Journal of Official Statistics
vol. 35, 2 (junio 2019)

sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 287–317, http://dx.doi.org/10.2478/JOS-2019-0014

Remarks on Geo-Logarithmic Price Indices

Jacek Białek[1]

As is known, all geo-logarithmic indices enjoy the axiomatic properties of being proportional, commensurable and homogeneous, together with their cofactors (Martini 1992a). Geo-logarithmic price indices satisfying the axioms of monotonicity, basis reversibility and factor reversibility have been investigated by Marco Fattore (2010), who has shown that the superlative Fisher price index does not belong to this family of indices. In this article, we discuss geo-logarithmic price indices with reference to the Laspeyres-Paasche bounding test and we propose a modification of the considered index family that satisfies this test. We also modify the structure of geo-logarithmic indices by using an additional parameter and, following the economic approach, we list superlative price index formulas that are members of the considered price index family. We obtain a special subfamily that approximates superlative price indices and includes the Fisher, Walsh and Sato-Vartia price indices.

Key words: Price index theory; geo-logarithmic price indices; superlative indices.

1. Introduction

The literature on the axiomatic index theory is very extensive (Krstcha 1988; Balk 1995; Von der Lippe 2007). From a theoretical point of view, a well-constructed index should satisfy a group of postulates (tests) arising from the axiomatic index theory. A system of minimum requirements for an index comes from Martini (1992b). According to the above-mentioned system, a price index should satisfy at least three conditions: *identity*, *commensurability* and *linear homogeneity* (see Appendix A, Subsection 8.1). German index theoreticians – Eichhorn and Voeller (1976) – introduced a more generally acceptable system (EV) of five, and later also of four, axioms: *strict monotonicity, price dimensionality*, *commensurability, identity* and (optionally) *linear homogeneity*. These five axioms imply other tests such as *proportionality* (*identity* plus *linear homogeneity*) or *quantity dimensionality* (*price dimensionality* plus *commensurability*) – see Von der Lippe (2007). In the literature, we can also encounter other systems – for example Olt (1996) examined several systems that provide less restrictive requirements than EV-systems. Moreover, some authors consider general price index formulas as having the above-mentioned desirable properties (Diewert 1976; Hill 2006; Fattore 2010; Białek 2012).

[1] Department of Statistical Methods, University of Lodz, ul. Uniwersytecka 3, 90-137, Lodz, Poland. Email: jbialek@uni.lodz.pl.

Acknowledgments: The article is financed by the National Science Centre in Poland (grant no. 2017/25/B/HS4/00387). I would like to thank some anonymous reviewers for their very helpful comments and inspiring remarks.

In this article, we discuss geo-logarithmic price indices, being a class of indices that contains several well-known indices and thus provides a useful framework for comparing properties of different index formulas (Fattore 2006, 2010). We analyse and modify geo-logarithmic price indices with reference to the *Laspeyres-Paasche bounding test,* which can be motivated in the 'economic approach'. According to this approach, upper and lower bounds for the index are provided by the Laspeyres and Paasche price index formulas. This follows from the choice of a cost of living index (COLI) as a target for the index, with an assumption about consumers' cost minimising behaviour. From the economic approach point of view, a "good" index should have a value between the above-mentioned bounds, that is, it should satisfy the *Laspeyres-Paasche bounding test*. We also modify the structure of geo-logarithmic indices by using an additional parameter and, following the economic approach, we list superlative price index formulas that are members of the considered price index family or obtained as the first-order approximation of the geo-logarithmic price index.

Our motivation has its genesis in the inflation measurement. The final report of the Boskin Commission begins with a recommendation that "the Bureau of Labour Statistics (BLS) should establish a cost of living index (COLI) as its objective in measuring consumer prices" (Boskin et al. 1996, 2). Further discussion on the theory of the COLI can be found in the following papers: Diewert (1993), Jorgenson and Slesnick (1983), and Pollak (1989). In practice, the Laspeyres price index is used to measure the Consumer Price Index (CPI) – see White (1999), Clements and Izan (1987). The Lapeyres formula does not take into account changes in the structure of consumption that occur as a result of price changes in a given time interval. It leads to the conclusion that the Laspeyres index can be biased due to the commodity substitution. Many economists consider the *superlative indices* (such as the Fisher index, the Walsh or the Törnqvist index) to be the best approximation of COLI (Von der Lippe 2007). Thus, any general classes of indices (such as the geo-logarithmic price index family) including these superlative index formulas seem to be especially interesting from the theoretical and practical point of view.

From a theoretical point of view, the feature of belonging to the geo-logarithmic price index family is a reason to consider the price index as good in the context of, for example, Martini's system of minimal requirements. The indices belonging to the discussed class have good properties, which is discussed in the further part of the article, although it cannot, of course, be said that an index outside this class does not have these properties. The author's modifications of the geo-logarithmic price index family proposed in the article yield indices that additionally fulfil the Laspeyres-Paasche bounding test, which is a desirable feature from the point of view of the economic approach. Moreover, it turns out that the relevant subclasses of one of the geo-logarithmic price index family modifications are notably close to the recognised superlative indices (the Fisher or Walsh indices), at the same time being their superset. From a practical point of view, the use of geo-logarithmic indices can also bring many benefits. If the world switches to the use of scanner data (e.g., in the CPI, HICP estimation, and so on) with "on-line" availability of data, then it could be possible to control x and y parameters (occurring in the class formula) to optimise that is, variance or mean square error in the geo-logarithmic index, used, for example, as the CPI (inflation) estimator. Thus, using, for example, a subclass where values fluctuate around superlative indices, it will be possible to select among those elements one that has

distinctive statistical properties. Finally, the issue of geo-logarithmic indices seems to be interesting in itself, as there are still a few open, scientific problems. For example, one can inquire whether the range of index variability of this class is wider/narrower in relation to the variability range of superlative indices or whether some subclass of the geo-logarithmic class generates only superlative indices.

The article is organised as follows: Section 2 introduces the geo-logarithmic price index family, Section 3 presents its axiomatic properties and its particular subfamily, Section 4 provides generalisations of this family and discusses their properties and particular cases, Section 5 is a simulation study of all the considered index families, Section 6 is an empirical study, Section 7 provides some final comments and points out some open issues needing further research, Appendix (Section 8) contain definitions of basic index axioms and some computational details needed in the article.

2. Geo-Logarithmic Price Index Family

Let us consider a group of N commodities observed at times s, t (the time moment s is considered as the *basis*) and let us denote:

$p_s = [p_{s1}, p_{s2}, \ldots, p_{sN}]'$ – a vector of prices at time s;
$p_t = [p_{t1}, p_{t2}, \ldots, p_{tN}]'$ – a vector of prices at time t;
$q_s = [q_{s1}, q_{s2}, \ldots, q_{sN}]'$ – a vector of quantities at time s;
$q_t = [q_{t1}, q_{t2}, \ldots, q_{tN}]'$ – a vector of quantities at time t.

Let us denote by $\tau(x,y)$ the logarithmic mean of two positive real numbers x and y, that is,

$$\tau(x, y) = \frac{x - y}{\ln(x) - \ln(y)}, \tag{1}$$

if $x \neq y$ and $\tau(x, y) = x$ otherwise (Carlson 1972).

For $x, y \in [0, 1]$, let q^x and q^y be two vectors whose components are defined as follows

$$q_i^x = q_{ti}^x q_{si}^{1-x}, \quad q_i^y = q_{ti}^y q_{si}^{1-y}, \quad \text{for} \quad i = 1, 2, \ldots, N \tag{2}$$

and let

$$w_{ti}^x = \frac{p_{ti} q_i^x}{\sum_{i=1}^N p_{ti} q_i^x}, \tag{3}$$

$$w_{si}^y = \frac{p_{si} q_i^y}{\sum_{i=1}^N p_{si} q_i^y}. \tag{4}$$

The geo-logarithmic, or the P_{xy}, family is the class of price indices defined by (Fattore 2006)

$$P_{xy}(q_s, q_t, p_s, p_t) = \prod_{i=1}^N \left(\frac{p_{ti}}{p_{si}}\right)^{v_i^{xy}}, \tag{5}$$

where weights v_i^{xy} are as follows

$$v_i^{xy} = \frac{\tau\left(w_{ti}^x, w_{si}^y\right)}{\sum_{j=1}^N \tau\left(w_{tj}^x, w_{sj}^y\right)}. \tag{6}$$

The following theorem (Fattore 2010) is the fundamental result for the P_{xy} parameterisation.

Theorem 1. The mapping associating the pair $(x, y) \in [0, 1] \times [0, 1]$ with the index P_{xy} is one to one, that is, if $(x, y) \neq (u, v)$, then $P_{xy} \neq P_{uv}$.

3. Axiomatic Properties of Geo-Logarithmic Price Indices

The geo-logarithmic family of price indices was proposed by the Italian statistician Martini (1992a). As was mentioned before, from a theoretical point of view, a well-constructed index should satisfy a group of postulates (tests) arising from the axiomatic index theory. Although there is no universal agreement on the axiomatic properties for a formula to be considered as an index (IMF 2004), one of commonly accepted systems of minimum requirements for the price index formula comes also from Martini (1992b). Obviously, each P_{xy} index satisfies *identity* and since Theorem 2 holds (Subsection 3.1), the geo-logarithmic price indices fulfil Martini's minimal requirements.

3.1. List of Axioms

In Fattore (2010), we can find proof of the following theorems.

Theorem 2. Geo-logarithmic price indices P_{xy} satisfy: (1) *proportionality*, (2) *commensurability* and (3) *homogeneity*. Moreover, the *basis reversibility* axiom holds if and only if $y = 1 - x$.

Theorem 3. An index from the P_{xy} class is *monotonic* if and only if $x = y$.

The immediate conclusion from Theorem 2 and Theorem 3 is the fact that the only monotonic geo-logarithmic price index being basis reversible is $P_{0.5\ 0.5}$ (Subsection 3.2). In Fattore (2010), it is proved that the only *factor reversible* element of the P_{xy} family is the Sato-Vartia index P_{10} (Von der Lippe 2007).

3.2. Special Subfamily P_{xx}

Since Theorem 3 holds and taking into consideration the *monotonicity* axiom from the EV-system, it seems interesting to consider a special subfamily P_{xx}. Let us note that for $x = y$ from (5) and (6) we obtain (Fattore 2010)

$$P_{xx} = \frac{\sum_{i=1}^N p_{ti}\, q_i^x}{\sum_{i=1}^N p_{si}\, q_i^x} = \frac{\sum_{i=1}^N p_{ti}\, q_{ti}^x\, q_{si}^{1-x}}{\sum_{i=1}^N p_{si}\, q_{ti}^x\, q_{si}^{1-x}}. \tag{7}$$

In particular, we obtain some known price index formulas. For instance, the Laspeyres (P_{La}), Paasche (P_{Pa}) and Walsh (P_{W}) price indices can be expressed as

$$P_{\mathrm{La}} = \frac{\sum_{i=1}^{N} p_{ti}\, q_{si}}{\sum_{i=1}^{N} p_{si}\, q_{si}} = P_{00}, \tag{8}$$

$$P_{\mathrm{Pa}} = \frac{\sum_{i=1}^{N} p_{ti}\, q_{ti}}{\sum_{i=1}^{N} p_{si}\, q_{ti}} = P_{11}, \tag{9}$$

$$P_{\mathrm{W}} = \frac{\sum_{i=1}^{N} p_{ti}\sqrt{q_{si}\, q_{ti}}}{\sum_{i=1}^{N} p_{si}\sqrt{q_{si}\, q_{ti}}} = P_{0.5\ \ 0.5}. \tag{10}$$

Example 1

Let us take into consideration a group of $N = 12$ commodities, where prices and quantities at time moments s and t are presented in Table 1. Figure 1 presents functions P_{xy} and P_{xx} for

Table 1. The values of prices and quantities at time moments s and t.

Commodity	q_s	q_t	p_s	p_t
1	350	200	900	1000
2	550	200	1600	1700
3	5000	3000	460	500
4	710	500	3	3.2
5	350	340	100	105
6	890	700	1000	1150
7	850	800	900	1000
8	600	500	1530	1600
9	5000	3000	480	500
10	700	500	4	4.2
11	550	340	100	110
12	800	700	1000	1100

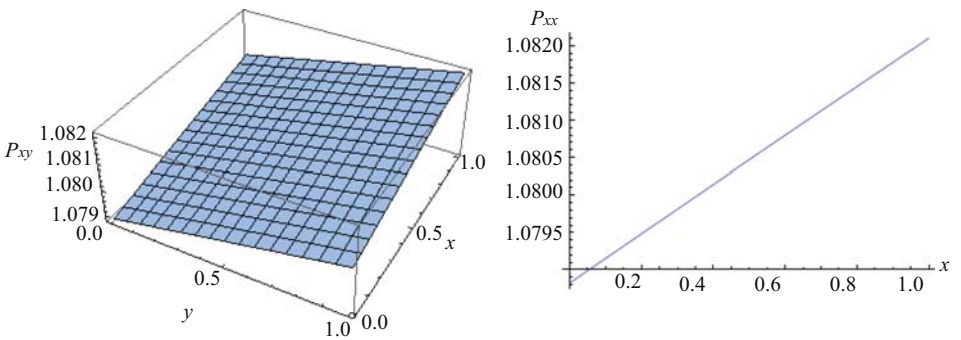

Fig. 1. Functions P_{xy} and P_{xx} depending on x and y for dataset described in Table 1.

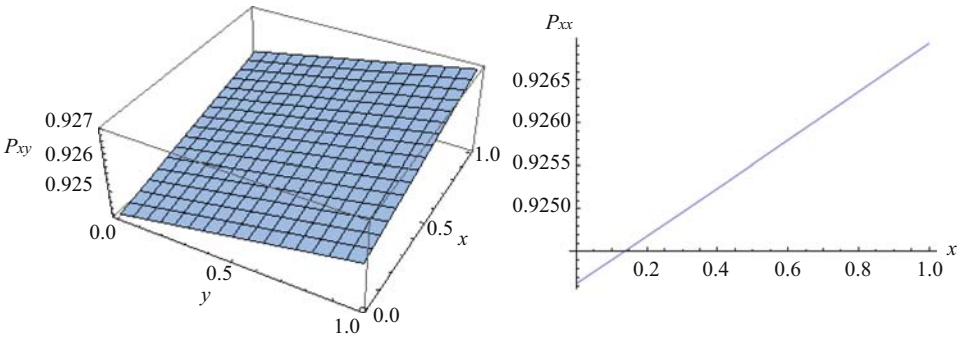

Fig. 2. Functions P_{xy} and P_{xx} depending on x and y for the reverse case (t is the base period).

$x, y \in [0, 1]$. Figure 2 presents functions P_{xy} and P_{xx} for the reverse case, that is, when the moment t is treated as the base period. It suggests that in the case of negative correlation between prices and quantities, the P_{xy} formula is a monotonic (here increasing) function of its arguments, that is, in our example the value of P_{xy} goes up if x or y increases. If the suggestion were true, from (8) and (9) we would have an immediate conclusion that P_{xy} satisfies the Laspeyres-Paasche bounding test. In fact, it is not generally true (see Subsection 3.3).

3.3. Geo-logarithmic Price Indices and the Laspeyres-Paasche Bounding Test

The Consumer Price Index approximates changes in costs of household consumption assuming constant utility, particularly in settings where COLI, Cost of Living Index, is chosen as a target for the index. In the so-called economic approach, the upper and lower bounds for the COLI are provided by the Laspeyres and Paasche price index formulas. If the price index value is within these bounds, then we say that this price index satisfies the Laspeyres-Paasche bounding test belonging to the group of mean value tests (Von der Lippe 2007).

Example 2

Let us take into consideration a group of $N = 4$ commodities where prices and quantities at time moments s and t are presented in Table 2. Figure 3 presents the function P_{xy} for $x, y \in [0, 1]$. Figure 4 presents the function P_{x1} for $x \in [0, 1]$. Figure 5 presents the function P_{xx} for $x \in [0, 1]$.

Table 2. The values of prices and quantities at time moments s and t.

Commodity	q_s	q_t	p_s	p_t
1	300	200	80	90
2	1200	900	500	550
3	2000	1	120	130
4	4.1	4	30000	31500

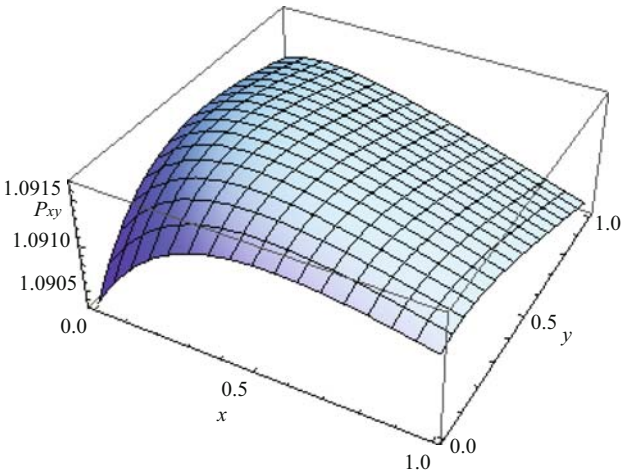

Fig. 3. Function P_{xy} depending on x and y for dataset described in Table 2.

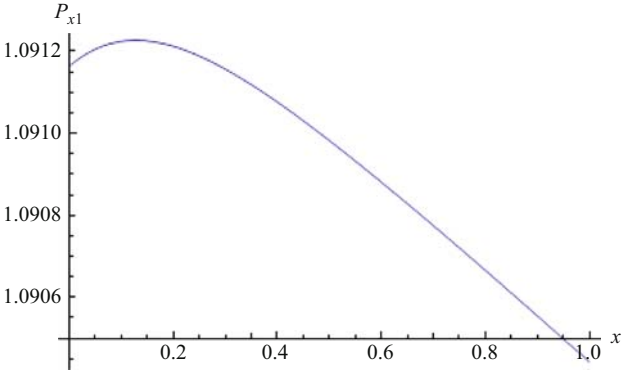

Fig. 4. Function P_{x1} depending on x for dataset described in Table 2.

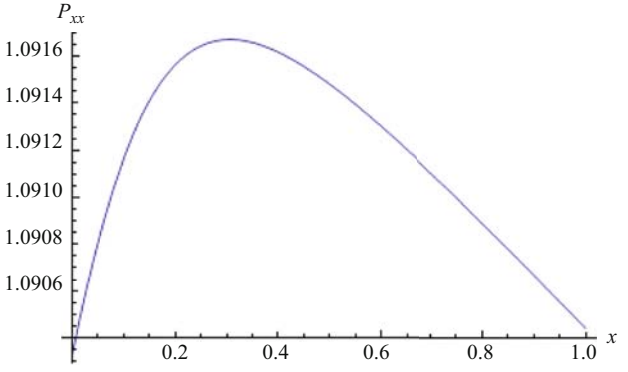

Fig. 5. Function P_{xx} depending on x for dataset described in Table 2.

Observing Figures 1, 2 and 3, we conclude that even if changes between prices and quantities are inversely related, the indices from P_{xy} or P_{xx} families may fail the Laspeyres-Paasche bounding test since $P_{00} = P_{\mathrm{La}}$ and $P_{11} = P_{\mathrm{Pa}}$. Moreover, the P_{xy} formula does not have to be a monotonic function of its arguments. Obviously, the quantity response to price changes is extremely strong in the case of commodity number 3 and it would not be observed in practice. Nevertheless, any considered and accepted test from the axiomatic price index theory must hold for any vectors of prices and quantities. The following question arises: what about the case when the quantity response is not so extreme (it is naturally limited) and still prices and quantities are inversely related? To answer this question, we run a simulation study (see Section 5) in which the parameter connected with the quantity changes is controlled.

3.4. *Geo-logarithmic Price Indices and Superlative Indices*

Following the economic approach to the price index theory, Diewert proposed the special family of indices that he called *superlative* (Diewert 1976). Although the axiomatic and the economic approaches differ from each other, connections between them are worth studying (Von der Lippe 2007). Fattore (2010) has proven that the only superlative index number belonging to the geo-logarithmic family is the Walsh index ($P_{0.5\ 0.5}$). Among superlative price indices, a very important role is played by the Törnqvist index:

$$P_{\mathrm{T}} = \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}} \right)^{\frac{w_{si}^0 + w_{ti}^1}{2}}, \tag{11}$$

which does not belong to the P_{xx} family (Fattore 2010). Nevertheless, in the same paper it is proved that the first-order approximation of the geo-logarithmic price index has a Törnqvist-like form. Similarly, the Fisher price index

$$P_{\mathrm{F}} = \sqrt{P_{\mathrm{La}} P_{\mathrm{Pa}}}, \tag{12}$$

is not a member of the geo-logarithmic price index family but since the superlative Fisher and Törnqvist indices approximate each other (Dumagan 2002), the Fisher price index also should approximate the geo-logarithmic price indices.

Example 3

Let us use data from Example 1. Figure 6 presents the function $|P_{xy} - P_{\mathrm{F}}|$ depending on $x, y \in [0, 1]$. Figure 7 presents the function $|P_{xx} - P_{\mathrm{F}}|$ depending on $x \in [0, 1]$.

We observe (See Figure 6) that the best Fisher index approximation that uses P_{xy} indices is obtained here for $y = 1 - x$. The $P_{x\,1-x}$ subfamily was investigated by Fattore (2010). He has proven that $P_{x\,1-x}$ indices satisfy the Martini's minimal requirements.

Let us also note that the *basis reversibility* axiom holds if and only if $y = 1 - x$ (see Theorem 2). Thus, further investigations on the $P_{x\,1-x}$ subfamily seem to be especially interesting. Observing Figure 7, we can see that the best Fisher index approximation that uses P_{xx} formulas is obtained for $x = 0.5$, which leads to the Walsh price index ($P_{0.5\ 0.5}$) being the only *monotonic* element of the $P_{x\,1-x}$ subfamily. It is not surprising since the superlative price indices approximate each other. However, this is not the best Fisher price

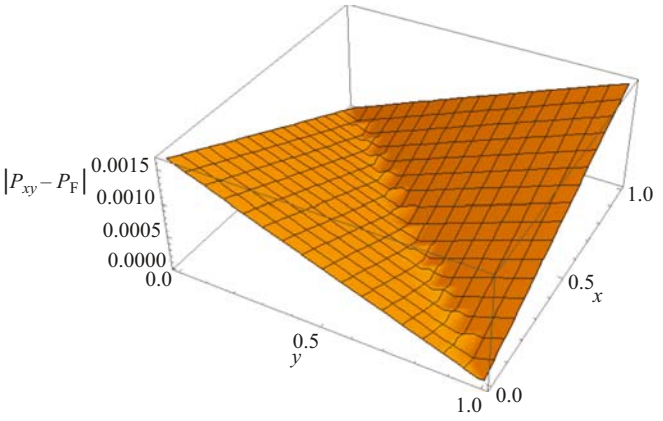

Fig. 6. Function $|P_{xy} - P_F|$ *depending on* $x, y \in [0, 1]$ *for dataset described in Table 1.*

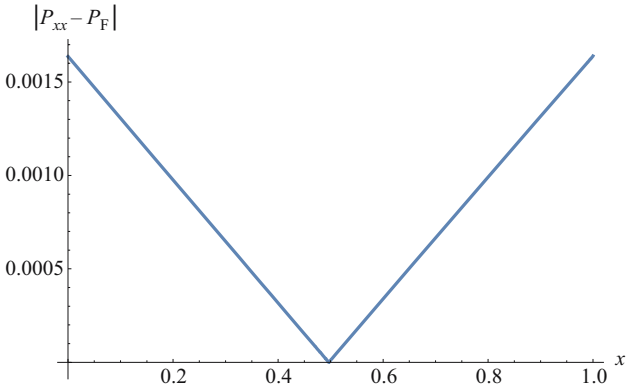

Fig. 7. Function $|P_{xx} - P_F|$ *depending on* $x \in [0, 1]$ *for dataset described in Table 1.*

index approximation in our study, that is, although $P_W = 1.08047 \approx P_F = 1.08046$, the index P_{01} seems to be a better proxy for the Fisher index value (See Figure 8). Please note that the P_{01} index is not the Sato-Vartia price index (it is easy to verify that, in general, values of P_{01} differ from values of P_{10}).

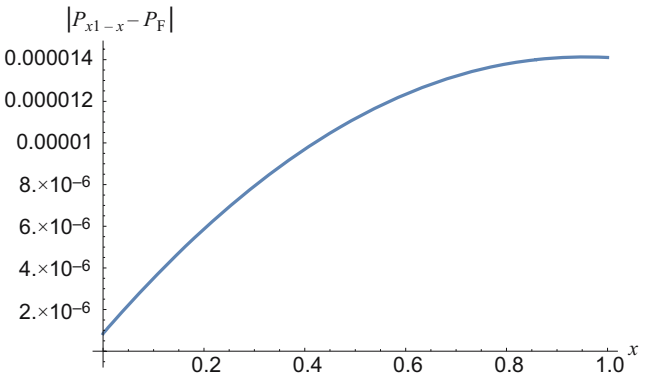

Fig. 8. Function $|P_{x1-x} - P_F|$ *depending on* $x \in [0, 1]$ *for dataset described in Table 1.*

4. Generalisation of the Geo-Logarithmic Price Index Family

We consider two problems here. Firstly, it would be interesting to modify the structure of the geo-logarithmic family to obtain the price index family $\tilde{P}_{xyz}$ including the Fisher index. Secondly, we intend to verify consequences of changing the weighted geometric mean into the weighted arithmetic mean of quantities in $\tilde{P}_{xxz}$ subfamily.

4.1. Generalisation Through an Additional Parameter

Similarly to (2), (3), (4) and (6), let us denote by

$$q_i^{Ax} = q_{ti}^{x}\, q_{si}^{1-x}, \quad q_i^{Ay} = q_{ti}^{y}\, q_{si}^{1-y}, \tag{13}$$

$$q_i^{Bx} = q_{ti}^{1-x} q_{si}^{x}, \quad q_i^{By} = q_{ti}^{1-y} q_{si}^{y}, \tag{14}$$

$$w_{ti}^{Ax} = \frac{p_{ti}\, q_i^{Ax}}{\sum_{i=1}^{N} p_{ti}\, q_i^{Ax}}, \quad w_{si}^{Ay} = \frac{p_{si}\, q_i^{Ay}}{\sum_{i=1}^{N} p_{si}\, q_i^{Ay}}, \tag{15}$$

$$w_{ti}^{Bx} = \frac{p_{ti}\, q_i^{Bx}}{\sum_{i=1}^{N} p_{ti}\, q_i^{Bx}}, \quad w_{si}^{By} = \frac{p_{si}\, q_i^{By}}{\sum_{i=1}^{N} p_{si}\, q_i^{By}}, \tag{16}$$

$$v_{Ai}^{xy} = \frac{\tau\left(w_{ti}^{Ax}, w_{si}^{Ay}\right)}{\sum_{j=1}^{N} \tau\left(w_{tj}^{Ax}, w_{sj}^{Ay}\right)}, \quad v_{Bi}^{xy} = \frac{\tau\left(w_{ti}^{Bx}, w_{si}^{By}\right)}{\sum_{j=1}^{N} \tau\left(w_{tj}^{Bx}, w_{sj}^{By}\right)}, \tag{17}$$

for $i = 1, 2, \ldots, N$, $x, y \in [0, 1]$.

Under significations (13)–(17), we define the new class of price indices ($\tilde{P}_{xyz}$) as follows

$$\tilde{P}_{xyz} = \left\{ \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}} \right)^{v_{Ai}^{xy}} \right\}^{z} \left\{ \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}} \right)^{v_{Bi}^{xy}} \right\}^{1-z}, \quad \text{for} \quad x, y, z \in [0, 1]. \tag{18}$$

Firstly, let us note that for fixed values of x, y and z the price index $\tilde{P}_{xyz}$ fulfils the Martini's minimal requirements since it can be expressed as a weighted geometric mean of two price indices (with weights z and $1 - z$), satisfying the Martini's minimal requirements (see Appendix B, Subsection 8.2). In fact, these two price indices (defined inside curly brackets in Equation 18) satisfy the Martini's minimal requirements. The first one (on the left side of Equation 18) is identical with P_{xy} index (for fixed values of x and y) and its axiomatic properties were proved by Fattore (2010). The proof of the same group of axioms in the case of the second price index (inside curly brackets on the right side of Equation18) would be analogous.

Secondly, let us note that the following relation holds

$$\tilde{P}_{xy1} = \tilde{P}_{1-x\ \ 1-y\ \ 0} = P_{xy}, \quad \text{for} \quad x, y \in [0, 1], \tag{19}$$

which means that the P_{xy} family is a special case of the $\tilde{P}_{xyz}$ family.

Moreover, $\tilde{P}_{101}$ and $\tilde{P}_{010}$ are the Sato-Vartia indices and also we obtain

$$\tilde{P}_{001} = \tilde{P}_{110} = P_{\mathrm{La}}, \tag{20}$$

$$\tilde{P}_{111} = \tilde{P}_{000} = P_{\mathrm{Pa}}, \tag{21}$$

$$\tilde{P}_{\frac{111}{222}} = P_{\mathrm{W}}, \tag{22}$$

and, what is more interesting, we have

$$\tilde{P}_{00\frac{1}{2}} = \tilde{P}_{11\frac{1}{2}} = P_{\mathrm{F}}. \tag{23}$$

Finally, the following approximation can be proved (see Appendix C, Subsection 8.3).

$$\forall i \in \{1, 2, \ldots, N\} \quad q_{si} \approx q_{ti} \ \wedge \ w_{si}^{x} \approx w_{ti}^{y} \ \Rightarrow \ \tilde{P}_{xyz} \approx P_{\mathrm{T}}. \tag{24}$$

Example 4

Let us use data from Example 1. Figure 9 presents the function $\tilde{P}_{xy\frac{1}{2}}$ depending on $x, y \in [0, 1]$.

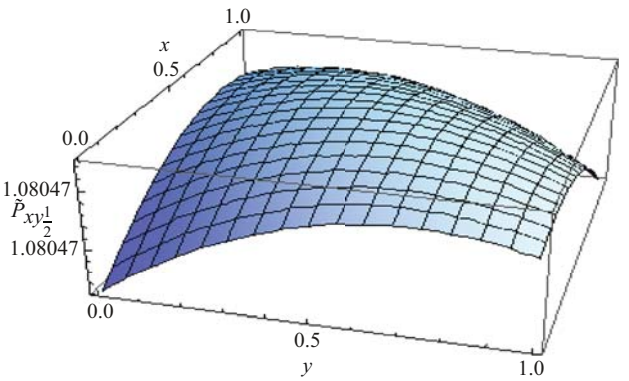

Fig. 9. Function $\tilde{P}_{xy\frac{1}{2}}$ depending on $x, y \in [0, 1]$ for dataset described in Table 1.

As we can see, the interval of values of indices from the considered family (for $z = 0.5$) is very narrow and they fluctuate around superlative index values ($P_{\mathrm{W}} = 1.08047, P_{\mathrm{F}} = 1.08046$).

Example 5

Let us take into consideration a group of $N = 5$ commodities where prices and quantities at time moments s and t are presented in Table 3. Figure 10 presents the function P_{xy} for $x, y \in [0, 1]$. Figure 11 presents the function $\tilde{P}_{xy\frac{1}{2}}$ for $x, y \in [0, 1]$. Similarly to the results

obtained in the Example 4, the interval of values of indices from the $\tilde{P}_{xy\frac{1}{2}}$ family is very narrow. We compare its range with the range obtained for a class of superlative price indices introduced by Diewert (1976). The Diewert's proposition of the above-mentioned class of indices is as follows

$$P_{\mathrm{D}}(r) = \left(\frac{\sum_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{\frac{r}{2}} \frac{p_{si}q_{si}}{p_s q_s}}{\sum_{i=1}^{N} \left(\frac{p_{si}}{p_{ti}}\right)^{\frac{r}{2}} \frac{p_{ti}q_{ti}}{p_t q_t}} \right)^{\frac{1}{r}}, \tag{25}$$

Table 3. The values of prices and quantities at time moments s and t.

Commodity	q_s	q_t	p_s	p_t
1	100	70	80	90
2	820	900	500	550
3	20000	15000	120	130
4	50	40	30000	31500
5	4000	3000	3	3.5

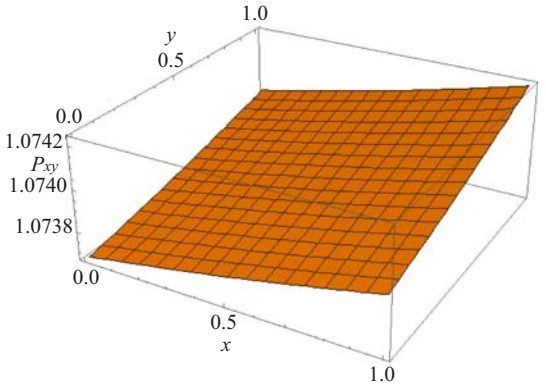

Fig. 10. Function P_{xy} depending on x and y for dataset described in Table 3.

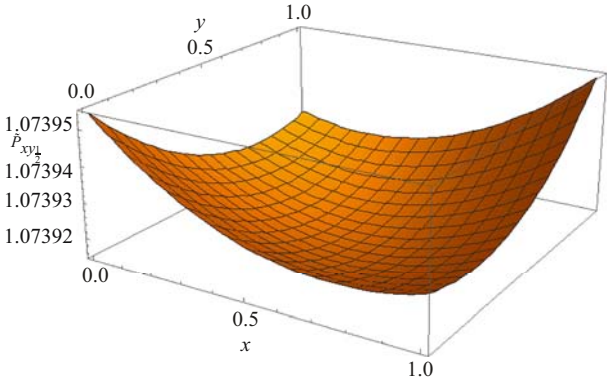

Fig. 11. Function $\tilde{P}_{xy\frac{1}{2}}$ depending on x and y for dataset described in Table 3.

where $r \in R\backslash\{0\}$ and

$$p_\tau q_\tau = \sum_{k=1}^{N} p_{\tau k}\, q_{\tau k}, \quad \text{for} \quad \tau = s, t. \tag{26}$$

Figure 12 presents the function $P_{\mathrm{D}}(r)$ for $r \in [-1000, 1000]\backslash\{0\}$. After optimisation of functions P_{xy}, $\tilde{P}_{xy\frac{1}{2}}$ and $P_{\mathrm{D}}(r)$ we obtain their following ranges: $P_{xy} \in [1.07367,\ 1.07424]$, $\tilde{P}_{xy\frac{1}{2}} \in [1.07392,\ 1.07395]$ and $P_{\mathrm{D}}(r) \in [1.07393,\ 1.10587]$. The length of the interval of possible index values is the smallest in the case of the family $\tilde{P}_{xy\frac{1}{2}}$. The open question is whether the above conclusion has a general character.

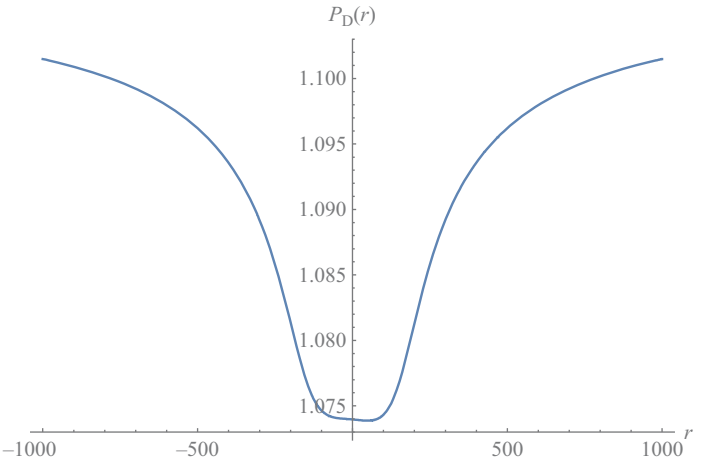

Fig. 12. Function $P_{\mathrm{D}}(r)$ for $r \in [-1000,1000]\backslash\{0\}$.

4.2. Modification Through Mean Change

Fattore (2010) shows that

$$\prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{v_{Ai}^{xx}} = \frac{\sum_{i=1}^{N} p_{ti} q_{ti}^{x} q_{si}^{1-x}}{\sum_{i=1}^{N} p_{si} q_{ti}^{x} q_{si}^{1-x}}, \tag{27}$$

and, by the analogy, we obtain

$$\prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{v_{Bi}^{xx}} = \frac{\sum_{i=1}^{N} p_{ti} q_{ti}^{1-x} q_{si}^{x}}{\sum_{i=1}^{N} p_{si} q_{ti}^{1-x} q_{si}^{x}}. \tag{28}$$

From (18), (27) and (28) we obtain

$$\tilde{P}_{xxz} = \left(\frac{\sum_{i=1}^{N} p_{ti} q_{ti}^{x} q_{si}^{1-x}}{\sum_{i=1}^{N} p_{si} q_{ti}^{x} q_{si}^{1-x}}\right)^{z} \left(\frac{\sum_{i=1}^{N} p_{ti} q_{ti}^{1-x} q_{si}^{x}}{\sum_{i=1}^{N} p_{si} q_{ti}^{1-x} q_{si}^{x}}\right)^{1-z}. \tag{29}$$

Let us note that if we change the geometric mean of quantities into the arithmetic mean of quantities in the $\tilde{P}_{xxz}$ formula, we obtain

$$\tilde{P}^{A}_{xxz} = \left(\frac{\sum_{i=1}^{N} p_{ti}(xq_{ti} + (1-x)q_{si})}{\sum_{i=1}^{N} p_{si}(xq_{ti} + (1-x)q_{si})} \right)^{z} \left(\frac{\sum_{i=1}^{N} p_{ti}((1-x)q_{ti} + xq_{si})}{\sum_{i=1}^{N} p_{si}((1-x)q_{ti} + xq_{si})} \right)^{1-z}, \quad (30)$$

This is still a quite general family of indices. In particular, we have

$$\tilde{P}^{A}_{00\frac{1}{2}} = \tilde{P}^{A}_{11\frac{1}{2}} = P_{\mathrm{F}}, \quad (31)$$

$$\tilde{P}^{A}_{000} = \tilde{P}^{A}_{111} = P_{\mathrm{Pa}}, \quad (32)$$

$$\tilde{P}^{A}_{001} = \tilde{P}^{A}_{110} = P_{\mathrm{La}}, \quad (33)$$

$$\tilde{P}^{A}_{\frac{1}{2}\frac{1}{2}\frac{1}{2}} = P_{\mathrm{ME}}. \quad (34)$$

where P_{ME} denotes the Marshal-Edgeworth price index (see Von der Lippe 2007).

What is more interesting, the following theorem can be proved (see Appendix D, Subsection 8.4).

Theorem 4. Each price index from the $\tilde{P}^{A}_{xxz}$ subfamily satisfies the *Laspeyres-Paasche bounding test*.

4.3. *Properties of Cofactors of Modified Geo-logarithmic Price Indices*

"Index numbers come in pairs in economic theory, one of price and the other a matching one of quantity. In economic practice they tend to be found paired off in this way (. . .). Such a pair may be designed to account for the variation in a value aggregate, as when movements in aggregate expenditure of consumers are analysed into the two components of changes in prices and in real consumption" (Allen 1975, 1).

According to the cited fragment and to ensure the joint consistency of both price and quantity comparisons it could be desirable in practice using such price indices which, together with their cofactors, satisfy fundamental tests from axiomatic index theory.

Let us note that for the given sets of prices and quantities, described by N-dimensional vectors p_s, p_t, q_s and q_t (see Section 2), the ratio

$$V(q_s, q_t, p_s, p_t) = \frac{\sum_{i=1}^{N} p_{ti}\, q_{ti}}{\sum_{i=1}^{N} p_{si}\, q_{si}} \quad (35)$$

is called the *value index* between time moments s and t. The aim of the price and quantity index theory is to decompose the value index as the product of two strictly positive functions

$$V(q_s, q_t, p_s, p_t) = P(q_s, q_t, p_s, p_t) \cdot Q(q_s, q_t, p_s, p_t), \quad (36)$$

where P and Q denote the well-defined price and quantity indices. The given price index formula $P(q_s, q_t, p_s, p_t)$ has its associated cofactor defined by

$$cof\ P(q_s, q_t, p_s, p_t) = \frac{V(q_s, q_t, p_s, p_t)}{P(q_s, q_t, p_s, p_t)}. \tag{37}$$

From (36) and (37) we have that the cofactor of a given price index is the associated quantity index. The geo-logarithmic price index family has the distinctive feature that the cofactors of its elements satisfy the proportionality and homogeneity axioms (see Appendix A, Subsection 8.1). From the axiomatic index theory (Balk 1995), we know that only the fulfilment of the factor reversibility axiom guarantees that the cofactor (with respect to quantities) satisfies all properties fulfilled by price index itself (with respect to prices). It can be easily explained since in that case the cofactor and the price index share the same functional form (Fattore 2010). As it is known, the factor reversibility test is very restrictive and it rules out most indices commonly used in practice, such as the Laspeyres index (Von der Lippe 2007). Many authors treat this axiom as a nonessential property. To ensure the joint consistency of both price and quantity comparisons, alternatively we can search for a class of price indices satisfying at least an important subset of fundamental axioms together with their cofactors. In this sense, such a class of indices can be considered "good". Motivated by looking for such a "good class", Martini (1992) proposed the geo-logarithmic price index family.

In the paper by Fattore (2010), it is proved that cofactors of geo-logarithmic price indices satisfy the proportionality and homogeneity axioms (see *Proposition 10* and its proof in this original work). Since the proportionality holds for any $x, y \in [0, 1]$ and for any positive real number k, we have

$$cof\ \ P_{xy}(q_s, q_t, p_s, kp_s) = k. \tag{38}$$

From (19) and (38) we conclude that

$$cof\ \ \tilde{P}_{xy1}(q_s, q_t, p_s, kp_s) = k. \tag{39}$$

Since the equality (39) holds for any $x, y \in [0, 1]$, we obtain as a consequence

$$cof\ \ \tilde{P}_{1-x\ \ 1-y\ \ 1}(q_s, q_t, p_s, kp_s) = k. \tag{40}$$

Let us note that any index from the $\tilde{P}_{xyz}$ family can be written as

$$\tilde{P}_{xyz} = (\tilde{P}_{xy1})^z (\tilde{P}_{1-x\ \ 1-y\ \ 1})^{1-z}. \tag{41}$$

From (41) we have

$$\begin{aligned} cof\ \ \tilde{P}_{xyz}(q_s, q_t, p_s, p_t) &= \frac{V(q_s, q_t, p_s, p_t)}{[\tilde{P}_{xy1}(q_s, q_t, p_s, p_t)]^z [\tilde{P}_{1-x\ \ 1-y\ \ 1}]^{1-z}} \\ &= \left[\frac{V(q_s, q_t, p_s, p_t)}{\tilde{P}_{xy1}(q_s, q_t, p_s, p_t)}\right]^z \left[\frac{V(q_s, q_t, p_s, p_t)}{\tilde{P}_{1-x\ \ 1-y\ \ 1}(q_s, q_t, p_s, p_t)}\right]^{1-z}, \end{aligned} \tag{42}$$

and it leads to the following conclusion

$$cof\ \ \tilde{P}_{xyz}(q_s, q_t, p_s, p_t) = [cof\ \ \tilde{P}_{xy1}(q_s, q_t, p_s, p_t)]^z \cdot [cof\ \ \tilde{P}_{1-x\ \ 1-y\ \ 1}(q_s, q_t, p_s, p_t)]^{1-z}. \tag{43}$$

From (39), (40) and (43) we obtain

$$cof\ \ \tilde{P}_{xyz}(q_s, q_t, p_s, kp_s) = k^z k^{1-z} = k. \tag{44}$$

Thus, cofactors of $\tilde{P}_{xyz}$ indices satisfy the proportionality axiom. The proof for the homogeneity could be done analogically. Let us note that the problem with these axioms appears in the case of the $\tilde{P}^A_{xxz}$ index family because weighting by arithmetic means of quantities makes the cofactors violating the proportionality axiom. In our opinion, it does not mean that such a choice of weights is wrong and cannot be accepted since indices from the $\tilde{P}^A_{xxz}$ family satisfy Martini's minimal requirements and they fulfil the Laspeyres-Paasche bounding test. Moreover, these indices remain quite stable even when prices are strongly fluctuated (see Simulation 2 in Section 5).

