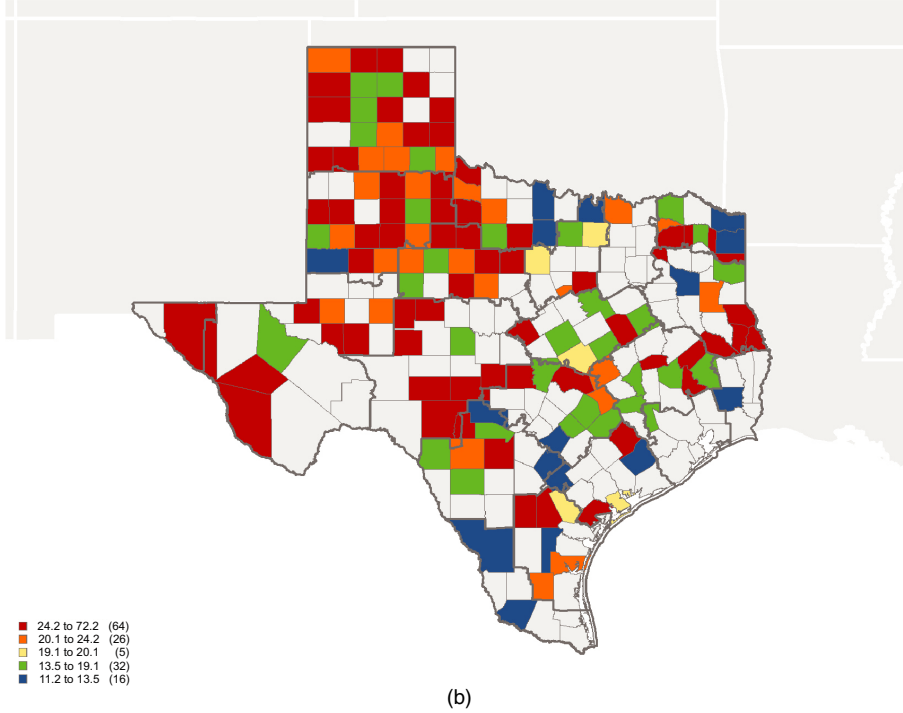
To help to understand the pattern of effectiveness of the state estimate as an estimate of the county, we found it useful to prepare a graphical representation of the pattern of bias. This spatial representation is designed to show two characteristics of the estimates at the same time: for each characteristic—bias and variance—it shows the pattern of severity across the state by county; and the pattern is presented in such a way that the proportion of the state's population affected to each extent can easily be seen.

We started by rank ordering the counties by the degree of bias (for Fig. 1) and variance (for Fig. 2). Then we divided this ordered distribution into quintiles by population (number of seventh-grade schoolchildren), so that each quintile contained the same number of students. We represented the lowest quintile (smallest bias or variance—best) in blue, the next quintile in green, the middle quintile in yellow, the second highest (second largest bias or variance) in orange and the highest quintile (largest bias or variance) in red. Our experience in other applied contexts suggests that practitioners are much more comfortable with a discretized than a continuous scale; the colour palette has also been successful. Fig. 1(a) shows a map of Texas, divided into counties, where the counties are shaded to reflect the quintiles of bias. The map reveals several features of the generalizability of the SimCalc treatment effect. Perhaps the most obvious is that the absolute bias of the treatment effect is in general rather small. For over 60% of the population (more than three quintiles), the estimated relative bias of the state average is less than 4.5%. The bias reaches 11% only in the worst quintile. Fig. 1(b) shows the worst quintile divided into quintiles itself to illustrate the variation in relative bias among this worst quintile. Fig. 1(b) reveals that, even in this quintile for which the statewide estimate is most biased, counties typically have bias less than about 20%. It is worth noting, however, that for 8% of the population of students the state estimate is substantially biased. We believe that this spatial (and geopolitical) representation of relative effects, in the maps in Fig. 1, provides a useful method of illustrating the generalizability of treatment effect estimates to education administrators and elected local officials.