5. Simulation Study

Simulation 1

Let us take into consideration a group of $N = 12$ components where prices and quantities are normally distributed as follows:

$$p_i^\tau \sim N\left(p_{i0}^\tau, v_i^\tau p_{i0}^\tau\right)$$

$$q_i^\tau \sim N\left(q_{i0}^\tau, u_i^\tau q_{i0}^\tau\right)$$

where $\tau = s, t$, $N(\mu, \sigma)$ denotes the normal distribution with the mean μ and the standard deviation σ, v_i^τ denotes the volatility coefficient (coefficient of variation) of the $i-$th price at time τ, i.e., $v_i^\tau = D\left(p_i^\tau\right)/p_{i0}^\tau$, u_i^τ denotes the volatility coefficient of the $i-$th quantity at time τ, i.e., $u_i^\tau = D\left(q_i^\tau\right)/q_{i0}^\tau$. Before generating prices and quantities, we generated values of volatility coefficients using uniform distributions, that is, $v_i^\tau \sim U(0, v^\tau)$ and $u_i^\tau \sim U(0, u^\tau)$. Expected values of prices and quantities are described by vectors from Example 1, that is,

$$P_0^t = [1000,\ 1700,\ 500,\ 3.2,\ 105,\ 1150,\ 1000,\ 1600,\ 500,\ 4.2,\ 110,\ 1100]';$$

$$P_0^s = [900,\ 1600,\ 460,\ 3,\ 100,\ 1000,\ 900,\ 1530,\ 480,\ 4,\ 100,\ 1000]';$$

$$Q_0^t = [200,\ 200,\ 3000,\ 500,\ 340,\ 700,\ 800,\ 500,\ 3000,\ 500,\ 340,\ 700]';$$

$$Q_0^s = [350,\ 550,\ 5000,\ 710,\ 350,\ 890,\ 850,\ 600,\ 5000,\ 700,\ 550,\ 800]'.$$

In our experiment, we are going to control values of volatility coefficients of prices and quantities by setting values of v^s, v^t, u^s, u^t and observe their influence on the discussed

general indices and their distance to the Laspeyres and Paasche formulas. We consider several cases, that is, Case 1 (the volatilities of price and quantity processes are low and the quantity response to price changes is quite normal – see Example 1), Case 2 (the volatilities of prices and quantities are large, the quantity response to price changes is strongly fluctuated), Case 3 (the volatility of prices is small but the volatility of quantities is large, that is, the quantity response to price changes may be strong), Case 4 (the volatility of prices is large but the volatility of quantities is small, that is, the quantity response to price changes is rather small). For each case, we generate values of price and quantity vectors in $n = 1000$ repetitions. Let us denote for fixed values of x and y and for each of kth repetition:

$$\Delta 1_k = (P_{xy} - \min(P_{\mathrm{La}}, P_{\mathrm{Pa}}))_k, \tag{45}$$

$$\Delta 2_k = (\max(P_{\mathrm{La}}, P_{\mathrm{Pa}}) - P_{xy})_k, \tag{46}$$

$$\Delta 3_k = (\tilde{P}_{xy\frac{1}{2}} - \min(P_{\mathrm{La}}, P_{\mathrm{Pa}}))_k, \tag{47}$$

$$\Delta 4_k = (\max(P_{\mathrm{La}}, P_{\mathrm{Pa}}) - \tilde{P}_{xy\frac{1}{2}})_k. \tag{48}$$

Selected histograms (for special values of x and y) for random variables defined by (45) – (48) and for $v^s = v^t = u^s = u^t = 0.1$ are presented in Figures 13, 14 and 15. The simulation results are presented in Tables 4, 5, 6 and 7.

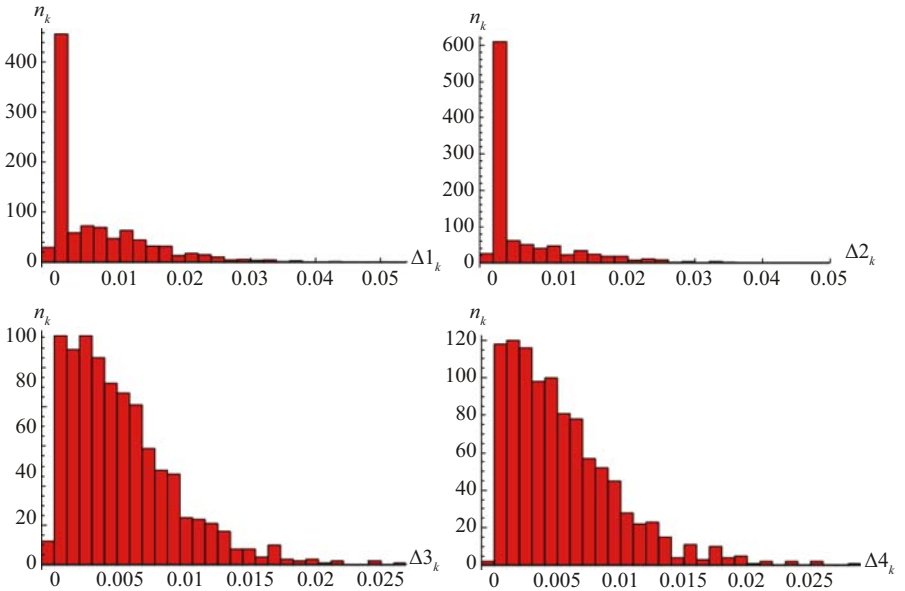

Fig. 13. Histograms for random variables $\Delta 1$, $\Delta 2$, $\Delta 3$ *and* $\Delta 4$ *and for* $x = y = 0.95$.

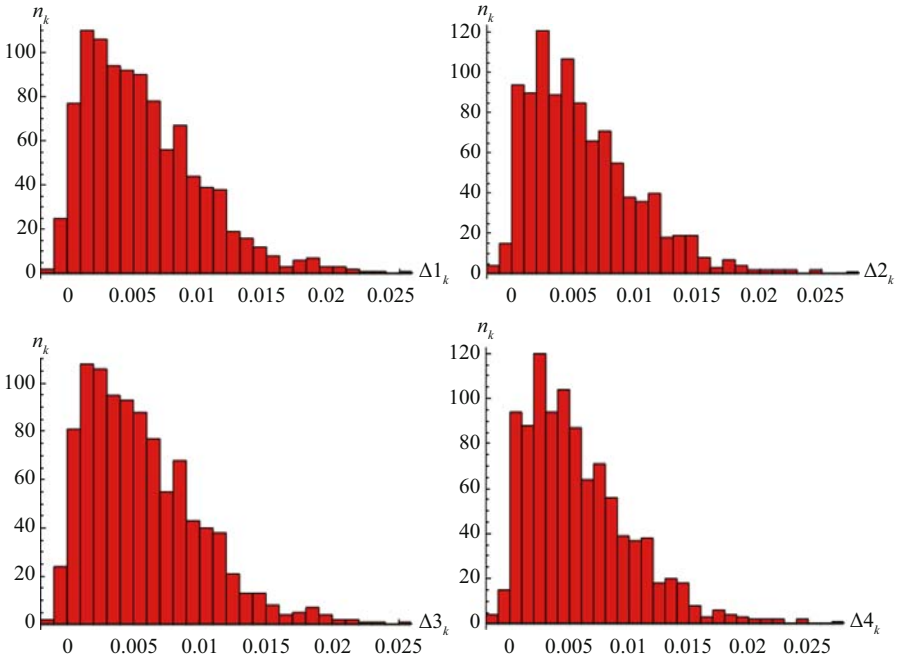

Fig. 14. Histograms for random variables $\Delta 1$, $\Delta 2$, $\Delta 3$ *and* $\Delta 4$ *and for* $x = 0.6$, $y = 0.4$.

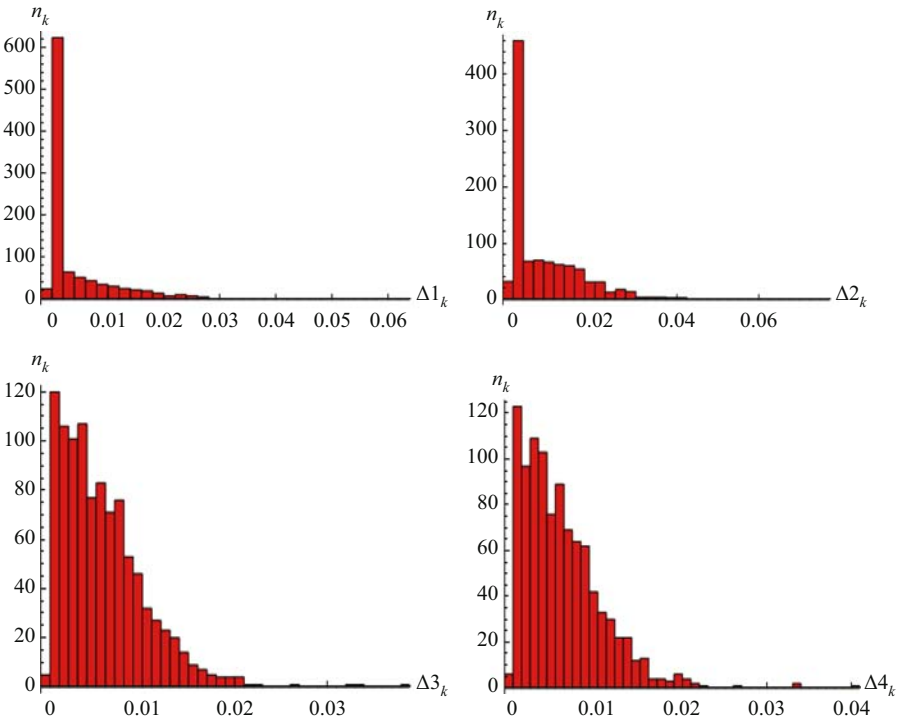

Fig. 15. Histograms for random variables $\Delta 1$, $\Delta 2$, $\Delta 3$ *and* $\Delta 4$ *and for* $x = y = 0.05$ *(*).(*) Expected values of Laspeyres and Paasche formulas equal respectively:* $P_{\mathrm{La}} = 1.054$, $P_{\mathrm{Pa}} = 1.044$.

Table 4. Verifying the Laspeyres-Paasche bounding test for P_{xy} and $\tilde{P}_{xy0.5}$ – Case 1.

Statistics	Case 1: $v^s = v^t = 0.05$; $u^s = u^t = 0.05$					
	$x = 0.05$ $y = 0.05$	$x = 0.25$ $y = 0.25$	$x = 0.5$ $y = 0.5$	$x = 0.75$ $y = 0.75$	$x = 0.95$ $y = 0.05$	$x = 0.95$ $y = 0.95$
Mean (P_{xy}) (Std. Dev.P_{xy})	1.101 (0.038)	1.097 (0.035)	1.102 (0.038)	1.087 (0.030)	1.086 (0.031)	1.103 (0.037)
Mean ($\tilde{P}_{xy0.5}$) (Std. Dev.$\tilde{P}_{xy0.5}$)	1.103 (0.038)	1.097 (0.034)	1.102 (0.038)	1.089 (0.031)	1.086 (0.031)	1.104 (0.038)
$card\{k : \Delta_{1k} < 0\}$	28	26	17	21	15	16
$card\{k : \Delta_{2k} < 0\}$	20	19	22	23	12	30
$card\{k : \Delta_{3k} < 0\}$	7	25	17	15	15	2
$card\{k : \Delta_{4k} < 0\}$	4	14	22	17	9	4

Table 5. Verifying the Laspeyres-Paasche bounding test for P_{xy} and $\tilde{P}_{xy0.5}$ – Case 2.

Statistics	Case 2: $v^s = v^t = 0.2$; $u^s = u^t = 0.2$					
	$x = 0.05$ $y = 0.05$	$x = 0.25$ $y = 0.25$	$x = 0.5$ $y = 0.5$	$x = 0.75$ $y = 0.75$	$x = 0.95$ $y = 0.05$	$x = 0.95$ $y = 0.95$
Mean (P_{xy}) (Std. Dev.P_{xy})	1.066 (0.127)	1.143 (0.140)	1.103 (0.126)	1.043 (0.130)	1.067 (0.134)	1.143 (0.134)
Mean ($\tilde{P}_{xy0.5}$) (Std. Dev.$\tilde{P}_{xy0.5}$)	1.082 (0.124)	1.148 (0.141)	1.103 (0.126)	1.023 (0.137)	1.067 (0.134)	1.131 (0.132)
$card\{k : \Delta_{1k} < 0\}$	28	28	39	44	21	34
$card\{k : \Delta_{2k} < 0\}$	22	25	32	30	33	31
$card\{k : \Delta_{3k} < 0\}$	3	26	39	34	19	3
$card\{k : \Delta_{4k} < 0\}$	7	20	32	21	27	5

Table 6. Verifying the Laspeyres-Paasche bounding test for P_{xy} and $\tilde{P}_{xy0.5}$ – Case 3.

Statistics	Case 3: $v^s = v^t = 0.05$; $u^s = u^t = 0.2$					
	$x = 0.05$ $y = 0.05$	$x = 0.25$ $y = 0.25$	$x = 0.5$ $y = 0.5$	$x = 0.75$ $y = 0.75$	$x = 0.95$ $y = 0.05$	$x = 0.95$ $y = 0.95$
Mean (P_{xy}) (Std. Dev.P_{xy})	1.068 (0.033)	1.092 (0.034)	1.047 (0.047)	1.107 (0.039)	1.113 (0.044)	1.110 (0.040)
Mean ($\tilde{P}_{xy0.5}$) (Std. Dev.$\tilde{P}_{xy0.5}$)	1.070 (0.032)	1.093 (0.034)	1.047 (0.047)	1.109 (0.041)	1.113 (0.044)	1.109 (0.041)
$card\{k : \Delta_{1k} < 0\}$	24	27	34	40	24	28
$card\{k : \Delta_{2k} < 0\}$	31	30	33	25	19	34
$card\{k : \Delta_{3k} < 0\}$	8	21	34	35	24	3
$card\{k : \Delta_{4k} < 0\}$	6	21	33	20	19	4

Table 7. Verifying the Laspeyres-Paasche bounding test for P_{xy} and $\tilde{P}_{xy0.5}$ – Case 4.

Statistics	Case 4: $v^s = v^t = 0.2$; $u^s = u^t = 0.05$					
	$x = 0.05$ $y = 0.05$	$x = 0.25$ $y = 0.25$	$x = 0.5$ $y = 0.5$	$x = 0.75$ $y = 0.75$	$x = 0.95$ $y = 0.05$	$x = 0.95$ $y = 0.95$
Mean (P_{xy}) (Std. Dev.P_{xy})	1.064 (0.128)	1.067 (0.132)	1.078 (0.126)	1.091 (0.128)	1.109 (0.128)	1.064 (0.127)
Mean ($\tilde{P}_{xy0.5}$) (Std. Dev.$\tilde{P}_{xy0.5}$)	1.065 (0.125)	1.062 (0.131)	1.078 (0.126)	1.094 (0.129)	1.109 (0.128)	1.065 (0.128)
$card\{k : \Delta_{1k} < 0\}$	21	38	26	26	30	18
$card\{k : \Delta_{2k} < 0\}$	28	30	39	39	27	29
$card\{k : \Delta_{3k} < 0\}$	2	24	26	22	25	4
$card\{k : \Delta_{4k} < 0\}$	7	25	39	27	16	6

Simulation 2

The presented simulation study is a continuation of the previous one but, it is done for 10 000 repetitions. For the given probability distributions of prices and quantities (see Simulation 1), we observe fluctuations of the following random variables: P_{La}, P_{Pa}, P_{F}, and P_{xx}, $\tilde{P}_{xx0.5}$, $\tilde{P}^{A}_{xx\frac{1}{2}}$ for different values of x. The results for Cases 1–4 are presented in Tables 8–11.

Table 8. Basic characteristics of the considered price indices for data from case 1.

Statistics:	Mean / (Standard deviation) / (Volatility coefficient) for Case 1					
Index	$x = 0.2$	$x = 0.3$	$x = 0.4$	$x = 0.6$	$x = 0.7$	$x = 0.8$
P_{La}	1.05422 / (0.03998) / (0.03792)					
P_{Pa}	1.05071 / (0.04375) / (0.04164)					
P_{F}	1.05246 / (0.04165) / (0.03958)					
P_{xx}	1.05345 (0.04068) (0.03862)	1.05308 (0.04105) (0.03898)	1.05272 (0.04141) (0.03934	1.05202 (0.04217) (0.04009)	1.05168 (0.04256) (0.04047)	1.05135 (0.04295) (0.04085)
$\tilde{P}_{xx0.5}$	1.05240 (0.04174) (0.03966)	1.05238 (0.04177) (0.03969)	1.05237 (0.04178) (0.03971)	1.05237 (0.04178) (0.03971)	1.05238 (0.04177) (0.03969)	1.05240 (0.04174) (0.03966)
$\tilde{P}^{A}_{xx0.5}$	1.05270 (0.04141) (0.03933)	1.05277 (0.04133) (0.03926)	1.05281 (0.04129) (0.03922)	1.05281 (0.04129) (0.03922)	1.05277 (0.04133) (0.03926)	1.05270 (0.04141) (0.03933)

Table 9. Basic characteristics of considered price indices for data from Case 2.

Statistics:	Mean / (Standard deviation) / (Volatility coefficient) for Case 2					
Index	$x = 0.2$	$x = 0.3$	$x = 0.4$	$x = 0.6$	$x = 0.7$	$x = 0.8$
P_{La}	1.09340 / (0.13236) / (0.12105)					
P_{Pa}	1.07245 / (0.12992) / (0.12114)					
P_{F}	1.08288 / (0.12911) / (0.11922)					
P_{xx}	1.08928 (0.13092) (0.12019)	1.08720 (0.13036) (0.11991)	1.08512 (0.12993) (0.11973)	1.08093 (0.12941) (0.11972)	1.07882 (0.12934) (0.11989)	1.07670 (0.12940) (0.12018)
$\tilde{P}_{xx0.5}$	1.08297 (0.12942) (0.11950)	1.08300 (0.12952) (0.11960)	1.08302 (0.12959) (0.11965)	1.08302 (0.12959) (0.11965)	1.08300 (0.12952) (0.11960)	1.08297 (0.12942) (0.11950)
$\tilde{P}^{A}_{xx0.5}$	1.08371 (0.12934) (0.11935)	1.08396 (0.12943) (0.11940)	1.08411 (0.12948) (0.11943)	1.08411 (0.12948) (0.11943)	1.08396 (0.12943) (0.11940)	1.08371 (0.12934) (0.11935)

Table 10. Basic characteristics of the considered price indices for data from case 3.

Statistics:	Mean / (Standard deviation) / (Volatility coefficient) for Case 3					
Index	$x = 0.2$	$x = 0.3$	$x = 0.4$	$x = 0.6$	$x = 0.7$	$x = 0.8$
P_{La}	1.04143 / (0.04959) / (0.04761)					
P_{Pa}	1.05054 / (0.04477) / (0.04261)					
P_{F}	1.04597 / (0.04688) / (0.04482)					
P_{xx}	1.04294 (0.04872) (0.04671)	1.04375 (0.04268) (0.04624)	1.04460 (0.04780) (0.04576)	1.04641 (0.04683) (0.04475)	1.04738 (0.04633) (0.04423)	1.04839 (0.04581) (0.04370)
$\tilde{P}_{xx0.5}$	1.04566 (0.04716) (0.04510)	1.04556 (0.04725) (0.04519)	1.04550 (0.04730) (0.04524)	1.04556 (0.04730) (0.04524)	1.04550 (0.04725) (0.04519)	1.04566 (0.04716) (0.04510)
$\tilde{P}^{A}_{xx0.5}$	1.04561 (0.04702) (0.04497)	1.04550 (0.04707) (0.04502)	1.04543 (0.04710) (0.04505)	1.04550 (0.04710) (0.04505)	1.04543 (0.04707) (0.04502)	1.04561 (0.04702) (0.04497)

Table 11. Basic characteristics of the considered price indices for data from case 4.

Statistics:	Mean / (Standard deviation) / (Volatility coefficient) for Case 4					
Index	$x = 0.2$	$x = 0.3$	$x = 0.4$	$x = 0.6$	$x = 0.7$	$x = 0.8$
P_{La}	1.07090 / (0.12855) / (0.12004)					
P_{Pa}	1.08411 / (0.12423) / (0.11459)					
P_{F}	1.07749 / (012567) / (0.11663)					
P_{xx}	1.07463 (0.12738) (0.11854)	1.07627 (0.12687) (0.11788)	1.07777 (0.12639) (0.11727)	1.08037 (0.12555) (0.11621)	1.08148 (0.12517) (0.11574)	1.08247 (0.12483) (0.11532)
$\tilde{P}_{xx0.5}$	1.07854 (0.12585) (0.11668)	1.07887 (0.12591) (0.11670)	1.07907 (0.12594) (0.11671)	1.07907 (0.12594) (0.11671)	1.07887 (0.12591) (0.11670)	1.07854 (0.12585) (0.11668)
$\tilde{P}^{A}_{xx0.5}$	1.07671 (0.12595) (0.11698)	1.07647 (0.12604) (0.11709)	1.07633 (0.12610) (0.11715)	1.07633 (0.12610) (0.11715)	1.07647 (0.12604) (0.11709)	1.07671 (0.12595) (0.11698)

6. Empirical Study

As it was mentioned earlier (see Subsection 3.3), the *Consumer Price Index* (CPI) is commonly used as a basic measure of inflation. The index approximates changes in the costs of household consumption assuming the constant utility (COLI, *Cost of Living Index*). Although in practice the Laspeyres price index is used to measure the CPI, many statisticians and economists consider the Fisher index to be the best approximation of COLI. Thus, in the following section we apply P_{xx}, $\tilde{P}_{xx0.5}$ and $\tilde{P}^{A}_{xx0.5}$ indices to verify their

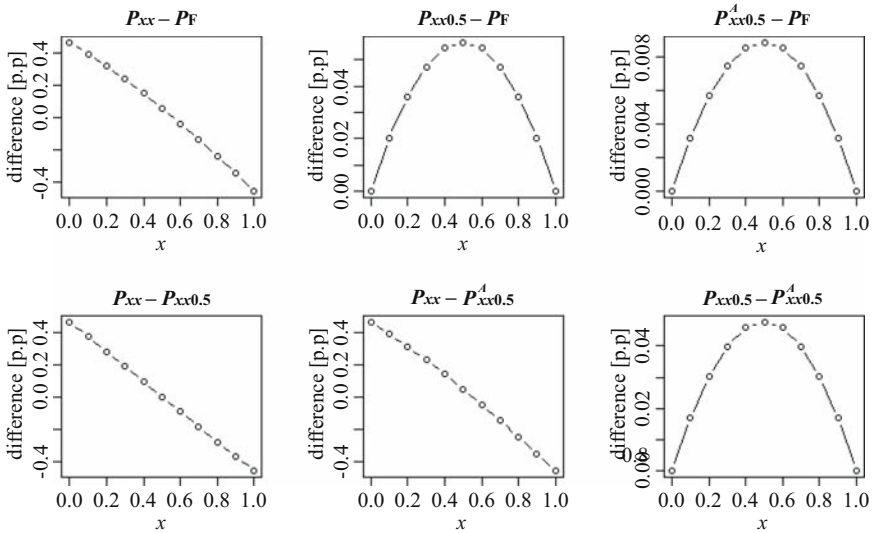

Fig. 16. Differences between indices from the considered subfamilies and the Fisher index∗ (Bulgaria, 2011)
(∗) $P_{La} = 1.0438$, $P_F = 1.0392$.

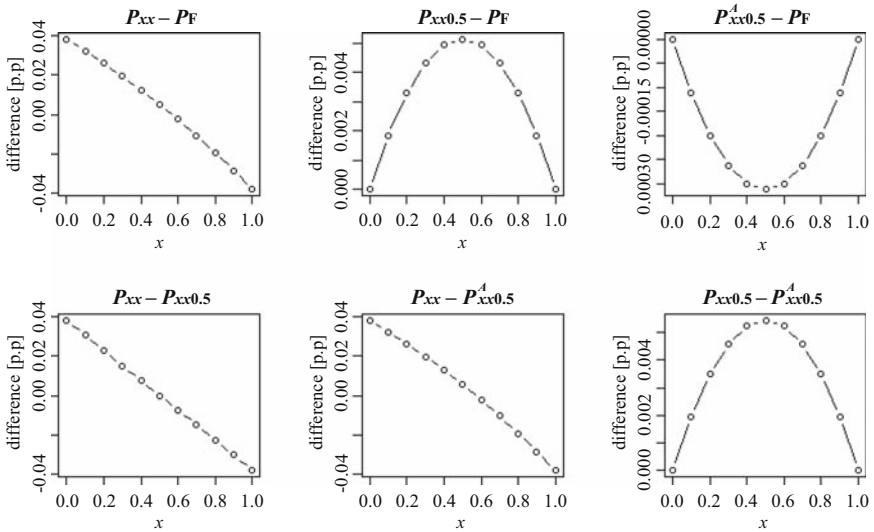

Fig. 17. Differences between indices from the considered subfamilies and the Fisher index∗ (Bulgaria, 2016)
(∗) $P_{La} = 0.9841$, $P_F = 0.9838$.

Fisher formula approximations and distances among them using CPI data from the United Kingdom and Bulgaria. Currently there are no differences between the CPI and the HICP (Harmonised Index of Consumer Prices) in the case of these countries. Thus, we use yearly data from Eurostat from the COICOP-4 digit level of aggregation and we calculate the above-mentioned price indices for different values of x and for years 2011 and 2016. The computed differences (in percentage points) between the proposed indices and the Fisher price index are presented in Figures 16–19.

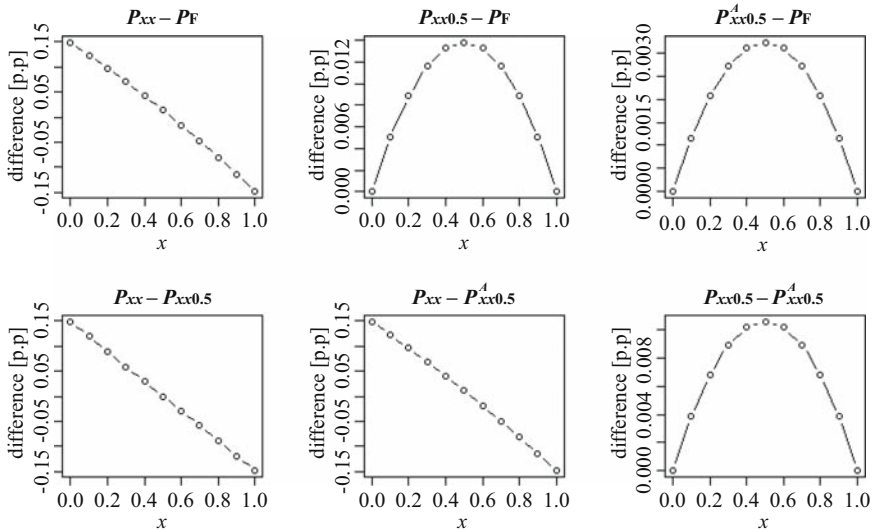

Fig. 18. Differences between indices from the considered subfamilies and the Fisher index (United Kingdom, 2011).* (*) $P_{La} = 1.0459$, $P_F = 1.0444$.

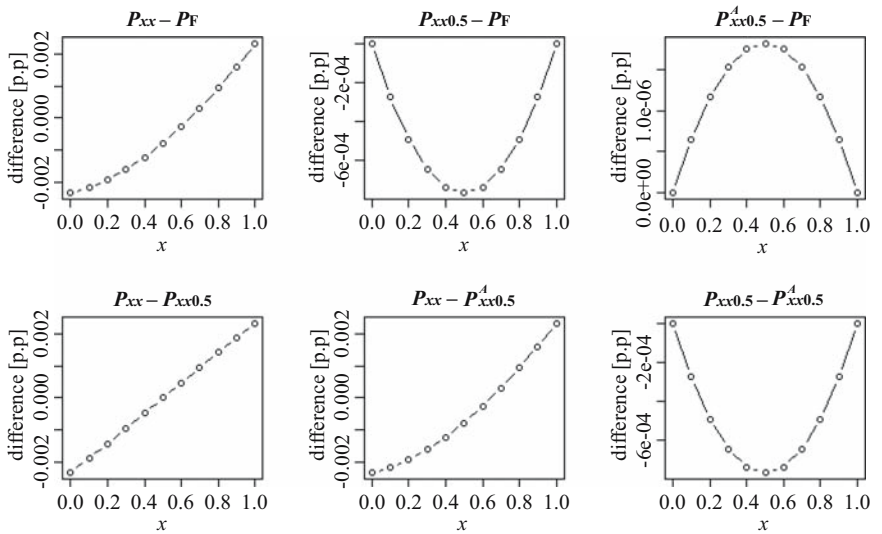

Fig. 19. Differences between indices from the considered subfamilies and the Fisher index (United Kingdom, 2016).* (*) $P_{La} = 0.99841$, $P_F = 0.99843$.

7. Conclusions

In **Simulation 1**, we observe that indices P_{xy} and $\tilde{P}_{xy0.5}$ provide identical results for $x = y = 0.5$ and quite similar results for other values of parameters x and y, that is, we observe small differences between expected index values (arithmetic means) calculated for their generated values. These expected values are nonmonotonic functions of x and y hence we cannot recommend such parameter values (x_0, y_0) that would lead to minimisation or maximisation of the considered general price indices P_{xy} and $\tilde{P}_{xy0.5}$. It is worth adding that values of these indices may strongly depend on parameters x and y, that is, indices belonging to this general class of price indices may differ substantially from each other. For instance, in Case 2 (see Table 5) means of generated P_{xy} values are as follows: 1.067 ($x = 0.95$, $y = 0.05$) or 1.143 ($x = 0.95$, $y = 0.95$) and analogical means of generated $\tilde{P}_{xy0.5}$ values equal: 1.067 ($x = 0.95$, $y = 0.05$) or 1.131 ($x = 0.95$, $y = 0.95$). The precision of estimation of P_{xy} and $\tilde{P}_{xy0.5}$ indices, that is, the standard deviations of their generated values, is comparable with respect to the size of the parameters and they do not seem to depend on x and y (see Tables 4–7). This is a practical conclusion: even if fluctuations of prices and quantities are large, we observe similar volatility among price indices from the same general class of indices. Nevertheless, comparing results from Tables 4, 6 and 7, we can conclude that rather price fluctuations than quantity fluctuations influence volatilities of P_{xy} and $\tilde{P}_{xy0.5}$ indices. Finally, the most crucial difference between the compared general class of indices is that the *probability* [1] of satisfying the Laspeyres-Paasche bounding test is bigger in the case of $\tilde{P}_{xy0.5}$ index (it is much bigger for small (near zero) and big (near one) values of x and y). The above-mention probability is estimated as a ratio of the number of generated cases when the considered price index fulfills the Laspeyres-Paasche bounding test and the total number of repetitions. In other words, we observe relatively fewer cases when the value of $\tilde{P}_{xy0.5}$ index is outside of the interval determined by the Laspeyres and Paasche price indices in comparison with analogical cases for the P_{xy} formula (see Tables 4–7 and also Figures 13–15).

In **Simulation 2**, we observe that the range of expected values of P_{xx} is relatively big (depending on x) in Cases 1 and 4, that is, when prices are strongly fluctuated (in Case 4 the maximum difference equals almost 0.8 p.p, see Table 11). In the same cases, expected (mean) values of generated indices from $\tilde{P}_{xx0.5}$ and $\tilde{P}^{A}_{xx0.5}$ classes remain stable and their changes are not bigger than 0.1 p.p (Table 8 and Table 11). Moreover, even if price fluctuations are really small (Case 3, see Table 10), generated values of P_{xx} indices may differ from each other by more than 0.5 p.p. The most important fact is that although volatilities of generated indices are comparable in each case (obviously volatility coefficients are higher in Cases 1 and 4 connected with high values of price dispersions), only values of $\tilde{P}_{xx0.5}$ and $\tilde{P}^{A}_{xx0.5}$ indices seem to approximate the mean of generated Fisher price indices effectively. Taking into consideration also (22), (23) and (24), it may seem likely that indices from the $\tilde{P}_{xx0.5}$ subclass are closest to superlative price indices.

The **Empirical study** confirms previously obtained results. Indices from the $\tilde{P}_{xx0.5}$ and $\tilde{P}^{A}_{xx0.5}$ subfamilies generate values that are very close to the superlative Fisher index and differences between them are very small. When the effect of substitution is observed, that is, when the difference between values of Laspeyres and Paasche indices rises, we can note

large differences between P_{xx} indices and the Fisher index, and between P_{xx} indices and indices from the $\tilde{P}_{xx0.5}$ and $\tilde{P}^{A}_{xx0.5}$ subfamilies (see Figures 16 and 18). When the CPI has no substitution bias ($P_{\mathrm{La}} \approx P_{\mathrm{F}}$), the values of indices from all the considered subfamilies approximate each other (see Figures 17 and 19). And one more remark – only the differences $P_{xx} - P_{\mathrm{F}}$, $P_{xx} - \tilde{P}_{xx0.5}$ and $P_{xx} - \tilde{P}^{A}_{xx0.5}$, as functions of $x \in [0, 1]$, seem to be monotonic and approximately linear.

7.1. Final Remarks

The proposed and wide class of price indices ($\tilde{P}_{xyz}$) has similar axiomatic properties as the geo-logarithmic price index family and, in particular, each index from this family satisfies the Martini's minimal requirements. It is worth adding that cofactors of $\tilde{P}_{xyz}$ indices satisfy the proportionality and homogeneity axioms (see Subsection 4.3). It is important from the perspective of the economic approach that there is a possibility of modification of the $\tilde{P}_{xxz}$ family to obtain such a general class of indices ($\tilde{P}^{A}_{xxz}$) that satisfies the Laspeyres-Paasche bounding test (Theorem 4). It should also be added that the particular element of the $\tilde{P}_{xyz}$ and $\tilde{P}^{A}_{xxz}$ families is the superlative Fisher price index, which is not an element of the geo-logarithmic price index class. Moreover, for any value of $x \in [0, 1]$ generated $\tilde{P}_{xx0.5}$ and $\tilde{P}^{A}_{xx0.5}$ indices seem to approximate the values of the Fisher price indices effectively. Thus, since for the superlative Walsh and Fisher price indices it holds that $\tilde{P}_{\frac{1}{2}\frac{1}{2}\frac{1}{2}} = P_{\mathrm{W}}$ and $\tilde{P}_{00\frac{1}{2}}$ $\tilde{P}_{11\frac{1}{2}} = P_{\mathrm{F}}$, the subfamily $\tilde{P}_{xx\frac{1}{2}}$ seems to be worth further studying. From the theoretical point of view, it would be interesting to consider an "average representative" of the above-mentioned subclass of indices, that is, the price index calculated for some x_0 which fulfils $\tilde{P}_{x_0\, x_0\, \frac{1}{2}} = \int_0^1 \tilde{P}_{xx\frac{1}{2}} dx$.

8. Appendix

8.1. Appendix A

Below we present formal definitions of major postulates (tests) arising from the axiomatic index theory and used in Theorem 1. Let us consider the price index formula $P(q^s, q^t, p^s, p^t)$. Let us also denote by λ any $N \times N$ diagonal matrix with positive elements $\lambda_1, \lambda_2, \ldots, \lambda_N$ and by k a positive, real number.

- *Identity* means that

$$P(q^s, q^t, p^s, p^s) = 1.$$

- *Proportionality* can be described by the following condition:

$$P(q^s, q^t, p^s, kp^s) = k.$$

- *Commensurability* can be expressed as follows:

$$P(\lambda^{-1}q^s, \lambda^{-1}q^t, \lambda p^s, \lambda p^t) = P(q^s, q^t, p^s, p^t).$$

- *Linear homogeneity* has the following form:

$$P(q^s, q^t, p^s, kp^t) = kP(q^s, q^t, p^s, p^t).$$

- *Price dimensionality* can be expressed as follows:

$$P(q^s, q^t, kp^s, kp^t) = P(q^s, q^t, p^s, p^t).$$

- *Strict monotonicity* is defined as follows:

$$P(q^s, q^t, p^s, \tilde{p}^t) > P(q^s, q^t, p^s, p^t), \quad \text{if} \quad \tilde{p}^t \geq p^t$$

and

$$P(q^s, q^t, \tilde{p}^s, p^t) < P(q^s, q^t, p^s, p^t), \quad \text{if} \quad \tilde{p}^s \geq p^s,$$

where $\tilde{p}^t \geq p^t$ means that at least one element of the nonnegative vector $\tilde{p}^t$ is greater than the corresponding element of the vector p^t (the relation $\tilde{p}^t \leq p^t$ is defined analogously).

- *Basis reversibility* means that

$$P(q^s, q^t, p^s, p^t)\ P(q^t, q^s, p^t, p^s) = 1.$$

- *Factor reversibility* can be expressed as

$$P(q^s, q^t, p^s, p^t)\ P(p^s, p^t, q^s, q^t) = \frac{\sum_{i=1}^{N} p_i^t q_i^t}{\sum_{i=1}^{N} p_i^s q_i^s}.$$

8.2. Appendix B

Observation. Let us assume that two price indices P_1 and P_2 satisfy the axioms from the Martini's minimal requirements. Then, the price index being the weighted geometric mean of P_1 and P_2 indices, i.e., $P = P_1^z P_2^{1-z}$ for $z \in [0, 1]$, also satisfies the Martini's minimal requirements.

Proof Let us assume that $P_1(q^s, q^t, p^s, p^t)$ and $P_2(q^s, q^t, p^s, p^t)$ satisfy *identity, commensurability* and *linear homogeneity* (see Appendix A, Subsection 8.1). Let us consider the price index $P(q^s, q^t, p^s, p^t) = P_1^z(q^s, q^t, p^s, p^t)\ P_2^{1-z}(q^s, q^t, p^s, p^t)$ for a real number $z \in [0, 1]$.

The price index $P(q^s, q^t, p^s, p^t)$ also satisfies:

- Identity
  since $P(q^s, q^t, p^s, p^s) = P_1^z(q^s, q^t, p^s, p^s)\ P_2^{1-z}(q^s, q^t, p^s, p^s) = 1^z\ 1^{1-z} = 1$;
- Commensurability
  since for any $N \times N$ diagonal matrix λ with positive elements $\lambda_1, \lambda_2, \ldots, \lambda_N$ we have

$$\begin{aligned} P(\lambda^{-1}q^s, \lambda^{-1}q^t, \lambda p^s, \lambda p^t) &= P_1^z(\lambda^{-1}q^s, \lambda^{-1}q^t, \lambda p^s, \lambda p^t)\ P_2^{1-z}(\lambda^{-1}q^s, \lambda^{-1}q^t, \lambda p^s, \lambda p^t) \\ &= P_1^z(q^s, q^t, p^s, p^t)\ P_2^{1-z}(q^s, q^t, p^s, p^t) = P(q^s, q^t, p^s, p^t); \end{aligned}$$

- *Linear homogeneity*

$$\begin{aligned} P(q^s, q^t, p^s, kp^t) &= P_1^z(q^s, q^t, p^s, kp^t)\ P_2^{1-z}(q^s, q^t, p^s, kp^t) \\ &= k^z\ P_1^z(q^s, q^t, p^s, p^t)\ k^{1-z}\ P_2^{1-z}(q^s, q^t, p^s, p^t) = k\ P(q^s, q^t, p^s, p^t). \end{aligned}$$

Thus, the price index $P(q^s, q^t, p^s, p^t)$ satisfies the axioms from the system of minimal requirements proposed by Marco Martini (1992b).

8.3. Appendix C. Heuristic Proof of Approximation (24)

Let us note that assuming $\forall i \in \{1, 2, ..., N\}$ $q_{si} \approx q_{ti}$ we obtain as a consequence

$$q_{si}^x \approx q_{ti}^x, \quad \text{for} \quad x \in [0, 1] \tag{C1}$$

and thus

$$q_i^{Ax} \approx q_{si}, \quad q_i^{Bx} \approx q_{si}, \quad \text{for} \quad i \in \{1, 2, \ldots, N\} \tag{C2}$$

From (C2) we obtain the following approximations

$$w_{si}^{Ax} \approx w_{si}^0 \quad \text{and} \quad w_{si}^{Bx} \approx w_{si}^0. \tag{C3}$$

Analogically, we can write that

$$w_{ti}^{Ax} \approx w_{ti}^1 \quad \text{and} \quad w_{ti}^{Bx} \approx w_{ti}^1. \tag{C4}$$

Repeating steps (C2)–(C4) for the approximation $q_{si}^y \approx q_{ti}^y$ we obtain

$$w_{si}^{Ay} \approx w_{si}^0, \quad w_{si}^{By} \approx w_{si}^0, \quad w_{ti}^{Ay} \approx w_{ti}^1, \quad w_{ti}^{By} \approx w_{ti}^1. \tag{C5}$$

It is proved by Fattore (2010) that within the limit $w_{si}^x \longrightarrow w_{ti}^y$ it holds that

$$P_{xy} = \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{v_i^{xy}} \approx \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{\frac{w_{si}^x + w_{ti}^y}{2}}. \tag{C6}$$

Since $w_{si}^x = w_{si}^{Ax}$, $w_{ti}^y = w_{ti}^{Ay}$ and consequently $v_i^{xy} = v_{Ai}^{xy}$ from (C6) and the assumption that $w_{si}^x \approx w_{ti}^y$ we have

$$\prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{v_{Ai}^{xy}} \approx \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{\frac{w_{si}^{Ax} + w_{ti}^{Ay}}{2}}. \tag{C7}$$

Analogically to the Fattore's way, it can be proved that

$$\prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{v_{Bi}^{xy}} \approx \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{\frac{w_{si}^{Bx} + w_{ti}^{By}}{2}}. \tag{C8}$$

Thus, from (18), (C7) and (C8) we obtain

$$\begin{aligned} \tilde{P}_{xyz} &= \left\{\prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{v_{Ai}^{xy}}\right\}^{z} \left\{\prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{v_{Bi}^{xy}}\right\}^{1-z} \\ &\approx \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{\frac{1}{2}\left\{z\left(w_{si}^{Ax} + w_{ti}^{Ay}\right) + (1-z)\left(w_{si}^{Bx} + w_{ti}^{By}\right)\right\}}. \end{aligned} \tag{C9}$$

From (C3), (C4), (C5) and (C9) we obtain the final conclusion that

$$\tilde{P}_{xyz} \approx \prod_{i=1}^{N} \left(\frac{p_{ti}}{p_{si}}\right)^{\frac{w_{si}^0 + w_{ti}^1}{2}} = P_{\mathrm{T}}, \tag{C10}$$

8.4. Appendix D (Proof of Theorem 4)

Lemma For any positive real values a, b, c, d and $x \in [0, 1]$ the following relation is true

$$\min\left\{\frac{a}{c}, \frac{b}{d}\right\} \leq \frac{ax + b(1-x)}{cx + d(1-x)} \leq \max\left\{\frac{a}{c}, \frac{b}{d}\right\}. \tag{D1}$$

Proof of the Lemma
Let us note that in the case of $x = 0$ or $x = 1$ the relation (D1) is obvious. Let us consider $x \in (0, 1)$ and, for instance, let us assume that

$$\frac{a}{c} \leq \frac{b}{d}. \tag{D2}$$

Suppose by contraposition that (D1) does not hold, that is, there exists some $x_0 \in (0, 1)$ that

$$\frac{ax_0 + b(1-x_0)}{cx_0 + d(1-x_0)} < \frac{a}{c}. \tag{D3}$$

The inequality (D3) can be written equivalently as

$$acx_0 + bc(1-x_0) < acx_0 + ad(1-x_0), \tag{D4}$$

and that leads to the false (with respect to the assumption (D2)) conclusion that

$$\frac{b}{d} < \frac{a}{c}. \tag{D5}$$

In an analogous way, we can prove that under the assumption (D2) it is impossible to obtain

$$\frac{ax_0 + b(1-x_0)}{cx_0 + d(1-x_0)} > \frac{b}{d}. \tag{D6}$$

Proof of Theorem 4
Firstly, let us consider any $(x, z) \in (0, 1) \times (0, 1)$. Let us signify by

$$\theta_1(x) = \sum_{i=1}^{N} p_{ti}(xq_{ti} + (1-x)q_{si}), \tag{D7}$$

$$\theta_2(x) = \sum_{i=1}^{N} p_{si}(xq_{ti} + (1-x)q_{si}), \tag{D8}$$

$$\theta_3(x) = \sum_{i=1}^{N} p_{ti}((1-x)q_{ti} + xq_{si}), \tag{D9}$$

$$\theta_4(x) = \sum_{i=1}^{N} p_{si}((1-x)q_{ti} + xq_{si}), \tag{D10}$$

From (28) and (D7)–(D10) we have that

$$\ln\left(\tilde{P}^A_{xxz}\right) = z(\ln\theta_1(x) - \ln(\theta_2(x)) + (1-z)(\ln\theta_3(x) - \ln(\theta_4(x)), \tag{D11}$$

and thus, according to the necessary condition for the existence of the local extreme, it must hold

$$\frac{\partial \ln\left(\tilde{P}^A_{xxz}\right)}{\partial z} = \ln\frac{\theta_1(x)\theta_4(x)}{\theta_2(x)\theta_3(x)} = 0, \tag{D12}$$

From (D6) we obtain immediately that

$$\theta_1(x)\theta_4(x) = \theta_2(x)\theta_3(x), \tag{D13}$$

and it leads to the following condition

$$\sum_{i=1}^{N} p_{ti}\, q_{ti} \sum_{i=1}^{N} p_{si}\, q_{si}[x^2 - (1-x)^2] = \sum_{i=1}^{N} p_{ti}\, q_{si} \sum_{i=1}^{N} p_{si}\, q_{ti}[x^2 - (1-x)^2]. \tag{D14}$$

Since in (D14) we take into consideration any prices and quantities, we conclude that it must hold that $x^2 - (1-x)^2 = 0$ or equivalently $x = 0.5$. Let us note that taking $x = \frac{1}{2}$ we obtain $\tilde{P}^A_{\frac{1}{2}\frac{1}{2}z} = P_{\mathrm{ME}}$ and this formula does not depend on the parameter z. In other words, since it holds that

$$\frac{\partial \ln\left(\tilde{P}^A_{0.5\ 0.5\ z}\right)}{\partial z} = 0. \tag{D15}$$

that is, each point on the plane $(0.5, z)$ is a stationary point for the function $\ln\left(\tilde{P}^A_{xxz}\right)$ (and thus, also for $\tilde{P}^A_{xxz}$) depending on (x, z). Thus, the potential local extreme of the function $\tilde{P}^A_{xxz}$ is obtained in such points and it equals P_{ME}.

Now, let us verify the behaviour of the function $\tilde{P}^A_{xxz}$ in the frontier of the set $[0,1]\times[0,1]$, here denoted by $D = Fr([0,1]\times[0,1])$. To reach this purpose, let us consider the following sets: $D_1 = \{(x,z) : x \in \{0,1\} \wedge z \in (0,1)\}$, $D_2 = \{(x,z) : x \in (0,1) \wedge z \in \{0,1\}\}$, $D_3 = \{(0,0),(0,1),(1,0),(1,1)\}$, where obviously $D = D_1 \cup D_2 \cup D_3$. Let us note that limiting the domain of the function $\tilde{P}^A_{xxz}$ to D_1 we obtain for $z \in (0,1)$

$$\tilde{P}^A_{00z} = P^z_{\mathrm{La}} P^{1-z}_{\mathrm{Pa}} \quad \text{or} \quad \tilde{P}^A_{11z} = P^{1-z}_{\mathrm{La}} P^z_{\mathrm{Pa}}, \tag{D16}$$

where obviously the price index being the geometric mean of the Laspeyres and Paasche price indices fulfils the Laspeyres-Paasche bounding test. Limiting the domain of the

function $\tilde{P}^A_{xxz}$ to D_2 we obtain for $x \in (0,1)$

$$\tilde{P}^A_{xx0} = \frac{\sum_{i=1}^N p_{ti}((1-x)q_{ti} + xq_{si})}{\sum_{i=1}^N p_{si}((1-x)q_{ti} + xq_{si})}, \tag{D17}$$

or

$$\tilde{P}^A_{xx1} = \frac{\sum_{i=1}^N p_{ti}(xq_{ti} + (1-x)q_{si})}{\sum_{i=1}^N p_{si}(xq_{ti} + (1-x)q_{si})}. \tag{D18}$$

For instance, taking $a = \sum_{i=1}^N p_{ti}q_{ti}$, $b = \sum_{i=1}^N p_{ti}q_{si}$, $c = \sum_{i=1}^N p_{si}q_{ti}$ and $d = \sum_{i=1}^N p_{si}q_{si}$ from the Lemma, we have the immediate conclusion that for any $x \in (0,1)$ it holds that

$$\min(P_{\mathrm{La}}, P_{\mathrm{Pa}}) \leq \tilde{P}^A_{xx1} \leq \max(P_{\mathrm{La}}, P_{\mathrm{Pa}}), \tag{D19}$$

since $P_{\mathrm{La}} = \frac{b}{d}$ and $P_{\mathrm{Pa}} = \frac{a}{c}$. The analogous conclusion from the Lemma is that

$$\min(P_{\mathrm{La}}, P_{\mathrm{Pa}}) \leq \tilde{P}^A_{xx0} \leq \max(P_{\mathrm{La}}, P_{\mathrm{Pa}}), \tag{D20}$$

and thus, similarly to (D16), limiting the domain of the function $\tilde{P}^A_{xxz}$ to D_2, we can write that

$$\tilde{P}^A_{xxz} = P_{\mathrm{La}}^{1-z_0} P_{\mathrm{Pa}}^{z_0}, \tag{D21}$$

for $x \in (0,1)$, $z \in \{0,1\}$, and some $z_0 \in [0,1]$.

Limiting the domain of the function $\tilde{P}^A_{xxz}$ to D_3, from (32) and (33) we can reduce the set of the function values to $\{P_{\mathrm{La}}, P_{\mathrm{Pa}}\}$, i.e.,

$$\left\{\tilde{P}^A_{xxz} : (x,z) \in D_3\right\} = \{P_{\mathrm{La}}, P_{\mathrm{Pa}}\}. \tag{D22}$$

Let us note that the function $\tilde{P}^A_{xxz}$ is continuous in the closed and bounded set $[0,1] \times [0,1]$ being a convex quadrangle. From (D15), (D16), (D21), (D22) and the Weierstrass extreme value theorem, we know that the minimum and maximum value of the function $\tilde{P}^A_{xxz}$ belongs to the following set: $\left\{P_{\mathrm{La}}, P_{\mathrm{Pa}}, P_{\mathrm{La}}^z P_{\mathrm{Pa}}^{1-z}, P_{\mathrm{La}}^{1-z} P_{\mathrm{Pa}}^z, P_{\mathrm{ME}}\right\}$ for a $z \in [0,1]$. Since the price index P_{ME} satisfies the Laspeyres-Paasche bounding test (it is an immediate consequence of the Lemma used for $x = 0.5$ and $a = \sum_{i=1}^N p_{ti}q_{ti}$, $b = \sum_{i=1}^N p_{ti}q_{si}$, $c = \sum_{i=1}^N p_{si}q_{ti}$ and $d = \sum_{i=1}^N p_{si}q_{si}$.), we have the final conclusion that the above-mentioned test is also satisfied in the case of any price index from the $\tilde{P}^A_{xxz}$ subfamily.

9. References

Allen, R.G.D. 1975. *Index Numbers in Theory and Practice*. London: Macmillan Press.

Balk, B.M. 1995. "Axiomatic Price Index Theory: A Survey." *International Statistical Review* 63: 69–93. Doi: https://doi.org/10.2307/1403778.

Białek, J. 2012. "Proposition of the general formula for price indices." *Communications in Statistics: Theory and Methods* 41(5): 943–952. Doi: https://doi.org/10.1080/03610926.2010.533238.

Boskin, M.J., E.R. Dulberger, R.J. Gordon, Z. Griliches, and D. Jorgenson. 1996. *Toward a More Accurate Measure of the Cost of Living*. Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index.

Carlson, B.C. 1972. The logarithmic mean, Amer. Math. Monthly. 79: 615–618. Doi: https://doi.org/10.1080/00029890.1972.11993095.

Clements, K.W. and H.Y. Izan. 1987. "The Measurement of Inflation: A Stochastic Approach." *Journal of Business and Economic Statistics* 5: 339–350. Doi: https://doi.org/10.1080/07350015.1987.10509598.

Diewert, W.E. 1976. "Exact and Superlative Index Numbers." *Journal of Econometrics* 4: 114–145. Doi: https://doi.org/10.1016/0304-4076(76)90009-9.

Diewert, W.E. 1993. *The economic theory of index numbers: a survey*, Essays in index number theory, vol. 1, Eds. W.E. Diewert and A.O. Nakamura: 177–221, Amsterdam.

Dumagan, J. 2002. "Comparing the Superlative Törnqvist and Fisher ideal indices." *Economic Letters* 76: 251–258. Doi: https://doi.org/10.1016/s0165-1765(02)00049-6.

Eichhorn, W. and J. Voeller. 1976. *Theory of the Price Index. Fisher's Test Approach and Generalizations*. New York: Springer-Verlag.

Fattore, M. 2006. On the monotonicity of the geo-logarithmic price indexes. In: Proceedings of the XLIII Scientific Meeting, Società Italiana di Statistica, Cleup, Padova.

Fattore, M. 2010. "Axiomatic Properties of Geo-Logarithmic Price Indices." *Journal of Econometrics* 156(2): 344–353. Doi: https://doi.org/10.1016/j.jeconom.2009.11.004.

Hill, R.J. 2006. "Superlative Index Numbers: Not All of Them Are Super." *Journal of Econometrics* 130: 25–43. Doi: https://doi.org/10.1016/j.jeconom.2004.08.018.

IMF. 2004. Producer Price Index Manual, International Monetary Fund.

Jorgenson, D.W. and D.T. Slesnick. 1983. *Individual and social cost of living indexes*, Price level measurement: proceedings of a conference sponsored by Statistics Canada, W.E. Diewert and C. Montmarquette: 241–336, Ottawa: Statistics Canada.

Krstcha, M. 1988. *Axiomatic characterization of statistical price indices*. Heidelberg: Physica-Verlag.

Martini, M. 1992a. I numeri indice in un approccio assiomatico. Giuffré, Milan.

Martini, M. 1992b. "General Function of Axiomatic Index Numbers." *J. Ital. Statist. Soc* 3: 359–376. Doi: https://doi.org/10.1007/bf02589086.

Olt, B. 1996. *Axiom und Struktur in der statistischen Preisindextheorie*. Germany: Peter Lang.

Pollak, R.A. 1989. *The theory of the cost-of-living index*. Oxford: Oxford University Press.

White, A.G. 1999. "Measurement Biases in Consumer Price Indexes." *International Statistical Review* 3: 301–325. Doi: https://doi.org/10.1111/j.1751-5823.1999.tb00451.x.

Von der Lippe, P. 2007. *Index Theory and Price Statistics*. Germany: Peter Lang.

Received February 2017
Revised August 2018
Accepted September 2018

sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 319–336, http://dx.doi.org/10.2478/JOS-2019-0015

Prospects for Protecting Business Microdata when Releasing Population Totals via a Remote Server

James Chipperfield[1], *John Newman*[1], *Gwenda Thompson*[1], *Yue Ma*[2], *and Yan-Xia Lin*[2]

Many statistical agencies face the challenge of maintaining the confidentiality of respondents while providing as much analytical value as possible from their data. Datasets relating to businesses present particular difficulties because they are likely to contain information about large enterprises that dominate industries and may be more easily identified. Agencies therefore tend to take a cautious approach to releasing business data (e.g., trusted access, remote access and synthetic data). The Australian Bureau of Statistics has developed a remote server, called TableBuilder, which has the capability to allow users to specify and request tables created from business microdata. The tables are confidentialised automatically by perturbing cell values, and the results are returned quickly to the users. The perturbation method is designed to protect against attacks, which are attempts to undo the confidentialisation, such as the well-known differencing attack. This paper considers the risk and utility trade-off when releasing three Australian Bureau of Statistics business collections via its TableBuilder product.

Key words: Business data; online access; perturbation; remote server; statistical disclosure control.

1. Introduction

While carrying out their role of collecting and disseminating data, statistical agencies generally need to determine effective ways of meeting two key objectives: to maintain the confidentiality of respondents and to provide its society with as much analytical value from the data as possible. The two most common types of data that require confidentialisation are data about people and data about businesses. Person-level data and business-level data have many aspects in common. However, there are some characteristics commonly associated with business data that may make confidentialising a more challenging problem.

Typically, some industries will be dominated by large businesses whose information is difficult to conceal without suppressing or altering the data to a large extent. For many business collections, continuous data items, such as turnover or profit, are of key interest to users. Some of these continuous data items, such as capital expenditure, may have many zero values and a few large values. Certain aspects of a business's operations can become

[1] Australian Bureau of Statistics, P.O. Box 10, Belconnen, Australian Capital Territory 2616, Australia. Emails: james.chipperfield@abs.gov.au, john.newman@abs.gov.au, and gwenda.thompson@abs.gov.au
[2] University of Wollongong, Wollongong, New South Wales 2522, Australia. Emails: mayue3588@gmail.com and yanxia@uow.edu.au
Acknowledgments: The views in this article do not necessarily represent those of the Australian Bureau of Statistics.

public knowledge, for example through the release of annual reports. Some users may also have access to administrative data related to the businesses. There are potentially high incentives for attackers to try to discover confidential information about businesses because this may lead to a competitive advantage. These issues can become even more problematic in countries with smaller economies, because of the limited number of businesses that operate in those countries.

For some statistical agencies, there are legislative differences between the treatment of person-level data and business-level data. For example, there can be opportunities to gain consent to publish business data. This may allow the release of more data, but can also make the process of applying confidentiality protection more complex. This is because there is a need to monitor which businesses provide consent and because confidentialisation is complicated in cases when consenting and nonconsenting businesses appear in the same cell of a table. Another example is where confidentialisation is required at multiple levels of business structure.

For the reasons listed above, release of detailed business micro-data by statistical agencies may allow attackers to discover confidential information about a business. This is why, at least in the case of the Australian Bureau of Statistics (ABS), the vast majority of its business data is still released in the form of broad-level tables.

A common approach for confidentialising tabular business data is suppression (González 2005; Tambay and Fillion 2013). This approach is easiest to apply when the full set of tables to be published is known in advance. If additional tables are requested, then any further suppression will need to take into account which cells were previously suppressed. In practice the full set of tables is rarely known in advance. This means that suppression is unlikely to be optimal and that the amount of information released with each set of additional tables will be increasingly suppressed. Some statistical agencies consider alternative approaches including accredited or "trusted" access (Abrahams and Mahony 2008), replacing sensitive data with synthetic data (Miranda and Vilhuber 2013) and various ways of perturbing micro-data.

The ABS has developed an approach that could be used to release confidentialised totals from business data through its remote server, called TableBuilder. A simple model for a remote server (Chipperfield and O'Keefe 2014; Chipperfield 2014; O'Keefe and Chipperfield 2013; Thompson et al. 2013) is: (1) an analyst submits a query (i.e., request for a table) to the remote server; (2) the remote server automatically modifies or restricts the query's output; (3) the server sends the modified output to the analyst. Tambay (2017) use the ideas of a remote server (specifically TableBuilder) while also perturbing the underlying micro-data. For reviews of remote servers in use or in development in national statistical agencies, see Lucero et al. (2011), Reuter and Museux (2010).

There are some key advantages of a remote server. First, the degree to which an estimate is modified depends upon the output itself. For example, modification of an estimate may be relatively high if a cell is dominated by a single business and relatively low if a table cell has many small businesses of roughly equal size. Second, because an analyst is restricted from viewing the micro-data, less modification is needed than would otherwise be the case. Third, it allows users to gain rapid access to estimates they request. Fourth, the modification algorithm assures a specified level of protection is guaranteed.

This article evaluates the prospect of allowing access to business survey data via TableBuilder. For a full description, see Part 1 of Thompson et al. (2013). Section 3 defines disclosure and utility and discusses how TableBuilder's perturbation settings could be chosen to optimise the trade-off between disclosure and utility. Sections 4, 5 and 6 evaluate utility of TableBuilder outputs, conditional on a certain level of disclosure risk, for two surveys and one administrative collection of the ABS. Section 7 makes some concluding remarks, including a discussion of the prospects of releasing business data in TableBuilder.

2. TableBuilder

2.1. Totals

Here we describe the essential perturbation algorithm, but for a more complete description see Part 1 of Thompson et al. (2013). Consider any particular cell in a table and let there be n sample units contributing to the cell, where the units are indexed by $i = 1, 2, \ldots, n$. Define a continuous valued characteristic (e.g., income or turnover) for the ith unit (e.g., business) by y_i so that $|y_1| \geq |y_2| \geq |y_3|. \ldots \geq |y_n|$. The absolute values are taken because it is the magnitude of y, not whether it is positive or negative, that has bearing on considerations of risk and utility. (Changing all y values from positive to negative in a cell would not affect the perturbation distribution P^*- this is as it should be because a large negative y value is just as sensitive as a large positive y value.) If we define the estimation weight for the ith unit in the cell by w_i the survey estimate of the total is $\hat{Y} = \sum_i w_i y_i$. Instead of releasing $\hat{Y}$, TableBuilder releases $\hat{Y}^* = \hat{Y} + P^*$, where $P^* = \sum_{i=1}^{K} \left(m_i d_i^* h_i^*\right) y_i w_i$ is the perturbation amount and:

- m_i is a positive constant parameter. This parameter moderates the magnitude of the perturbation relative to the value y_i. In particular, the parameter $m_1 (i = 1)$ is the most important of the parameters as it plays an important role in protecting the largest contributor's value, y_1. The optimal value of m_i depends upon the distribution of y within the cell, the risk measure and the utility measures. Given the complexity of these dependencies, the optimal values are calculated in simulation (see Subsection 3.4).
- d_i^* is a random variable taking the value -1 and 1 with equal probability and so determines the direction of the perturbation.
- K is the number of top contributors in the cell that are used in calculation of P^*. We found that there was little value in allowing $K > 4$ since the main aim here is to protect the largest contributor's value (see Subsection 3.1).
- h_i^*, for purposes of this evaluation, was a random value drawn from a symmetric triangular distribution with lower limit $1 - b = 0.7$ to $1 + b = 1.3$ and the mode of 1. h_i^* has mean $E\left(h_i^*\right) = 1$ and variance $Var\left(h_i^*\right) = b^2/12$ which, with $b = 0.3$, is equal to 0.075. The bimodal distribution generated by $d_i^* h_i^*$ is symmetric round zero, $\sigma_*^2 = Var\left(d_i^* h_i^*\right) = 1 + b^2/12$ and has little mass around 0. This avoids unacceptable small values while also ensuring that the perturbation has mean zero. Exploring other distributions would likely be a fruitful line of research (Krsinich and Piesse 2002; Evans et al. 1998; Tambay 2017).

The form of P^* is intuitive in the sense that its magnitude is in proportion to the size of the K largest, and most at risk, contributors. Allowing $K > 1$ allows more degrees of freedom to specify the perturbation distribution, P^*, and so will allow it to better approximate the optimum distribution.

There is no constraint in this procedure to ensure consistency between the perturbed estimates. This means a perturbed total for Australia will not exactly equal the perturbed totals for each state summed over all states.

The current functionality of TableBuilder is such that K, the m_is and the distributions of d_i^* and h_i^* are essentially fixed for a given business collection. This means, for instance, that it is not possible to allow P^* to depend upon whether or not a cell is known to be sensitive and that it is not possible to allow the value of K to vary across cells. In Section 7 we discuss the benefits of relaxing this constraint.

We can see that $E^*(P^*) = 0$, where '$*$' represents the perturbation process. We did consider generating P^* from the Laplace distribution so as to achieve ε-differential privacy (Dwork et al. 2006), but the utility loss was far too great.

Table 1 gives an example of the perturbation of a cell total. We set $K = 4$, $\mathbf{m} = (m_1, m_2, m_3, m_4) = (0.6, 0.4, 0.3, 0.2)$ and there are $n = 8$ businesses in this cell. The estimator of total $\hat{Y} =$ USD 263,719 is perturbed by $P^* = -$ USD 18,278 so that the released estimate is $\hat{Y}^* =$ USD 245,441. The particular choice of values for K and $\mathbf{m}$ in Table 1 are for illustration only.

A Unit Key is a positive integer less than 2^{32} that is permanently and randomly assigned to each unit. The Unit Key is the random seed used to generate the value of d_i^* for $i = 1, \ldots, n$. This means, once generated, all d_i^*s are fixed in all calculations of P^*. It also means that a unit's contribution to P^*, when applicable, is either always positive or always negative – this was to reduce the perturbation variance of differences between cell totals, where the cells had some units in common (for more discussion on this see Subsection 3.4 and in Section 7).

A Cell Key is calculated by summing the Unit Keys for all the units contributing to the cell and then dividing by a large prime number. This essentially means that the Cell Key depends upon the exact set of n records that belong to the cell. The random seed for h_i^*

Table 1. Example of perturbation ($K = 4$).

ID	Turnover (USD) y_i	Estimation weight w_i	Magnitude m_i	Direction d_i^*	Noise h_i^*	Weighted turnover $y_i w_i$	Perturbation amount $P_i^* = m_i d_i^* h_i^* y_i w_i$
1	72.1	458.2	0.6	1	0.95	33,036	18,831
2	65.3	185.7	0.4	−1	1.02	12,126	−4,947
3	65.3	752.7	0.3	−1	1.54	49,151	−22,708
4	50.1	612.6	0.2	−1	1.54	30,691	−9,453
5	49.2	977.5				48,093	
6	45.4	458.7				20,825	
7	36.9	896.3				33,073	
8	36.9	995.2				36,723	
Sum						$\hat{Y} = 263{,}719$	$P^* = -18{,}278$

depends upon the Unit Key for unit *i and* its associated Cell Key – this means adding a unit to a cell will generate a new and independent value of h_i^* for each unit in the cell.

It follows that P^* and so $\hat{Y}^*$ will take the same value for any cell containing the same set of n units,- that is, TableBuilder will release the same estimate for logically equivalent cells because they will have the same set of contributors. This means that it is not possible to average over the effect of perturbation by requesting the same logically defined cell count in different tables.

Cells with a small number, say fewer than H, of contributing businesses are typically suppressed in the publications of many statistical agencies. TableBuilder effectively does the same thing. If $n \leq H$ then $\hat{Y}^* = 0$. If $H = 2$ then this provides protection against attacks on cells with sample counts of '1' or '2'.

The estimate of the count of units in the population belonging to a cell is $\hat{N} = \sum_{i=1}^{n} w_i$. Instead of releasing the ratio, $\hat{T} = \hat{Y}/\hat{N}$, TableBuilder releases $\hat{T}^* = \hat{Y}^*/\hat{N}$. (We do not discuss the perturbation of $\hat{N}$ here. For details see Chipperfield et al. 2016.)

2.2. *Confidence Intervals*

It is straightforward to derive 95% confidence intervals around each cell estimate, $\hat{Y}^*$. It would be slightly more difficult to derive confidence intervals for a linear combination (e.g., the difference) of perturbed cell estimates. TableBuilder does not release confidence intervals. Instead TableBuilder releases the variance of $\hat{Y}^*$, given by

$$\sigma^2 = Var_{s^*}(\hat{Y}^*) = Var_s[E_*(\hat{Y}^* \mid s)] + E_s[Var_*(\hat{Y}^* \mid s)],$$

where the first term represents the variation due to the sampling process, denoted by s, and the second term is the variation due to perturbation process, denoted by '$*$'. TableBuilder estimates the first term using the standard Jackknife. TableBuilder calculates the second term by $\sum_{i=1}^{K} m_i^2 \sigma_*^2 w_i^2 y_i^2$, where σ_*^2 is defined earlier. As mentioned, the variance cannot be used to construct 95% confidence intervals confidence intervals using $(\pm 1.96\sigma^2)$ because the perturbations are not approximately normally distributed. As the ratio of the perturbation variance to the sampling variance increases, the more inaccurate the coverage rates based on the normality assumption would become. One option would be to suppress a cell estimate if this ratio is above some threshold value. This is a topic for further research.

3. Measuring Disclosure Risk and Utility

3.1. *Attack Scenarios*

We measure the disclosure risk with respect to three 'attack scenarios'. In each scenario, the target is the largest contributor to the cell ($i = 1$), the target value is therefore y_1, and the attacker knows that the weight of the largest and second largest contributors is equal to one ($w_1 = w_2 = 1$). The largest contributor is chosen to be the target because it, of all the units in the cell, has the highest associated risk of disclosure. In Scenario 1 and 2, the attacker does not know the value of y for any of the contributors to the cell. However, in Scenario 3, the attacker is the second largest contributor to the cell ($i = 2$) and so is able to

use its known contribution, y_2, to improve upon the accuracy of Attack 1. This means that the disclosure risk of Scenario 3 will always be at least as high as Scenario 1.

Attack Scenario 1: The value of $\hat{Y}^*$ is used as an estimate of y_1. The attacker's estimate of y_1 under this scenario is $\hat{y}_1^{(1)} = \hat{Y}^*$

Attack Scenario 2: The attacker uses the difference between two cell estimates, $\hat{y}_1^{(2)} = \hat{Y}^* - \hat{Y}^*(i = 1)$ as an estimate of y_1, where $\hat{Y}^*(i = 1)$ is the same as $\hat{Y}^*$ except that the largest contributor, $i = 1$, is dropped from the cell.

Attack Scenario 3: This is the same as Attack Scenario 1 except that the attacker is also the second highest contributor to the cell ($i = 2$). The attacker can use its known contribution, y_2, to improve its estimate of y_1. The estimate of y_1 under this scenario is $\hat{y}_1^{(3)} = \hat{Y}^* - y_2$.

Scenario 2 is an example of a differencing attack. Differencing attacks can be effective because any two tables on their own may have low disclosure risk but, when differenced, may have a high disclosure risk. They can be particularly effective when used via a remote server since, at least in the case of TableBuilder, the attacker is relatively free to request tables of their choice (Thompson et al. 2013). More detailed discussions about differencing attacks using remote servers can be found in O'Keefe and Chipperfield (2013).

In order for an attack to succeed the attacker needs:

1. To know that the target is in the sample. It is well known that statistical agencies typically select large businesses with a higher probability than smaller businesses. For smaller businesses, sampling may provide some protection since an attacker will not know if a particular business is selected in the sample. Since the underlying micro-data are not observed, it would be necessary to conduct a series of attacks in order to confirm whether or not a small business is actually in the sample (Chipperfield and O'Keefe 2014).
2. In the case of Attack Scenario 2, to know how to uniquely identify the target in terms of a set of quasi-identifiers. This allows the attacker to "drop" the target business from the cell in a table. To conduct Attack Scenario 1 and 3, the business does not have to be uniquely identified, often referred to as *identification*, only that the target business dominates the cell.
3. To circumvent TableBuilder's confidentiality protections and disclosing an attribute of the business.

TableBuilder gives users a high degree of flexibility in choosing a table's dimensions and scope. There is often considerable information about large businesses in the public domain which may in turn make identification likely (e.g., there may only be one private hospital in a small area). Accordingly, we conservatively assume that 1. and 2. occur with certainty. So for large businesses at least, the only protection available in TableBuilder is perturbation. Consequently, perturbation is the focus of how disclosure risk is measured (see Subsection 3.2).

3.2. *What is Disclosure*

In many organisations, disclosure is considered to occur for a business if published estimates can be used to accurately infer an attribute (e.g., total turnover) of a business.

It is not necessary for the attribute to be inferred exactly – the degree of (or threshold for) accuracy required for disclosure must be determined by the Statistical Agency.

We say that the disclosure risk from Attack Scenario s, τ_s, is acceptable if the probability that the estimate $\hat{y}_1^{(s)}$ is within $V_s\%$ of the true value y_1 is less than R_s. This is a stochastic generalisation of the P% Rule (Tambay and Fillion, 2013). Therefore the disclosure risk from Attack s, τ_s, is acceptable if,

$$\tau_s = \Pr\left(\frac{\left|\hat{y}_1^{(s)} - y_1\right|}{y_1} \leq V_s\%\right) \leq R_s \tag{1}$$

We can say that for attack s, V_s is the threshold value that draws the line between what does and what does not constitute a disclosure and R_s is the *acceptable disclosure risk*. Different values of (R_s,V_s) in different scenarios could be justified on the basis of whether the attack scenario is likely to occur in the first place (e.g., level of sophistication and prior knowledge required to carry out the attack) and the level of the business structure (e.g., business, enterprise, employee) that is attacked.

To illustrate the rule, consider the following example. Consider three businesses in a cell that have weights of 1 and Income USD 98, USD 1 and USD 1. Following Attack 1, a user could guess that the Income of the largest contributor is equal to the cell estimate of USD 100. This guess would be wrong by only 2% (USD 100−USD 98)/USD 98. TableBuilder would not release the unperturbed estimate of USD 100; it would instead release a perturbed estimate. Consider if the possible perturbed estimates (each equally likely) were 60, 70, 100, 130, 140, 150, 160. Again following Attack 1, if a user now guessed that the Income of the largest respondent (USD 98) is equal to the cell's perturbed estimate, the guesses would be wrong by -39, -29%, -2%, 33%, 43%, 53%, and 63%. The guesses using perturbed estimates would be within 18% about 15% of the time. The risk associated with Attack 1 would be acceptable if disclosure and the disclosure risk were $V_1 = 18$ and $R_1 = 0.15$, respectively.

3.3. Defining Utility Loss

We measure utility loss associated with the perturbed estimate $\hat{Y}^*$ by

$$L = |P^*|/\hat{Y}. \tag{2}$$

The magnitude of the perturbation, $|P^*|$, depends upon K and the 'magnitude values' m_i for $i = 1, \ldots, K$. The utility loss measure is also used by Yancey et al. (2002) in assessing utility loss of a sample mean. There are other useful measures of utility loss, including the mean square error and the mean absolute error (Domingo-Ferrer and Torra 2001).

3.4. Optimal Magnitude Values

The optimal value of $\mathbf{m}$ minimises L, given by (2), subject to the constraint given by (1) for $s = 1$, 2, and 3, where $(R_1,V_1) = (0.15,18)$, $(R_2,V_2) = (0.15, 11)$ and $(R_3,V_3) = (0.15,11)$. That is, the optimal value of $\mathbf{m}$ minimises utility loss subject to having an acceptable disclosure risk from Attacks 1, 2 and 3. Below we describe how the optimal values of $\mathbf{m}$ were obtained.

It is important to note that the *scale* of the distribution of y in the cell does not affect the optimal value of $\mathbf{m}$ – what is important is the relative size of y for the contributors to the cell. Table 2 shows a variety of distributions for y. The distributions are made up of between two and four units (other units could well belong to cell but, if they do, we assume they make a negligible contribution). For each of these distributions, we had to choose a value of K. We found limited additional benefit from allowing $K = 4$ and so we decided to set $K = 3$ for all distributions. This means where a distribution was made up of four units, only the top three contributing units were used in calculating the perturbation, P^*. The exception to this was Distribution 5, which was made up of only two contributors (of relative size 60 and 40), and so we set $K = 2$.

For each distribution of y, Table 2 gives the optimal value of $\mathbf{m}$ for $K = 3$. For a given distribution of y, the optimal value was found by:

(i) measuring the average value of L and measuring the disclosure risks, τ_s for $s = 1, 2$ and 3, for a range of different values of $\mathbf{m}$**.** For a given value of $\mathbf{m}$**,** these measures were calculated by simulating the perturbation distribution, P^*, 500 times and conducting Attacks 1, 2 and 3.

(ii) identifying the value of $\mathbf{m}$ from (i) that minimised L while also meeting the constraint on the risk from Attacks 1, 2 and 3 as described in the first sentence of Subection 3.4.

Table 2 shows that, as the distribution of y becomes more uniform, the optimal values in the vector $\mathbf{m}$ increase in size.

Figure 1 illustrates the risk-utility trade-off with respect to only Attack 2. *Utility Loss* is the average value of L and *Risk* $= \tau_2$ is the proportion of times condition (1) was met from Attack 2, over 500 independently generated values of P^*. Figure 1 plots *Utility Loss* by *Risk* for Attack Scenario 2 for a range of values of $\mathbf{m}$ and for two disclosure thresholds ($V_2 = 11$, 18). Recall that $V_2 = 11$ means that disclosure occurs when $\hat{y}_1^{(2)}$ is within 11% of y_1.

Table 2. Magnitude values that meet constraints on the disclosure risk and maximise utility for different contributor values.*

Distribution number	Distribution of y (relative size of contributors)				Optimal values (K = 3)		
	1st	2nd	3rd	4th	m_1	m_2	m_3
1	90	5	5	0	0.15	0.1	0.1
2	80	10	5	5	0.15	0.1	0.1
3	70	20	10	0	0.15	0.1	0.1
4	60	20	10	10	0.2	0.1	0.1
5	60	40	0	0	0.25	0.15	n/a
6	50	20	20	10	0.25	0.15	0.1
7	40	30	30	0	0.3	0.2	0.1
8	30	30	30	10	0.4	0.3	0.2
9	25	25	25	25	0.5	0.4	0.3

*The constraints on the disclosure risk that are imposed by Attacks 1, 2 and 3 are (R_1, V_1) = (0.15, 18), (R_2, V_2) = (0.15, 11) and (R_3, V_3) = (0.15, 11).

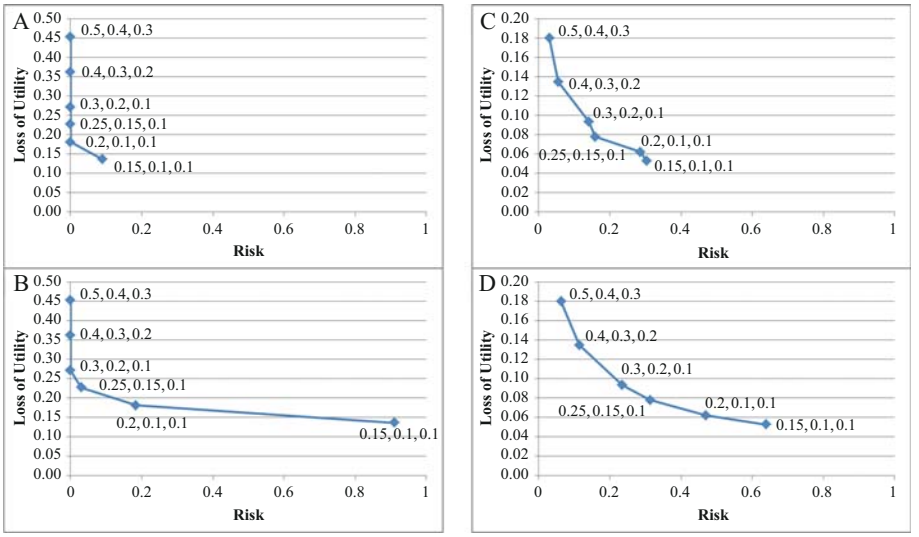

Fig. 1. Risk vs utility loss under attack scenario 2. Figures A and B have y values of (90, 5, 5, 0), Figures C and D have y values of (30, 30, 30, 10). Figures A and C define disclosure by $V_2 = 11$, Figures B and D define disclosure by $V_2 = 18$.

Figure 1A shows if the (relative) y values were (90, 5, 5), $\mathbf{m} = (0.15, 0.1, 0.1)$ and the disclosure threshold was $V_2 = 11$, that *Utility Loss* = 13% and *Risk* = 10%. Figure 1B shows that if disclosure was instead defined by $V_2 = 18$, *Risk* would ris dramatically to 90%.

Figure 1C shows that if the relative y values were (30, 30, 30, 10), $\mathbf{m} = (0.15, 0.1, 0.1)$ and the disclosure threshold $V_2 = 11$ that *Utility Loss* = 5% and *Risk* = 30%. Figure 1D shows that if the disclosure threshold was instead $V_2 = 18$ that *Risk* = 64%.

Ideally, TableBuilder would allow the choice of $\mathbf{m}$ to depend upon on the *actual* distribution of y in each cell (as per Table 2). As TableBuilder does not have this capability, we must choose a single value of $\mathbf{m}$ that guarantees an acceptable disclosure risk for *all* distributions of y in Table 2. The resulting optimal value would be $\mathbf{m} = (0.5, 0.4, 0.3)$. However, we did not use $\mathbf{m} = (0.5, 0.4, 0.3)$ because the utility loss was too high. The compromise value of $\mathbf{m} = (0.4, 0.3, 0.2)$, that we used in all empirical studies below, does not strictly have an acceptable disclosure risk for Distribution 9 in Table 2. (Note: because the disclosure risk is somewhat contextually free in the way it is defined here, we focus on measuring utility loss in the empirical studies).

Work on the optimal distribution for d_i^* is currently being investigated by some of the authors of this article. Consider using $q^* d_i^*$ instead of d_i^*, where $q^* = 1$ if n is odd and is equal to -1 if n is even. This change would reduce the disclosure risk of a differencing attack (Attack Scenario 2) while having no effect on the disclosure risk for other attacks. We see that, for any two cells that differ by a single target unit, the direction of the perturbation, $q^* d_i^*$, will be positive for one of the cells and will be negative for the other cell. This change would increase the perturbation variance of the difference, $\hat{Y}^* - \hat{Y}^*(i = 1)$, while having no impact on the perturbation variance of the individual totals, $\hat{Y}^*(i = 1)$ or $\hat{Y}^*$.

4. Evaluation of Employees Earnings and Hours

Employee Earnings and Hours (EEH) is a two-yearly survey of employing organisations in Australia. EEH uses a two-stage sample selection approach. The first stage involves selecting a probability sample of employer units, from the ABS Business Register. The statistical unit for the first stage comprises all activities of an employer in a particular state or territory based on the Type of Activity Unit (TAU). The sampling unit for the second stage is employee. Employees are in scope of the second stage selection if they earned pay during the reference period. Data collected in the survey are used to estimate the composition and distribution of average weekly earnings, hours worked, and the methods of setting pay (e.g., award only, collective agreement, and individual agreement). EEH currently applies suppression to protect respondents against disclosure where a 'respondent' can be an employee, TAU, or at the highest level of Enterprise Group.

4.1. Utility at Employee Level

Tables 3 and 4 summarise the utility loss resulting from perturbing estimates with $\mathbf{m} = (0.4, 0.3, 0.2)$. Here we measure the utility of typical EEH estimates.

Table 3 shows, in most cases, that perturbation changes the estimates by less than 1%. When perturbation makes larger changes (6–7%) to estimates, the associated sampling errors (not provided in Table 3) are also high, due to small sample sizes. For example, the estimate for Community and Personal Service Workers in Owner Manager of Incorporated Enterprises was perturbed by 7.3% and has a Relative Standard Error (RSE) in the range 25–50%.

Table 4 shows, again, that the percentage impact of perturbation is often less than 1%. As in Table 3, the larger differences are for estimates with RSEs between 25% and 50% (RSEs not provided in Table 4). For example, on the one hand, the estimate for Mining and Award Only is perturbed by −9.7% and has a standard error of 10–25%, whereas the estimate for Manufacturing and Award Only is perturbed by only 0.1% and has a standard error of 5%. Since these changes are significantly less than the RSEs the loss of utility would be minimal. Feedback from users of the EEH is that this level of utility loss is acceptable.

4.2. Protection of TAUs

The three attack scenarios are also possible at the TAU level. TableBuilder does not recognise the TAU hierarchy in any way and so its perturbation settings cannot manage disclosure risk at the TAU level. For example, TableBuilder does not recognise if all employees in a cell belong to one TAU. The question is whether, nevertheless, there is acceptable disclosure risk at the TAU level given perturbation is only designed to have acceptable disclosure risk at the employee level.

To illustrate, Table 5 summarises the data collected from a realistic but hypothetical sample of 25 employees who were themselves selected from three different TAUs and who all belong to a single cell of a table. We assume the attacker knows that the cell contains only the three selected TAUs and that the TAUs were selected with certainty. The inverse of the within-TAU employee sampling fraction is used to weight its sample of employees, thus giving the TAU contribution to the cell estimate. Table 5 shows the

Table 3. Estimates of average weekly total cash earnings (USD) by occupation and agreement type and percentage impact of perturbation (%).

Occupation	Award only (%)	Collective agreement (%)	Individual arrangement (%)	Owner incorporated enterprise (%)	All (%)
Managers	1127.1 (1.1)	1982.7 (−0.1)	2083.2 (0.2)	1304.2 (−2.0)	1928.9 (0.1)
Professionals	1147.7 (−0.9)	1383.9 (−0.3)	1507.5 (−0.5)	2036.8 (−0.3)	1436 (−0.1)
Technicians and trades workers	695.8 (−0.6)	1544.4 (−0.2)	1272.5 (0.6)	1219.6 (3.4)	1250.2 (0.3)
Community and personal service	546.5 (0.9)	841.2 (0.5)	587.8 (0.3)	932.1 (7.3)	709.1 (0.3)
Clerical and adminstration	708.4 (0.3)	1063.3 (−0.2)	962 (−0.2)	831.9 (−3.9)	970.2 (−0.1)
Sales workers	419 (−0.7)	485.1 (−0.1)	935.9 (−0.1)	949.7 (0.7)	606.7 (0.0)
Machinery operators	863.8 (0.2)	1509.7 (0.1)	1154.9 (−0.2)	1095.9 (6.7)	1284.7 (0.1)
Labourers	496.5 (0.8)	958.5 (1.3)	780.9 (−1.2)	1321.4 (6.5)	784.3 (0.7)
All	632.7 (−0.2)	1150.7 (0.0)	1277.9 (0.1)	1325.6 (−0.9)	1122.8 (0.0)

Table 4. Estimates of average weekly total cash earnings (USD) by industry and agreement type and percentage Impact of perturbation (%).

Occupation	Award only (%)	Collective agreement (%)	Individual arrangement (%)	Owner incorporated enterprise (%)	All (%)
Mining	1285.3 (−9.7)	2237 (0.7)	2538.6 (0.2)	1639 (−13.3)	2388.8 (0.0)
Manufacturing	614.6 (0.1)	1292.8 (−0.9)	1307.7 (1.1)	1270 (4.8)	1221.7 (0.3)
Electricity, gas, water and waste services	917.9 (0)	1745.4 (−0.3)	1851.5 (−0.5)	966.9 (8.8)	1735.3 (−0.3)
Construction	811.6 (−2.4)	2110.4 (−0.2)	1333.2 (−0.1)	1086.8 (1.6)	1440.2 (0.0)
Wholesale trade	706.2 (2.7)	1110.3 (−0.4)	1352.4 (0.1)	1152.3 (−1.9)	1258.5 (0.0)
Retail trade	479.4 (0.8)	489 (−0.3)	965 (−0.8)	992.6 (−2.5)	640.2 (−0.5)
Accommodation and food services	477.6 (0.5)	398.9 (0.2)	739.8 (0.5)	700.6 (−6.1)	539.3 (0.1)
All industries	633.2 (−0.1)	1150.8 (0.0)	1278.3 (0.1)	1352.6 (1.1)	1122.7 (0.0)

Table 5. Perturbation of weekly earnings.

	Total number of employees	Number of sampled employees	Contribution to unperturbed estimate	Sample RSE (%)
TAU 1	1700	11	2,639,240	13.1
TAU 2	630	8	996,000	4.4
TAU 3	140	6	198,660	14.3
Total unperturbed estimate	2480	25	3,834,000	9.1
Total perturbed estimate	2700	23	3,925,000	9.1

unperturbed estimates of Total Earnings of USD 3,834,000. Given that TAU 1 and 2 dominate the cell, it is likely that such a cell would be suppressed.

Consider if the unperturbed estimate of the cell was released. Under Attack Scenario 3, TAU 2 subtracts their contribution to the unperturbed estimate in order to estimate TAU 1's contribution to Total Earnings. The estimate would be $\hat{y}_1^{(3)} = 3{,}834{,}000 - 996{,}000 = 2{,}838{,}000$ or 7% higher than the actual contribution of TAU 1.

By way of an aside, it is important to note that disclosing TAU 1's contribution to the cell estimate is not by itself disclosure. TAU 1's contribution to the cell estimate, which is based on only 11 out of 1700 of its employees, will differ from TAU 1's true Total Earnings (obtained by summing the Total Earnings of each of its 1700 employees) due to sample error. Based on the sample of eleven of its employees, the RSEs of TAU 1's Total Earnings is 13%. Within the framework of Section 3 we can say, assuming a normal distribution, that there is a 95% chance that TAU 1's contribution to the cell estimate is within 26% of TAU 1's true Total Earnings; or equally we could say that TAU 1's contribution to the estimate is within 18% of its true Total Earnings about 83% of the time. So even if the attacker was able to exactly calculate TAU 1's contribution to the estimate, the sampling of employees provides some protection against disclosing TAU 1's true Total Earnings. Here we conservatively assume that a TAU's contribution to Total Earnings and its true Total Earnings are the same.