Essentially the same procedure can be applied to studying the variance (uncertainty) of treatment effect estimates. As in the case of bias, the range of the variance of the estimated treatment effect in each of the 254 counties relative to the variance for the whole state estimate also depends on how well the sample represents the particular type of schools in that county. If the county type is particularly well represented, the variance may be less than for the whole state estimate. If the county type is not well represented, the variance may be considerably more. We found that the variance of the county-specific estimates ranged from 0.60 to 9.5 times that of the whole state estimate, corresponding to standard errors that ranged from 0.151 to 0.60 (the standard error for the state estimate was 0.195). To help to understand the pattern of variance in the state estimate as an estimate of the county, we found it useful to prepare a graphical representation of the pattern of relative variance (the variance of the county estimate divided by the variance of the state average estimate). Fig. 2(a) shows a map of Texas, divided into counties, where the counties are shaded to reflect the quintiles of relative variance. The map reveals several features of the generalizability of the SimCalc treatment effect. Once again, perhaps the most obvious

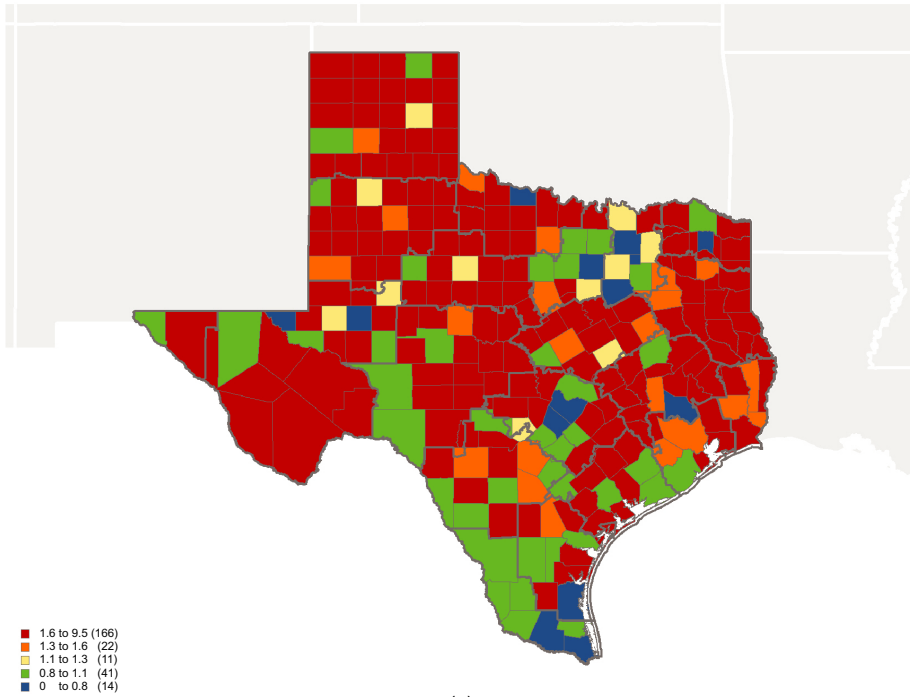


(a)

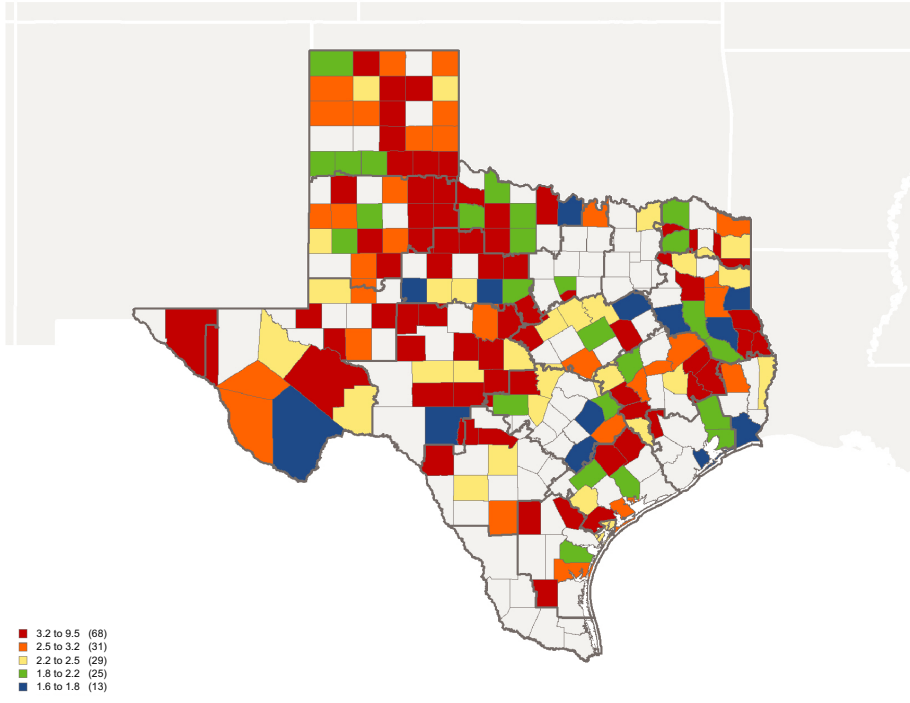


(b)