If we repeat Attack Scenario 3 using perturbed, rather than unperturbed, estimates we see that the estimate of TAU 1's contribution would be $\hat{y}_1^{(3)} = 3{,}925{,}000 - 996{,}000 = 2{,}929{,}000$ and would be 11.5% larger than the actual contribution of TAU 1. Over repeated perturbations, we showed (details not given) for this example that TableBuilder would not provide sufficient protection (using $R_3 = 11$, $V_3 = 18\%$) of 'TAU 1's contribution' from Attack 3. Furthermore, we could equally have constructed an alternative example whereby a cell only contains the eleven employees from TAU 1. In this alternative example, the risk from disclosing TAU 1's contribution to the cell estimate would be higher still.

In conclusion, while the TableBuilder perturbation settings guarantee a minimum level of disclosure risk at the employee level, they have little control over the disclosure risk for TAUs that are selected with certainty (typically TAUs with more than 50 employees). However, TAUs that are sampled without certainty may well have sufficient protection if the protections of sample error were to be taken into account.

5. Utility of Releasing International Trade in Goods via TableBuilder

International Trade in Goods is a monthly administrative by-product collection of all in-scope imports and exports to/from Australia. ABS policy is that these commodity values must be protected only if that business officially requests (i.e., 'self-select') to be protected against disclosure. When such a 'self-selected' business contributes to a cell, it is determined whether or not the value of the commodity associated with that business breaks confidentiality rules – if it does then the cell is suppressed. Staff who work on this collection describe the current process of suppression as "involved and time-consuming".

Possible alternative approaches to managing disclosure risk:

i. Perturb the commodity values for only self-selected businesses prior to releasing the data as a public use file. In the Australian context, there is a certain level of public sensitivity to releasing even a perturbed commodity amount for a self-selected business. For this reason, this option was not considered further.
ii. Perturb all cell estimates as described in Section 3. This assumes that all businesses self-select and so will result in more perturbation than is strictly required.
iii. Perturb commodity values for self-selected businesses so that, even if they belonged to a cell on their own, the disclosure risk from Attack 1 is acceptable (using the criteria $(R_1,V_1) = (0.15,18)$). The values for businesses that do not 'self-select' are not perturbed. Users can access the micro-data via TableBuilder with all its perturbation routines turned off. In theory, this would give the same estimates as Approach I, but avoids releasing business-level micro-data.

Tables 6 gives published estimates for merchandise exports by state and from the International Trade in Goods and Services, Australia (ABS cat. No 5368.0). We see that under approach II, some estimates are significantly changed by perturbation. A large perturbation is always due to a small number of dominant businesses in a cell. Table 6

Table 6. Merchandise exports (USD M) by state/territory.

	Published estimate	Perturbed estimates-protect all businesses under approach II (percentage impact of perturbation %)	Perturbed estimates-protect only self-selected businesses under approach I and III (percentage impact of perturbation %)
NSW	2822	2910 (3.1)	2837 (0.5)
VIC	1406	1409 (0.2)	1400 (−0.4)
QLD	2927	2971 (1.5)	2927 (0.0)
SA	728	765 (5.1)	724 (−0.6)
WA	9205	8877 (−3.6)	9253 (0.5)
TAS	207	192 (−7.4)	207 (0.0)
NT	444	512 (15.4)	444 (0.0)
ACT	4	6 (35.4)	4 (0.0)
AUS	17743	17642 (−0.6)	17796 (0.3)

shows that the utility loss under the approach III is quite small by comparison. This is because, at least in the estimates of Table 6, large dominant businesses often do not self-select. Feedback from users is that the loss of utility under approach III is acceptable at a high level and further work is planned to consider whether this would also be the case for estimates at fine levels. For cells in which dominating businesses self-select, the perturbation applied by TableBuilder may be unacceptably high. In the next section we see a situation where the perturbation is, in fact, unacceptably high.

6. Utility of Releasing Land Management Practices via Tablebuilder

Land Management Practices Survey (LaMPS) estimates are released every financial year. LaMPS selects a sample of agricultural businesses in Australia above a minimum cut-off size. LaMPS aggregates are released in the form of tables. Suppression is then applied to table cells that are considered to have an unacceptable disclosure risk. Often, estimates in the cell of a LaMPS table will contain a small number of dominant contributors. Next we briefly show the utility of key estimates after they have been perturbed via TableBuilder.

For eight Australian states and territories, Table 7 shows the (RSE) of the published estimate of Total Nitrogen Fertiliser and the impact of perturbation. Using the magnitude values (0.4, 0.3, 0.2), the impact of perturbation is under 5%, with the exception of the Australian Capital Territory (ACT). In a few cases, the impact of perturbation is comparable to the RSE associated with the estimate (e.g., in NT, the RSE was 3.7% and the impact of perturbation was 3.2%).

For the ACT, the impact of perturbation was 42%. This estimate has low utility after perturbation. The impact of perturbation is high because the ACT has a comparatively low number of contributing businesses and some dominant contributors. While LaMPS would not appear to be suitable for release via TableBuilder, the next section discusses some ways forward.

7. Final Remarks

The three case studies in this article discuss the challenges of allowing access to business data via TableBuilder. For some estimates, TableBuilder can provide an effective level

Table 7. RSE and percentage impact of perturbation: Total nitrogen fertiliser applied by state/territory.

State	RSE (%)	Impact of perturbation (%)
NSW	2.7	0.6
Vic	5.6	0.6
Qld	2.9	0.3
SA	4.1	1.8
WA	2.1	0.3
Tas	4.8	3.4
NT	3.7	3.2
ACT*		42.4
Australia	1.4	−0.1

*Published value for ACT was incorporated into NSW.

of protection against disclosure without noticeably affecting utility of the estimates. However, there are certain cell estimates that would not seem to be suitable for release using TableBuilder – these include cells containing a small numbers of businesses that are dominant contributors. More work would be required before the ABS would consider allowing access to its business data via TableBuilder.

However, we believe that the work and findings here will be applicable to other statistical agencies. This is because the features of the business data considered in this paper are common across the world: dominating businesses (e.g., monopoly and duopolies); the need to protect against disclosure at multiple levels of a business hierarchy; and data collected from samples and from administrative sources. Below, we discuss possible applications our work and future work that will improve the disclosure risk-utility trade-off of a remote server approach.

A practical application of our work would be to release as much data as possible through TableBuilder, but to exclude certain subsets of businesses (large businesses). Other methods could be explored for releasing these data subsets – for example, users with a particular research need for the excluded data could apply for access through a special user request, and other methods (such as suppression) could be applied to protect the data. This approach would allow the release of a wide range of business data in a cost-effective way, while still retaining the flexibility to release specific estimates via means other than TableBuilder.

There are some interesting areas for further work:

1. The attacker does not know the target's estimation weight (i.e., it is always assumed to be equal to one). The extent to which this reduces the disclosure risk has not been measured here. A way of taking this into account would be to allow the magnitude values m_i in P^* to depend upon the weight of the K largest contributors, w_i, for $i = 1, \ldots, K$. It is likely that a unit with a high weight would require a much smaller (possibly equal to zero) magnitude value than a unit with a small weight.
2. Sampling (e.g., sampling of employees in Section 4) reduces the risk of disclosure because the attacker does not know if the target unit is selected in the sample. This is important since a benefit of the remote server is that, unlike the release of micro-data, attacks may be required to even establish whether the target is selected in the sample. Chipperfield and O'Keefe (2014) showed even establishing whether or not a target is in the sample can require a significant number of attacks. The reduction in disclosure risk due to sampling could be off-set by a reduction in the degree of perturbation, leading to an increase in utility.
3. Sampling error can reduce the disclosure risk (e.g., in Section 4 we ignored the protection provided to a TAU due to selecting only a sample, rather than all, of its employees). It would be interesting to allow the magnitude of the perturbation to depend upon the degree of protection already provided by sample error (e.g., if sampling employees within a TAU provides sufficient protection, is there a need to perturb the TAU's contribution to estimates)?
4. Preventing a differencing attack from occurring in the first place. This would mean supressing a cell if it, together with a previously released cell, met the condition of a differencing attack.

5. The optimal magnitude parameters (Subsection 3.4) assumed y took only positive values. This could be extended to allow for negative values of y.
6. As mentioned, the current functionality of TableBuilder fixes K, m_i for i $= 1, \ldots,$ K and the distributions of d_i^* and h_i^*. Further work could consider allowing these to depend upon the perceived sensitivity of y and the distribution of y in the cell (e.g., if the top three contributors' relative values of y were approximately -10, 20 and 90).
7. Should the agency release the perturbation parameters underlying P^*? Releasing the parameters would, under any attack scenario, allow an attacker to put a bound on y_1 for each cell total that contained unit 1. The risk and the utility of releasing the parameters would need to be measured. Instead, an indication (perhaps in ranges) of the size of the perturbation or the MSE of the perturbation would be released but, again, any impact on risk and utility should be measured.

8. References

Abrahams, C. and K. Mahony. 2008. "2New Policy and Procedures Governing the Release of Microdata Derived from ONS Social Surveys." *13th GSS Methodology Conference*, London, June 23, 2008. Available at: https://www.ons.gov.uk/ons/media-centre/events/past-events/thirteenth-gss-methodology-conference–23-june-2008 (accessed January 2018).

Chipperfield, J.O. 2014. "Disclosure-Protected Inference with Linked Micro-data using a Remote Analysis Server." *Journal of Official Statistics* 30: 123–146. Doi: http://dx.doi.org/10.2478/jos-2014-0007.

Chipperfield, J.O. and C. O'Keefe. 2014. "Disclosure-Protected Inference using Generalised Linear Models." *International Statistical Review* 82: 371–391. Doi: https://doi.org/10.1111/insr.12054.

Chipperfield, J.O., D. Gow, and B. Loong. 2016. "The Australian Bureau of Statistics and releasing frequency tables via a remote server." *Statistical Journal of the IAOS* 1: 53–64. Doi: https://doi.org/10.3233/SJI-160969.

Domingo-Ferrer, J. and V. Torra. 2001. "Disclosure Protection Methods and Information Loss for Microdata." In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz, 91–110. Amsterdam: North-Holland.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography TCC*, edited by S. Halevi and R. Rabin, 265–284. Heidelberg: Springer.

Evans, T., L. Zayatz, and J. Slanta. 1998. "Using Noise for Disclosure Limitation of Establishment Tabular Data." *Journal of Official Statistics* 4: 537–551. Available at: https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/using-noise-for-disclosure-limitation-of-establishment-tabular-data.pdf (accessed January 2019).

González, J.J.S. 2005. "A Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Methods." *Operations Research* 53: 819–829. Doi: https://doi.org/10.1287/opre.1040.0202.

Krsinich, F. and A. Piesse. 2002. "Multiplicative Microdata Noise for Confidentialising Tables of Business Data." *Statistics New Zealand*. Available at: http://archive.stats.govt.nz/browse_for_stats/businesses/business_characteristics/multiplicative-microdata-noise-for-business-data.aspx (accessed January 2019).

Lucero, J., L. Zayatz, L. Singh, J. You, M. DePersio, and M. Freiman. 2011. "The Current Stage of the Microdata Analysis System at the U.S. Census Bureau." *Proceedings of the World Congress of the International Statistical Institute*, 3115–3133. Dublin. Available at: http://2011.isiproceedings.org/papers/650103.pdf (accessed January 2019).

Miranda, J. and L. Vilhuber. 2013. "Looking back on three years of Synthetic LBD Beta." Cornell University. Available at: http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1013&context=ldi (accessed January 2019).

O'Keefe, C. and J. Chipperfield. 2013. "A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems." *International Statistical Review* 81: 426–455. Doi: https://doi.org/10.1111/insr.12021.

Reuter, W.H. and J.M. Museux. 2010. "Establishing an Infrastructure for Remote Access to Microdata at Eurostat." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and E. Magkos, 249–257. Berlin, Heidelberg: Springer.

Tambay, J. 2017. "A layered perturbation method for the protection of tabular outputs." *Survey Methodology* 43: 31–40. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017001/article/14818-eng.pdf?st=qzA3QL0u (accessed January 2019).

Tambay, J.-L. and J.M. Fillion. 2013. "Strategies for processing tabular data using the G-Confid cell suppression software." *Proceedings of the Survey Research Methods Section*. American Statistical Association Joint Statistical Meetings, Montreal, August 3–8, 2013. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/7_gconfid.pdf (accessed January 2019).

Thompson, G., S. Broadfoot, and D. Elazar. 2013. "Methodology for the Automatic Confdentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics." *UNECE Work Session on Statistical Data Confidentiality*, Ottawa, October. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf (accessed January 2019).

Yancey, W.E., W.E. Winkler, and R.H. Creecy. 2002. "Disclosure Risk Assessment in Perturbative Micro-data Protection." In *Inference Control in Statistical Databases*, edited by J. Domingo-Ferrer, 135–151. New York: Springer.

Received September 2016
Revised September 2018
Accepted January 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 337–352, http://dx.doi.org/10.2478/JOS-2019-0016

Enhancing Survey Quality: Continuous Data Processing Systems

Karl Dinkelmann[1], *Peter Granda*[1], *and Michael Shove*[1]

Producers of large government-sponsored surveys regularly use Computer-Assisted Interviewing (CAI) software to design data collection instruments, monitor fieldwork operations, and evaluate data quality. When used in conjunction with responsive survey designs, last-minute modifications to problems in the field are quickly addressed. Complementing this strategy, but little discussed, is the need to implement similar changes in the post data collection stage of the survey data life cycle. We describe a continuous data processing system where completed interviews are carefully examined as soon as they are collected; editing, recode, and imputation programs are applied using CAI tools; and the results are reviewed to correct problematic cases. The goal: provide higher quality data and shorten the time between the conclusion of data collection and the appearance of public use data files.

Key words: Data quality; curation; tools; dissemination.

1. Introduction

Many survey research projects depend heavily on Computer-Assisted Interviewing (CAI) to program the design of data collection instruments, improve error checking, closely monitor fieldwork operations to counter nonresponse, increase response rates, and evaluate completed cases almost immediately after they are collected. More recently, several commentators have focused on post data collection issues, particularly with correcting nonsampling errors as an essential component to improve overall survey quality (De Waal 2013; Thalji et al. 2013). Even before CAI became a standard method of conducting many large national and cross-national surveys, the connections between data collection and data processing had become more collaborative (Biemer 2010). Principal investigators have a great incentive to process and analyze their data as quickly as possible in order to publish their results and to meet data sharing requirements now demanded by many funding agencies.

CAI added a very powerful dimension to this connection. It permitted storage of the variable-level metadata: variable names and labels, question text, universe statements, interviewer instructions, missing data definitions, and so on within the actual data collection instrument. Although not an early priority, CAI systems could repurpose this metadata for such things as public use documentation or to reuse the material when creating project reports.

Certainly, one of the main features of CAI systems is to perform data checking during the interview process itself. The programming logic built into the survey instrument

[1] University of Michigan, USA, Institute for Social Research, 330 Packard Street, Ann Arbor, MI, 48104, U.S.A. Emails: karldi@umich.edu, peterg@umich.edu, and mshove@umich.edu

prevents impossible or improbable responses and often carefully controls acceptable answers for demographic questions. For many years, national statistical agencies have developed internal controls to monitor and edit incoming data to standardize workflows and improve data quality (Bethlehem 1997). However, certain types of complex surveys, such as ones that collect family histories and have lengthy questionnaires, which severely test respondent recall, present significant challenges for any automated checking system. Respondents can easily misstate or fail to remember the dates of important events that may become evident only when the entire interview is completed. Data producers must also balance the quest for accuracy with the need to complete interviews within available budgets. Surveys with these characteristics often require considerable checking and editing after the data collection period ends.

Under such conditions, we suggest treating the post-data collection process in the same way as we now treat the planning and conduct of field operations. This article proposes a "*continuous data processing system*" to routinely evaluate inconsistent or illogical responses and make appropriate corrections. The model described below does not require new tools or systems, but uses the features of the original CAI data collection program to perform automated data checking, cleaning, and processing tasks at the same time that interviews are completed in the field.

The initial implementation of this system grew from data processing tasks connected with producing public use files for the National Survey of Family Growth (NSFG), a nationally representative survey conducted by the National Center for Health Statistics (NCHS) in the United States that gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health (http://www.cdc.gov/nchs/nsfg.htm). The NSFG uses CAI systems for all aspects of data collection and the transmission of completed interviews to a central project database. Internal communications protocols review incoming cases from the field daily to verify that each case has sufficient information to qualify as "completed" based on agreed project parameters. After verification, completed interviews are ready for the checking and editing operations.

Our goals for testing the system with the NSFG included addressing the following questions: (1) whether or not it was feasible to review completed records immediately after they were collected in the field; (2) was the new CAI programming application successful in correcting all responses affected by changes in the values of erroneous entries; and (3) how much time and effort would the implementation of this system save for data producers.

We begin by describing why continuous data processing systems are a useful tool for complex surveys, provide a comprehensive description of the system used for the NSFG, how successful we believe the system worked in this initial application, and finish with an assessment of the many data quality implications such systems may provide in the production of public use data files and documentation.

2. Continuous Data Processing Systems

2.1. Why is it Necessary to Change Post-data Collection Procedures?

While it is true that CAI software facilitates the collection and checking of data in the field, it is also the case that CAI programmed instruments are focused on completing interviews

as quickly as possible. Transforming raw data elements into public use files that secondary analysts can use effectively often requires several processing steps that may include:

- Preliminary consistency checking of completed interviews,
- Creation of new recoded/derived variables to facilitate analytic use,
- Imputation of item missing values,
- Generation of several types of weights dependent on the survey population,
- Variance estimation,
- Disclosure review to decide which variables are appropriate for public release,
- Changes to the data (e.g., top and bottom coding, swapping, perturbation) to further protect respondent confidentiality, and
- Creation of extensive documentation on the entire data life cycle to facilitate use by secondary analysts.

These processing steps are often lengthy and time-consuming because of the complexity of CAI-generated interviews. However, the CAI software makes it possible to create a systematic approach to a continuous data processing design that can contribute significantly to expediting processing tasks and satisfying the needs of funders, data producers, and interested researchers at the same time.

A continuous data processing system would perform many of the tasks listed above on a regular schedule so that interviewing and processing occur almost simultaneously. Such a system could conceivably edit and check cases immediately after completion, create recodes (i.e., derived variables calculated from raw data variables) and sampling error codes, calculate weights, and build the basic documentation files. Some tasks, such as imputation and disclosure review would take place only after enough data was collected to permit secondary analyses.

The creation and success of any continuous data processing system would depend upon close collaboration between the collector of the data and those who produce the public use or analytic data files. This collaborative effort must adhere to one of the basic principles about case editing: disturb the original data as little as possible.

One of the earliest attempts to set rules for case editing and apply them to a system for survey data appeared in a seminal article by I.P. Fellegi and D. Holt entitled "A Systematic Approach to Automatic Edit and Imputation" (Fellegi and Holt 1976). Their objective was to design an automated procedure for editing and imputing data that would alter the fewest possible values, maintain the frequency structure of the data file, and derive imputation rules directly from the editing rules. Believing that designing separate computer programs to edit and correct records would be costly and error-prone (perhaps as true today as it was in 1976!) they suggested an approach based on simple, logical rules created by subject matter experts.

In theory, it is now the case, some 40 years later, that the advent and continuous development of CAI software has made it possible to avoid a large amount of post-data collection processing and systematically improve data quality by simplifying data capture and editing tasks. This becomes possible when the CAI software encompasses both data collection and data cleaning operations.

Edit checks are routinely built directly into the software to reduce interviewer entry errors and to require respondents to rethink and correct erroneous or questionable answers. Common patterns of quality checking have emerged with these CAI systems. Completed

interviews are sent from the field to the coordinating center or survey headquarters on a daily basis. Each interview undergoes some type of automated review, and, if problems arise, interviewers and survey managers are in immediate communication to resolve them. Programmers can produce tables that check the values of key indicators in the survey.

Another method used in CAI programming to check recorded values is to create 'computed' variables (essentially recodes built right into the CAI programming structure) based on responses to original questions which can be transferred into final output files and serve as summary variables, saving time and effort in secondary analyses. For example, respondents may answer a series of questions about their race and/or ancestry that would then be condensed into a single 'computed' variable, which is stored and subsequently transferred to the output data file.

These CAI programming structures are especially valuable when a survey collects extensive respondent and family histories regarding work patterns, educational attainments, family formation, and health issues over extended time periods. In such surveys when reporting key family events, interviewers can expect that specific dates might not always be accurate, particularly when the event occurred many years before the date of the interview. Immediate checking of anomalous dates can often be incorporated into the CAI software programs through "hard" edit checks that force interviewers to review problematic or impossible responses with the respondent and correct the information before completing the remainder of the interview. However, survey designers also consider keeping such "hard" edit checks to a minimum so as not to increase the time it takes to collect the interview, cause a refusal, or increase respondent burden. The tradeoff often involves using "soft" edit checks that permit the interviewer to review a particular response but move on to other questions if the respondent does not provide adequate clarification. (Soft checks are also used when a given response is unlikely/improbable yet could still be possible).

Recode programs provide a third opportunity to check possible reporting errors. Post processing recodes can either use raw and/or CAI-generated 'computed' variables in their creation. Once the actual code for generating these post-collection recode variables is complete and thoroughly checked, any cases not meeting the specified conditions may indicate some discrepancy with the data as it was originally collected.

However, the costs of all of these computer-assisted checks may be considerable in both programming effort (as well as testing that they all work correctly) and in the extensive subsequent reviews necessary to ascertain the nature and extent of the problems that they might uncover. Testing of such programs can begin when sample cases are input into the CAI program or if a formal pretest is part of the data collection process. Even with such rigorous testing, it is not always possible to collect a broad enough range of responses to guarantee that the CAI programs are error-free. Having respondents recall events, which happened many years earlier, may present formidable obstacles to ascertain the validity of CAI checking programs.

2.2. *What Steps are Necessary to Implement a Continuous Data Processing System and What Implications Would it Have on Data Quality and Data Dissemination?*

This approach, involving both human and machine interaction, permits completed interviews to be carefully examined as soon as they are collected; identifies problematic

cases; determines resolutions; and, most importantly, *applies data edits directly in the CAI software*. As described below, these data editing and cleaning operations, because they work in close connection with the logic and rules programmed into the CAI instrument, will reduce errors and improve the efficiency of subsequent programs to create derived variables and imputed values, particularly with regard to correcting erroneous date and time values.

The NSFG consists of two main data collection applications, one for women and another for men. The female instrument has more than 8,200 internal consistency checks, programmed within the application to assist the interviewer with inconsistencies found during the course of the interview. The male instrument has more than 3,700. Routing for both instruments is highly dependent on respondents' reporting of events over time. Time and date calculations made within the CAI application use the system time of the data collection laptop during the interview. These date calculations are then used with programmed consistency checks to create new variables throughout the instruments. To facilitate working with data coded in months and years within the application, the concept of the "century month codes" was used. A century month is based on a coding system where the value of 1 is assigned to the month of January 1900 and increments by one for each succeeding month. The following formula translates actual months and years into century months:

$$\text{Century Month} = 12\ (\text{Year} - 1900) + \text{Month}$$

For example, February 2018 would equate to century month 1417.

Accurate reporting of events is necessary for proper routing though the instruments. In addition to internal consistency checks, the female instrument attempts to assist respondents by using a life history calendar to anchor key events to aid in the recall of dates of pregnancies and contraceptive usage to answer them more accurately.

However, despite having more than 11,000 internal consistency checks to aid the interviewer and a life history calendar to assist the respondent with capturing dates correctly, errors happen. To allow us to apply edits to the instrument after data collection finished, we had to turn off the dynamic nature of looking at the computer's system date. This was done by adding additional code to the date processing portion of the CAI system logic to ensure date calculations during the data editing process would be based on the date the interview was completed (instead of dynamically looking at the computer's system date). This allows us to programmatically apply edits to the survey data and systematically reprocess the rules of the instrument. When these edits are applied, it forces downstream rules within the CAI application to update any other areas that would be involved within a given edit. This can sometimes result in 20 or more constructed variables updated from one variable edit applied.

3. System Implementation

The overall model proposed for continuous data processing is illustrated in Figure 1.

The development of a practical continuous processing system should commence even before the start of data collection. Principal investigators often hire survey organizations to collect, clean, and process data for them. As questionnaire specifications are prepared

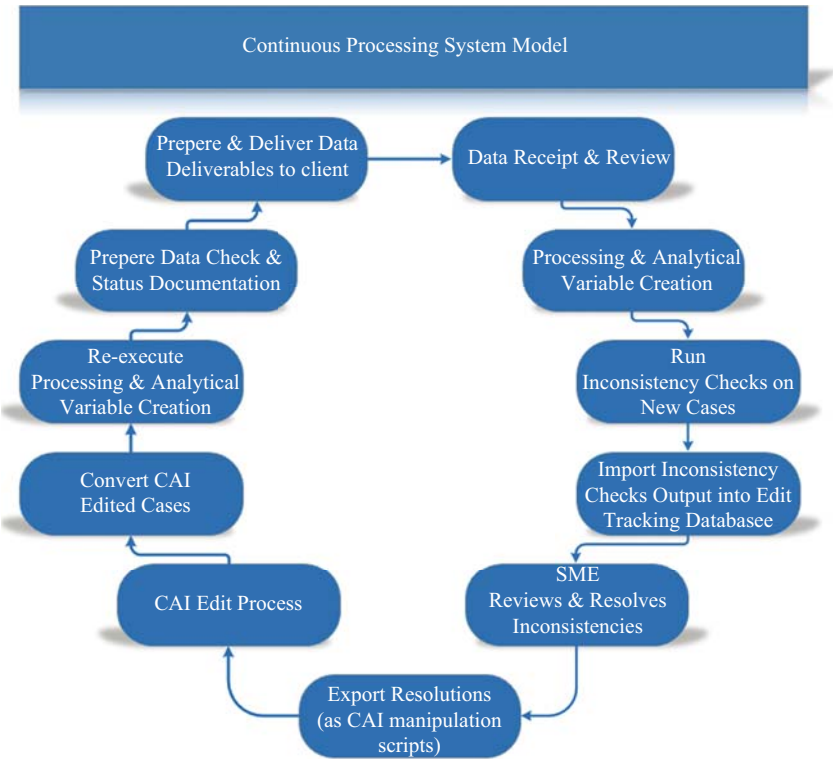

Fig. 1. Continuous processing system model.

and tested, the two key players who design and conduct the survey should meet and decide which types of quality checks they will perform on an initial set of cases. Ideally, these decisions would be in place before any data is collected. Data processing teams could conceivably modify quality checking routines as completed interviews arrive back from the field and they learn which sections of the questionnaire require enhanced review. After a relatively short period, both researchers and data processors would finalize quality checking procedures and methods for dealing with any unexpected anomalies.

At the same time, the data processing team would write and test the post-processing recode programs on this same set of initial cases. The research team would view the results and suggest alterations to recode specifications if additional conditions needed to be included in the programs to cover unexpected reporting situations. A set of system processing rules would follow this procedure. The goal of this initial, potentially intense set of interactions between researchers and data processors would be to integrate both data editing and recoding into a single system that would operate automatically as each new batch of cases arrived in the coordinating or data center. After a time, researchers would only need to review those cases that did not fit the set of agreed upon rules that they had created with the data processors.

Once these preliminary steps are completed and rules and procedures established, the continuous data processing system would operate in production mode with real cases from

Fig. 2. Edit tracking database template.

the field. Checks are then done on an ongoing basis when cases can be evaluated soon after their collection providing the best opportunity for evaluation and resolution.

The heart of this continuous data processing system is the CAI instrument itself. It is the foundation upon which the data editing process is built and consists of the following elements, steps, and procedures that are integrated within the CAI environment.

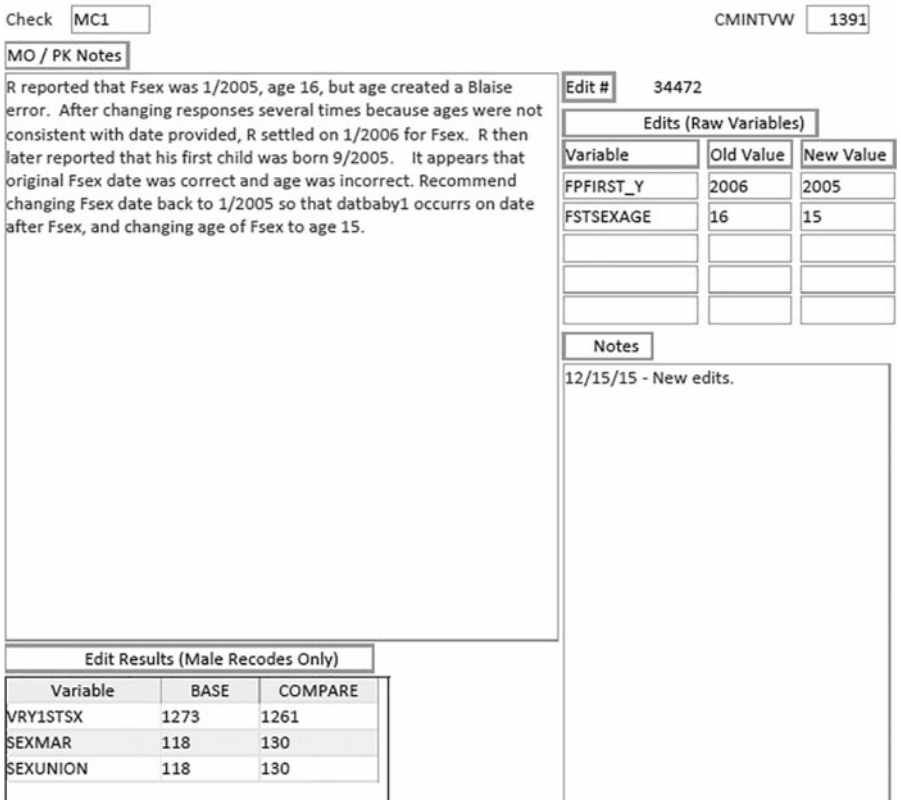

Fig. 3. Subject-matter Expert editing recommendations.

Data Edit Step		Resulting Outcome
1.	Isolate problem cases.	Database with cases to edit.
2.	Apply data edit(s) & re-execute the CAI software rules.	Database with edits applied; with off-route data removed & updated constructed variables.
3.	Read the cases from step 2 to determine newly on-route but empty items.	List of variables placed on-route and are empty.

Fig. 4. Editing steps and outcomes.

A separate "Edit Tracking Database" exists to both review and resolve inconsistencies in overall record logic or individual variables. Subject-Matter Experts (SME) examine problematic cases using both available data and paradata including interviewer comments, case notes, the review of specific interviews through an instrument keystroke playback, and, in some cases, re-contacting the interviewer as quickly as possible for additional information. Solutions are captured in the Edit Tracking Database using a series of forms. Figure 2 shows an example of one of these forms with some of the values for certain variables expressed in century months as described earlier.

Figure 3 shows an interview flagged for editing, the recommendation for editing, and the pre-editing and post-editing values for both raw and computed/recoded variables. In this case, the respondent reported inconsistent information about month of first sex and

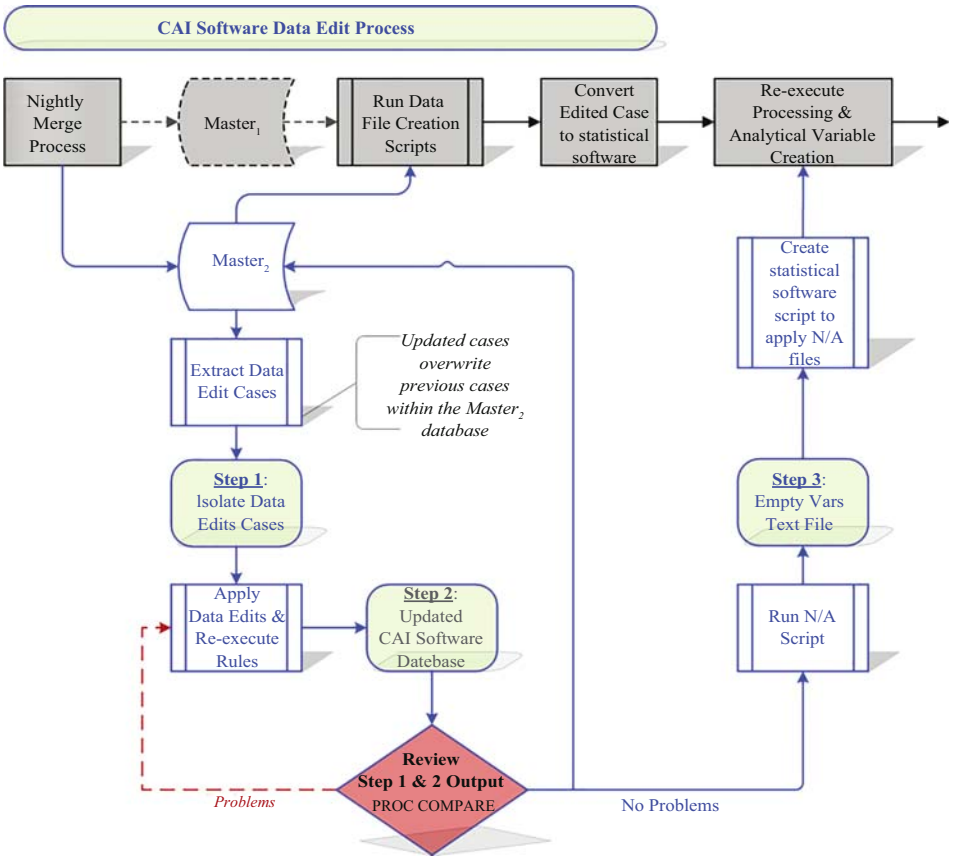

Fig. 5. Capturing the entire process.

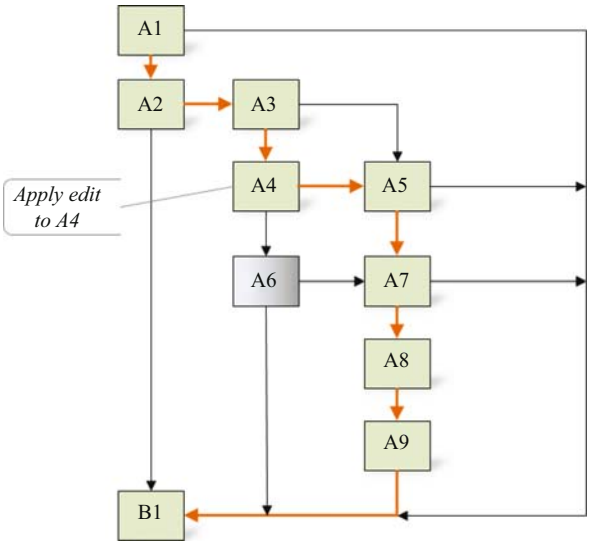

Fig. 6. Sample case question flow.

month of birth of first child. After review by a SME, the date of first sex was corrected as indicated in Figure 3.

The Edit Tracking Database is then queried to dynamically create a series of scripts used in the editing process within the CAI software which not only corrects the original inconsistency, but also values for all other variables affected by the change as noted in Figure 4.

The entire process can be diagrammed as shown in Figure 5.

The 'Nightly Merge Process' captures all cases that interviewers completed that day. These cases are exported to two identical data files: Master_1 and Master_2. Cases identified for review because of the editing checks are extracted and placed into a separate file for adjudication (Step 1). Project staff reviews each case, dynamically exports edit scripts, applies the necessary edits, and re-executes the CAI software program rules in order that the logical flow of the questionnaire is maintained. Re-executing the rules of the CAI

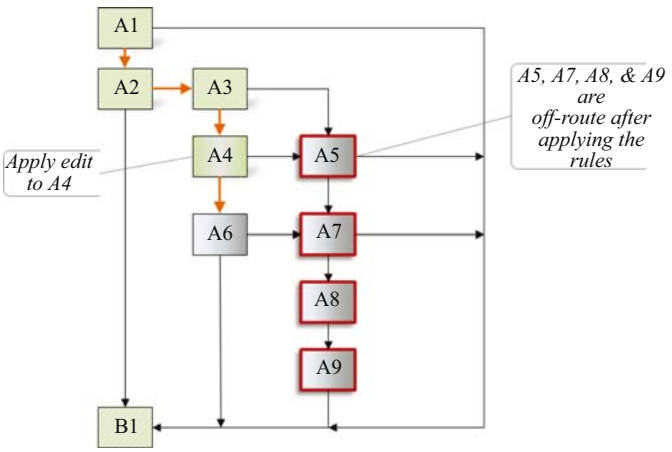

Fig. 7. Sample case rerouting.

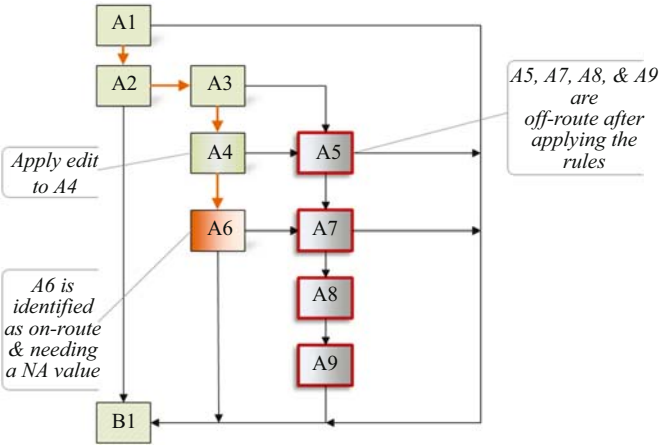

Fig. 8. Sample case editing.

software is done programmatically within the edit scripts. This insures unnecessary calculations or questions that have become off-route are removed and downstream calculations and/or questions that have become on-route are recalculated or identified as missing variables and assigned "not ascertained" (shown in Figure 5 as "N/A") in the process. For example, let us examine the case when a respondent originally answered affirmatively to Question A4. She is then routed to Question A5 and is not asked Question A6, as shown in Figure 6.

If the review process determines that she intended to answer Question A4 negatively, she would be routed to Question A6 instead. Because of this change, the CAI software programming rules would put A6 on-route while A5 and A7 to A9 would be placed off-route. This is illustrated in Figure 7.

The edit would change the value of A4 from affirmative to negative, alter the value of A6 to "not ascertained" since it is now on-route but has no value since it was not asked during the interview, and change the values for A5, A7-A9 to missing since these questions are now off-route.

This procedure is captured in Steps 2 and 3 of Figure 5. Step 2 could be repeated if the review of the new edits indicated a mistake in the code correcting any original values.

Once the review steps are completed, the edited cases are copied back into the Master_2 file and subsequently output from the CAI software and into an ASCII data file with accompanying syntax files that will read the data in such proprietary statistical software packages as SAS and SPSS. It is important to note that the Master_1 file is untouched in this process. It continually collects all of the *original* raw data from the field. This permits data managers to refer back to the original data whenever necessary, should questions arise later about any of the cases that have been edited in Master_2.

Storage of multiple databases should not be a problem for most surveys. While both the Master_1 and Master_2 files are updated at regular intervals, they are cumulative. Earlier versions do not require permanent backup and can be deleted. Once data collection is complete, one copy of the Master_1 data file and one copy of the Master_2 data file will be permanently archived.

4. Data Quality Implications

This editing system will enhance data quality in several ways:

4.1. Enable More Rapid Corrections in the Survey Instrument

Performing checks as cases come in from the field will not only catch potential errors in data collection but might also uncover potential errors in the instrument itself. A consistent pattern of questionable or erroneous values for a particular variable could indicate an error in programming or routing. Correcting such errors as early as possible will minimize the number of cases that must be adjudicated when the final data files are constructed. This illustrates how having a continuous data processing system as part of the normal workflow of a project can also improve data collection activities as such processing checks can actually affect how the instrument is implemented in the field.

4.2. Provide More Consistent Responses in Cases Where the Data Collection Instrument is Particularly Complex

CAI programming allows the construction of very complex instruments that often contain large numbers of "calculation intensive" computed variables – variables actually created by the instrument itself to record information, such as dates reported by the respondent which cover a long period of time. For example, when a survey collects extensive respondent and family histories regarding work patterns, educational attainments, family formation, and health issues over extended time periods the reporting of key family events interviewers and data producers can expect that specific dates might not always be accurate, particularly when the event occurred many years before the date of the interview.

Edit checks would identify and correct probable inconsistent records quickly avoiding the need to do so at the end of the data collection period, when it might be more difficult to uncover details about particular cases.

4.3. Place Active Data Processing Work as a Central Element of the Overall Survey Data Life Cycle

Opportunities to edit and clean data become less effective as the time span between the collection of a case and its review grows. Data collection and data processing should not be two separate stages that occur at very different times, but should occur simultaneously to quickly adjudicate problematic cases. A continuous data processing design will permit comprehensive descriptions of all data checking and cleaning operations from the start of data collection, providing secondary analysts with additional information for them to judge the quality of the data at their disposal.

4.4. Minimize "over Editing" of Data

After data collection ends, data producers often have a tendency to review any suspicious values not caught by the CAI instrument itself during post-collection checking and cleaning operations. The review may involve thousands of interviews, some going back several months or even a year or more. Retrospective editing from this time perspective is

very labor-intensive and is filled with uncertainty, especially when the interviews were collected much earlier. No matter what editing procedures are used, the review of all problematic cases at one time often encourages reviewers to feel that they must address every single case and make editing decisions, even if they do not have enough information to do so. In such cases, the question of whether or not the quality of the data is improved is open to debate (De Waal 2013). Checking cases soon after they are collected for consistent and logical reporting of life history events provides better and less costly opportunities to resolve them since more information is available to make informed decisions. In some instances, interviewers themselves can be recontacted to take advantage of their knowledge since they would have recently completed the case (Seiss et al. 2014). Even if one considers accuracy to be the most important aspect of data quality, survey researchers agree that timeliness and accessibility are also equally key components of quality (Biemer and Lyberg 2003).

4.5. Maximize the "cost-error Optimization" Ratio

The overall quality of the data depends significantly on how project resources are spent. Sufficient resources are necessary for all aspects of the survey data life cycle. Doing extensive data processing work after data collection ends might result in the expenditure of excessive funds and resources on data cleaning operations. If there is a large number of problematic cases to resolve, even if they might only involve a single variable or two, the result could be an unnecessary delay in the release/dissemination of public use files for the research community.

The costs and time involved in editing must always be balanced by the perceived improvements made to the statistical integrity of the final data file itself. Often referred to as "cost-error optimization", data producers should seek a balance in editing operations that seek out systemic problems, but avoid the temptation to check all values for all cases in hopes of producing a dataset devoid of error. Such a goal, of course, is never possible, but the power of modern survey instruments and technologies may make it difficult to decide where the "trade off" occurs. The data file may have 99% of all cases reviewed and cleaned, but the remaining 1% could easily take an inordinate amount of time to resolve. With limited resources, projects often must determine how to deal best with the cost-error optimization ratio. When is the best time to terminate cleaning procedures? Using CAI as part of a continuous processing operation allows projects to determine the kinds of consistency checks they will do. Data managers can concentrate on resolving only those cases. In effect, the system decides where the "trade off" occurs based on a specific set of rules developed by project researchers.

Project staff must always consider the "cost-error optimization" factor when performing these investigations. Test interviews entered by project staff or by real interviewers in a pretest should produce a set of rules and procedures that will determine which areas of the survey instrument and/or key variables are checked when the survey moves into full field production.

A key issue in this process is to determine the involvement of interviewers in the overall editing process. When a particular completed case exhibits unusual anomalies, field supervisors can contact interviewers directly as soon as possible to investigate and correct

possible errors. Recent research has indicated that those most closely involved in the data collection process are more likely to resolve inconsistencies with greater accuracy than other members of the survey research team (Sana and Weinreb 2008).

Project staff must balance the costs involved in having interviewers recheck cases against the potential loss of collecting additional interviews. A continuous data processing system requires a set of clear rules that determines when an interviewer becomes involved in a case, when the case is adjudicated in the main office or coordinating center, and when the inconsistency should remain on the data file. An effective system is not predicated on identifying and seeking to resolve every error or inconsistency. Its objective is to define which anomalies should receive further investigation and to provide a means of doing so at the least cost that will preserve as much of the original data as possible.

4.6. Encourage Faster Data Processing Times

Making the data checking and cleaning operations a continuous process will enable data producers of such complex surveys as NSFG to adjudicate interviews as they emerge from the field. If performed on a regular schedule (e.g., weekly or monthly or even quarterly), many cleaning operations could be completed before the data collection period ends. In a typical two-year data collection period for NSFG, there are an average of 59 female and three male interviews per month flagged for post-collection edits. Completing the editing process at this early stage can also result in reduced errors overall and improved efficiency in subsequent programs, that is, the production of derived variables (recodes), variable modifications due to disclosure review, and imputation. These additional processing steps can proceed more quickly, resulting in quicker turnaround times for the appearance of public use files and happier secondary analysts.

The implementation of a continuous data processing system with the NSFG rests on an ongoing collaboration between the survey organization that collects and processes the data and the principal investigators at the National Center for Health Statistics (NCHS). Subject matter experts at NCHS receive error-checking reports on a regular basis, evaluate proposed solutions that the survey organization provides based on a review of each case including any comments provided by the interviewer, and make final decisions on all edits. This process, which takes place while data collection is ongoing, lessens the amount of time devoted to this task after data collection ends. If the system is implemented in the same time as the collection of data is monitored, it can result in the release of public use data files several months earlier than originally anticipated.

A continuous data processing system also provides more flexibility in scheduling new releases of data. Since new cases are consistently reviewed, checked, and updated, they can be maintained in a single data repository. This facilitates the creation of different types of data files, for example, for different time periods or for specific kinds of respondent groups as data accumulates sufficiently to encourage analyses of new topics or subpopulations.

5. Total Survey Error (TSE)

Any quality enhancements derived from implementing a continuous data processing system directly relate to such nonsample aspects of the total survey error paradigm as usability/interpretability, relevance, accessibility, and timeliness/punctuality (Groves and

Lyberg 2010). Biemer (2010) cites the existence of quality reports and profiles as evidence that these concepts are attracting greater attention by survey managers. Yet they only exist for relatively few major surveys and focus more on discussions of response and imputation rates, but very seldom on other components of TSE, largely because guidelines and requirements do not yet exist (Groves and Lyberg 2010).

Increasingly, major surveys such as the European Social Survey (ESS) provide formal reports when data files and documentation are released for public use. Yet these documents still focus primarily on such topics as coverage, sampling, and nonresponse adjustment. The authors of the ESS quality report for Round 6 recognize this emphasis in their own work and go on to state that "the equivalent and comprehensive report for future ESS rounds should cover all or at least more aspects of the survey life cycle: from translation and sampling to data cleaning and processing. This extension is necessary to assess the overall quality of the produced data" (Beullens et al. 2014). A developed continuous data processing component that creates comprehensive documentation throughout the data collection process can become an integral part of the TSE evaluation and provide data users with a fuller understanding of the survey's "fitness for intended use".

6. Summary

This article has argued that the production of public use data files from complex surveys that rely on computer-assisted data collection software would benefit from the implementation of continuous data processing systems. Testing such a system with the National Survey of Family Growth allowed us to investigate some key questions about how successful it might work and what obstacles it might encounter.

Our first goal for testing the system focused on the feasibility of reviewing records soon after interviews were completed. It is common practice in survey research that all records are automatically checked for completeness, as well as plausible values on certain key variables. It was relatively easy to expand these checks to search for more subtle inconsistencies that would normally be resolved much later during the post-collection period. This work did involve additional time and effort, but became part of a regular monthly error-checking routine. We believe that implementing this enhanced review was successful, but it required full cooperation between the data producer and project investigators to adjudicate problematic cases on a timely basis.

Our second goal was to test the validity of using the CAI program, created to collect the data as efficiently as possible, to correct errors uncovered *after* interviews were sent back from the field. We considered this process as a novel development in CAI programming uses. Would the program correct erroneous values and make appropriate changes to values on subsequent questions if necessary? Our examination of all altered values suggested that the programming changes worked as intended. In particular, the program successfully created "inapplicable" or "not ascertained" values based on changes made to key variables that affected the routing of the questionnaire into different paths.

Finally, and perhaps the most difficult outcome to measure, were the costs and benefits of implementing this continuous processing system. The costs included the CAI programming changes, the monthly checks of all records, determination of which records to change and assigning appropriate values to each item, checking the results, and

replacing records with corrected values when the data producer and project investigators agreed on the change. The benefits included saving time in the post-data collection phase by adjudicating problematic records as early as possible, simplifying the recoding and imputation processes by eliminating inconsistent inputs, and focusing all staff on the importance of thinking about the creation of public use files as an integral component of the project from its inception. The key overall factor in measuring the success of this system may very well be the degree of cooperation and commitment to work on the task continuously. Since, in most cases, resources are always stretched, it is often easier to decide to pursue this kind of checking when the project is focused solely on producing public use files. We believe the system implemented for the NSFG improved the quality of the end product, but every survey with similar characteristics may decide differently.

While they are not an integral part of responsive survey design, we suggest that continuous data processing systems may add a new component to recent examinations of the effectiveness of such designs (Tourangeau et al. 2016) and shares similar characteristics. It allows data producers to administer the post-data collection process in the same manner as the planning and conduct of field operations. Just as principal investigators review the data coming in from the field and make adjustments to rework existing questions or formulate new ones and as survey managers follow sampling strategies and constantly review interviewer assignments to maximize response rates, so too data managers and processors should review and, where appropriate, correct erroneous data values. When continuous data processing happens while field operations are ongoing, we begin to mesh the survey production and data processing environments, moving them away from their long history of separation and closer to a unified process.

Utilizing the advantages of CAI programming as an integral part of a continuous data processing system can have significant advantages: the production of higher quality data, expedited availability to the research community and greater flexibility in addressing topics that are more timely and relevant to current research agendas.

7. References

Bethlehem, J. 1997. "Integrated Control Systems for Survey Processing." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. De leeuw, C. Dippo, N. Schwarz, and D. Trewin, 371–392. New York: Wiley and Sons, Inc.

Beullens, K., H. Matsuo, G. Loosveldt, and C. Vandenplas. 2014. *Quality report for the European Social Survey, Round 6*. London: European Social Survey ERIC. Available at: http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_quality_report.pdf (accessed September 2018).

Biemer, P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5): 817–848. Doi: https://doi.org/10.1093/poq/nfq058.

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons.

Biemer, P., D. Trewin, H. Bergdahl, and Y. Xie. 2017. "ASPIRE." In *Total Survey Error in Practice*, edited by P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West, 359–385. Hoboken, New Jersey: John Wiley & Sons, Inc. Doi: https://doi.org/10.1002/9781119041702.ch17.

De Waal, T. 2013. "Selective Editing: A Quest for Efficiency and Data Quality." *Journal of Official Statistics* 29(4): 473–488. Doi: https://doi.org/10.2478/jos-2013-0036.

Fellegi, I.P. and D. Holt. 1976. "A Systematic Approach to Automatic Edit and Imputation." *Journal of the American Statistical Association* 71: 17–35. Doi: https://doi.org/10.1080/01621459.1976.10481472.

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. Doi: https://doi.org/10.1093/poq/nfq065.

Groves, R.M., W.D. Mosher, J. Lepkowski, and N.G. Kirgis. 2009. *Planning and Development of the Continuous National Survey of Family Growth* . National Center for Health Statistics. Vital Health Stat 1(48). Available at: https://www.ncbi.nlm.nih.gov/pubmed/20141029 (accessed May 2019).

Sana, M. and A.A. Weinreb. 2008. "Insiders, Outsiders, and the Editing of Inconsistent Survey Data." *Sociological Methods Research* 36: 515–541.

Seiss, M., E.A. Vance, and R.P. Hall. 2014. "The Importance of Cleaning Data During Fieldwork: Evidence from Mozambique." *Survey Practice* 7(4). E-ISSN: 2168-0094. Available at: http://www.surveypractice.org/article/2864-the-importance-of-cleaning-data-during-fieldwork-evidence-from-mozambique. (accessed September 2018).

Thalji, L., C.A. Hill, S. Mitchell, R. Suresh, H. Speizer, and D. Pratt 2013. "The General Survey System Initiative at RTI International: An Integrated System for the Collection and Management of Survey Data." *Journal of Official Statistics* 29(1): 29–48. Doi: https://doi.org/10.2478/jos-2013-0003.

Tourangeau, R., J.M. Brick, S. Lohr, and J. Li. 2016. "Adaptive and Responsive Survey Designs: A Review and Assessment." *Journal of the Royal Statistical Society, Series A* 180(1): 203–223. Doi: https://doi.org/10.1111/rssa.12186.

Received August 2016
Revised July 2018
Accepted September 2018

sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 353–386, http://dx.doi.org/10.2478/JOS-2019-0017

Measuring Trust in Medical Researchers: Adding Insights from Cognitive Interviews to Examine Agree-Disagree and Construct-Specific Survey Questions

Jennifer Dykema[1], *Dana Garbarski*[2], *Ian F. Wall*[3], *and Dorothy Farrar Edwards*[4]

While scales measuring subjective constructs historically rely on agree-disagree (AD) questions, recent research demonstrates that construct-specific (CS) questions clarify underlying response dimensions that AD questions leave implicit and CS questions often yield higher measures of data quality. Given acknowledged issues with AD questions and certain established advantages of CS items, the evidence for the superiority of CS questions is more mixed than one might expect. We build on previous investigations by using cognitive interviewing to deepen understanding of AD and CS response processing and potential sources of measurement error. We randomized 64 participants to receive an AD or CS version of a scale measuring trust in medical researchers. We examine several indicators of data quality and cognitive response processing including: reliability, concurrent validity, recency, response latencies, and indicators of response processing difficulties (e.g., uncodable answers). Overall, results indicate reliability is higher for the AD scale, neither scale is more valid, and the CS scale is more susceptible to recency effects for certain questions. Results for response latencies and behavioral indicators provide evidence that the CS questions promote deeper processing. Qualitative analysis reveals five sources of difficulties with response processing that shed light on under-examined reasons why AD and CS questions can produce different results, with CS not always yielding higher measures of data quality than AD.

Key words: Agree-disagree questions; questionnaire design; cognitive interviewing; response processes; data quality; construct-specific questions.

[1] University of Wisconsin Survey Center (UWSC), 4308 Sterling Hall, 475 N. Charter St. Madison, WI 53706, U.S.A. Email: dykema@ssc.wisc.edu
[2] Loyola University Chicago, Coffey Hall 440, 1032 W. Sheridan Rd. Chicago, IL 60660, U.S.A. Email: dgarbarski@luc.edu
[3] Steelcase, 901 44th Street SE, Grand Rapids, MI, 49508, U.S.A. Email: ianfwall@gmail.com
[4] University of Wisconsin-Madison, 2176 Medical Science Center, 1300 University Avenue Madison, WI 53706, U.S.A. Email: dfedwards@education.wisc.edu

Acknowledgments: This study was funded by NIMHD (National Institute on Minority Health and Health Disparities) grant P60MD003428 (PD: A. Adams). Project: Increasing Participation of Underrepresented Minorities in Biomarker Research (PI: D. Farrar Edwards). Additional support was provided by the University of Wisconsin Survey Center (UWSC), which receives support from the College of Letters and Science at the University of Wisconsin-Madison, and the facilities of the Social Science Computing Cooperative and the Center for Demography and Ecology (NICHD core grant P2C HD047873). The authors thank: Tara Piché, Kenneth D. Croes, and Nathan Jones for their contributions in the development and analysis of the cognitive interviews; Nadia Assad, Jesus Renteria, and Ray Garza for research assistance; and Steven Blixt, Nora Cate Schaeffer, and John Stevenson for their comments on earlier drafts. Opinions expressed here are those of the authors and do not necessarily reflect those of the sponsors or related organizations.

1. Introduction

Questions that measure subjective constructs or evaluations historically have used an agree-disagree (AD) response format that presents respondents with a statement and asks them to indicate whether they agree or disagree with the statement or to rate their level of agreement. For example, the following question, administered for decades in the General Social Survey (GSS) is part of a scale designed to measure political efficacy: "The average citizen has considerable influence on politics. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?" (Smith et al. 2013).

While researchers have advocated for the positive psychometric properties of AD questions (see Willits et al. 2016), the ubiquity of these items primarily stems from their ease of use. Scales comprised of AD items are "easy to write" and efficient to administer; the same response categories can be used for each statement included in a battery of questions regardless of the content or complexity of the statement (Krosnick and Presser 2010). However, these positive features may be offset by increased burden for respondents and interviewers, which may ultimately lead to reductions in data quality. For example, AD questions may be more subject to response effects like acquiescence (the tendency to agree) or extreme responding (the tendency to select the lowest or highest response categories) (Krosnick and Presser 2010; Liu et al. 2015).

In recent writing, questionnaire designers eschew AD formats and advocate for construct-specific (CS) response formats (Fowler and Cosenza 2009; Krosnick and Presser 2010; Saris et al. 2010). Instead of asking participants to rate their level of agreement, CS questions directly ask about the item's underlying response dimension and provide construct-specific response categories. For example, a CS version of the GSS political efficacy question would be: "How much influence does the average citizen have on politics: none, a little, some, quite a bit, or a great deal?" The direct method of questioning offered by the CS format is argued to yield more reliable and valid data because it is less cognitively burdensome, less likely to be misinterpreted, and less likely to be associated with response effects.

In the current study, we use a mixed methods approach to evaluate the measurement properties of questions about trust in medical researchers using AD or CS questions. Trust is a central concept in the social and medical sciences because of its effect on decision-making and association with behavior. Trust is also a key component in social exchange theory, which posits that individuals are more likely to respond positively to a request to participate in research when they trust the originator of the request and perceive the ratio of rewards to costs to be personally acceptable (Dillman et al. 2014). Many have suggested that challenges recruiting and retaining research participants from underrepresented groups, such as racial and ethnic minorities, is rooted in a general distrust of medical researchers (Corbie-Smith et al. 2002; Scharff et al. 2010). Indeed, individuals with lower levels of trust indicate being less willing to participate in a future research study (Hall et al. 2006; Mainous et al. 2006; Braunstein et al. 2008). To better understand the public's trust in medical researchers, researchers need to measure the construct with sufficient reliability and validity. However, most scales use AD questions (e.g., Hall et al. 2006), which may lower data quality. Thus, we sought to improve on the measurement of trust in medical researchers by using CS questions.

1.1. Cognitive Processing of AD and CS Questions

Tourangeau et al. (2000) discuss four stages that a respondent progresses through in constructing an answer to a survey question, including comprehension of the question, retrieval of relevant information from memory to answer the question, use of retrieved information to make judgments, and selection and reporting of an answer. Researchers have expanded on this model to describe the unique cognitive steps required to answer an AD question (see Carpenter and Just 1975; Fowler and Cosenza 2009; Saris et al. 2010; Höhne and Lenzner 2018; Dykema et al. 2019).

Consider the response process embarked on by a respondent answering the AD question: "Medical researchers work very hard to make sure the participants in their studies are safe. Do you strongly agree, agree, neither agree nor disagree, disagree, strongly disagree." To answer this question, the respondent must first comprehend the literal and pragmatic meaning of the statement "Medical researchers work very hard to make sure the participants in their studies are safe" (Comprehension). Next, the respondent has to identify the question's underlying response dimension (Identification), which is the intensity of "working hard," (i.e., how hard medical researchers work to ensure the safety of research participants). Identification is accomplished by understanding the meaning of the statement as well as attending to any threshold words (e.g., "very") in the statement (Saris et al. 2010). Threshold words are those often included in AD questions that establish a threshold without presenting the full range of scale options. These include intensifiers (e.g., "extremely"), frequency markers (e.g., "rarely"), and quantifiers (e.g., "most"). After they identify the underlying response dimension, respondents must generate their own response or internal value to this response dimension (Generation). Here, our fictional respondent generates an internal value of "pretty hard" to the response dimension "how hard medical researchers work" and places that internal value on the underlying response dimension (Placement). The ensuing steps encompass a set of complicated cognitive processes in which the respondent evaluates the distance between their internal value of "pretty hard" and the threshold value of "very hard" (Threshold evaluation), and then assesses whether the distance between their internal value and the threshold value indicates "agreement," "disagreement," or "neutrality" (Polarity evaluation). Finally, guided by their evaluations of thresholds and polarity, the respondent must map their internal value onto one of the discrete categories offered in the "agreement" or "disagreement" range or select the midpoint if offered (Mapping).

The cognitive processing steps undertaken by a respondent answering the same item formatted in a construct-specific manner – that is, "How hard do medical researchers work to make sure that the participants in their studies are safe: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?" – is greatly simplified and predicted to be less burdensome. As with the AD version, the respondent must first comprehend the question (Comprehension). Next, they determine the underlying response dimension (Identification), which is reinforced by both the wording and ordering of the response categories (e.g., "not at all hard," "a little hard," etc.). Similar to processing with the AD question, the respondent generates an internal value of "pretty hard" (Generation), but placement of the internal value is done directly by mapping the internal value onto one of

the discrete categories offered (Mapping), thereby circumventing the steps of Placement, Threshold evaluation, and Polarity evaluation.

Of course, the model for processing AD questions assumes respondents are optimally engaged with the task of responding and attentively progress through the steps. Höhne and colleagues (Höhne and Krebs 2018; Höhne and Lenzner 2018; Höhne et al. 2017) argue this may not be the case. Because AD questions are usually presented in multi-item batteries in which the wording of the statements vary but the response categories remain the same – always some form of agreement to disagreement – they encourage superficial processing. In contrast, when multiple CS questions are grouped together, they will likely use different construct-specific response categories, encouraging deeper processing and motivating effort. In support of this proposition, researchers demonstrated that respondents in an eye-tracking study attended to CS response categories more when they varied from question to question (Höhne and Lenzner 2018), but there were no differences in processing times between AD and CS questions when the questions were presented in grids in which the response categories did not vary for either question format (Höhne et al. 2017).

1.2. Experimental Evidence Comparing AD and CS Questions

Despite strong recommendations among questionnaire designers to use CS questions in lieu of AD questions, only a handful of experimental studies demonstrate CS questions yield higher data quality (Lelkes and Weiss 2015). The most compelling evidence is provided by Saris et al. (2010), who compared AD and CS response formats for items using split-ballot multitrait-multimethod (MTMM) designs. The experiments were conducted in face-to-face interviews that used show cards or self-administered questions across multiple countries in the European Social Survey (ESS). Overall, CS questions yielded higher estimates of reliability and validity. These findings were replicated by Revilla and Ochoa (2015), who also used split-ballot MTMM experiments and found much lower quality for AD than CS questions in data collected in Mexico and Columbia.

Dykema et al. (2012) assessed the measurement properties of items about political efficacy from the General Social Survey. Findings indicated a trend such that CS items were associated with higher internal consistency reliability (see also Hanson (2015) who reported higher test-retest reliability for CS items). This study also examined whether behaviors produced by interviewers and respondents varied by the format of the question, focusing on behaviors that were associated with lower data quality in prior research (e.g., interviewers misreading questions and respondents qualifying responses or saying "don't know") (Dijkstra and Ongena 2006; Dykema et al. 1997; Schaeffer and Dykema 2011). The authors found that AD items yielded more instances of interviewers misreading questions and more disfluency tokens (e.g., "um"). Kuru and Pasek (2016) used confirmatory factor analyses (CFA) and structural equation modeling within an experimental design about Facebook use and demonstrated a methodological bias due to AD response formats. Controlling for the method effect reduced reliability and validity estimates for the AD items, but not the CS items, indicating AD items may inflate reliability and regression estimates more than CS items. In validity tests, the criterion relationships yielded stronger relationships with the CS items. In addition, Höhne and

Krebs (2018) reported the AD response format was more susceptible to response scale direction effects than the CS format for internally-focused, self-administered questions about achievement and intrinsic job motivation, but not for externally-focused questions.

Other studies, however, have not reported greater data quality for CS questions. Lelkes and Weiss (2015) and Liu et al. (2015) both analyzed an experiment comparing AD and CS questions about political efficacy embedded in the American National Election Study. Lelkes and Weiss (2015) reported no differences in reliability or concurrent validity for the response formats, and neither format was more valid among those respondents susceptible to acquiescence. In addition, Liu et al. (2015) reported extreme response style was present for both AD and CS formats based on latent class factor analysis. Finally, in a web survey comparing questions presented stand-alone or in grids, Höhne et al. (2017) reported no differences between AD and CS questions for data quality as indicated by non-differentiation and dropping out of the survey before completion.

1.3. Limitations of Past Comparisons between AD and CS Items

Although many prior experimental studies provide evidence in support of CS items yielding more reliable and valid responses, most of the analyses are limited to a comparison of how AD and CS items differ with regard to the closed-ended responses they yield, leaving out potentially crucial information about the *response process* respondents undertake when answering questions. This information would be useful when designing and testing new questions. Dykema et al. (2012) began to look at the response process by examining interviewer and respondent behaviors. However, they were limited in their ability to provide clear insights about what characteristics of items may be most difficult or what aspects of the response process may cause cognitive difficulties because they could only examine behaviors produced during the process of answering standardized survey questions.

Comparing responses to AD and CS items and evaluating the response process with cognitive interviews in which respondents are asked to describe what they are thinking about while answering survey questions may prove fruitful for developing targeted approaches to improve data quality. In the current study, we evaluate both close-ended response tendencies as well as aspects of the response process using cognitive interviewing techniques. This approach allows us to incorporate quantitative and qualitative data in order to provide insight about *why* differences in response tendencies occur between AD and CS items.

1.4. Current Study

The goal of the current study is to provide an in-depth, mixed-methods analysis of a scale designed to measure the general public's trust in medical researchers. We randomized participants to either an AD or CS version of the scale, and conducted cognitive interviews that included follow-up questions designed to identify problems during the response process. This study is motivated by two main questions: (1) how do closed-ended survey responses differ between the versions of the scale, and (2) what aspects of the response process might explain differences in response tendencies? Based on the empirical research reviewed above, which suggests the CS response categories may be more demanding to

process, more likely to encourage deeper processing, and often are associated with higher data quality, we predict:

H_1: The CS scale will yield closed-ended responses with higher reliability than the AD scale.
H_2: The CS scale will yield closed-ended responses with higher validity than the AD scale.
H_3: The CS scale will be associated with greater recency effects than the AD scale.
H_4: Responses to CS questions will yield longer processing times than AD questions.
H_5: Responding to CS items will involve more instances of behavioral indicators of response difficulty (e.g., respondents providing uncodable responses).

We further explore the motivations for these hypotheses in Subsection 2.5 where we describe our measures and analytic strategy. Immediately following the analysis of data based on these quantitatively-focused predictions, we explore and leverage the qualitative data generated during cognitive testing to help illuminate why the AD and CS items behaved in predicted or unpredicted ways. The qualitative portion of the study is exploratory and not grounded in previous research, and we do not put forth hypotheses for these analyses.

2. Methods

2.1. Overview of the Cognitive Interviewing Phase of the Voices Heard Survey Development

We conducted cognitive interviews to evaluate questions for inclusion in the Voices Heard Survey, a telephone survey targeting members of minority groups underrepresented in biomedical research (Edwards 2015). The primary goal of the survey was to measure perceptions of the barriers and facilitators to participating in medical research studies that collect biomarkers (e.g., saliva and blood), and to document whether there were important differences among groups identified by their race and ethnicity. To develop questions for the survey, we first conducted key informant interviews to identify major themes around which to write questions (e.g., mistrust of medical researchers, logistical constraints, fear of discomfort and pain). Next, we tested the questions in two rounds of cognitive interviewing.

2.2. Sample for the Cognitive Interviews

We conducted 64 cognitive interviews, 32 in two rounds, from 2012 to 2013 using a community-based, quota sampling strategy to recruit participants. Members of the project team recruited participants through connections with leaders in specific racial and ethnic communities, by visiting churches and community centers, by attending events sponsored by groups (e.g., pow-wows held by several American Indian tribes), and by posting flyers at targeted locations in communities. We confined recruiting to southern Wisconsin. The quota strategy yielded equal numbers of African American, American Indian, Latino, and white participants, distributed nearly uniformly by gender (male versus female), age (between 30–55 years versus 56 years or more), and education (high school or less versus

Table 1. Descriptive statistics of participant characteristics and participation measures for the agree-disagree (AD) and construct-specific (CS) experimental groups.

	AD		CS			
Panel A: Participant characteristics	Proportion or Mean (S.D.)	n	Proportion or Mean (S.D.)	n	Test	p-value
Age						
30–55 years	0.47	15	0.53	17	$\chi^2 = 0.25$	0.80
56 years or more	0.53	17	0.47	15		
Female	0.50	32	0.50	32	$\chi^2 = 0.00$	1.00
Race/Ethnicity						
African American	0.25	8	0.25	8	$\chi^2 = 0.00$	1.00
American Indian	0.25	8	0.25	8		
White	0.25	8	0.25	8		
Latino	0.25	8	0.25	8		
Education						
High school or less	0.47	15	0.50	16	$\chi^2 = 0.06$	1.00
Some college or more	0.53	17	0.50	16		
Panel B: Participation measures						
Participated in medical research in past	0.28	32	0.35	31	$\chi^2 = 0.39$	0.53
Expressed likelihood to participate in research involving						
Answering questions	3.34 (0.79)	32	3.09 (0.86)	32	t = 1.22	0.23
Providing saliva	3.10 (1.19)	31	2.78 (1.24)	32	t = 1.03	0.31
Providing blood	2.97 (1.11)	31	2.35 (1.43)	31	t = 1.89	0.06
Providing tissue	2.27 (1.28)	30	1.84 (1.46)	31	t = 1.13	0.23
Providing cerebrospinal fluid	1.10 (1.25)	31	0.91 (1.35)	32	t = 0.85	0.56

Note: p-values for Chi-squared tests are from Fisher's exact tests.

some college or more) (see Panel A, Table 1). Interviews were conducted at locations that were convenient for the participants, including homes, libraries, and offices. Participants were remunerated for their time and effort.

2.3. *Interviewing and Transcription*

Following a format commonly employed in cognitive interviewing (Willis 2005; Fortune-Greeley et al. 2009; Willis and Miller 2011), interviewers asked participants a question being tested for use in the survey interview, and then after participants provided their response to the closed-ended survey question, interviewers administered a series of structured, open-ended probes and follow-up questions. We designed the probes and follow-up questions to uncover how participants formulated their answers to the survey questions, to reveal any problems they had with comprehension of specific terms or retrieval of information from memory, and to document issues participants faced in mapping their responses onto the response categories.

Interviewers received a full day of training on cognitive interviewing tailored for the study, and they were required to complete a practice interview before being certified.

We matched interviewers and participants on race/ethnicity and, for all cases except one, on gender. While interviews were primarily conducted in English, seven participants (distributed nearly evenly between the AD and CS conditions) elected to be interviewed in Spanish by a Spanish-speaking interviewer. On average, interviews took approximately an hour to complete (M = 61.10 minutes, SD = 20.17). Interviews were audio-recorded and digital files were created. In order to facilitate coding and analysis, interviews were transcribed verbatim on a question-by-question basis into Excel.

2.4. *Trust/Mistrust Scale Development and Experimental Design*

We examine respondents' answers and their response processes during administration of an 11-item scale measuring trust in medical researchers (see Appendix A, Subsection 6.1), for the exact wording of the items, which varied slightly between the two rounds of cognitive interviewing). We randomly assigned participants to the AD or CS scale using a between-subjects design. Items in a given scale appeared in the same sequence (i.e., they were not randomized), roughly 30 minutes into the interview.

To develop the scale, we conducted a literature review that identified approximately 100 questions from 12 studies about trust in medical care providers and researchers (Anderson and Dedrick 1990; Hayman et al. 2001; Corbie-Smith et al. 2002; Hall et al. 2002a; Hall et al. 2002b; Zheng et al. 2002; Thompson et al. 2004; Hall et al. 2006; Mainous et al. 2006; Egede and Ellis 2008; Henderson et al. 2008; Williams et al. 2010). The majority of the questions used AD response formats. From the pool of candidate questions, we modified 11 for the AD scale. We generated items for the CS scale by rewriting the AD version of the question to ask about the underlying response dimension implied by the question. For example, the AD item about informed consent ("hide information") asked, "*Medical researchers never hide information about the possible risks of participating. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?"* To translate to the CS format, we used the threshold word "never" to identify "frequency" as the underlying dimension: "*How often do medical researchers hide information about the possible risks of participating: never, rarely, sometimes, very often, or extremely often?*" Thus, the CS item directly asked participants for their evaluation of the relative frequency with which medical researchers' hide information, rather than having participants rate their level of agreement with a statement about medical researchers "never" hiding information.

AD items used the same five response categories for each item ("strongly agree" to "strongly disagree") but the CS items had response categories that varied depending on the underlying dimension (e.g., "never" to "extremely often" for frequency). The final scale was balanced and included a roughly equal number of positively and negatively valenced items. Positively valenced items are those in which a higher valued response category (e.g., "strongly agree" for AD questions and "a great deal" for CS) indicated most trust; negatively valenced items are those in which a higher valued response category indicated least trust. Thus, when the item is positively valenced, the direction of the response scale is ordered from most to least trust for the AD questions and least to most trust for the CS questions; when the item is negatively valenced, the scale is ordered from least to most trust for AD questions and most to least trust for CS questions.

2.5. Measures and Analytic Strategy

We present three sets of analysis using a mixed-methods approach (Johnson and Onwuegbuzie 2004): (1) measures of the quality of survey responses (e.g., reliability, validity, and recency); (2) response latencies and behavioral indicators of response difficulty; and (3) sources of response difficulties captured during the cognitive interviewing response process.

2.5.1. Trust Scale Summary Statistics and Reliability Measures

We examine trust scale summary statistics, including item nonresponse, mean trust scale scores, and reliability estimates. We assess the effect of item-missing data and other measures described below by estimating aggregate-level regression models that evaluate the items in a scale collectively by treating each question answered by the participant as a separate observation. Models estimate robust standard errors to correct for the fact that individual observations are independent across participants but dependent within a given participant (Rogers 1994). For item-missing values, estimates are from a logistic regression model with response format coded "1" for CS, "0" for AD. We score trust items from 0 to 4, with lower scores indicating less trust in medical researchers (e.g., depending on whether the item is positively or negatively valenced, "strongly disagree" may be coded as 0 or 4; see Appendix A, Subsection 6.1) and compute scale values by summing across the items. We impute cases with missing values with the median value for the non-missing cases on an item-by-item basis, separately for the AD and CS response formats (Hall et al. 2006). We evaluate internal consistency reliability using Cronbach's alpha (Streiner et al. 2015), and test for significance by treating the alpha coefficients as correlations and applying Fisher's r-to-z transformation (Tourangeau et al. 2004).

2.5.2. Concurrent Validity

Past research demonstrates a strong association between the public's trust in medical researchers and their actual or expressed likelihood of participating in a medical research study (Hall et al. 2006; Mainous et al. 2006; Braunstein et al. 2008). We assess concurrent validity by examining whether the relationship between trust and participation is stronger for the AD or CS response format. Questions assess whether participants ever participated in medical research (coded "1" if "yes;" "0" if "no") and their expressed likelihood of participating in medical research studies involving answering questions or providing samples of saliva, blood, tissue, and cerebrospinal fluid.

Because we were testing these questions for inclusion in a larger survey, their wording and response categories varied between rounds of interviewing, particularly for the expressed likelihood of participating measures. For example, Round 1 used the response categories "not at all likely, a little likely, somewhat likely, pretty likely, and very likely," while Round 2 used the response categories "very likely, somewhat likely, neither likely nor unlikely, somewhat unlikely, and very unlikely." We conducted a series of exploratory analyses to determine whether responses could be combined across rounds. First, treating the measures as continuous, we converted the raw scores into z-scores. Next, we tested for measurement invariance of the items using correlations between the standardized scores

(tested using the *sem* command in Stata 14). Results (not shown) indicated that the response set from each round performed as a parallel measure, supporting the use of combining them across rounds. In analysis, we present values for the expressed likelihood to participate measures by scoring the items 0 to 4 for least to most likely to participate.

We regress each of the participation measures on trust scores separately for the AD and CS experimental groups and then for a model that includes the trust score, response format (coded "1" if CS; "0" if AD), and the interaction between these. We use logistic regression when the dependent variable is dichotomous (past participation) and ordinary least squares (OLS) regression with the continuous (expressed likelihood to participate) dependent variables.

2.5.3. Recency Effects

A recency effect refers to the tendency for respondents to be more likely to select a response category when it appears later in the list regardless of their true answer (Krosnick and Presser 2010). Recency effects are more likely when questions are presented orally (i.e., by an interviewer). According to satisficing theory, recency effects are also more likely when questions are more cognitively demanding: respondents will be more likely to select the last category if it seems reasonable (Holbrook 2008). Because the response categories vary from question to question for the CS questions, we predicted they would be more demanding to process and recall, and respondents would be more likely to select the final response category. However, we expected this effect would be more pronounced for the positively valenced items for which the final category for the CS questions indicates a higher level of trust and may be perceived as a more "reasonable" or agreeable answer.

We assess recency by examining whether the proportion of responses selecting the final category in the list is higher by response format for each question and aggregating across the positively and negatively valenced items with aggregate-level logistic regression models.

2.5.4. Response Latencies

Response latencies (RLs) capture the length of time participants spend processing while they are formulating answers to survey questions (Bassili and Scott 1996; Draisma and Dijkstra 2004). We predicted longer latencies for the CS questions because they use variable construct-specific response categories, encouraging deeper processing (Höhne and Lenzner 2018; Höhne et al. 2017).

Coders timed RLs using audio recordings of the interviews and the visual waveform functionality in the audio software Audacity (Audacity Developer Team 2008). Audacity provides a visual representation of the sound wave, on which coders were able to highlight sections of audio precisely, timing RLs to the thousandths of seconds. Coders began timing after interviewers read the last word of the question during their initial reading of the question. Timing continued through all utterances, including any subsequent readings of the question, and ended when participants uttered the first sound of a word that unambiguously answered the question (e.g., by providing a response category offered by the question). We code interruptions (where a codable response was offered before the

entire question was read) and final dispositions that do not accompany a codable response (i.e., don't knows, refusals, and uncodable responses) as missing RLs.

Because they generally have a skewed distribution and outliers, we followed recommendations and top- and bottom-coded values at the 95th and 5th percentiles within each item and use logged values (Yan and Tourangeau 2008). We top- and bottom-coded at the 95th and 5th percentiles and the not the 99th and 1st percentiles because our sample size is small and there are no observations at the 99th and 1st percentiles. ICC between raters using the transformed data is 0.89, which is considered excellent reliability (Landis and Koch 1977). We examine differences for individual questions using t-tests and aggregated across all of the items using OLS regression and aggregate-level tests.

2.5.5. Behavioral Indicators of Response Difficulty (BIRDs)

We predicted the CS questions would be associated with higher levels of BIRDs. As noted, we expected the varying construct-specific response categories to encourage a more elaborated cognitive processing of the questions and the AD questions to encourage a more superficial processing of the questions. Within the context of the cognitive interview, the behavioral indicators of response difficulty offer evidence of a more elaborated processing.

For each question-response interaction (one per participant for each of the trust questions), we tallied the occurrence or non-occurrence of behavioral indicators of response processing difficulties among participants (described below). These indicators are not necessarily final dispositions: a participant may initially say they do not know how to respond to a question, which would be coded as an occurrence of "don't know/refusal," but the interviewer may repeat the question and obtain a codable response. One team member coded all interactions and another member independently coded two thirds of the interactions. All behaviors yielded kappa values with good to excellent agreement (Fleiss 1981).

- *Codable response with qualification* (kappa = 0.83). Participant provides a codable answer (one of the response categories), but qualifies it by adding "probably," "I guess," "maybe," "depends," etc.
- *Codable response with elaboration* (kappa = .87). Participant provides a codable answer, but also provides additional information during their initial response. If the additional information contradicts the codable answer, the response is coded as an "uncodable response."
- *Uncodable response* (kappa = 0.81). Answer does not answer the question or cannot be coded into the response categories (e.g., unrelated report or ambiguous answer).
- *Seeks clarification* (kappa = 0.96). Participant asks for clarification of all or part of a survey question, asks that all or part of a question or response categories be repeated, or repeats part of the question in a way that sounds like a question. These are coded whether or not a codable response is part of the utterance.
- *Question repeated* (kappa = .70). Coded any time the full question and/or response categories are read more than one time before a final disposition (codable response or otherwise) is achieved. This code supplements "seeks clarification," because interviewers sometimes repeat questions without a formal request by the participant.

- *Don't know/refusal* (kappa = 0.83). Instead of or in addition to providing an answer, the participant says "don't know" (or the equivalent) and/or refuses to answer the question.

We examine differences between the AD and CS response format for each behavioral indicator aggregated across questions within a scale using logistic regression and aggregate-level tests.

2.5.6. Cognitive Interviewing Data

Lastly, we incorporate qualitative data from the cognitive interviews to explore differences between response processes resulting from AD and CS items. Answers to questions and probes were analyzed qualitatively. The goal was to identify problems that arose, potentially involving misunderstandings of terms, interpreting the intent of the question in different ways, and issues mapping responses onto the categories provided (Tourangeau et al. 2000; Willis 2005). Preliminary codes were based on potential sources of problems during the response process and more specific codes arose during preliminary assessment of the transcripts (Ryan and Bernard 2003). Once the coding scheme was finalized, each interview was coded.

3. Results

3.1. Baseline Comparison of Experimental Groups

As anticipated, the randomization of participants to experimental groups was effective: there are no significant differences ($p < .05$) between the experimental groups based on participants' characteristics (Table 1, Panel A), or by the participation measures used in the validity analysis (Table 1, Panel B), although participants in the CS group reported slightly higher levels of expressed likelihood of providing blood ($p = .06$).

3.2. Trust Scale Summary Statistics and Reliability Measures

Panel A in Table 2 presents summary statistics and reliability coefficients for the scales. Aggregating across 704 question administrations (64 participants × eleven questions), we find the CS scale is associated with significantly higher levels of missing data than the AD scale. Mean trust scores, however, do not significantly differ between the response formats, regardless of whether we impute for missing values. Contrary to expectations, the alpha coefficient, a measure of internal consistency reliability, is significantly higher for the AD than the CS scale.

3.3. Concurrent Validity

We predicted a positive association between trust and participation. Results are in the expected direction for four of the participation measures for the AD scale and five of the participation measures for the CS scale (Table 3). These bivariate associations, however, are only significant for providing tissue and cerebrospinal fluid for the AD scale. Further, the interaction term is only significant for providing tissue: participants' level of expressed

Table 2. Proportion and mean level of trust scale summary statistics, reliability estimates, recency measures, response latencies, and behavioral indicators of response difficulty by AD and CS response formats.

	AD		CS				
Data quality outcomes	Proportion or mean (S.D.)	n	Proportion or mean (S.D.)	n	Difference	Test	p-value
Panel A: Trust scale summary statistics and reliability							
Item-missing aggregate-level	0.03	352	0.10	352	−0.07	b = 1.23; s.e. = 0.44	0.01
Scale score							
List-wise deletion of missing	24.66 (7.31)	32	24.84 (7.35)	32	−0.19	t = −0.10	0.92
Missing imputed	27.53 (4.93)	32	25.63 (7.27)	32	1.91	t = 1.23	0.22
Cronbach's alpha							
List-wise deletion of missing	0.85	32	0.60	32	−0.26	z = 2.09	0.02
Missing imputed	0.84	32	0.58	32	−0.26	z = 2.13	0.03
Panel B: Recency							
Positive valence							
General trust	0.03	31	0.10	31	−0.07	$\chi^2 = 1.07$	0.30
Participants' interest	0.07	30	0.21	29	−0.14	$\chi^2 = 2.47$	0.15
Participants' safety	0.03	32	0.38	26	−0.35	$\chi^2 = 11.65$	<0.01
Tell about risks	0.03	30	0.37	30	−0.34	$\chi^2 = 10.42$	<0.01
Treat fairly	0.03	31	0.36	28	−0.33	$\chi^2 = 10.24$	<0.01
Protect privacy	0.03	32	0.50	30	−0.47	$\chi^2 = 17.77$	<0.00
Aggregate-level results	0.04	186	0.32	174	−0.28	b = 2.47; s.e. = 0.92	0.01
Negative valence							
Researchers' interest	0.19	31	0.45	29	−0.26	$\chi^2 = 4.49$	0.05
Select minorities	0.17	29	0.13	32	0.04	$\chi^2 = 0.27$	0.72
Hide information	0.06	31	0.04	26	0.02	$\chi^2 = 0.19$	1.00
Treat like guinea pig	0.16	32	0.11	28	0.05	$\chi^2 = 0.31$	0.71
Know more	0.09	32	0.29	28	−0.20	$\chi^2 = 3.38$	0.09
Aggregate-level results	0.14	155	0.20	143	−0.07	b = 0.48; s.e. = 0.41	0.24

Table 2. Continued.

Data quality outcomes	AD Proportion or mean (S.D.)	AD n	CS Proportion or mean (S.D.)	CS n	Difference	Test	p-value
Panel C: Response latency							
General trust	1.29 (1.43)	31	0.64 (1.16)	31	0.65	t = 1.97	0.05
Participants' interests	1.15 (1.36)	31	1.46 (1.73)	29	−0.31	t = −0.77	0.45
Participants' safety	0.82 (1.60)	31	1.21 (1.28)	27	−0.38	t = −1.00	0.32
Tell about risks	0.60 (1.53)	29	1.14 (1.29)	30	−0.54	t = −1.47	0.15
Treat fairly	0.90 (1.49)	31	1.26 (2.05)	27	−0.35	t = −0.76	0.45
Protect privacy	0.43 (1.71)	31	1.33 (1.80)	30	−0.90	t = −2.01	0.05
Researchers' interest	1.44 (1.47)	32	2.07 (1.42)	29	−0.63	t = −1.70	0.09
Select minorities	1.07 (1.46)	26	2.09 (1.60)	32	−1.03	t = −2.52	0.01
Hide information	1.02 (1.48)	30	1.24 (1.65)	25	−0.22	t = −0.52	0.61
Treat like guinea pig	1.46 (1.60)	31	1.59 (1.71)	28	−0.13	t = −0.31	0.76
Know more	1.20 (1.54)	32	0.99 (1.60)	29	0.21	t = 0.51	0.61
Aggregate-level results	1.04 (1.53)	335	1.37 (1.62)	317	−0.33	b = 0.33; s.e. = 0.19	0.09
Panel D: BIRDs							
Aggregate-level results							
Codable + qualification	0.08	352	0.17	352	−0.09	b = 0.89; s.e. = 0.35	0.01
Codable + elaboration	0.24	352	0.12	352	0.12	b = −0.85; s.e. = 0.32	0.01
Uncodable	0.07	352	0.13	352	−0.06	b = 0.74; s.e. = 0.33	0.02
Seeks clarification	0.20	352	0.20	352	0.00	b = 0.02; s.e. = 0.22	0.94
Question repeated	0.19	352	0.24	352	−0.05	b = 0.30; s.e. = 0.22	0.17
"Don't know" or refusal	0.07	352	0.11	352	−0.04	b = 0.46; s.e. = 0.35	0.19

Notes: Aggregate-level tests assess the effect of the measure (e.g., item-missing responses) evaluated collectively across questions treating each question answered by the participant as a separate observation. p-values for Chi-squared tests are from Fisher's exact tests.

Table 3. Concurrent validity analysis: Regression results using trust scores to predict participation for various types of medical research.

	AD only		CS only		AD and CS	
	b	(S.E.)	b	(S.E.)	b	(S.E.)
Past participation						
Trust score	−0.050	(0.056)	0.014	(0.077)	−0.050	(0.056)
Response format					−1.307	(2.606)
Trust score x response format					0.064	(0.095)
Answering questions						
Trust score	−0.008	(0.024)	0.022	(0.037)	−0.008	(0.025)
Response format					−1.114	(1.214)
Trust score x response format					0.030	(0.044)
Providing saliva						
Trust score	0.013	(0.025)	0.013	(0.037)	0.013	(0.025)
Response format					−0.237	(1.223)
Trust score x response format					−0.000	(0.044)
Providing blood						
Trust score	0.024	(0.022)	0.032	(0.038)	0.024	(0.024)
Response format					−0.673	(1.190)
Trust score x response format					0.008	(0.043)
Providing tissue						
Trust score	0.069**	(0.021)	−0.046	(0.036)	0.069**	(0.023)
Response format					2.806*	(1.128)
Trust score x response format					−0.115**	(0.041)
Providing cerebrospinal fluid						
Trust score	0.065**	(0.021)	0.023	(0.038)	0.065**	(0.024)
Response format					0.894	(1.155)
Trust score x response format					−0.042	(0.042)

Notes: Regression coefficients are from logistic regression for past participation and OLS regression for answering questions and providing saliva, blood, tissue, and cerebrospinal fluid. Trust scale scores are computed by summarizing the z-scores across questions within experimental group.
* $p < .05$; ** $p < .01$

likelihood to participate in research by providing tissue is significantly lower with the CS scale. Overall, the scales appear equally valid in predicting participation.