Fig. 1. Relative bias (per cent) by county: (a) quintiles; (b) quintiles of worst quintile



(a)



(b)

Fig. 2. Relative variance by counties: (a) quintiles; (b) quintiles of worst quintile

is that the relative variance of the county estimates of treatment effect is usually close to 1. For about 20% of the population (one quintile) the estimated variance of the county estimates is more than 60% greater than that of the state average. Fig. 2(b) shows the worst quintile divided into quintiles itself to illustrate the variation in relative bias among this worst quintile. Fig. 2(a) reveals that, in this quintile for which generalization of the state average treatment effect is worst, the relative variance exceeds 2.5 for half the population. These counties are rural counties with very low population densities (averaging 14 per square mile) compared with an average of 670 per square mile for counties in quintile 1; these rural counties also have lower proportions of minority populations (African-American and Latino). This analysis of relative variance demonstrates that variance, more than bias, may in this case be a limiting factor in the usefulness of state average treatment effects for smaller units. The maps that are presented in Fig. 2 provide a useful tool for illustrating how this uncertainty (variance) can compromise the generalizability of treatment effect estimates and where in the population this effect is greatest, even when there is relatively little bias.

4. Conclusions

We have described one way in which considerations about the generalizability of treatment effects of randomized experiments can be formalized. This begins with identification of a well-defined inference population, and the inference problem is to estimate the average treatment effect in that population. If a set of covariates can be identified that explain variation in the treatment effect, the study sample can be stratified according to propensity scores. If treatment effects can be estimated for every stratum of the propensity score distribution, then the weighted average (by using inference population weights) of the stratum-specific treatment effects is an estimate of the PATE. The key concept here is ignorability of sampling given the covariate set.

Several consequences follow from this development. One is that claims about generalizability are ill formed without specification of the inference population to which generalization is desired. To put it another way, claims that a particular experiment has external validity make no sense without specifying an inference population. A second is that special methods to address generalization are necessary only if there is variation in treatment effects. If treatment effects are constant, the estimate from any experiment is an estimate of the PATE (in any population). If there is variation in treatment effects, even with probability sampling from a population, balanced experimental designs do not necessarily yield unbiased estimates of PATEs, unless appropriate weighting is used. Finally, even when unbiased estimation of PATEs is possible, the experimental sample may be too poorly matched to the inference population to provide estimates with sufficiently small variance to support practical inference.

An extreme of poor matching is the case in which the variance of the PATE is (in essence) infinite. This occurs whenever there are covariate values in the inference population that have no corresponding treatment and control group units in the sample from which to estimate a (conditional) treatment effect. Following the method that is suggested in this paper, this occurs when a propensity score stratum does not have at least one treatment and one control group unit. This corresponds to undercoverage in a survey. In such cases the experimental sample provides no empirical basis for estimation of the average treatment effect in the inference population without additional assumptions. We argue that the ability to identify undercoverage is one of the advantages of the approach that is described in this paper.

When coverage issues limit inferences, we might impute treatment effect values for the strata in which they are missing by using a model of the relationship between prognostic covariates and treatment effects. Alternatively, values for the missing conditional treatment effects might

be generated *a priori* or by bracketing the plausible range of treatment effects for sensitivity analysis. Another possibility is to redefine the inference population to be the population in which the experimental sample can support inferences with reasonable precision. The principal difficulty here is to describe the population for which inferences are possible. Because they will be defined in terms of propensity score strata, they need not correspond to meaningful geographic or social subgroups of the original inference population. However, the new inference population can be enumerated and therefore described.