3.4. Recency

For positively valenced items, we predicted the proportion of responses using the last category for the CS scale (which varied by question but for which the last category indicates more trust) would be higher than the proportion of responses using the last categories for the AD scale (which is always "strongly disagree" such that the last value indicates less trust). Indeed, the CS scale yields more responses using the last category (Panel B, Table 2)

for all of the positively valenced items, and the difference is significant for four of the items and for the aggregate-level test. In contrast, for the negatively valenced items, the difference is only significant for one item and the aggregate-level test is not significant.

3.5. *Response Latencies (RLs)*

We predicted RLs would be longer for the CS scale because the changing construct-specific response categories encourage deeper processing. Overall, nine of the eleven CS items have longer mean RLs than the parallel AD item (Panel C, Table 2); two of these ("protect privacy" and "select minorities") are statistically significant and the aggregate-level test is marginally significant ($p < .09$).

Interestingly, RLs are significantly longer for the AD scale for the first question administered as part of the scale ("general trust"). Here respondents are hearing the question and response categories read for the first time, and the longer response time could be evidence for the more cognitively burdensome response task offered by the AD response format and distinct from the effect of grouping AD questions in a battery.

3.6. *Behavioral Indicators of Response Difficulty (BIRDs)*

We predicted the CS questions would be associated with higher levels of BIRDs. Panel D in Table 2 presents aggregate-level logistic regression tests for the BIRDs indicators (see Appendix B, Subsection 6.2, for question-by-question results). We find significantly higher levels of codable answers with qualifications and uncodable answers for the CS questions versus higher levels of codable answers with elaborations for the AD questions; there are no differences between response formats for seeking clarification, asking to have a question repeated, or providing a don't know or refusal response. These results help interpret the longer response latencies found with the CS items. In almost all cases, indicators of response difficulty, such as providing a qualification or uncodable response require more interactional time to resolve and result in longer response latencies. In contrast, elaborations tend to follow codable answers and so would not be part of the latency.

3.7. *Cognitive Interviewing Data*

Quantitative analyses demonstrate the AD and CS scales differ on several of the data quality indicators examined, but these analyses do not provide insight about *why* this is so. A strength of the current study is that we embedded the AD-CS comparison in cognitive interviews in which participants first answered the survey questions that comprised the quantitative analysis, and then answered open-ended follow-up questions about their answers. We incorporate participants' qualitative responses to further examine *why* the response formats produced different results. The qualitative analysis revealed five potential sources of difficulties:

1. Understanding the intent of questions: interpreting questions about opinions as knowledge questions,
2. Understanding the intent of questions: managing comparisons between target objects in a question,

3. Difficulty mapping: dealing with a lack of knowledge or ambivalence,
4. Difficulty mapping: remembering CS categories, and
5. Difficulty mapping: mismatched vocabulary.

We assess how these difficulties with the response task varied depending on whether the participant received the AD or CS scale, and we provide excerpts from the cognitive interviews to illustrate how these potential sources of error manifested.

3.7.1. Understanding the Intent of Questions: Interpreting Questions About Opinions as Knowledge Questions

Trust is a subjective evaluation that may or may not depend on facts about events and behaviors related to the trustee, in this case medical researchers (Hall et al. 2001). Several participants said they did not have enough information to answer the trust questions, indicating they interpreted questions as asking about their knowledge rather than for their evaluation (Excerpt 1, Table 4). The intent of the question about "participants' interests" was to gauge each participants' **attitudes** about the relative frequency with which medical researchers have the best interests of participants from their racial/ethnic group in mind. The intent was not to measure objectively how often medical researchers actually engage in the behavior. This participant's responses, however, indicated she felt we were asking her to "project [her] thoughts into another person," While we observed this issue for both AD and CS formats, the higher proportion of administrations yielding "don't knows" for the CS scale (11%, compared to 7% for the AD scale) indicates this may have been more of a problem for CS questions.

3.7.2. Difficulty Interpreting Intent of Questions: Managing Comparisons between Target Objects in a Question

During follow-up questioning, we documented several participants unintentionally reversing the direction of a comparison, providing a codable answer incongruent with their reasoning. The question about "researchers' interests" asked participants to evaluate how much they believe medical researchers care about their research **compared** to the participants in their studies. In Excerpt 2, the interviewer realizes the participant's reasoning did not match her initial response of "a great deal." While the interviewer catches the incongruence, it is not possible to tally how often this type of mismatch occurred and went unnoticed by the participant or the interviewer, especially if the participant provided vague reasoning for their answer. Not only did respondents flip the direction of the comparison so that their answers were about researchers caring more about their participants than their research, but they needed more time to provide a codable answer. The CS version of the "researchers' interests" question had the fourth longest response latency, suggesting comparisons using CS may be particularly burdensome.

3.7.3. Difficulty Mapping: Dealing With a Lack of Knowledge or Ambivalence

Participants reported feeling uninformed or ambivalent about matters related to medical researchers. They dealt with this ambivalence differently depending on whether they were attempting to map responses onto AD or CS categories. With AD items, when a participant expressed that they needed more knowledge on the topic in order to answer the question,

Table 4. Excerpts from the Cognitive Interviews.

Actor	Text
Excerpt 1:	**"Participants' interests," CS Response Format**
Interviewer	When they are conducting research, how often do medical researchers have the best interests of participants from your racial or ethnic group in mind: never, rarely, sometimes, very often, or always?
Participant	There would be no way for me to know the answer to that question.
Interviewer	Tell me why.
Participant	You're, you're asking me to, um, project my thoughts into another person. I, I, I can't do that. I can only answer questions that directly involve me, I guess.
Excerpt 2:	**"Researchers' interests," CS Response Format**
Interviewer	To what extent do medical researchers care more about their research than they do about the participants in their studies: not at all, a little, somewhat, quite a bit, or a great deal?
Participant	A great deal.
Interviewer	And can you tell me more about why you answered a great deal?
Participant	Because you need the participants to even figure out what's going on in the study that they're performing. So I would think they would care a lot, just as much as they do for the study.
Interviewer	Okay. Um, this question is a little confusing, so I'm going to reread it to you. Um, to what extent do medical researchers care more about their research than they do about the participants in their studies? And you said a great deal. Um, but then you were.
Participant	Oh, you were saying do they care more about the research than they do the peoples that are participating in it.
Interviewer	Yeah. I believe that's what the question is asking.
Participant	No, I don't think.
Interviewer	Okay. So would you say not at all, a little, somewhat, quite a bit?
Participant	Not at all.
Interviewer	Okay. And you said that, that you, I'm sorry. Could
Participant	I said that because, um, they need the participants to figure this stuff out for the study. So I would think they would, it would be equal, even.
Excerpt 3:	**"Participants' interests," AD Response Format**
Participant	Well, see, that, that, I'm going to either, you know, agree and disagree at the same time or whatever, neither disagree or disagree, because, again, I, I'm ignorant. I don't have that much knowledge on, uh, people that do these kinds of things, these studies and everything. So I don't know if they, they're more interested in a certain ethnic group or a certain category or age or whatever. I, I have no knowledge on, uh, why a person does medical interviews, you know, so I, I don't agree or disagree. I have no, no knowledge.

Excerpt 4:	**"Treat like a guinea pig," CS Response Format**
Interviewer	How often do medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies: never, rarely, sometimes, very often, or extremely often?
Participant	Well, let me put never, because, in reality, I don't know.
Interviewer	Okay. And the question following up is tell me more about why you answered never for this question.
Participant	It's because I don't know. I'm not informed in that aspect.
Excerpt 5:	**"Know more," CS Response Format**
Interviewer	How often do medical researchers want to know more than they need to know: never, rarely, sometimes, very often, or extremely often?
Participant	Hmm, the categories again are what?
Interviewer	Never, rarely, sometimes, very often, or extremely often?
Participant	And now I forgot the question [L].
Interviewer	How often do medical researchers want to know more than they need to know?
Participant	Okay. Um, rarely.
Excerpt 6:	**"Researchers' interests," CS Response Format**
Participant	Oh, I'm sorry, can, can you repeat the question one more time?
Interviewer	The question or the responses?
Participant	The question.
Interviewer	Okay. To what extent do medical researchers care more about their research than they do about the participants in their studies: not at all, a little, somewhat, quite a bit, or a great deal?
Participant	Well, for me, the, the answers you're giving me are, are difficult to use to w-, to make this response.
Interviewer	Okay. Can you tell me a little bit more about why it's difficult?
Participant	Sure. It, it's not the type of vocabulary that I use. I don't use quite a bit or a great deal a lot for any, for anything, so I don't have any, any, any f-, any meaning in anything that I say.
Interviewer	Okay. And if you had to choose one, what answer would you give me?
Participant	Well, the, the questions, the answers you gave me before, those were easier to use.

they were often able to provide a codable answer by responding with "neither agree nor disagree." Because they are bipolar, AD items include a middle category that appeared to be a reasonable option for participants who did not feel they had enough knowledge or who were not inclined to answer one way or the other (Excerpt 3). In contrast, the CS items are unipolar and lack a clear "neutral" (middle) category. In order for an uncertain or ambivalent participant to conclude the interaction, she could either pick a category that was not an exact match to her "true" state or provide an uncodable response (e.g., "don't know"). There is some evidence that at least two of these behaviors occurred during responses to CS items; the participant in Excerpt 4 does both.

3.7.4. Difficulty Mapping: Remembering CS Categories

One clear difference between the AD and CS items is the format of the response categories: AD questions use the same response categories for each item while the CS categories vary by question. A potential source of difficulty is that the CS categories were harder for participants to remember, an issue exacerbated by the aural presentation of the items and the number of items in the scale (the scale included 11 items). In Excerpt 5, the participant requested to hear the response categories a second time, but by the time the participant had a handle on the categories, they had forgotten the content of the question.

3.7.5. Difficulty Mapping: Mismatched Vocabulary

In a few cases participants reported the CS categories did not use their common language. In Excerpt 6, the participant indicated difficulty using the categories and elaborated that particular phrases like "quite a bit" or "a great deal" are not in his usual vocabulary, so he does not have "meaning in anything" he says. This participant was Latino and the interview was conducted in Spanish, which could have further complicated category interpretability. If participants are uncomfortable or unfamiliar with vocabulary and interpret and use the categories differently, reliability may be lower. This problem may be particularly relevant in cross-cultural research and research with diverse samples, such as this study.

4. Discussion

Based on past research, we formulated several hypotheses about how the closed-ended survey responses would differ between the AD and CS versions of the scale. We expected the CS scale would yield higher reliability and validity than the AD scale. Results, however, indicated higher reliability for the AD scale and neither scale appeared more valid in predicting participation. While these results seem to favor the AD scale, AD responses may be more internally consistent, as indicated by the significantly higher value of coefficient alpha, because factors like acquiescence and the use of the same response categories increases common method variance and not because of the scale's ability to reliably measure the underlying construct. Unfortunately, the small sample sizes in our study precluded more sophisticated analyses, such as using structural equation models to account for potential method effects that may have biased estimates of reliability and validity, particularly for the AD scale (e.g., Kuru and Pasek 2016). In addition, we were limited in the availability of criterion measures for the validity analysis. We selected the past participation and expressed likelihood to participate measures because of their demonstrated relationship with trust in past research, but they are not ideal; their wording varied somewhat between rounds of interviewing, their response format was more similar to the CS than AD format, and they were not strongly associated with trust scores in this study.

In developing scales, experts recommend that they be balanced with half of the items measuring one direction of the construct (in our case high trust with the positively valenced items) and half measuring the other direction (in our case low trust with the negatively valenced items) (Streiner et al. 2015). Because the CS response categories vary from question to question, they are likely to be more demanding to process and recall (Höhne and Lenzner 2018), and we hypothesized that the CS scale would be associated

with greater recency effects than the AD scale, particularly for the positively valenced CS questions for which the final category indicated a higher level of trust and possibily a more "reasonable" or agreeable answer. Consistent with our expectations, we found a higher proportion of responses in the last category for the CS questions overall.

That participants often struggled to remember the wording of the variable CS categories was also observed in the qualitative data. Much of the past research comparing AD and CS questions has been conducted using self-administered questions or with visual aids. We did not provide showcards because the cognitive interviews were the first step in a study with the purpose of developing a telephone survey. In this unfamiliar context, participants may have been unable or unwilling to dedicate cognitive resources to remembering the variable CS categories, often necessitating a reread of CS items. Further, the trust scale was quite long – it included 11 items. Most previously tested comparisons of AD and CS questions involve many fewer items. When CS scales are long and contain many questions with variable response categories, they may be more problematic when presented aurally. It is possible that the variable nature of the categories coupled with the length of the scale contributed to the lower reliability we documented with the CS scale.

We also predicted that responses to CS questions would yield longer processing times than AD questions and that responding to CS items would involve more instances of behavioral indicators of response difficulty. Overall, results match our expectations. Aggregating across questions, the CS scale is associated with a marginally significant higher mean response latency and higher levels of qualifications and uncodable answers, behaviors that tend to increase response latencies. We often interpret longer response latencies and the presence of the behavioral indicators we measured as signs of cognitive response difficulty (e.g., Bassili and Scott 1996; Schaeffer and Dykema 2011). However, recent theorizing and evidence suggests these outcomes may be desirable when comparing AD and CS questions (Höhne and Lenzner 2018; Höhne et al. 2017). Because CS questions use construct-specific response categories, which will likely vary on a question-by-question basis in a battery, they encourage deeper processing, which will increase processing time and likely result in respondents producing other behaviors when attempting to respond in an optimal manner.

Being able to turn to qualitative data from our cognitive interviews added substantially to our understanding of these mixed quantitative results. They revealed several sources of difficulties for respondents that varied by the AD or CS questioning format. For example, the CS items may have confused the intent of the question. Even if AD items generally are more cognitively burdensome than CS items, the response dimension "agreement" is a reminder to respondents that the question seeks their evaluation about the topic. For CS items, depending on the response dimension (e.g., intensity, frequency, or quantity), the intent of the question can be less clear. For example, several items seemed to ask about knowledge or facts related to the target object (e.g., "how hard do medical researchers work to ensure participants in their studies are safe") rather than respondents' evaluations about the target object (e.g., "how confident are you that medical researchers work hard to ensure participants in their studies are safe"). They also focused on evaluations of external objects (e.g., "medical researchers") and to a lesser extent on internal or self-focused objects (e.g., "you"). Although others have reported higher validity for CS questions that

focus on evaluations of the characteristics or qualities of external objects such as Facebook or doctors (e.g., Kuru and Pasek 2015; Saris et al. 2010), we recommend future research explore whether AD or CS questions yield more desirable data quality outcomes for questions of the type examined here.

Experiments evaluating data quality for the inclusion of middle categories for bipolar questions has been mixed (see Krosnick and Presser 2010), and another important finding from the qualitative analysis was a description of participants' use of the middle "neither agree nor disagree" category to deal with a lack of knowledge and express ambivalence with the bipolar AD questions. In contrast, the unipolar CS response categories do not include a clear middle or "neutral" category. In this unfamiliar context, participants sought ways to express ambivalence, but struggled to do so with the CS response categories, often resulting in significantly higher level of item-nonresponse and uncodable answers. In contrast, participants who were asked AD items often selected "neither agree nor disagree" to express uncertainty or ambivalence. Our findings are consistent with those of Sturgis et al. (2014) who probed respondents selecting the "neither/nor" middle category during the administration of three attitudinal questions to determine why respondents selected that category. Overwhelmingly respondents reported selecting the middle category because they did not have an opinion on the issue. Further, this strategy was employed more often among respondents who indicated more interest in the topic under consideration, possibly as a way to "save face" and avoid having to say "don't know" outright. From a measurement perspective, respondents use of the "neither/nor" middle category is highly problematic: while respondents may reliably select this middle option, their response is not a valid measure of the construct being assessed. Researchers have noted problems with the interpretation of the middle category with AD questions and often suggest that responses using this category should be analyzed separately and not as the middle value between agree and disagree (Willits et al. 2016).

5. Conclusions

For survey methodologists, one important consideration is that AD and CS items seem to demand different levels of cognitive effort, which may vary depending on characteristics of the questions and the mode of administration including: (1) valence (whether the question is positively or negative valenced); (2) offered response dimension (whether the offered response dimension measures intensity, frequency, or quantity; the offered response dimension for an AD question is by definition intensity – the intensity of agreement – but the offered response dimension – the dimension that is explicit with a CS question – will likely vary); (3) number of response categories (most comparisons use five categories, but some experiments use seven or eleven categories); (4) labeling of categories (whether categories are fully labeled versus end-point-only labeled); (5) direction of response categories (whether the categories increase in value – "not at all" to "extremely" – or decrease in value – "extremely" to "not at all"); and (6) polarity (whether the question is bipolar or unipolar; AD questions are always bipolar, CS questions can vary). In addition, if questions are bipolar, an important feature is the inclusion (or exclusion) of a middle category (e.g., "neither agree nor disagree") and how it is labeled (Dykema et al. 2019).

In our experiment, the AD and CS questions varied based on their offered response dimension (the response dimensions for the CS questions were construct-specific by design and tapped into the dimensions of intensity, frequency, and quantity), the direction of the response categories (the AD response categories were ordered from high to low – "strongly agree" to "strongly disagree" – while the CS categories were ordered from low to high – "not at all" to "a great deal," "never" to "always"), and their polarity (the AD questions were bipolar; the CS were unipolar). While our design does not allow us to estimate the unique effects of these characteristics, we encourage future work using multifactorial designs that will provide researchers with the ability to estimate the effects of particular characteristics.

With regard to the mode of administration, one critical difference between interviewer-administered and self-administered modes is that respondents need to encode and recall the response categories in order to map their response. Providing showcards for all CS items during in-person interviews may reduce cognitive burden on respondents. However, this solution is not easily applicable to telephone interviews and CS scales that include many items with variable response categories may be problematic. Another possibility is to select response options that vary less dramatically from question to question and that use the everyday language of respondents, which may introduce an additional challenge if the item is to be used in cross-cultural research. These issues are likely to receive increased scrutiny as surveys that mix interviewer- and self-administration grow and researchers continue to explore methods to measure and reduce mode effects (De Leeuw and Berzelak 2016).

We note several other limitations of this study. First, while prior research indicates AD questions are more problematic for respondents with lower education (e.g., Schuman and Presser 1996), our analytic sample was small, precluding subgroup analyses based on education or other socio-demographic variables. Second, if particular aspects of trust are more salient for certain groups (such as those defined by race/ethnicity) or groups use response categories differently, measurement nonequivalence may help explain a portion of the current results (Davidov et al. 2014). Future work investigating whether AD or CS items yield higher levels of measurement equivalence is needed, especially for cross-cultural research and research involving diverse samples, such as the current study.

Third, our data were collected in a unique situation, that of a cognitive interview. While it is possible that the semi-structured nature of this interviewing situation affected outcomes in ways that would not generalize to a standardized survey interview, the format allowed us to collect data on how participants process AD and CS questions, a major contribution of this study. Given that this study yielded some unexpected findings, pretesting of new CS items remains crucial, particularly if the sample is diverse, the target object being evaluated is unfamiliar to participants, and if the evaluation being solicited is complex (e.g., trust, evaluations of others).

6. Appendix

6.1. Appendix A

Appendix A. Exact wording of the agree-disagree (AD) and construct-specific (CS) questions from the trust in medical researchers scale, by round of cognitive interviewing.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Positive valence (most positively valenced category – e.g., "strongly agree" for AD and "a great deal" for CS – indicates most trust)				
General trust	All things considered, you trust medical researchers a great deal. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	All things considered, how much do you trust medical researchers: none, a little, some, quite a bit, or a great deal?	All things considered, you trust medical researchers a great deal. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	All things considered, how much do you trust medical researchers: none, a little, some, quite a bit, or a great deal?
Participants' interests	Medical researchers always have the best interests of participants from your racial or ethnic group in mind. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How much of the time do medical researchers have the best interests of participants from your racial or ethnic group in mind: none of the time, a little of the time, some of the time, most of the time, or all of the time?	**When they are conducting research,** medical researchers always have the best interests of participants from your racial or ethnic group in mind. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	**When they are conducting research**, how **often** do medical researchers have the best interests of participants from your racial or ethnic group in mind: **never, rarely, sometimes, very often, or always**?
Participants' safety	Medical researchers work hard to make sure that the participants in their studies are safe. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure that the participants in their studies are safe: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?	Medical researchers work hard to make sure that the participants in their studies are safe. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure that the participants in their studies are safe: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?

Appendix A. Continued.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Tell about risks	Medical researchers always tell participants everything they need to know about the risks of participating in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers tell participants everything they need to know about the risks of participating in their studies: never, rarely, sometimes, very often, or extremely often?	Medical researchers always tell participants everything they need to know about the risks of participating in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers tell participants everything they need to know about the risks of participating in their studies: never, rarely, sometimes, very often, or **always**?
Treat fairly	Medical researchers treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups: never, rarely, sometimes, very often, or extremely often?	Medical researchers **always** treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group the same as participants from other racial or ethnic groups: never, rarely, sometimes, very often, or **always**?
Protect privacy	Medical researchers work hard to make sure they keep information from participants private and secure. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure they keep information from participants private and secure: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?	Medical researchers work **extremely** hard to make sure they keep information from participants private and secure. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How hard do medical researchers work to make sure they keep information from participants private and secure: not at all hard, a little hard, somewhat hard, very hard, or extremely hard?

Appendix A. Continued.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Negative valence (most positively valenced category – e.g., "strongly agree" for AD and "a great deal" for CS – indicates least trust)				
Researchers' interests	Medical researchers care more about their research than they do about the participants in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	Compared to their research, do medical researchers care a lot less, somewhat less, about the same, somewhat more, or a lot more about the participants in their studies?	Medical researchers care more about their research than they do about the participants in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	**To what extent do medical researchers care more about their research than they do about the participants in their studies: not at all, a little, somewhat, quite a bit, a great deal?**
Select minorities	Medical researchers are more likely to select minorities for their most risky studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How likely are medical researchers to select minorities for their most risky studies: not at all likely, a little likely, somewhat likely, very likely, or extremely likely?	**When selecting participants for their most risky studies**, medical researchers are more likely to select minorities. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	**When selecting participants for their most risky studies**, how likely are medical researchers to select minorities: not at all likely, a little likely, somewhat likely, very likely, or extremely likely?
Hide information	Medical researchers never hide information about the possible risks of participating. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers hide information about the possible risks of participating: never, rarely, sometimes, very often, or extremely often?	Medical researchers **often** hide information about the possible risks of participating **in medical research studies**. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers hide information about the possible risks of participating **in medical research studies**: never, rarely, sometimes, very often, or extremely often?

Appendix A. Continued.

	Round 1		Round 2	
Question label	AD version	CS version	AD version	CS version
Treat like guinea pig	Medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies: never, rarely, sometimes, very often, or extremely often?	Medical researchers **often** treat participants from your racial or ethnic group like guinea pigs in their studies. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers treat participants from your racial or ethnic group like guinea pigs in their studies: never, rarely, sometimes, very often, or extremely often?
Know more	Medical researchers often want to know more than they need to know. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers want to know more than they need to know: never, rarely sometimes, very often, or extremely often?	Medical researchers often want to know more than they need to know. Do you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree?	How often do medical researchers want to know more than they need to know: never, rarely, sometimes, very often, or extremely often?

Notes: Differences in question wording between rounds are shown in bold.

6.2. Appendix B

Appendix B. Proportion of participants exhibiting a given behavioral indicator of response difficulty (BIRD), by question and experimental group.

	Codable response + qualification				Codable response + elaboration				Uncodable Response				Seeks Clarification			
	AD	CS	Difference	p-value	AD	CS	Difference	p-value	AD	CS	Difference	p-value	AD	CS	Difference	p-value
Positive valence																
General trust	0.03	0.03	0.00	1.00	0.09	0.00	0.09	0.24	0.06	0.09	−0.03	1.00	0.19	0.03	0.16	0.10
Participants' interests	0.00	0.13	−0.13	0.11	0.19	0.09	0.10	0.47	0.03	0.19	−0.16	0.10	0.31	0.19	0.12	0.39
Participants' safety	0.09	0.22	−0.13	0.30	0.28	0.16	0.12	0.37	0.03	0.06	−0.03	1.00	0.13	0.19	−0.06	0.73
Tell about risks	0.00	0.25	−0.25	0.01	0.25	0.13	0.12	0.34	0.03	0.06	−0.03	1.00	0.22	0.06	0.14	0.15
Treat fairly	0.09	0.09	0.00	1.00	0.38	0.13	0.25	0.04	0.03	0.22	−0.19	0.05	0.22	0.31	−0.09	0.57
Protect privacy	0.13	0.13	0.00	1.00	0.06	0.13	−0.07	0.67	0.03	0.03	0.00	1.00	0.19	0.25	−0.06	0.76
Negative valence																
Researchers' interest	0.13	0.31	−0.18	0.13	0.22	0.25	−0.03	1.00	0.13	0.13	0.00	1.00	0.25	0.13	0.12	0.29
Select minorities	0.16	0.28	−0.12	0.37	0.31	0.09	0.22	0.06	0.09	0.19	−0.10	0.47	0.16	0.28	−0.12	0.37
Hide information	0.03	0.22	−0.19	0.05	0.28	0.09	0.19	0.11	0.16	0.19	−0.03	1.00	0.16	0.09	0.07	0.71
Treat like guinea pig	0.13	0.13	0.00	1.00	0.41	0.13	0.28	0.02	0.13	0.16	−0.03	1.00	0.19	0.25	−0.06	0.76
Know more	0.09	0.13	−0.04	1.00	0.19	0.13	0.06	0.73	0.03	0.16	−0.13	0.20	0.19	0.16	0.03	1.00

Appendix B. Continued.

	Question Repeated				Don't Know or Refusal			
	AD	CS	Difference	p-value	AD	CS	Difference	p-value
Positive valence								
General trust	0.19	0.03	0.16	0.10	0.06	0.00	0.06	0.49
Participants' interests	0.25	0.38	−0.13	0.42	0.19	0.09	0.10	0.47
Participants' safety	0.16	0.19	−0.03	1.00	0.00	0.16	−0.16	0.05
Tell about risks	0.25	0.16	0.09	0.54	0.09	0.06	0.03	1.00
Treat fairly	0.25	0.31	−0.06	0.78	0.09	0.19	−0.10	0.47
Protect privacy	0.22	0.28	−0.06	0.77	0.09	0.09	0.00	1.00
Negative valence								
Researchers' interest	0.22	0.44	−0.22	0.11	0.03	0.06	−0.03	1.00
Select minorities	0.16	0.38	−0.22	0.09	0.16	0.06	0.10	0.43
Hide information	0.19	0.03	0.16	0.10	0.00	0.22	−0.22	0.01
Treat like guinea pig	0.19	0.31	−0.12	0.39	0.00	0.13	−0.13	0.11
Know more	0.06	0.19	−0.13	0.26	0.06	0.13	−0.07	0.67

Notes: For each question, n = 32 participants for the AD scale and n = 32 participants for the CS scale. p-values are from Chi-squared tests and Fisher's exact tests.

7. References

Anderson, L.A. and R.F. Dedrick. 1990. "Development of the Trust in Physician Scale: A Measure to Assess Interpersonal Trust in Patient-physician Relationships." *Psychological Reports* 67: 1091–1100. Doi: https://doi.org/10.2466/pr0.1990.67.3f.1091.

Audacity Developer Team. 2008. Audacity (Version 1.2.6) [Computer Software]: Available at: http://www.audacityteam.org/download/ (accessed April 2019).

Bassili, J.N. and B.S. Scott. 1996. "Response Latency as a Signal to Question Problems in Survey Research." *Public Opinion Quarterly* 60: 390–399. Doi: https://doi.org/10.1086/297760.

Braunstein, J.B., N.S. Sherber, S.P. Schulman, E.L. Ding, and N.R. Powe. 2008. "Race, Medical Researcher Distrust, Perceived Harm, and Willingness to Participate in Cardiovascular Prevention Trials." *Medicine* 87: 1–9. Doi: https://doi.org/10.1097/MD.0b013e3181625d78.

Carpenter, P.A. and M.A. Just. 1975. "Sentence Comprehension: A Psycholinguistic Processing Model of Verification." *Psychological Review* 82: 45–73. Available at: http://psycnet.apa.org/doi/10.1037/h0076248 (accessed April 2019).

Corbie-Smith, G., S.B. Thomas, and D.M.M. St. George. 2002. "Distrust, Race, and Research." *Archives of Internal Medicine* 162: 2458–2463. Doi: https://doi.org/10.1001/archinte.162.21.2458.

Davidov, E., B. Meuleman, J. Cieciuch, P. Schmidt, and J. Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40: 55–75. Doi: https://doi.org/10.1146/annurev-soc-071913-043137.

De Leeuw, E. and N. Berzelak. 2016. "Survey Mode or Survey Modes?" In *The SAGE Handbook of Survey Methodology*, edited by C. Wolf, J. Dominique, T.W. Smith, and F. Yang-chih, 142–156. Los Angeles: SAGE Publications Ltd. Available at: https://books.google.com/books?hl=en&lr=&id=g8OMDAAAQBAJ&oi=fnd&pg=PA142&dq=survey+mode+or+modes+berzelak&ots=DyqMiBT1oS&sig=hGg7pa80-bI535N5GgSUwvLmLfY#v=onepage&q=survey%20mode%20or%20modes%20berzelak&f=false (accessed April 2019).

Dijkstra, W. and Y. Ongena. 2006. "Question-Answer Sequences in Survey-Interviews." *Quality & Quantity* 40: 983–1011. Doi: https://doi.org/10.1007/s11135-005-5076-4.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th edition). Hoboken, NJ: John Wiley.

Draisma, S. and W. Dijkstra. 2004. "Response Latency and (Para)linguistic Expression as Indicators of Response Error." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer, 131–148. New York: Springer-Verlag. Doi: https://doi.org/10.1002/0471654728.ch7.

Dykema, J., J.M. Lepkowski, and S. Blixt. 1997. "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 287–310. N.Y: Wiley-Interscience. Available at: https://onlinelibrary.wiley.com/doi/10.1002/9781118490013.ch12 (accessed April 2019).

Dykema, J., N.C. Schaeffer, and D. Garbarski. 2012. "Effects of Agree-Disagree Versus Construct-Specific Items on Reliability, Validity, and Interviewer-Respondent Interaction." Presented at the American Association for Public Opinion Research, May 17–20. 2012. Orlando, Florida, U.S.A.

Dykema, J., N.C. Schaeffer, and D. Garbarski. 2019. "Towards a Reconsideration of the Use of Agree-Disagree Questions in Measuring Subjective Evaluations." Unpublished manuscript, University of Wisconsin-Madison, Madison-WI.

Edwards, D.F. 2015. "Voices Heard." Presented at the Health Equity Leadership Institute, Madison, WI.

Egede, L.E. and C. Ellis. 2008. "Development and Testing of the Multidimensional Trust in Health Care Systems Scale." *Journal of General Internal Medicine* 23: 808–815. Doi: https://doi.org/10.1007/s11606-008-0613-1.

Fleiss, J.L. 1981. *Statistical Methods for Rates and Proportions*, 2nd edition. New York: Wiley.

Fortune-Greeley, A.K., K.E. Flynn, D.D. Jeffery, M.S. Williams, F.J. Keefe, R.B. Reeve, G.B. Willis, and K.P. Weinfurt. 2009. "Using Cognitive Interviews to Evaluate Items for Measuring Sexual Functioning Across Cancer Populations: Improvements and Remaining Challenges." *Quality of Life Research* 18: 1085–1093. Doi: https://doi.org/10.1007/s11136-009-9523-x.

Fowler, F.J. and C. Cosenza. 2009. "Design and Evaluation of Survey Questions." In *The Sage Handbook of Applied Social Research Methods*, edited by L. Bickman and D.J. Rog, 375–412. Thousand Oaks, CA: Sage.

Hall, M.A., F. Camacho, E. Dugan, and R. Balkrishnan. 2002a. "Trust in the Medical Profession: Conceptual and Measurement Issues." *Health Services Research* 37: 1419–1439. Doi: https://doi.org/10.1111/1475-6773.01070.

Hall, M.A., F. Camacho, J.S. Lawlor, V. DePuy, J. Sugarman, and K. Weinfurt. 2006. "Measuring Trust in Medical Researchers." *Medical Care* 44: 1048–1053. Available at: http://www.jstor.org/stable/41219560 (accessed April 2019).

Hall, M.A., E. Dugan, B. Zheng, and A.K. Mishra. 2001. "Trust in Physicians and Medical Institutions: What is It, Can It be Measured, and Does It Matter?" *Milbank Quarterly* 79: 613–639. Doi: https://doi.org/10.1111/1468-0009.00223.

Hall, M.A., B. Zheng, E. Dugan, F. Camacho, K.E. Kidd, A. Mishra, and R. Balkrishnan. 2002b. "Measuring Patients' Trust in their Primary Care Providers." *Medical Care Research and Review* 59: 293–318. Doi: https://doi.org/10.1177/1077558702059003004.

Hanson, T. 2015. "Comparing Agreement and Item-Specific Response Scales: Results from an Experiment." *Social Research Practice* 1: 17–25. Available at: http://the-sra.org.uk/wp-content/uploads/social-research-practice-journal-issue-01-winter-2015.pdf (accessed April 2019).

Hayman, R.M., B.J. Taylor, N.S. Peart, B.C. Galland, and R.M. Sayers. 2001. "Participation in Research: Informed Consent, Motivation and Influence." *Journal of Paediatrics and Child Health* 37: 51–54. Available at: https://doi.org/10.1046/j.1440-1754.2001.00612.x (accessed April 2019).

Henderson, G., J. Garrett, J. Bussey-Jones, M.E. Moloney, C. Blumenthal, and G. Corbie-Smith. 2008. "Great Expectations: Views of Genetic Research Participants Regarding

Current and Future Genetic Studies." *Genetics in Medicine* 10: 193–200. Doi: https://doi.org/10.1097/GIM.0b013e318164e4f5.

Höhne, J.K. and D. Krebs. 2018. "Scale Direction Effects in Agree/Disagree and Item-Specific Questions: A Comparison of Question Formats." *International Journal of Social Research Methodology* 21: 91–103. Doi: https://doi.org/10.1080/13645579.2017.1325566.

Höhne, J.K. and T. Lenzner. 2018. "New Insights on the Cognitive Processing of Agree/Disagree and Item-Specific Questions." *Journal of Survey Statistics and Methodology* 6: 401–417. Doi: https://doi.org/10.1093/jssam/smx028.

Höhne, J.K., S. Schlosser, and D. Krebs. 2017. "Investigating Cognitive Effort and Response Quality of Question Formats in Web Surveys Using Paradata." *Field Methods* 29: 365–382. Doi: https://doi.org/10.1177/1525822x17710640.

Holbrook, A.L. 2008. "Recency Effect." In *Encyclopedia of Survey Research Methodology*, edited by P.J. Lavrakas, 695–696. Newbury Park, CA: Sage.

Johnson, R.B. and A.J. Onwuegbuzie. 2004. "Mixed Methods Research: A Research Paradigm Whose Time Has Come." *Educational Researcher* 33: 14–26. Doi: https://doi.org/10.3102/0013189X033007014.

Krosnick, J.A. and S. Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, Second Edition, edited by P.V. Marsden and J.D. Wright, 263–313. Bingley, UK: Emerald Group Publishing Limited.

Kuru, O. and J. Pasek. 2016. "Improving Social Media Measurement in Surveys: Avoiding Acquiescence Bias in Facebook Research." *Computers in Human Behavior* 57: 82–92. Available at: https://doi.org/10.1016/j.chb.2015.12.008 (accessed April 2019).

Landis, J.R. and G.G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33: 159–174. Doi: https://doi.org/10.2307/2529310.

Lelkes, Y. and R. Weiss. 2015. "Much Ado about Acquiescence: The Relative Validity and Reliability of Construct-Specific and Agree-Disagree Questions." *Research and Politics* 2: 1–8. Doi: https://doi.org/10.1177/2053168015604173.

Liu, M., S. Lee, and F.G. Conrad. 2015. "Comparing Extreme Response Styles between Agree-Disagree and Item-Specific Scales." *Public Opinion Quarterly* 79: 952–975. Doi: https://doi.org/10.1093/poq/nfv034.

Mainous, A.G., D.W. Smith, M.E. Geesey, and B.C. Tilley. 2006. "Development of a Measure to Assess Patient Trust in Medical Researchers." *Annals of Family Medicine* 4: 247–252. Doi: https://doi.org/10.1370/afm.541.

Revilla, M. and C. Ochoa. 2015. "Quality of Different Scales in an Online Survey in Mexico and Columbia." *Journal of Politics in Latin America* 7: 157–177. Available at: https://journals.sub.uni-hamburg.de/giga/jpla/article/view/903/910 (accessed April 2019).

Rogers, W. 1994. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 13. Available at: https://ideas.repec.org/a/tsj/stbull/y1994v3i13sg17.html (accessed April 2019).

Ryan, G.W. and H.R. Bernard. 2003. "Techniques to Identify Themes." *Field Methods* 15: 85–109. Doi: https://doi.org/10.1177/1525822x02239569.

Saris, W.E., M. Revilla, J.A. Krosnick, and E.M. Shaeffer. 2010. "Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response

Options." *Survey Research Methods* 4: 61–79. Doi: https://ojs.ub.uni-konstanz.de/srm/article/view/2682/3971.

Schaeffer, N.C. and J. Dykema. 2011. "Response 1 to Fowler's Chapter: Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions." In *Question Evaluation Methods: Contributing to the Science of Data Quality*, edited by J. Madans, K. Miller, A. Maitland, and G. Willis, 23–39. Hoboken, NJ: John Wiley & Sons, Inc. Available at: https://doi.org/10.1002/9781118037003.ch3.

Scharff, D.P., K.J. Mathews, P. Jackson, J. Hoffsuemmer, E. Martin, and D. Edwards. 2010. "More than Tuskegee: Understanding Mistrust about Research Participation." *Journal of Health Care for the Poor and Underserved* 21: 879–897. Doi: https://doi.org/10.1353/hpu.0.0323.

Schuman, H. and S. Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage Publications, Inc.

Smith, T.W., P.V. Marsden, and M. Hout. 2013. General Social Survey, 1972–2010 [Cumulative File]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2013-02-07. Doi: https://doi.org/10.3886/ICPSR31521.v1.

Streiner, D.L., G.R. Norman, and J. Cairney. 2015. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford, UK: Oxford University Press.

Sturgis, P., C. Roberts, and P. Smith. 2014. "Middle Alternatives Revisited: How the neither/nor Response Acts as a Way of Saying "I Don't Know"?" *Sociological Methods & Research* 43: 15–38. Doi: https://doi.org/10.1177/0049124112452527.

Thompson, H.S., H.B. Valdimarsdottir, G. Winkel, L. Jandorf, and W.W. Redd. 2004. "The Group-Based Medical Mistrust Scale: Psychometric Properties and Association with Breast Cancer Screening." *Preventive Medicine* 38: 209–218. Doi: https://doi.org/10.1016/j.ypmed.2003.09.041.

Tourangeau, R., M.C. Couper, and F. Conrad. 2004. "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68: 368–393. Doi: https://doi.org/10.1093/poq/nfh035.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.

Williams, M.M., D.P. Scharff, K.J. Mathews, J.S. Hoffsuemmer, P. Jackson, J.C. Morris, and D.F. Edwards. 2010. "Barriers and Facilitators of African American Participation in Alzheimer Disease Biomarker Research." *Alzheimer Disease & Associated Disorders* 24: S24–S29. Available at: https://journals.lww.com/alzheimerjournal/Fulltext/2010/07001/Barriers_and_Facilitators_of_African_American.6.aspx (accessed April 2019).

Willis, G.B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.

Willis, G.B. and K. Miller. 2011. "Cross-Cultural Cognitive Interviewing: Seeking Comparability and Enhancing Understanding." *Field Methods* 23: 331–341. Doi: https://doi.org/10.1177/1525822x11416092.

Willits, F.K., G.L. Theodori, and A.E. Luloff. 2016. "Another Look at Likert Scales." *Journal of Rural Social Sciences* 31: 126–139. Available at: http://journalofruralsocialsciences.org/pages/Articles/JRSS%202016%2031/3/JRSS%202016%2031%203%20126-139.pdf (accessed April 2019).

Yan, T. and R. Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22: 51–68. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1331 (accessed April 2019).

Zheng, B., M.A. Hall, E. Dugan, K.E. Kidd, and D. Levine. 2002. "Development of a Scale to Measure Patients' Trust in Health Insurers." *Health Services Research* 37: 185–200. Doi: https://doi.org/10.1111/1475-6773.00145.

Received October 2017
Revised December 2018
Accepted January 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 387–408, http://dx.doi.org/10.2478/JOS-2019-0018

Item Response Rates for Composite Variables

Jonathan Eggleston[1]

Item response rates frequently serve as indicators of data quality and potential nonresponse bias. However, key variables from surveys, such as total household income or net worth, are often composite variables constructed from several underlying components. Because such composite variables do not have clearly identifiable response rates, inference on the data quality of these key measures is more difficult. This article proposes three new methods for aggregating data on response rates across questions to create a measure of item response for composite variables. To compare the three methods and illustrate how they can be used (both individually and collectively) to investigate data quality, I analyze item response for net worth in the Survey of Income and Program Participation (SIPP) and the Survey of Consumer Finances (SCF). These new measures provide detailed information about net worth estimates that would be difficult to assess without an item response aggregation method. Overall, these new item response rate methods provide a new way of describing data quality for key measures in surveys and for analyzing changes in data quality over time.

Key words: Response rates; nonresponse.

1. Introduction

Unit and item response rates are widely used tools in survey research to measure the potential impact of nonresponse bias (e.g., Bollinger et al. 2015). While low response rates do not necessarily lead to high nonresponse bias (Groves and Peytcheva 2008), constructing an alternative measure based on validation studies is often difficult or infeasible for many surveys. For example, evaluating nonresponse bias in wealth data from the United States is extremely difficult because the United States lacks comprehensive administrative data on wealth, due to the absence of a wealth tax. Because of this absence of validation data, response rates are seen as one of the primary indicators of data quality, and are sometimes the only indicator available.

Considerable attention has been focused on how to construct unit response rates properly (e.g., American Association for Public Opinion Research 2016), analyzing item response rates for individual survey questions (e.g., Ferber 1966), and studying nonresponse bias for some topics in which administrative data is available (e.g., Bound and Krueger 1991 for earnings; Bee and Mitchell 2017 for retirement income). However, little attention has been given to item response rates for composite variables that are created from several survey questions, despite the fact that these composite variables are

[1] U.S. Census Bureau, 4600 Silver Hill Road, Washington DC, 20233, U.S.A. Email: Jonathan.S.Eggleston@census.gov
Acknowledgments: The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

often the key measures in surveys and receive the most attention. Such variables are also called recode or summary variables. Examples of such variables include net worth from the Survey of Income and Program Participation (SIPP) and the Survey of Consumer Finances (SCF), employment status from the Current Population Survey (CPS), and household income from the CPS Annual Social and Economic Supplement.

While item response rates are available for the underlying components, this information may be hard for data users to synthesize. The SCF, for example, has over 140 assets and liabilities questions, and response rates can vary substantially across questions. For example, the question on home value in the 2013 SCF has an item response rate of 90.6%, while the question on the cash value of life insurance has a response rate of only 58.9%. Without aggregating the response rates in a meaningful way, it is extremely difficult to assess how key estimates from these surveys are impacted by item nonresponse.

Because researchers often look at response rates to gauge data quality, having such information for composite variables could be useful for individuals evaluating the trustworthiness of key estimates, evaluating the effects of a major survey redesign, or deciding upon which dataset to use for a research project. To the best of my knowledge, there has been no paper that focuses on how to construct item nonresponse rates for composite variables. In a paper on item nonresponse bias, Hokayem et al. (2017) briefly describe total item response rates for household income in the CPS Annual Social and Economic Supplement. However, their paper is not focused on item response rate calculations, nor do they discuss alternative ways of constructing such a measure.

This article proposes several new methods for aggregating data on response rates across questions to create a measure of item response for composite variables. These methods provide a useful way to summarize item response rates for key measures, and offer a feasible way of comparing item response rates across surveys and over time. The three proposed measures of response rates are

1. The percent of observations without any missing component,
2. A sum-weighted formula, which is the sum of reported values divided by the sum of all values, and
3. A median-weighted formula, in which the response rate for each question is weighted by the percent of respondents with a non-missing/non-zero value times the median value.

I present three methods instead of just one for several purposes. First, there is no single "correct" way to construct response rates, so some users may have a preference for one over the other. Second, it is useful to present the comparisons across surveys using multiple methods as a robustness check, in order to make sure that the relative differences in response rates between surveys is not solely due to idiosyncratic features of one composite rate formula. Third, one formula may be preferable over the other depending on the key statistic or composite variable of interest. Some factors to take into consideration are

- How many variables are used to construct the composite variable,
- The type of key statistic of interest (e.g., mean vs. median),
- Differences in response rates across questions, and
- Relative importance of each variable for the key statistic.

For example, if the key statistic is composed from only a few items, such the Positive and Negative Affect Schedule (PANAS), (Watson et al. 1998) or disability indicators in US surveys, then Method #1 may be sufficient. However, if the key statistic is composed of over 100 items, then Method #1 may give a misleading picture of the prevalence of item nonresponse, as someone would be coded as an item nonresponder even if they failed to answer just one question. For income and wealth statistics, Method #2 may be more useful if the key statistic is total wealth or income for the country, while Method #3 may be more useful for analyzing medians. Thus, similar to the numerous AAPOR unit response rate calculations (American Association for Public Opinion Research (AAPOR) 2016), presenting multiple formulas can help accommodate researchers who may have varying needs or reasons for using one method over another.

To illustrate these methods, I focus on the key estimates of household wealth in the United States (e.g., mean, median, total), which are available from the SIPP and the SCF. Within an individual survey, the three proposed formulas yield different answers for the level of item nonresponse. For example, when looking at all assets and debt in 2014 SIPP, only 27.7% of households have no imputed asset or debt value (Method #1). However, when using the median-weighted formula (Method #3), asset and debt questions have an average response rate of 77.3%. This discrepancy is driven by the fact that home values have a high response rate of 87.1%, and that home values are given a large weight in Method #3. The median-weighted formula gives primary residences a high weight because both the ownership rate and median values is high for this asset. Thus, Method #3 may give a better reflection of the degree of item response in a households' typical wealth portfolio than Method #1.

However, while these three formulas yield different values for the level of item response in a survey, all formulas present similar answers for how item response varies across surveys. To illustrate this point, I compare item response in wealth estimates between the 2008 SIPP, 2014 SIPP, and 2013 SCF. The U.S. Census Bureau redesigned SIPP starting with the 2014 Panel, at which time numerous changes were made to the survey. The asset section underwent a major revision; new assets were added and asset income and values were asked together rather than in separate sections. Thus, comparing the 2014 SIPP wealth data to other surveys serves as a useful case study for these new item response calculations. Using the three proposed composite response rate methods, the results in general show that overall item response rates increased for wealth questions in the new panel, but are still lower than in the SCF. Given this trend is found in all the methods, the results suggest that these findings are due to real trends found in the data and not due to idiosyncratic features of any one formula. In addition, I also present results that incorporate unit response rates into the calculations. Once unit response rates are incorporated, 2014 SIPP has a higher combined response rate for wealth than the SCF.

In summary, these three proposed aggregation methods 1) provide a concise summary of item response for key estimates and 2) help compare item response rates for composite measures across surveys. While a researcher could compare item response rates for individual questions between the two surveys, this task would be difficult given the large number of questions and issues with comparability of question text and concepts. Each of the three proposed methods have strengths and weaknesses for their inferential ability. Thus, recommendations for which method to use depends on the key statistic of interest.

For median wealth, Method #3 may be the most useful as it offers the benefits of being less sensitive to outliers than Method #2 and being more reflective of average item response tendencies than Method #1. Method #2 may be more useful for looking at total wealth, and Method #1 may be more useful for key statistics composed from a small number of discrete-choice questions. However, the case study in this article suggests that all methods will yield similar conclusions when comparing response rates across surveys.

The remainder of the article proceeds as follows. Section 2 gives a description of the SIPP and SCF data and highlight response rates for some key variables. Section 3 discusses the new methods for aggregating data on response rates. Section 4 presents a comparison of the three methods using 2014 SIPP wealth data. Section 5 presents the results comparing 2008 SIPP, 2014 SIPP, and 2013 SCF to demonstrate how the methods may be compared across surveys. Section 6 discusses how these methods can be modified for non-continuous variables. Finally, Section 7 concludes.

2. Background

Before presenting the methods for aggregating item response rates, I first present an overview of the SIPP and SCF data and highlight response rates for a few variables. This discussion will show that comparing question-level response rates over time and across surveys is difficult in light of questionnaire differences, suggesting the need for an alternative measure for comparing item response rates over time.

2.1. Description of Datasets

The SIPP is a longitudinal survey conducted by the U.S. Census Bureau that collects information about the income, assets, labor market activity, and participation in government programs of U.S. households (U.S. Census Bureau 2016). Information on a wide variety of assets and debt is collected and includes financial variables such as the value of savings accounts, checking accounts, retirement accounts, real estate, and credit card debt. SIPP interviews households for a period of about 2.5 to 5 years (depending on the panel), with each panel containing a new set of households. Households are interviewed primarily in person, with some interviews conducted via phone. The survey attempts to interview individually every adult in the household, although some personal interviews are "proxy interviews" in which another household member provides the information about the respondent. In order to improve estimates of receipt from government programs, SIPP oversamples low-income areas.

In the 2014 panel, the U.S. Census Bureau made many changes to the survey. One substantial change is that SIPP now interviews respondents less frequently in order to reduce costs (U.S. Census Bureau 2016). In earlier panels, interviews occurred every four months, but in the 2014 panel, interviews occurred once per year. However, interviews with questions on asset values occurred about once a year in previous panels, with some gaps across time. Therefore, SIPP collects wealth data with roughly the same frequency in the 2014 Panel as before. Another change was the introduction of the event history calendar (EHC), which is a visual method of collecting retrospective data on the timing of events, such as the loss of a job or health insurance coverage. This change was made to help reduce any negative impacts of the longer recall period on data quality.

The asset section also underwent a major revision. SIPP now collects data on additional assets that were not asked about previously, such as annuities, trusts, student loans, and education savings accounts. Moreover, questions on asset income and asset values are now asked concurrently rather than in separate sections of the interview. Wording for many questions changed as well. Because so much of the survey was modified, the quality of the net worth data may have changed substantially. These survey changes appear to have resulted in higher estimates of median wealth in 2014 SIPP compared to 2008 SIPP (Eggleston and Gideon 2017).

To evaluate the effects of the SIPP redesign on item response rates, a useful comparison survey is the SCF. The SCF is a survey sponsored by the Federal Reserve studying the income and wealth of US families (Bricker et al. 2017). Because SCF has more detailed wealth questions and its interviewers receive more training about types of assets, it has been labeled as the "gold standard" for wealth data, and thus can serve as a basis for comparison to any survey aiming to measure wealth in the United States. The National Research Council (2009) is one among many sources that have applied this label in reference to SCF. Historically, SIPP has had lower estimates of wealth than the SCF, although these discrepancies decreased in the 2014 Panel (Eggleston and Gideon 2017).

While both SIPP and SCF draw their samples from the general U.S. population and have similar content, there are several methodological differences between the surveys that might explain differences in their estimates of wealth and their item response rates. The SCF oversamples high-income individuals based on data from the Internal Revenue Service (IRS) in order to improve its estimates of income and wealth among high-wealth families. This sampling is in contrast with SIPP, which oversamples low-income areas. While weighting should theoretically result in both surveys producing similar estimates, the surveys may differ in the precision of their estimates for certain subpopulations.

In addition, SCF has a much smaller sample size than SIPP. The 2013 SCF contains about 6,000 families, while wave 1 of the 2014 SIPP Panel contains about 30,000 households, and the first wave of the 2008 Panel contains about 42,000 households. Both surveys primarily interview households in person. However, the SCF only interviews one person in the household, which is in contrast with SIPP that attempts to interview every adult in the household. Finally, the 2013 SCF offered a conditional USD 50 incentive (Hsu et al. 2017), while the 2008 and 2014 SIPP offered a USD 0, USD 20, or USD 40 incentive depending on a household's assigned experimental condition (Westra et al. 2015).

2.2. *Item Response Rates for Individual Questions*

To introduce how item response varies across the surveys and across questions, Table 1 presents item response rates for some variables in 2008 and 2014 SIPP panels and 2013 SCF. In this table, I define item response rates as the proportion of people who gave an answer (numeric value) to the respective wealth question. This classification excludes people who gave an answer of "don't know" or "refuse" to a question, or who drop out of the survey before reaching the particular asset question. This table contain standards errors for the response rates to account for sampling error. Sampling error could cause one of the surveys to have more item response among respondents, even if the questionnaire was identical. For SIPP, the individual providing information could either be the person

Table 1. Response rates for individual wealth variables.

Value variable	SIPP 2014	SIPP 2008	SCF 2013
Primary residence	87.1 (0.28)	75.1 (0.48)	90.6 (0.81)
IRA/Keogh Retirement account	59.2 (0.67)		
IRA retirement account		46.2 (0.64)	
Keogh retirement account		28.5 (1.87)	64.2 (14.4)
Regular IRA account			74.9 (2.69)
Roth IRA account			76.8 (3.3)
Trusts	47.2 (2.58)		70.6 (7.52)

Source: 2008 SIPP Panel (Wave 4), 2014 SIPP Panel (Wave 1), and 2013 SCF. Table presents item response rates for a small set of value variables. Replicate weights used to construct standard errors in all three surveys, and imputation implicates were used to construct the standard error for SCF. Standard errors shown in parenthesis.

in question or another household member providing a proxy interview. Thus, proxy responses are counted in the pool of item responders. Households that have unit nonresponse are dropped from the sample, so they are not included in the item response rate calculations for this table. In SIPP, if no one in the household responds, then the household is considered to be a unit nonresponder. In some other households, one person is interviewed but other people are unable to be interviewed (either in person or through a proxy interview). These noninterviewed people are treated as item nonresponders, so they are included in the item response rate calculations.

For some variables, such as the value of primary residences, there is a directly comparable variable in all three surveys. Table 1 shows that the item response rate for primary residences increased in SIPP from 75.1% to 87.1% after the redesign, but is still lower than the 2013 SCF rate of 90.6%. These differences are surprising, given the question text for the home value question is very similar amongst all the surveys (Eggleston and Gideon 2017).

For other variables, comparisons between the 2008 and 2014 SIPP Panels and 2013 SCF are not clear cut. For example, the 2008 SIPP and 2013 SCF have separate questions on Individual Retirement Account (IRA) and Keogh retirement account balances (types of tax-advantaged investment accounts available in the United States), but the 2014 SIPP Panel has one combined question. In addition, the SCF has a separate question on Roth IRA accounts and Regular IRA accounts, but SIPP does not make this distinction. Because of these differences, it is unclear how to compare response rates for IRA accounts across these surveys. Another complication arises for when a new question is added to the survey. For example, a direct question on trusts was added to SIPP in the 2014 Panel, and this item has a low item response rate of 52.8%. In some ways, data quality should be improved with the addition of this question because information on trust values is now gathered in its own question rather than through a catch-all question that asks respondents to report any other asset that haven't been asked about already. On the other hand, data quality might have decreased because this new question has a low response rate. Because new questions have no analogues from prior surveys, evaluating the effects of new variables on overall data quality needs to be addressed in a different way than variable-to-variable comparisons in response rates.

3. Equations for Item Nonresponse for Composite Variables

Because of the difficulty in comparing item response rates question by question, as described above, there needs to be a way of aggregating item nonresponse data across variables in order to generate statistics which can be used to compare response rates across surveys or over time within one survey. This section presents three different ways of aggregating response rates over several variables. In this section, the focus is on continuous numeric variables, such as income and the monetary value of assets, but I discuss at the end of the article in Section 6 how the results can be generalized to other types of variables.

One feature of continuous numeric variables is that the data can exhibit varying degrees of item nonresponse. For some questions about monetary amounts, respondents who give an initial answer of "don't know" or "refuse" are asked a range follow-up question (e.g., "is the value less than USD 500, between USD 500 and USD 1,000, or more than USD 1,000"). To denote a person's type of item response, let individual i's response to question q be indicated by $r_{i,q}$, where $r_{i,q} = 1$ if the individual responds to the numeric question, $r_{i,q} = 1/2$ if the individual does not answer the numeric question but gives an answer to a range follow-up question, $r_{i,q} = 0$ if the individual does not respond to either the numeric question or the range follow-up question, and $r_{i,q} = NIU$ if the person is not in universe (i.e., not on path) for the question.

In addition, weights are important to include in the response rate formulas because of potential survey design effects. For example, SIPP oversamples low-income areas and SCF oversamples high-income households. If item response propensities vary by income or income is correlated with other factors that affect response propensities (such as household size), then ignoring weights would cause item response comparisons between the surveys to be conflated with these sampling factors. Because of this factor, I use the variable w_i to denote the survey (adjusted) weight for individual i. Using adjusted weights accounts for over-sampling and nonresponse of certain populations, allowing the item response rates to reflect the odds that a random person from the population would respond to the question. However, if a researcher is only interested in calculating the prevalence of item nonresponse in a single survey (without any cross-survey comparisons), then excluding weights may be beneficial, as this would give the percent of the sample with item nonresponse.

3.1. Percent with no Missing Value

The first aggregation method calculates the proportion of individuals who give a response to every question used to create the composite variable c, given by the expression

$$1 - \frac{\sum_{i=1}^{n} w_i \max\left(d_{c,1}1(r_{i,1} \in \{0, 1/2\}), \ldots, d_{c,Q}1(r_{i,Q} \in \{0, 1/2\})\right)}{\sum_{i=1}^{n} w_i \max\left(d_{c,1}1(r_{i,1} \neq NIU), \ldots, d_{c,Q}1(r_{i,Q} \neq NIU)\right)}, \#M1(c)$$

where $d_{c,q} = 1$ if the survey or researcher uses question number q to construct composite variable number c, and zero otherwise. One advantage of Method #1 is that it is straightforward to explain and understand. Method #1, unlike Methods #2 and #3 below, does not require imputed values for when the variable is missing. In addition, as some

researchers drop imputed values from their analyses, this statistic gives an indicator of the proportion of the sample that would be kept if imputed values are excluded. The formula for Method #1 does not incorporate any range follow-up information into the response rate calculation, which is done for simplicity in this article. However, the express $r_{i,1} \in \{0, 1/2\}$ in the numerator in the formula for Method #1 could be modified to $r_{i,1} \in \{0\}$ to give the rate of people who gave at least a range follow-up response to every question.

The main disadvantage of Method #1 is that it does not capture varying degrees of item response across questions that may be meaningful. For example, Method #1 does not describe whether households tended to give an answer of "don't know" or "refuse" only for a small number of questions or for most questions in the survey. Also, the more questions that are used to create the composite variable, the more likely it is that the household will not give an answer to at least one question, which may create a false sense of inaccuracy for composite variables that are created for a large set of detailed questions.

3.2. *Sum-Weighted Response Rates*

The second aggregation method is calculated by taking the weighted sum of values (using survey weight w_i) with a given response type k (e.g., gave a range follow-up response) divided by the weighted sum of all values. The formula for Method #2 is given by the expression

$$\sum_{i=1}^{n} w_i \left(\sum_{q=1}^{Q} d_{c,q} 1(r_{i,q} = k) x_{i,q} \right) \Bigg/ \sum_{i=1}^{n} w_i \left(\sum_{q=1}^{Q} d_{c,q} x_{i,q} \right), \#M2(c,k)$$

in which $x_{i,q}$ is the actual response or the imputed value for individual i for question number q. Because imputed data are used for item nonresponders, this method can only be used in surveys that impute data for people with missing data.

For $k = 1$, as an example, Method #2 represents the percent of all values for composite variable c that consist of reported values. Because this method is separated by response type, there are three rates generated by this method (Reported, Complete Nonresponse, and Range Follow-up Response). Hokayem et al. (2017) use such a formula to describe response rates for household income in the CPS Annual Social and Economic Supplement. One advantage of Method #2 over Method #1 is that it gives more weight to assets that have larger values, on average. For example, if home equity constitutes a larger proportion of household net worth than bank accounts, then the sum-weighted formula (Method #2) would put more weight on imputed home values than bank account values. Method #1 gives equal weight to home value and bank account values if the household owns both assets, which may be undesirable if a researcher is concerned about assets that constitute a larger proportion of wealth portfolios.

One potential disadvantage of the sum-weighted formula (Method #2) is that respondents with larger values are given more weight than other respondents. To consider a trivial example, suppose a dataset consists of two observation with values of USD 1 and USD 1 million, and only one of the values is imputed. The calculated response rate (the rates of reporting a value) for Method #2 would be nearly 100% (1,000,000/1,000,001) if the USD 1 observation was the imputed value, but the response rate would be nearly 0%

(1/1,000,001) if the USD 1 million observation was imputed (whereas, for Method #1, both cases would result in a calculated response rate of 50%).

This effect also comes into play for wealth data. If high-wealth households have lower response rates than other respondents, then this facet of the data will cause the rate for Method #2 to be higher. Lillard et al. (1986), for example, find that high-income respondents are less likely to answer the wage question in the CPS. The influential effect from respondents with large values may be desirable when looking at statistics based on a mean or sum, such as aggregate income or net worth, as such statistics are influenced heavily by outliers. However, if a researcher is focused on median net worth instead, then this item response statistic may not properly reflect the behavior of respondents in the middle of the distribution.

3.3. Median-Weighted Response Rates

The final formula is computed by taking a weighted average of response rates from each question. This "importance" weight is the percent of individuals who have a non-missing and non-zero value for question q, denoted by o_q, times the median of non-missing and non-zero responses, denoted by $med(x_q)$. Both o_q and $med(x_q)$ can be constructed with survey weights. For assets, o_q represents the ownership rate for a given asset while for income, o_q represents the proportion of people who are receiving a particular source of income.

With this notation, the median-weighted item response rate for composite variable c is denoted by

$$\left(\sum_{q=1}^{Q} d_{c,q} o_q med(x_q) p_{q,k}\right) \Big/ \left(\sum_{q=1}^{Q} d_{c,q} o_q med(x_q)\right), \#M3(c,k)$$

in which $p_{q,k}$ is the item response rate for question q (constructed with survey weights), and the product $o_q med(x_q)$ is the importance weight given to each question. In Method #3, more weight is given to variables with a higher ownership (non-missing and non-zero) rate, and higher median values. Similar to Method #2, Method #3 gives more weight to more "important" variables that constitute a larger proportion of the composite variable. As with Method #2, Method #3 generates three rates for the three different types of response outcomes, and can only be used for surveys that impute data for missing values. However, one benefit of Method #3 over Method #2 is that the statistic is less influenced by outliers, which could be useful in order to gauge the nonresponse behavior of the "typical" respondent, or for analyzing the impact of item nonresponse on median estimates.

To help provide context for the median-weighted formula, Table 2 presents ownership rates and median values in 2014 SIPP for select variables, as well as the importance weight for these variables (before they are normalized to sum to one). This table shows that primary residences are given a large importance weight because 58.9% of households own a home, and the median home value is USD 180,000, which generates an importance weight of $o_q{*}med(x_{i,q}) = 188{,}000{*}58.9 = 10{,}602{,}000$ (these statistics exclude mobile homes, which are captured through a separate variable). The importance weight for 401k and thrift accounts is 17.2% of the weight for primary residence, as the ownership rates are

Table 2. Ownership rates and median values for individual wealth variables (2014 SIPP).

Variable	Ownership rate	Median value	Weight (in thousands)	Weight as a percentage of the primary residence weight
Primary residence	58.9	180,000	10,602	100.0
401k/Thrift retirement account	40.2	45,300	1,821	17.2
IRA/Keogh retirement account	28.1	40,000	1,124	10.6
Savings account (own name)	48.4	1,400	68	0.6
Interest checking (own name)	22.5	2,000	45	0.4
Education savings account (first account)	4.0	10,000	40	0.4
Trusts	1.5	100,000	150	1.4

Table presents ownership rates and median values for select variables from SIPP. In addition, to help explain the median-weighted formula for allocation, this table also shows the weight of the ownership rates times the median value, and the percentage the weight is of the weight for primary residences. Because SIPP is a household survey, I construct the ownership rates and the median values at the household level, even for questions asked to every adult in the household.
Source: 2014 SIPP Panel (Wave 1).

lower and the median value is about USD 45,000. IRAs follow a similar pattern with an importance weight that is 10.6% of the importance weight for primary residence. Other assets presented in this table have a much lower importance weight. Because savings accounts held in someone's own name have a median value of only USD 1,400, the importance weight is 0.6% of the weight for primary residences. Even though trusts have a median value of USD 100,000, the importance weight for trust is 1.4% of the importance weight for primary residences since the ownership rate is 1.5%. In summary, this table shows that when applying the median-weighted formula for all assets and debt in 2014 SIPP, the item response rates for home values and retirement account balances are given a relatively high importance weight, while bank accounts and trusts have a much lower importance weight.

To show how the methods presented in Section 3 can be used to summarize item response rates concisely for key aggregate measures in a survey, the next two sections apply these methods to wealth data. Section 4 provides a comparison of the three methods using 2014 SIPP for a variety of wealth categories. Section 5 presents the results comparing 2008 SIPP, 2014 SIPP, and 2013 SCF to demonstrate how the methods may be compared across surveys.

4. Comparison of Composite Response Rate Methods

Table 3 presents composite response rates for wealth questions in the 2014 SIPP. This table organizes assets into broad categories, such as bank accounts. Figure 1 also presents the same numbers graphically for the response rates for the three methods.

Table 3. Composite response rates for 2014 SIPP.

Category	Percent with no imputed value (Method #1)	Sum-weighted rates (Method #2)			Median-weighted rates (Method #3)		
			Imputed			Imputed	
		Reported	Complete Nonresponse	Range Follow-up Response	Reported	Complete Nonresponse	Range Follow-up Response
All assets and debt	27.7 (0.29)	69.9 (1.66)	18.2 (0.90)	11.9 (1.09)	77.3 (0.30)	17.3 (0.26)	5.3 (0.12)
Retirement assets	52.4 (0.52)	66.3 (0.98)	11.4 (0.44)	22.2 (0.73)	57.2 (0.53)	19.4 (0.35)	23.4 (0.37)
Financial assets (Non-retirement)	42.4 (0.38)	49.7 (1.60)	24.6 (1.15)	25.7 (1.58)	53.4 (0.72)	25.1 (0.62)	21.5 (0.51)
Bank accounts	56.2 (0.37)	52.6 (1.85)	22.0 (1.13)	25.4 (1.92)	60.2 (0.63)	22.0 (0.60)	17.8 (0.46)
Stocks	45.8 (0.81)	48.2 (2.33)	26.3 (1.98)	25.4 (2.17)	47.5 (0.93)	27.7 (0.83)	24.8 (0.69)
Bonds	49.8 (1.34)	42.9 (4.00)	28.0 (3.77)	29.1 (3.24)	47.5 (2.59)	28.1 (1.81)	24.3 (1.84)
Real estate	75.7 (0.34)	83.2 (0.61)	14.5 (0.54)	2.2 (0.39)	83.8 (0.29)	15.1 (0.29)	1.1 (0.07)
Vehicles	63.3 (0.39)	78.0 (0.33)	22.0 (0.33)		79.3 (0.28)	20.7 (0.28)	
Other assets	42.4 (0.50)	50.7 (7.21)	26.2 (4.12)	23.1 (4.72)	52.3 (0.99)	31.0 (1.10)	16.7 (0.79)
Business	52.5 (0.82)	52.1 (10.00)	24.2 (5.56)	23.7 (6.34)	54.7 (1.48)	22.2 (1.36)	23.1 (1.18)
Unsecured debt	68.7 (0.42)	76.1 (0.91)	23.9 (0.91)		75.1 (0.45)	24.9 (0.45)	

Source: 2014 SIPP Panel (Wave 1). Table presents allocation rates across aggregated wealth categories. For many wealth questions, respondents that give an initial answer of "don't know" or "refuse" are asked a range follow-up question (e.g., "is the value less than USD 500, between USD 500 and USD 1,000, or more than USD 1,000"). Because of this, response outcomes are coded as 1) "Reported," in which the respondent gave an answer to the initial question, 2) "Complete Nonresponse," in which the respondent didn't give an answer to either the initial question or the range follow-up, and 3) "Range Follow-Up Response," in which the respondent gave an answer to the range follow-up question. The sum-weighted allocation rates are the ratio of the sum of all values with a given allocation flag divided by the sum of all values for a given wealth category. The median-weighted allocation rates are a weighted average of the allocation rates from the underlying variables, where the weights are the ownership rate times the median value conditional on ownership. Replicate weights were used to construct standard errors, which are shown in parenthesis.

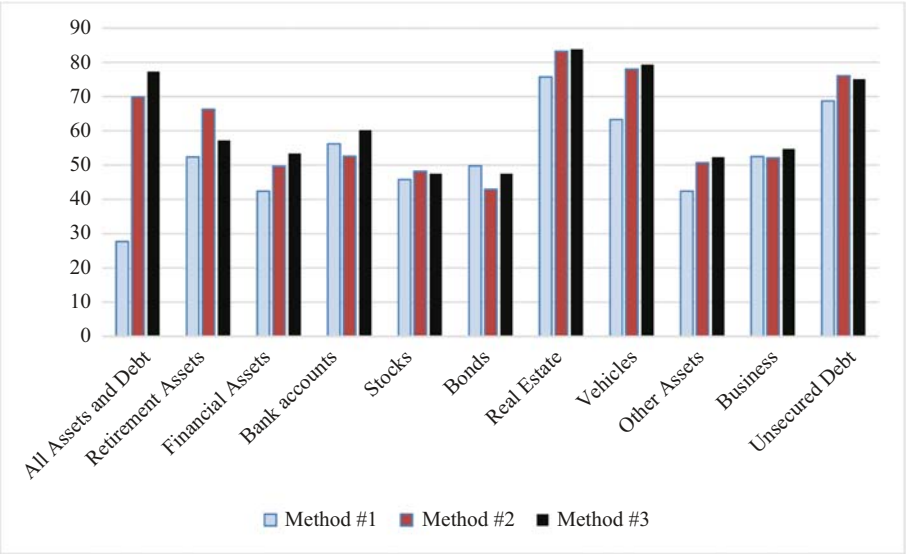

Fig. 1. Response rates from Table 3.
Source: 2014 SIPP Panel (Wave 1). Figure presents the response rates from Table 3 for the three item response rate formulas for composite variables. For Method #2 and #3, the estimate displayed in this figure is the "reported" rate shown in Table 3.

Details about these categories are presented in Appendix (Section 8). Because SIPP is a household survey and some of the composite variables are at the household level, I let the indicator of having no imputed values equal zero if anyone in the household has an imputed value. In addition, for the median-weighted response rate, I construct the ownership rates and the median values at the household level, even for questions asked to every adult in the household. In SIPP, some questions, such as home values, are asked to only one member in the household, while other questions are asked to every individual over 15.

The first column of numbers in Table 3 presents the rates for having no imputed values (Method #1). When looking at any asset or debt variable, 27.7% of households in the 2014 panel have no imputed value. Rates vary across assets, with the rates being high for real estate (75.7%), but lower for stocks (45.8%).

The second set of columns presents the sum-weighted response rates (Method #2). Note that for many wealth questions, respondents who give an initial answer of "don't know" or "refuse" are asked a range follow-up question (e.g., "is the value less than USD 500, between USD 500 and USD 1,000, or more than USD 1,000"). Because of this distinction, response outcomes are coded as 1) "Reported", in which the respondent gave an answer to the initial question, 2) "Complete Nonresponse", in which the respondent didn't give an answer to either the initial question or the range follow-up, and 3) "Range Follow-Up Response", in which the respondent gave an answer to the range follow-up question. For example, the sum-weighed numbers for bank accounts indicate that respondents gave a response to the question 52.6% of the time. For the other outcomes, respondents gave a response to the range follow-ups for bank account 25.4% of the time, and had complete nonresponse 18.2% of the time.

For the sum-weighted response rates (Method #2), the relative pattern between assets in response rates is similar to the no-imputed-value method. For example, the sum-weighted response rates for real estate is higher than the response rate for stocks, which is the same pattern that is found for Method #1. For stocks, even though about half of the data is imputed, about a quarter of the data consists of imputed data from range follow-up responses, so a sizable proportion of respondents are still providing some information about their stock values. While the relative patterns between assets is similar, the overall response rate for all wealth items is very different. Method #2 yields a response rate of 69.9%, even though only 27.7% (Method #1) of households have no imputed asset or debt item. These estimates suggest that while the majority of households have at least one imputed wealth item, the majority of household wealth consists of reported values.

One inherent characteristic of the sum-weighted formula (Method #2) is that more weight is given to observations with larger values. To elaborate on this point, Table 4 shows how the overall sum-weighted reporting rate (Method #2) for asset and debt items varies by net worth quintile. This table shows that for observations in the bottom fifth of the wealth distribution, the overall weighted response rate is 83%, but the response rate drops to 67.5% for the top fifth of the wealth distribution. This result could be driven either by high-wealth individual being less likely to respond for any given question, or because high wealth households hold assets like trusts, which have lower item response rates. To describe the relative weight each quintile has on the sum-weighted response rate (Method #2), Table 4 also shows the percent of aggregate assets that are held by each quintile of the wealth distribution. Aggregate assets are shown because, unlike net worth, assets are greater than or equal to zero for every household, which prevents the percent aggregate statistics from being negative for any quintile. These results show that the top quintile of the wealth distribution holds 89.5% of total assets, suggesting that the item response behavior of the top quintile is given a large weight when looking at item response rates for the entire sample.

In contrast with the sum-weighted formula (Method #2), the median weighted formula (Method #3) is much less sensitive to outliers given the inherent nature of medians. Table 3 shows the median-weighted response rates (Method #3) in the last set of columns. For all assets and debt variables, the response rate is 77.3% with Method #3, compared to 69.9% for the sum-weighted formula (Method #2). The response rate for

Table 4. Response rates and percent of aggregate assets by net worth quintile (Method #2).

Net worth quintile	Percent wealth reported, sum weighted	Percent aggregate assets
1st (Lowest)	83.0 (0.91)	2.1 (0.14)
3rd (Middle)	77.1 (0.63)	8.4 (0.41)
5th (Highest)	67.5 (2.40)	89.5 (0.53)

Table presents allocation rates across all wealth variables by net worth quintile, and the percent of aggregate assets held by each net worth quintile. The allocation rates are the ratio of the sum of all values with a given allocation flag divided by the sum of all values for a given wealth category. Replicate weights were used to construct standard errors, which are shown in parenthesis.
Source: 2014 SIPP Panel (Wave 1).

some other asset groups are higher as well with the median-weighted formula, such as the rate for bank accounts (60.2% versus 52.6%). However, the response rate for retirement assets is actually lower with the median-weighted formula (Method #3) at 57.2%, compared to the rate of 66.3% for the sum-weighted method (Method #2). One explanation for this result could be that people who are retired or close to retirement may be more aware of their retirement account balances than younger workers who are decades away from retirement. If awareness is related to item response rates and if older individuals have higher account balance, these factors could result in people with higher retirement account balances having higher item response rates for this asset. These differences in item response rates between groups would then cause the sum-weighted (Method #2) formula to have higher response rates for retirement accounts, as the people with higher values are the ones with higher response rates in this case. These examples illustrate that there is no fixed relationship between the response rate method and the value of the calculated rate. Each method may produce a higher or lower rate than another depending on the specific characteristics of the composite measure and respondent behavior.

All together, the results from Table 3 suggest that the majority of wealth data consists of reported rather than imputed data, even though the majority of households are missing data for at least one asset or debt question. To help provide more intuition for this point, Table 5 presents the relative weight each of the asset and debt categories have when calculating the median-weight response rate (Method #3) for all asset and debt items. As suggested by earlier results, Table 5 shows that the response rate for real estate (which includes home values) make up 70.3% of the overall response rate, and retirement accounts make up 12.0%. Bonds, which are a more uncommon asset, only make up 0.3% of the overall response rates. The weight real estate is given in Method #3 is reflective of how home equity constitutes a large proportion of many people's net worth. For example, in 2014 SIPP (Wave 1) median net worth is USD 80,039, but when home equity is excluded, the median drops by 68.6% to USD 25,116 (U.S. Census Bureau 2017). While there are some

Table 5. Decomposition of median-weighted formula (Method #3).

Category	Percentage total weight
Retirement assets	12.0 (0.33)
All financial assets not in retirement accounts	5.5 (0.16)
Bank accounts	2.5 (0.09)
Stocks	2.7 (0.13)
Bonds	0.3 (0.06)
Real estate	70.3 (0.46)
Vehicles	6.0 (0.07)
Other assets	4.0 (0.35)
Business	1.3 (0.30)
Unsecured debt	2.2 (0.07)

Table presents the percentage contribution each group's item response rate has when constructing the median-weighted response rate (Method #3). Replicate weights were used to construct standard errors, which are shown in parenthesis.
Source: 2014 SIPP Panel (Wave 1).

assets like trusts that have a response rate of under 50%, many of the assets which constitute a larger proportion of many people's net worth, such as home values, have a much higher response rate. Overall, Table 5 shows that for these particular measures the assets with lower response rates also happen to contribute less to net worth and thus are given less weight in the median-weighted response rate formula (Method #3). Though this pattern may not always be the case, it is likely that an item's importance and its item response rate would often be positively correlated due to increased salience to respondents or increased effort on the part of the survey sponsor to measure it well.

5. Comparison Across Surveys

The previous section focused on how the proposed item response formulas can be used to summarize item response for a key measure in one survey concisely. However, another use of these formulas is to compare item response rates across surveys. Table 6 presents results comparing response rates between 2008 SIPP, 2014 SIPP, and 2013 SCF using the three different methods. For expositional purposes, Table 6 only presents the results for all combined asset and debt categories.

Overall, this table shows that response rates went up in SIPP after the 2014 redesign, but the rates are still lower than SCF rates. For the median-weighted response rate (Method #3), 2008 SIPP has a rate of 68.8%, 2014 SIPP has a rate of 77.3, and SCF has a rate of 85.0%. Of all respondents with some degree of item nonresponse, SCF also has more individuals whose values are imputed from a range (as opposed to being imputed without any respondent reported value information) than SIPP. In 2014 SIPP, 23.3% of respondents who didn't respond initially have a value that is imputed from a range (as computed from the estimates displayed in Table 6), but this rate is 78% for SCF respondents. In this statistic on range follow-ups, the 23.3 estimate for SIPP is not shown in Table 6 but rather is computed as the imputed within range estimate (5.3) divided by the total imputation rate (100-77.3). The SCF estimate is constructed analogously. In addition, SCF allows respondents to give their own bounds for a range follow-up response, which may result in SCF having more people in this response category.

A similar relative pattern between surveys is found for the sum-weighted response rate (Method #2). For the rates of having no imputed values (Method #1), the relative pattern is slightly different. 2014 SIPP has the lowest rate at 27.7%, while 2008 SIPP is in the middle at 29.7%. This difference could be the result of 2014 SIPP having a greater number of asset and debt questions, which would give respondents more opportunity to item nonrespond to at least one question. But overall, the relative differences in response rates are fairly consistent across the three methods, suggesting that the results are not based on idiosyncratic features of any one formula.

Though the nature of the differences in rates across surveys is fairly clear, the reasons for these differences is more uncertain. For example, item response rates for primary residences are different amongst all three surveys (see Table 1), even though the question text for primary residences is similar between the surveys (Eggleston and Gideon 2017). Hard-to-quantify characteristics, like context effects or variation in interviewer training, could also explain the differences between the three surveys.

Table 6. SIPP and SCF composite response rates for all asset and debt questions.

Survey	Percent with no imputed value (Method #1)	Sum-weighted rates (Method #2)			Median-weighted rates (Method #3)		
			Imputed			Imputed	
		Reported	Complete Nonresponse	Range follow-up Response	Reported	Complete Nonresponse	Range follow-up Response
SIPP 2014	27.7 (0.29)	69.9 (1.66)	18.2 (0.90)	11.9 (1.09)	77.3 (0.30)	17.3 (0.26)	5.3 (0.12)
SIPP 2008	29.7 (0.36)	63.9 (1.41)	28.6 (1.52)	7.5 (1.59)	68.8 (0.43)	27.7 (0.41)	3.5 (0.11)
SCF 2013	48.0 (1.30)	75.3 (1.91)	3.51 (0.64)	21.2 (1.96)	85.0 (0.72)	3.36 (0.23)	11.7 (0.73)

Source: 2008 SIPP Panel (Wave 4), 2014 SIPP Panel (Wave 1), and 2013 SCF. Table presents aggregated allocation rates for all asset and debt questions. For many wealth questions, respondents that give an initial answer of "don't know" or "refuse" are asked a range follow-up question (e.g., "is the value less than USD 500, between USD 500 and USD 1,000, or more than USD 1,000"). Because of this, response outcomes are coded as 1) "Reported," in which the respondent gave an answer to the initial question, 2) "Complete Nonresponse," in which the respondent didn't give an answer to either the initial question or the range follow-up, and 3) "Range Follow-Up Response," in which the respondent gave an answer to the range follow-up question. The sum-weighted allocation rates are the ratio of the sum of all values with a given allocation flag divided by the sum of all values for a given wealth category. The median-weighted allocation rates are a weighted average of the allocation rates from the underlying variables, where the weights are the ownership rate times the median value conditional on ownership. Replicate weights used to construct standard errors in all three surveys, and imputation implicates were used to construct the standard error for SCF.

Table 7. Median-weighted response rates (Method #3).

Category	Unit response rate	Median-weighted item response rate (Method #3)	Combined response rate (Product of previous two columns)
SIPP 2014	68.8	77.3 (0.30)	53.2 (0.21)
SIPP 2008	67.6	68.8 (0.43)	46.5 (0.29)
SCF 2013	56.5	85.0 (0.72)	48.0 (0.41)

Source: 2008 SIPP Panel (Wave 4), 2014 SIPP Panel (Wave 1), and 2013 SCF. Table presents survey response rates and allocation rates across aggregated wealth categories. Last column is the multiplication of the two previous columns. The median-weighted allocation rates are a weighted average of the allocation rates from the underlying variables, where the weights are the ownership rate times the median value conditional on ownership. Replicate weights were used to construct standard errors, which are shown in parenthesis.

Another potential reason for the differences in *item* response rates is differences in *unit* response rates. SIPP unit response rates (AAPOR RR6) are 80.6% in the 2008 Panel (Wave 1) but only 68.8% in the 2014 Panel. This decline is consistent with the downward trend in unit responses rates for federal surveys in the United States. For example, the unit response rates in the National Health Interview Survey decreased from 84.9% in 2008 to 73.8% in 2014 (Czajka and Beyler 2016). The 2013 SCF has a response rate (AAPOR RR1) of about 65% for the general population sample, but a rate of about one-third for the high wealth oversample (Bricker et al. 2017). Yan et al. (2010) find that decreasing unit response rates in the Survey of Consumers has been associated with higher income item response rates, potentially suggesting that people who are not interviewed for the survey would also be more likely to have item nonresponse if they responded. This finding would suggest that part of the reason for differences in item response rates is differences in unit response rates.

To account for the effect of unit response when comparing response rates for composite variables across surveys, Table 7 presents the median-weighted response rates (Method #3) multiplied by the unit response rate for the survey. The estimates for 2008 SIPP use the cumulative unit response rate, which is the initial unit response rate at Wave 1 times the total attrition rate as of Wave 4, which is the first wave in which wealth data are collected. The unit response rate for SCF is somewhat difficult to construct since the Federal Reserve doesn't publish one combined unit response rate for the SCF. To construct one, I weight the response rate for the general population sample and the high-wealth sample by their sample sizes, which is listed online in the SCF codebook (Board of Governors of the federal Reserve System 2017) to construct a combined rate of 56.5%. To construct a combined response rate, the ideal method would be to weight by the inverse of the probability of selection rather than sample size. However, given the former information is not available on the publicly available dataset, using the sample size is the next best method.

In this combined rate, the relative standing of SCF changes compared to what is observed in Table 6. The 2008 SIPP still has the lowest rate of 46.5%, but the 2013 SCF is now lower than 2014 SIPP (48.0% versus 53.2%). Thus, when comparing response rates for composite variables across surveys, differences in unit response rates are important to take into consideration as well. If unit response is low, item response may be high, but both factor likely influence data quality and potential for nonresponse bias.

6. Non-Continuous Variables

Most of the article has focused on continuous variables such a household income or wealth. However, some measures, such as the unemployment rate, are based on a series of yes/no and discrete-choice questions, rather than on continuous variables. For such variables, aggregated measures of item response can also be constructed by modifying some of the formulas already presented in Section 3. The statistic of having no imputed values (Method #1) is easy to extend to non-continuous variables, as this method does not rely on the underlying variables being continuous. In addition, the equation for the median-weighted response rate (Method #3) can be modified to remove the median function from the weight to generate a statistic for non-continuous variables. In this revised formula for Method #3, more weight is given to questions that are on-path/in-universe for a larger set of respondents, which similarly gives more weight to more "important" questions. Since Method #2 relies on continuous amounts, this method cannot be modified for discrete-choice questions.

As a simple example, consider a composite response rate for whether someone is disabled. In SIPP, this indicator equals one if any of the six underlying core disability questions (Hearing, Seeing, Cognitive, Ambulatory, Self-Care, Independent Living) has a response of "yes". For the no-imputed-value formula (Method #1), this rate would be the percent of people who item respond to *all* of the disability question. For the modified Method #3, this rate would be the *average* of the response rates to the six disability questions, since the universe of these variables is the same. These two methods may yield slightly different numbers if individuals who item nonrespond to some of the disabilities questions but not others. For example, a proxy respondent who answers questions about another household member may know whether the person has difficulty hearing but be unsure whether the person has any cognitive problems, such as difficulty concentrating. In this case, the no-imputed-value formula (Method #1) would generate slightly lower response rates than the modified third method. As a simplified mathematical example, suppose 10% of the sample item nonresponds to all the disability questions, and a different 10% of the sample item nonresponds to just the cognitive question. In this case, the no imputed value measure would be 80%, but the modified Method #3 would be $(80 + 5*90)/6 = 88.33\%$. In summary, this method of aggregating response rates across questions is applicable to a variety of key measures, including measures based on non-continuous variables.

7. Conclusion

This article presents new methods for aggregating data on item response rates across questions in order to generate statistics of item response for composite measures, such as household income and wealth. These methods 1) provide a concise summary of item response for key estimates and 2) help compare item response rates across surveys, similar to American Association for Public Opinion Research (2016) unit response rates. Each of the three proposed methods have strengths and weaknesses for their inferential ability. Thus, recommendations for which method to use depends on the key statistic of interest. For median wealth, Method #3 may be the most useful, while Method #1 may be sufficient for composite variables composed from a smaller set of items. Table 8 summarizes the

Table 8. Strengths and limitations of each method.

	Percent with no imputed value (Method #1)	Sum-weighted rates (Method #2)	Median-weighted response rates (Method #3)
Strengths	A) Simple to understand B) Useful when only small number of questions are involved	A) Captures variation in item response rates across questions B) Useful when key statistic is a mean or total	A) Captures variation in item response rates across questions B) Less sensitive to outliers than Method #2 C) Useful when key statistics is a median
Limitations	A) Doesn't capture variation in item response rates across questions B) Not as useful when a large number of questions are involved	A) Sensitive to outliers B) Requires imputed values for missing data C) More complicated than Method #1	A) Requires imputed values for missing data B) More complicated than Method #1

advantages and disadvantages of each method. After applying these methods to wealth data, I find item response rates went up in SIPP after the 2014 redesign, but the rates are still lower than in the SCF. However, the 2014 SIPP has higher response rates than the 2013 SCF once unit response rates are incorporated.

This comparison in item response rates amongst surveys is greatly facilitated by my methods for aggregating item response rates. The SCF, for example, has over 140 assets and liabilities questions, so analyzing item nonresponse rates for each of these questions would be tedious and burdensome for researchers and data users. In addition, many of these questions do not have a direct correspondence in SIPP, so direct comparison by question is infeasible. These aggregation methods potentially alleviate these difficulties by combining all asset and debt questions into categories that are comparable across surveys. These methods can be applied to other composite measures, such as the unemployment rate, allowing for a new way of evaluating data quality in key measures.