Although we have illustrated the use of these ideas for evaluation and generalization of an experiment, they may be more valuable prospectively in planning experiments to improve generalizability. If an inference population and a covariate set **x** can be identified in advance, it is possible to select the study sample to match the inference population. For example, consider an experiment that will assign treatments within schools. Education field experiments often recruit samples over time, adding schools or districts as they can be persuaded to join the experiment. One possible strategy is to allow the recruitment to proceed as usual until some fixed proportion (say 50%) of the sample has been obtained. Then estimate the propensity to be in the study sample for each member of the inference population. Stratifying the population propensity scores into a few (e.g. five) strata will reveal which strata are sparsely populated by study sample. Moreover, it will provide a list of schools that, if recruited, would improve the match between experimental schools and the inference population. This permits recruitment efforts to be targeted to schools that would improve generalization. In the case of cluster-randomized experiments that assign schools to treatments, it might also be advantageous to make treatment assignments within strata to assure that all stratum-specific treatment effects could be estimated. In our experience, even a small amount of targeted recruitment can greatly reduce the variance of the estimated PATE.

The spatial representation of these results (as in Figs 1 and 2) has proved particularly useful in communicating the results to users and planners (an earlier example can be found in Lalonde *et al.* (2005)). In our example we used the county as the unit of interest; in other contexts electoral or administrative areas specific to the subject matter can be used. The choice of quintiles as the quantiles has proved to be easy to explain to non-quantitative users, and the spatial representation is particularly attractive to administrators and politicians. Clearly the results can be presented in parallel in tables or with another choice of quantile.

Acknowledgements

The authors thank Paki Reid-Brossard, for his invaluable work on the analysis for this paper, and Becki Curtis and Ned English from the National Opinion Research Center for producing the figures. They also thank the Joint Editor and Guest Associate Editor for advice on the balance between method and applications in the paper, and the editors and reviewers for comments that greatly improved the clarity of the presentation.

This paper is based in part on work supported by the US National Science Foundation under grant 08515295 and grant 1118978. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily represent the views of the National Science Foundation.

References

Abadie, A. and Imbens, G. W. (2009) Matching on the estimated propensity score. *Working Paper 15301*. National Bureau of Economic Research, Cambridge.

Cochran, W. G. (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 295–313.
Deming, D. (2009) Early childhood intervention and life-cycle skill development: evidence from Head Start. *Am. Econ. J. Appl. Econ.*, **1**, no. 3, 111–134.
Groves, R., Dillman, D., Eltinge, J. and Little, R. J. A. (2002) *Survey Nonresponse*. New York: Wiley.
Hedges, L. V. and O'Muircheartaigh, C. A. (2010) Improving inference for population treatment effects from randomized experiments. *Working Paper*. Northwestern University Institute for Policy Research, Evanston.
Kalton, G. (1968) Standardization: a technique to control for extraneous variables. *Appl. Statist.*, **17**, 118–136.
Kitagawa, E. M. (1964) Standardized comparisons in population research. *Demography*, **1**, 296–315.
Lalonde, R., O'Muircheartaigh, C. and Perkins, J. (2005) Mapping cultural participation in Chicago. Irving B. Harris Graduate School of Public Policy Studies, University of Chicago, Chicago. (Available from `culturalpolicy.uchicago.edu/publications/MappingCPICweb.pdf`.)
Little, R. J. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley.
Myers, D. and Schirm, A. (1999) The impacts of upward bound: final report for the phase I of the national evaluation. *Report*. Mathematica Policy Research, Washington DC.
Oaxaca, R. (1973) Male-female wage differentials in urban labor markets. *Int. Econ. Rev.*, **14**, 693–709.
Roschelle, J., Shechtman, N., Tatar, D. and Hegedus, S. (2010) Integration of technology, curriculum, and professional development for advancing middle school mathematics. *Am. Educ. Res. J.*, **47**, 833–878.
Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rosenbaum, P. R. and Rubin, D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Ass.*, **79**, 516–524.
Rosenberg, P. (1962) Test factor standardization as a method of interpretation. *Socl Forces*, **41**, 53–61.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D. B. (1977) Assignment to treatment on the basis of a covariate. *J. Educ. Statist.*, **2**, 2–16.
Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.
Rubin, D. B. (1980) Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *J. Am. Statist. Ass.*, **75**, 591–593.
Rubin, D. B. (1986) Which ifs have causal answers? *J. Am. Statist. Ass.*, **83**, 961–962.
Rubin, D. B. (2006) *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.
Tipton, E. (2013) Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *J. Educ. Behav. Statist.*, **38**, 239–266.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Table of results for all 254 counties of Texas'.