8. Appendix

In this Appendix, I describe the asset and debt groupings presented in the tables.

1. **Bank Accounts**: Consists of money in checking, savings, money market accounts, and Certificates of Deposit (CDs).
2. **Bonds**: Consists of U.S. Treasury securities, municipal bonds, and corporate bonds held outside of retirement accounts, as well as U.S. savings bonds. For SCF, this category also includes foreign and mortgage-backed bonds.

3. **Stocks**: Consists of shares of stocks and mutual funds held outside of retirement accounts.
4. **Financial Assets**: Consists of all assets in the bank account, bonds, and stock categories.
5. **Business**: Consists of the value and debt of businesses. SIPP asks respondents the percent of the business that they own. I use this variable to construct the business value for the household, but I do not incorporate the item nonresponse status of the percent owned variable when calculating the response rate for businesses.
6. **Other Assets**: For both 2014 SIPP Panel and SCF, this consists of the cash value of life insurance policies, annuities, trusts, and the value of all other assets captured in a catchall question. For 2008 SIPP, the measure consists of only values from a catchall question, as the annuity and trust questions were added in the 2014 Panel. The 2008 measures also excludes the cash value of life insurance policies, as this variable was excluded from net worth calculations because many respondents conflated cash value and face value of life insurance (Gottschalck and Moore 2007). The 2013 SCF measure also includes money owed to the respondent by friends, family, or businesses.
7. **Retirement Assets**: Consists on money in Individual Retirement Accounts (IRAs), Keogh accounts, and 401k/thrift accounts.
8. **Real Estate**: Consists of the value of primary residences; rental property; and other real estate, such as timeshares and vacation properties. The SCF has a question about the percent of the other real estate the respondent owns. I use this variable to construct the real estate value for the household, but I do not incorporate the item nonresponse status of the percent owned variable when calculating the response rates for real estate.
9. **Vehicles**: Consists of cars, trucks, SUVs, and recreational vehicles such as motorcycles, boats, and RVs.
10. **Unsecured Debt**: For 2014 SIPP, this consists of credit cards, student loans, and other debt. For 2008 SIPP, there was no separate question on student loans, but student loans should be included with "other debt." The SCF measure includes everything collected in the 2014 SIPP as well as data on other consumer loans and lines of credit.

9. References

American Association for Public Opinion Research. 2016. "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition." AAPOR. Available at: https://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx. (accessed April 2019).

Bee, A. and J. Mitchell. 2017. "Do Older Americans Have More Income Than We Think?" *SESHD Working Paper* 2017, No. 39. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2017/demo/SEHSD-WP2017-39.pdf. (accessed April 2019).

Board of Governors of the Federal Reserve System. 2017. Codebook for 2016 Survey of Consumer Finances. Available at: https://www.federalreserve.gov/econres/files/codebk2016.txt. (accessed April 2019).

Bollinger, C.R., B.T. Hirsch, C.M. Hokayem, and J.P. Ziliak. 2015. "Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch." *University of Kentucky Center for Poverty Research Discussion Paper Series*, no. 120. Available at: https://uknowledge.uky.edu/ukcpr_papers/120/. (accessed April 2019).

Bound, J. and A.B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9: 1–24. Doi: http://dx.doi.org/10.1086/298256.

Bricker, J., L.J. Dettling, A. Henriques, J.W. Hsu, L. Jacobs, K.B. Moore, S. Pack, J. Sabelhaus, J. Thompson, and R.A. Windle. 2017. "Changes in U.S. Family Finances from 2013 to 2016: Evidence from the Survey of Consumer Finances." *Federal Reserve Bulletin* 100(4). Available at: https://www.federalreserve.gov/publications/2017-September-changes-in-us-family-finances-from-2013-to-2016.htm. (accessed April 2019).

Czajka, J.L. and A. Beyler. 2016. *Background Paper Declining Response Rates in Federal Surveys: Trends and Implications*. Washington: Mathematica Policy Research. Available at: https://aspe.hhs.gov/system/files/pdf/255531/Decliningresponserates.pdf. (accessed April 2019).

Eggleston, J. and M. Gideon. 2017. "Evaluating Wealth Data Quality in the Redesigned 2014 Panel of the Survey of Income and Program Participation." *SIPP Working Paper*, no. 278. Available at: https://www.census.gov/library/working-papers/2017/demo/SEHSD-WP2017-35.html. (accessed April 2019).

Ferber, R. 1966. "Item Nonresponse in a Consumer Survey." *Public Opinion Quarterly* 30(3): 399–415. Doi: https://doi.org/10.1086/267432.

Gottschalck, A.O. and J.C. Moore. 2007. "Evaluation of Questionnaire Design Changes on Life Insurance Policy Data." *U.S. Census Research Report Series, Survey Methodology* 2007(14). Available at: https://census.gov/library/working-papers/2007/adrm/rsm2007-14.html. (accessed April 2019).

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias a Meta-Analysis." *Public Opinion Quarterly* 72(2): 167–189. Doi: https://doi.org/10.1093/poq/nfn011.

Hokayem, C., T. Raghunathan, and J. Rothbaum. 2017. "Ignorable Nonresponse? Improved Imputation and Administrative Data in the CPS ASEC." Paper presented at the *Association for Public Policy Analysis and Management Fall Research Conference, November 2–4*, 2017. Available at: https://appam.confex.com/appam/2017/webprogram/Paper23281.html. (accessed April 2019).

Hsu, J.W. M.D. Schmeiser, C. Haggerty, and S. Nelson. 2017. "The Effect of Large Monetary Incentives on Survey Completion: Evidence from a Randomized Experiment with the Survey of Consumer Finances." *Public Opinion Quarterly* 81(3): 736–747. Doi: https://doi.org/10.1093/poq/nfx006.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What do we really know about wages? The importance of nonreporting and census imputation." *Journal of Political Economy* 94, no. 3, Part 1: 489–506. Doi: https://doi.org/10.1086/261386.

National Research Council. 2009. Reengineering the Survey of Income and Program Participation. Constance F. Citro and John Karl Scholz, eds. Committee on National Statistics, Division of Behavioral and Social Sciences Education. Washington, DC: The National Academies Press.

U.S. Census Bureau. 2016. *Survey of Income and Program Participation 2014 Panel Users' Guide*. Washington, D.C., U.S.A. Available at: https://www.census.gov/programs-surveys/sipp/guidance/users-guide.html. (accessed April 2019).

U.S. Census Bureau. 2017. *Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2013*. Washington, D.C. U.S.A. Available at: https://www.census.gov/topics/income-poverty/wealth/data/tables.html. (accessed April 2019).

Watson, D., L.A. Clark, and A. Tellegen. 1998. "Development and Validation of Brief Measures of Positive and Negative Affect: the PANAS scales." *Journal of Personality and Social Psychology* 54(6): 1063–1070. Doi: http://doi.apa.org/journals/psp/54/6/1063.html.

Westra, A., M. Sundukchi, and T. Mattingly. 2015. "Designing a Multipurpose Longitudinal Incentives Experiment for the Survey of Income and Program Participation." In Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference, December 1–3, 2015. Washington, DC: Federal Committee on Statistical Methodology. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2016/03/E3_Westra_2015FCSM.pdf. (accessed April 2019).

Yan, T., R. Curtin, and M. Jans. 2010. "Trends in Income Nonresponse Over Two Decades." *Journal of Official Statistics* 26(1): 145–164. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/trends-in-income-nonresponse-over-two-decades.pdf. (accessed April 2019).

Received April 2018
Revised November 2018
Accepted December 2018

sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 409–460, http://dx.doi.org/10.2478/JOS-2019-0019

Validation of Two Federal Health Insurance Survey Modules After Affordable Care Act Implementation

Joanne Pascale[1], *Angela Fertig*[2], *and Kathleen Call*[3]

This study randomized a sample of households covered by one large health plan to two different surveys on health insurance coverage and matched person-level survey reports to enrollment records. The goal was to compare accuracy of coverage type and uninsured estimates produced by the health insurance modules from two major federal surveys – the redesigned Current Population Survey Annual Social and Economic Supplement (CPS) and the American Community Survey (ACS) – after implementation of the Affordable Care Act. The sample was stratified by coverage type, including two types of public coverage (Medicaid and a state-sponsored program) and three types of private coverage (employer-sponsored, non-group, and marketplace plans). Consistent with previous studies, accurate reporting of private coverage is higher than public coverage. Generally, misreporting the wrong type of coverage is more likely than incorrectly reporting no coverage; the CPS module overestimated the uninsured by 1.9 and the ACS module by 3.5 percentage points. Other differences in accuracy metrics between the CPS and ACS are relatively small, suggesting that reporting accuracy should not be a factor in decisions about which source of survey data to use. Results consistently indicate that the Medicaid undercount has been substantially reduced with the redesigned CPS.

Key words: Health insurance; validation; Affordable Care Act; marketplace.

1. Introduction

Surveys are the only source of data on the uninsured rate in the United States. The Affordable Care Act (ACA) introduced a certain amount of federal monitoring of insurance status through standardized Internal Revenue Service (IRS) forms (the 1095), but the potential for estimating insurance status from IRS data is in the early stages of exploration (Lurie and Pearce 2018). Thus, surveys remain the only source, and they are not without measurement error. For example, studies from the 1990s found the US uninsured rate ranged from a low of about 8% up to a high of almost 18% depending on the source (Bennefield 1996; Lewis et al. 1998; Rosenbach and Lewis 1998). Surveys generally derive the estimate of the uninsured by asking about coverage through a range of different sources or types of coverage, and then designating those with no reported coverage as uninsured. Therefore, to assess the accuracy of the uninsured estimate, misreporting of a broad range of plan types needs to be considered collectively. Put

[1] U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20233. U.S.A. Email: joanne.pascale@census.gov
[2] University of Minnesota, Humphrey School of Public Affairs, 130 Humphrey School, 301 19th Ave S, Minneapolis, 55455, U.S.A. Email: arfertig@umn.edu
[3] State Health Access Data Assistance Center, 2221 University Ave SE #345, Minneapolis, MN 55414. U.S.A. Email: callx001@umn.edu

another way, ". . .the uninsured are a residual group by definition. They are the people who fall in the cracks left by public and private insurance programs. . . As a result, one cannot produce or make sense of statistics about the uninsured without first producing or making sense of statistics about the insured." (Farley-Short 2001, 4)

Challenges in measuring health insurance in surveys have been well-documented since the 1980s (Blewett and Davern 2006; Lewis et al. 1998; Pascale 2008; Swartz 1986). For example, Medicaid is a major public insurance program for low-income families, and numerous studies have documented consistent and persistent under-reporting across a range of surveys (Blumberg and Cynamon 1999; Call et al. 2013; Czajka and Lewis 1999; Eberly et al. 2009; Klerman et al. 2009; Pascale et al. 2009; Rosenbach and Lewis 1998). Employer-sponsored insurance (ESI) dominates private coverage, and those without access to coverage through an employer, group or association often opt to purchase coverage directly from the insurer, which is known as non-group coverage. There is some evidence that reporting of private coverage is fairly accurate overall (Hill 2007; Nelson et al. 2000). However, other evidence suggests that non-group coverage is over-reported (Cantor et al. 2007), and that non-comprehensive non-group coverage (e.g., dental and vision plans) is often reported in tandem with another comprehensive plan, most often ESI coverage (Mach and O'Hara 2011).

In spite of the extensive research on public coverage, Medicaid has been studied in relative isolation from other plan types, and it is not entirely clear how misreporting of Medicaid affects estimates of other plan types, or whether over-reporting of Medicaid among those who are not enrolled may offset some Medicaid under-reporting. Studies that explore reporting accuracy of *both* public and private coverage, and how misreporting of one affects the other, are extremely rare. We know of only two (Davern et al. 2008; Nelson et al. 2000), and results are provocative. For example, ESI is by far the most prevalent source of coverage in the United States, so if even a small percentage of ESI enrollment is misreported as, say, public, this artificially inflates the public coverage estimate and offsets, to a large extent, the under-reporting of Medicaid, as was demonstrated by Davern et al. (2008). While results from these studies are highly valuable, both precede implementation of the ACA and neither examined the question series employed in major federal government surveys.

Indeed, the ACA added considerable complexity to the already complicated task of accurately categorizing health insurance coverage from surveys (Pascale 2016). One factor was the introduction of the "marketplace." This term has come to have a dual meaning. It is both a portal (aka: healthcare.gov) through which people can shop for and enroll in a range of coverage options – both public and private – and the term is commonly used to describe the coverage itself: non-group/direct-purchase coverage for which many enrollees receive a subsidy for the monthly premium. The ACA also further blurred the line between public and private coverage. Public and private coverage are often distinguished from each other by the party responsible for paying the monthly premium; if individuals and/or employers pay, the coverage is considered private, and if a government entity pays, it is considered public. However, even before the ACA, many states offered public programs (e.g., Children's Health Insurance Program (CHIP) plans) that required individuals to pay a monthly premium. Post-ACA, Medicaid eligibility was expanded in many states, but required premium contributions in some cases. To muddy the waters

further, marketplace coverage is sometimes fully subsidized by the government (i.e., the monthly premium is USD 0), but is still considered private. Another complicating factor is the "no-wrong-door" design of the portal. One objective of the portal was to make it easy for those seeking coverage to explore and obtain coverage anywhere on the spectrum from fully subsidized public coverage to unsubsidized marketplace coverage, depending on their eligibility. Thus, enrollees could begin their search for coverage expecting to be eligible for, say, subsidized private coverage, but end up qualifying for public coverage. All these issues – the dual meaning of the term marketplace, the blurry line between public and private coverage, and the no-wrong-door design of the portal – complicate the task of categorizing coverage type from survey data (Pascale et al. 2013).

Two major federal surveys that researchers and policymakers rely on for estimates of health insurance coverage are the Current Population Survey Annual Social and Economic Supplement (CPS) and the American Community Survey (ACS). In response to many of the measurement error issues noted above, after more than a decade of research and testing, the CPS was redesigned beginning with calendar year 2013 estimates (Pascale 2016; Pascale et al. 2016). In preparation for full implementation of the ACA in 2014, research was conducted to adapt the newly-redesigned CPS for marketplace coverage (Pascale et al. 2013). Research on adapting the ACS is ongoing and no ACA-specific changes have yet been made to the questionnaire; it is expected that respondents with marketplace coverage will report it as private, non-group coverage.

This study extends past research by measuring and comparing reporting accuracy of coverage type and the overall uninsured rate in the CPS and ACS in a post-ACA era. This is important, given the role of these two surveys in the research and policy arenas, the gaps in the literature on measurement error discussed above, and the relatively uncharted territory of reporting accuracy post-ACA. Two key aspects of the study are: (1) it uses survey data matched to enrollment records as a "truth source," and (2) the enrollment records cover multiple types of coverage, both public and private. Specifically, we examine two types of public coverage (Medicaid and a program called MinnesotaCare – a state-specific program for low-income families that charges a sliding-fee premium) and three types of private coverage (ESI, and non-group coverage within and outside the marketplace). This is a rare opportunity and gives us the chance to examine multiple dimensions of misreporting. The study extends research on reporting accuracy beyond Medicaid to address multiple types of public and private coverage. It also allows us to explore how misreporting of one type affects another. Specifically, most prior research has focused on the question of under-reporting: among those enrolled in coverage type X according to records, how many fail to report coverage type X in a survey? This design allows us to go beyond that question and examine, for example, if coverage type X was not reported, what coverage type, if any, *was* reported?

Data for this research come from the CHIME study (Comparing Health Insurance Measurement Error), a reverse record check study in which enrollment records were used to sample households with individuals known to be enrolled in various types of private and public coverage. Phone numbers associated with these households were randomly assigned to either the CPS or ACS health insurance module, and a brief split panel household-level telephone survey was conducted in the spring of 2015. Person-level matching was conducted to assess agreement between the survey data and the enrollment

records for individuals in the household. In terms of time period of coverage, both the CPS and ACS ask about coverage on the day of the interview, rendering a point in time (PIT) estimate. (The CPS also collects data on coverage from the beginning of the prior calendar year up to the interview date, but because the ACS is limited to point-it-time, this analysis focuses only on PIT estimates).

The ultimate objective of the current analysis is to use enrollment records from a private health plan as a "truth source" to evaluate and compare reporting accuracy of both coverage type and the net uninsured estimate at a point in time in the CPS and ACS. Three different reporting accuracy metrics were analyzed: under-reporting (enrollment records indicated coverage type X, but coverage type X was not reported in the survey); over-reporting (coverage type X was reported in the survey, but it could not be verified in the enrollment records); and prevalence (the estimate of coverage type X from the enrollment records compared to the estimate from the survey).

2. Methods

The CHIME study was multi-faceted and addressed several research questions, only some of which are the focus of this article. Below are highlights of the methodology relevant to this analysis, and complete study design details are documented in Fertig et al. (2018). As was noted in that paper, a common critique of record linkage studies is that administrative records come with their own sources of error. To mitigate this, we worked in close collaboration with informatics staff affiliated with the health insurer to maximize the veracity of the records data (e.g., by carefully examining and resolving duplicate records). Thus, we label the records as the "gold standard" and use terms like "accuracy" and "truth." However, we emphasize the quotes around these terms and we invite skeptical readers to interpret the results as simply a comparison of two data sources.

2.1. Sample

The study surveyed a stratified random sample of households known to have health insurance through one large regional insurer in the Midwest. At the time of data collection, the private health insurer offered all the major categories of private and public coverage: ESI, non-group outside the marketplace, marketplace coverage, and two types of public coverage: Medicaid and MinnesotaCare. The health insurer provided a sample of Minnesota households from each of these five coverage types or strata, as well as a "transition" strata of policyholders who switched from ESI to public or vice versa in 2014. Households were included in the sample if the home address was in Minnesota, the enrollment records included a phone number, and at least one eligible policyholder resided in the household. Eligible policyholders were under age 65 and belonged to one of the coverage type strata in December 2014, when the sample was drawn. At the time of the sample draw, there were just under 700,000 individual members across the six strata (see Table 1), and of those, roughly 270,000 were eligible policyholders. Among these policyholders, roughly 175,000 had a telephone number, and after removing duplicate addresses there were about 130,000 unique eligible households from which to sample.

To determine total sample size we began with the budget, which we estimated would support data collection yielding 5,000 completed household interviews. We assumed a

Table 1. Completed matched household- and person-level sample size by strata.

Strata[1]	Insurer population[2] (in thousands)		Households matched[3]		People matched[4]					
					Total		CPS		ACS	
	N	%	N	Match Rate	N	%	N	%	N	%
ESI	463	66.7%	309	83%	561	14.7	313	15.7	248	13.5
Medicaid	181	26.1%	481	83%	908	23.8	432	21.7	476	26.0
MinnesotaCare	26	3.7%	447	88%	635	16.6	336	16.9	299	16.3
Marketplace	1.7	0.2%	249	93%	330	8.6	178	8.9	152	8.3
Non-group	22	3.2%	698	88%	1,178	30.8	640	32.2	538	29.3
Transition	3	0.4%	122	90%	211	5.5	90	4.5	121	6.6
TOTAL	696.7	100%	2,306	87%	3,823	100%	1,989	100%	1,834	100%

[1]ESI refers to employer sponsored insurance; MinnesotaCare is a state-specific program for low-income families that charges a sliding-fee premium; Marketplace is non-group/direct-purchase coverage available on the portal for which many enrollees receive a subsidy for the monthly premium; Non-group is insurance that is purchased directly, not through an employer group or association and not on the portal; the Transition strata is comprised of policyholders who switched from ESI to public or vice versa in 2014.
[2]Distribution of people insured by the health plan in each strata at the time of the sample draw, in December 2014.
[3]Match rate reflects the percent of surveyed households in that strata in May/June 2015 that had at least one matched individual.
[4]Distribution of people per strata matched at the time of data collection, in May/June 2015.

response rate of about 30% and calculated we would need an initial sample size of 16,000 phone numbers. To determine how to allocate the sample across strata, we made assumptions about average household size and rates of under-reporting to conduct a power analysis with a threshold of 0.80. We aimed for a minimum detectable difference of about 2.5 percentage points in each stratum, but in two strata (marketplace and MinnesotaCare) the number of available households in the universe was insufficient to meet this goal. Thus, we sampled the entire universe for these two strata, but the minimum detectable difference for them was somewhat higher than ideal (about 5 percentage points for each).

The health insurer required that an advance letter be mailed informing eligible households that they were partnering with the Census Bureau on a study. The letter invited members who did not wish to participate to opt-out by calling in or writing to the health insurer's call center. Based on assumptions about opt-outs and bad address rates, the health insurer mailed a total of 22,000 advance letters with a goal of achieving 16,000 phone numbers of eligible and willing policyholders. We allowed about a month between the mail date of the letter and delivery of the final sample of members to Census; less than 6% of the letters were returned as a bad address or resulted in an opt-out. The final sample of 16,000 phone numbers was delivered to the Census Bureau in December 2014, for processing and preparation for data collection.

2.2. *Data Collection*

All interviews were conducted by Census Bureau telephone interviewers at the Hagerstown, Maryland, facility. Average administration time was 17 minutes. Data collection occurred during two distinct but consecutive three-week field periods from May 20 until June 28, 2015. In order to minimize interviewer effects, interviewers were assigned to one of two groups: each interviewer group was initially trained on either the ACS or CPS health insurance module and worked exclusively on that version during the first field period. At the end of the first field period, the interviewers switched questionnaire treatments and received a brief training on the new health insurance module and worked exclusively on that version during the second field period. We collected data from 2,660 households representing 6,644 people and a response rate of 22% using an adapted version of AAPOR RR 4 (American Association of Public Opinion Research 2016). Specifically, the RR4 reduces the denominator by including only a proportion of households with unknown eligibility (i.e., "unknown if occupied/household" and "other/unknown"). In the CHIME data collection, households of unknown eligibility included "unknown if occupied/household", "other/unknown", noncontacts, and "other" dispositions because these four groups were comingled.

After completion of the survey, in August 2015, the health insurer sent the Census Bureau a second file with data on every individual insured by the health insurer ($n = 35{,}591$) in the 16,000 households from the original sample, including enrollment data reflecting coverage in May and June of 2015. This ensured that the time period of coverage asked about in the survey was perfectly aligned with the time period indicated in the records.

We used a computer-match algorithm to link the enrollment person-record to its corresponding survey person-record for several reasons. First, there was some lag time

between the date the sample was selected and the interview date, so the original phone numbers could have been reassigned to a different household and/or the insured member(s) could have moved out of the household. Second, among phone numbers that matched at the household level, it is possible that not all household members were insured by this health insurer. Thus the person-level computer match was conducted using variables on both datasets: phone number, name, sex, date of birth and address. Clerical review of borderline matches was also conducted to ensure accurate matches. The number of matched households and people by strata are shown in Table 1. We were able to match at least one person in 87% of surveyed households. Fifty-eight percent of individuals with survey data were matched to an enrollment record. However, as members of one household may be covered by different health plans (or some may be covered and others not), many of the individuals in the survey data may not have a match in the enrollment records from this health plan. All households with at least one matched individual ($n = 2{,}306$) were included in the CHIME study.

2.3. Nonresponse Analysis

To assess whether the matched households were different from non-matched and nonrespondent households, we compared characteristics from the enrollment records in households with at least one matched person ($n = 2{,}306$) to households where no members were matched ($n = 13{,}694$) – either because no one in the household responded to the survey ($n = 13{,}340$) or because there was a completed interview for the household but no person-record matched to the enrollment records ($n = 354$).

As detailed in Fertig et al. (2018), compared to non-matched households, in matched households the policyholder was older (41.6 versus 34.5 years old, $p < 0.001$), was more likely to have moderate health risk (47% versus 43%, $p < 0.001$) and less likely to have low health risk or be a healthy user (27% versus 33%, $p < 0.001$), and there were fewer children (0.5 versus 0.7, $p < 0.001$) enrolled with the insurance company. The percent of female policyholders (51%) and the number of adult members of the household (1.4) was the same for both groups.

2.4. Demographics Across Treatment Groups

Demographic characteristics of matched individuals were compared across treatments and for most characteristics there were no significant differences (see Appendix A, Subsection 7.1). The exceptions were that, compared to CPS individuals, ACS individuals were more likely to reside in households with five or more persons, were slightly more likely to be Hispanic or other race, and were more likely to have a family income that is 139–199% of the federal poverty level (FPL). We adjust for these demographic differences across treatment arms in our analysis. Specifically, we run logistic regression models to determine whether the difference in reporting accuracy for CPS and ACS respondents was statistically significant when controls for family size, race/ethnicity, and family income were included in the model. We also used the coefficient estimates from these models to predict the likelihood of accurate reporting for ACS respondents if they had the same characteristics as CPS respondents (Tables 2 and 3 present ACS adjusted results; see Appendix B (Subsection 7.2) for CPS and unadjusted ACS results).

Table 2. Under-[1] and over-reporting,[2] prevalence estimates[3] and differences across CPS and ACS[4], standard sample.

Coverage type in records[5]	Under-reporting				Over-reporting				Prevalence estimates[6]							
	%		Difference		%		Difference		%		Difference		%		Difference	
	CPS	ACS	CPS-ACS		CPS	ACS	CPS-ACS		Recs	CPS	CPS-Recs		Recs	ACS	ACS-Recs	
Private	1.2	3.5	−2.3	***	2.3	6.7	−4.3	***	71.8	72.6	0.9	**	69.8	70.7	1.0	***
ESI	1.9	4.4	−2.5	***	2.8	5.4	−2.7	***	67.9	68.5	0.6		66.0	65.4	−0.7	***
NongMkt	22.3	15.0	7.2	***	44.5	40.6	3.9	***	3.8	5.4	1.6	***	3.7	6.4	2.6	***
Public	16.8	16.8	0.0		2.1	8.6	−6.5	***	28.4	24.1	−4.3	***	30.3	27.4	−2.9	***
Insured	1.9	3.5	−1.6	***	n/a	n/a	n/a	n/a	100.0	98.1	−1.9	***	100.0	96.5	−3.5	***

*** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$; n/a = not applicable.

[1]Under-reporting = false negatives or the % of those known to have Coverage Type X for whom Coverage Type X is not reported in the survey.

[2]Over-reporting = false positive or the % of those for whom Coverage Type X is reported, but who (a) cannot be validated in the enrollment records to have Coverage Type X and (b) can be validated in the enrollment records to have Coverage Type Y.

[3]Prevalence = survey estimates of Coverage Type X versus prevalence of Coverage Type X indicated in the enrollment records.

[4]ACS estimates are predicted based on adjustments for varying demographics across treatments (see Subsection 2.4). Appendix B (Subsection 7.2) shows unadjusted estimates.

[5]Private coverage is the aggregate of employer sponsored insurance (ESI), non-group insurance purchased outside the marketplace (Nong) and within the marketplace (Mkt); NongMkt is non-group and marketplace coverage combined; Public insurance is Medicaid in the Standard Sample; Insured includes both private and public insurance, but does not include health insurance provided through the military, the Indian Health Service, or Medicare.

[6]The prevalence of public and private coverage indicated in the enrollment records may sum to more than 100 percent because some individuals may have both private and public coverage. The estimated prevalence of public and private coverage sums to less than the estimated insured prevalence because insurance provided through the military, the Indian Health Service, or Medicare may have been reported as the insurance type in the survey but was not categorized as public or private coverage in this analysis, but the estimated insured prevalence includes all individuals who were not reported as uninsured.

Table 3. Under-[1] and over-reporting,[2] prevalence estimates[3] and differences across CPS and ACS[4], augmented sample.

Coverage type in records[2]	Under-reporting				Over-reporting				Prevalence estimates[6]							
	%		Difference		%		Difference		%		Difference		%		Difference	
	CPS	ACS	CPS-ACS		CPS	ACS	CPS-ACS		Recs	CPS	CPS-Recs		Recs	ACS	ACS-Recs	
Private	1.2	3.8	−2.6	***	4.2	9.3	−5.1	***	68.1	70.3	2.1	***	65.8	68.5	2.7	***
ESI	1.9	4.6	−2.7	***	3.0	6.1	−3.2	***	65.5	65.2	−0.3	*	62.3	61.9	−0.4	***
NongMkt	22.3	15.0	7.2	***	54.9	54.2	0.6	**	3.7	6.3	2.6	***	3.5	7.7	4.2	***
Public	19.2	22.0	−2.8	***	1.8	7.0	−5.2	***	32.0	26.3	−5.7	***	34.6	29.0	−5.6	***
Insured	2.0	3.8	−1.8	***	n/a	n/a	n/a	n/a	100.0	98.0	−2.0	***	100.0	96.2	−3.8	***

*** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$; n/a = not applicable.

[1]Under-reporting = false negatives or the % of those known to have Coverage Type X for whom Coverage Type X is not reported in the survey.

[2]Over-reporting = false positive or the % of those for whom Coverage Type X is reported, but who (a) cannot be validated in the enrollment records to have Coverage Type X and (b) can be validated in the enrollment records to have Coverage Type Y.

[3]Prevalence = survey estimates of Coverage Type X versus prevalence of Coverage Type X indicated in the enrollment records.

[4]ACS estimates are predicted based on adjustments for varying demographics across treatments (see Subsection 2.4). Appendix B (Subsection 7.2) shows unadjusted estimates.

[5]Private coverage is the aggregate of employer sponsored insurance (ESI), non-group insurance purchased outside the marketplace (Nong) and within the marketplace (Mkt); NongMkt is non-group and marketplace coverage combined; Public insurance is Medicaid and MinnesotaCare in the Augmented Sample; Insured includes both private and public insurance, but does not include health insurance provided through the military, the Indian Health Service, or Medicare.

[6]The prevalence of public and private coverage indicated in the enrollment records may sum to more than 100 percent because some individuals may have both private and public coverage. The estimated prevalence of public and private coverage sums to less than the estimated insured prevalence because insurance provided through the military, the Indian Health Service, or Medicare may have been reported as the insurance type in the survey but was not categorized as public or private coverage in this analysis, but the estimated insured prevalence includes all individuals who were not reported as uninsured.

2.5. Weights

Our sample distribution across strata was driven by the goal of maximizing the ability to detect differences across treatments in reporting accuracy. Thus, by design, the distribution of sample across strata does not reflect any particular population. For the analysis dataset to be a useful reflection of a given population, we followed the only precedent we know of (Davern et al. 2008) and created weights to make the coverage type distribution match that of the original sampling frame – that is, the distribution of the total population of the health insurer (Table 1, second column). Because distributions were not identical in the CPS and ACS, we created separate weights for each arm. All results are presented as weighted percentages of the population.

2.6. Questionnaires

To set the context for the health insurance series of questions, the CHIME survey instrument began with a subset of items included in both the CPS and ACS on demographics, labor force and unearned income. The question wording of these three modules was identical across treatments, and after the unearned income module, half the respondents were randomly assigned to the CPS health insurance module and the other half to the ACS health insurance module.

Under both the CPS and ACS survey designs, a single household respondent is asked to answer health insurance questions for all household members. However, there are some key differences in the modules. First is with regard to structure. The CPS begins with general questions on source or type of coverage and then narrows down to capture the needed detail, while the ACS asks directly about discrete types of coverage. See Figure 1 for an abbreviated version of the questions, and see Appendix C (Subsection 7.3) for the complete health insurance modules. A second key difference is with regard to detail. The CPS includes questions that enable non-group coverage obtained outside the marketplace to be distinguished from marketplace coverage (see items 11–13 in Figure 1), and it includes questions to distinguish Medicaid from MinnesotaCare (see items 9–13 in Figure 1), while the ACS questions do not capture these details. Thus for all ACS-CPS comparisons, we aggregate non-group and marketplace coverage into a single category.

2.7. Categorizing Coverage Type

While categorizing a respondent's source of coverage is straightforward in the ACS given the module's structure, the CPS is considerably more complicated. A separate analysis of the CPS exploited the enrollment records in the CHIME study to guide an algorithm for classifying coverage type. Answers to questions about features of the coverage (such as source, type of government/state plan and name of government/state program), and questions about the marketplace, premiums, and subsidization carefully evaluated (Pascale et al. 2018b) and used to classify the coverage into ESI, non-group, marketplace or public coverage (Pascale et al. 2018a). Once we had these disaggregated categories for the CPS, we created semi-aggregated categories in the CPS to match the ACS categories, and finally created aggregated private and public categories for a comparative analysis on the following individual and aggregated categories of coverage:

1. Private (ESI and/or non-group and/or marketplace coverage)
2. Public coverage (Medicaid and/or MinnesotaCare)

3. Employer-sponsored insurance (ESI)
4. Non-group and/or marketplace coverage
5. Uninsured

2.8. Analysis Samples and Monthly Premium Contributions

The MinnesotaCare program provided us the opportunity to explore reporting accuracy for a public program that requires enrollees to contribute to the monthly premium.

CPS

Logic Check 1: If disabled or age=65+ ➔1; else ➔2

1. Are you covered by Medicare?
   - Yes ➔14
   - No ➔2
2. Are you NOW covered by any type of health plan?
   - Yes ➔ 3
   - No ➔Qs on Medicaid and other public plans; verify currently uninsured ➔18
3. Is it provided thru a job, govt, or other way?
   - Job ➔6
   - Government ➔4
   - Other way ➔7
4. Is that plan related to a JOB with the government?
   - Yes ➔6
   - No ➔5
5. Is that Medicaid/CHIP, Medicare, military, other?
   - Medicaid/CHIP/other/DK ➔9
   - Military ➔[type of military plan] ➔10
   - Medicare ➔14
6. Is the plan related to military service in any way? [if yes, type of military plan] ➔10
7. How is it provided – parent/spouse, direct, other?
   - Parent/spouse/direct ➔10
   - Other ➔8
8. Is it thru former emp, union, group, assn, school?
   - Former emp/union/group/assn/school ➔10
   - Other ➔ 9
9. What do you call the program?
   - Medicaid
   - Medical Assistance
   - Indian Health Service
   - MinnesotaCare
   - Minnesota Comprehensive Health Association
   - PMAP
   - Healthcare.gov
   - Plan through MNsure
   - Other government plan
   - Other (please specify)

   ➔11
10. Who is the policyholder? [If direct in Q7 ➔ 11; else ➔ 14]
11. Is that coverage thru the marketplace?
12. Is there a monthly premium? [if yes ➔ 13; else ➔14]
13. Is the premium subsidized based on family income?
14. [Questions on past months of coverage]
15. Any [other] coverage Jan 2014 till now?
    - Yes ➔loop thru series again, starting with 3
    - No ➔Logic Check 2 for next person on roster

Logic Check 2: For this next person, if any coverage was already reported, start with Q15; else start with Logic Check 1; If no more people on roster ➔END

ACS

1. Are you currently covered by health insurance through a current or former employer or union?
   - Yes
   - No
2. Are you currently covered by health insurance purchased directly from an insurance company?
   - Yes
   - No
3. Are you currently covered by Medicare, for people age 65 or older or people with certain disabilities?
   - Yes
   - No
4. Are you currently covered by Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability?
   - Yes
   - No
5. Are you currently covered by TRICARE or other military health care?
   - Yes
   - No
6. Are you currently covered through the Veteran's Administration**?**
   - Yes
   - No
7. Are you currently covered through the Indian Health Service?
   - Yes
   - No
8. Are you currently covered by any other health insurance or health coverage plan?
   - Yes ➔ (specify name of health care plan)
   - No

Fig. 1. Abbreviated CPS and ACS health insurance modules.

MinnesotaCare began in 1992 as a state-subsidized public health insurance program, where low-income households that do not qualify for Medicaid pay a subsidized monthly premium based on their income. As such, MinnesotaCare could function as a kind of proxy for public programs that require premium contributions in other states (such as the Children's Health Insurance Program and some Medicaid expansion participants). There is wide variation in the number of public programs within a given state, and complex rules for many of these programs regarding eligibility and premium contribution requirements. Fertig et al. (2018) includes a table of each state's public programs and premium contribution requirements as of August 2016. At that time, a total of 16 states offered only public programs that do not require a monthly premium contribution; 21 states offered at least one program that required a monthly premium contribution for at least some enrollees; and 14 states offered at least one public program that required a monthly premium contribution for all enrollees. In total, more than 69% of states offered one or more public program that required a monthly premium contribution for at least some, if not all, enrollees.

To gain insight into this diverse landscape, we exploit the presence of MinnesotaCare enrollees in the CHIME sample by presenting all results for two different analytic samples. The "Standard" sample excludes individuals with only MinnesotaCare (n = 657) and the "Augmented" sample includes those with MinnesotaCare. Appendix D (Subsection 7.4), displays the sample size and distribution for both samples. While we cannot predict reporting accuracy for all states with this study, we offer results from these two samples as a reasonable approximation of upper and lower bounds of reporting accuracy across states, depending on the structure and complexity of public programs within the state. In other words, the Standard sample results are meant to approximate reporting accuracy in states where Medicaid and other public program offerings do not require premium contributions. The Augmented sample is meant to represent reporting accuracy in states with a nontrivial number of individuals enrolled in public programs that require a premium contribution.

Finally, we omitted any individuals from the analytic sample for the current study if they did not have coverage at the time of the interview according to the enrollment records (n = 130). Because our sample was selected in December 2014 but the interview was conducted in May/June 2015, we cannot discern whether an individual with no coverage in the enrollment records at the time of the interview was uninsured or insured with another company. Thus, our Standard sample contained 3,036 person-records and the Augmented sample contained 3,693 person-records.

2.9. Reporting Accuracy Metrics

We use three different metrics to evaluate reporting accuracy. First is under-reporting (aka false negatives): the percent of people known to have Coverage Type X (according to enrollment records) for whom Coverage Type X is not reported in the survey. Second is the other side of the coin or over-reporting (aka false positives): the percent of people for whom Coverage Type X is reported, but who could not be validated in the enrollment records to have Coverage Type X. For the third metric, we compare the survey estimate of Coverage Type X to the prevalence of Coverage Type X indicated in the enrollment records.

The over-reporting metric is somewhat compromised by our study design. Because we have enrollment records from only a single insurer, we cannot say with certainty that a report of Coverage Type X that cannot be validated in *our* records of Coverage Type X is truly an over-report. It could be a false positive, or it could be a legitimate report of Coverage Type X from a *different* insurer. However, one strength of our study design is that we have enrollment records on a broad range of coverage types. Therefore, among those for whom Coverage Type X was reported but could not be validated as Coverage Type X in our records, we can examine how often it could be validated as Coverage Type Y in our records.

3. Results

As discussed in the methods section, this study was conducted as an experiment, using only a subset of the data collection and processing systems used to produce the official CPS and ACS estimates. We use the term "survey" as a convenient shorthand to mean "health insurance questionnaire module." That is, all results presented reflect the impact of only the questionnaires; the effects of editing, imputation and other aspects of the processing system are not assessed in this study.

3.1. Standard Sample

We begin with results for the three metrics for all four categories of individual and aggregated coverage types, as well as the uninsured (see Table 2). The left-most panel shows results for under-reporting. For example, the first row indicates that among those with any kind of private coverage according to the records, no private coverage was reported for 1.2% of those in the CPS treatment and 3.5% of those in the ACS treatment. In both survey treatments, levels of under-reporting varied by coverage type and were fairly low (below 5%) for ESI, private and insured, and higher (15–22.3 %) for non-group/marketplace and public. For public coverage the under-reporting was identical across surveys (at 16.8%). For other coverage types the differences across surveys were generally small but statistically significant and varied by coverage type. For ESI, private and insured, under-reporting was lower in the CPS than the ACS, by 1.6–2.5 percentage points. For non-group/marketplace the ACS under-report was lower than the CPS by 7.2 percentage points. Among those with any kind of private or public coverage according to the records, no coverage at all was reported for 1.9% of CPS enrollees, and 3.5% of ACS enrollees.

Turning to over-reporting in the center panel, the first row indicates that among those for whom private coverage was reported, 2.3% could not be validated in the CPS records to have private coverage, and 6.7% could not be validated in the ACS records. Generally, over-reporting ranged from 2.1% to 8.6% across coverage types with the exception of non-group/marketplace, which was dramatically higher – over 40% in both surveys. Across coverage types, CPS-ACS differences were still fairly small in magnitude but statistically significant. Within type of private coverage, over-reporting of non-group/marketplace was 44.5% and 40.6% in the CPS and ACS, and over-reporting of ESI was only 2.8% and 5.4% in the CPS and ACS (respectively). Among those for whom public coverage was reported, over-reporting in the ACS was higher than in the CPS – 8.6% and 2.1%, respectively.

In terms of overall prevalence (right-most panel) – how close the survey estimate came to the population prevalence – estimates varied across coverage types and surveys, but all

were within about one to four percentage points of population prevalence. Private and non-group/marketplace coverage were slightly over-estimated in both surveys, and public coverage was slightly under-estimated in both. Regarding the uninsured, in both surveys people known to have some type of coverage were reported as uninsured – 1.9% in the CPS treatment and 3.3% in the ACS treatment.

3.2. *Augmented Sample*

Results on under-reporting in the Augmented sample (which includes MinnesotaCare enrollees in the public coverage category) map closely to the Standard sample results in terms of overall levels and CPS/ACS differences, with the exception of public coverage (see Table 3, the left-most panel). Overall levels of under-reporting for public coverage were higher in both surveys in the Augmented compared to the Standard sample. Also, while under-reporting was the same across surveys in the Standard sample (at 16.8%), in the Augmented sample the CPS resulted in less under-reporting (19.2%) than the ACS (22.0%).

With regard to over-reporting (center panel), the most notable difference between the Standard and Augmented samples was in non-group/marketplace coverage, which increased by more than ten percentage points in both surveys – from 44.5 to 54.9 percentage points in the CPS and from 40.6 to 54.2 percentage points in the ACS. This shift reduced the CPS-ACS differential; in the Standard sample, CPS over-reporting was 3.9 percentage points higher than ACS but in the Augmented sample, CPS over-reporting was only 0.6 percentage points higher than ACS. Over-reporting of public coverage decreased in both surveys but more so in the ACS than the CPS. In the ACS, over-reporting went from 8.6% in the Standard sample down to 7.0% in the Augmented sample, and in the CPS over-reporting went from 2.1% to 1.8%.

In terms of prevalence (right-most panel of Table 3), across both surveys, patterns were similar when moving from the Standard to the Augmented sample. Specifically, private coverage was over-estimated more in the Augmented than the Standard sample, and this was driven by non-group/marketplace coverage (not ESI) across both surveys. For public coverage both surveys underestimated coverage in both samples, but the gap widened in the Augmented compared to the Standard sample, and more so in the ACS than the CPS. In the CPS, the under-estimate of public coverage went from 4.3 percentage points in the Standard sample to 5.7 percentage points in the Augmented sample. In the ACS, the under-estimate went from 2.9 percentage points in the Standard sample to 5.6 percentage points in the Augmented sample, putting it on par with the CPS over-estimate.

3.3. *Non-Group/Marketplace Coverage*

Results for non-group/marketplace coverage were something of an anomaly. Levels of under- and over-reporting were lower in the CPS than ACS for all coverage types except this one, and levels of over-reporting for non-group/marketplace were markedly higher than all the other coverage types in both surveys. To explore this further we break down the under-reporting results into more detail. Because the enrollment records indicate non-group versus marketplace coverage, regardless of which survey respondents were assigned to, we can examine under-reporting results separately for non-group and marketplace enrollees (see Table 4, left and right panels, respectively) for both CPS and ACS. We also

examine not just under-reporting but presumed misreporting – that is, among non-group and marketplace enrollees whose coverage was not reported as non-group/marketplace, how often was a different type of coverage or no coverage reported? In Table 4 we show the percentage of enrollees for whom the correct coverage type was reported and label that the "Target" row. Note this is simply a different expression of under-reporting; rather than show the percentage who did NOT report the known coverage type (as in Tables 2 and 3), in Table 4 we show the percent who DID report the known coverage type. The next three rows indicate that non-group/marketplace coverage was NOT reported, but a different coverage type (ESI, public or other) was reported. The final row indicates how often no coverage of any type was reported.

Results for the ACS show that levels of reporting the Target coverage type are roughly equivalent among non-group and marketplace enrollees (85.6% and 83.6%, respectively). However, in the CPS, levels of reporting the Target coverage type are much lower among marketplace enrollees (62.9%) than non-group enrollees (78.5%). With regard to Non-Target reporting, in the ACS it is roughly evenly split between ESI and public among both non-group and marketplace enrollees (e.g., among non-group enrollees, ESI and public coverage reporting is 6.4% and 5.4%, respectively). In the CPS, however, among non-group enrollees the most common Non-Target coverage is by far ESI (15.3%), and public and other/unspecified are roughly evenly split (3.2% and 2.4%). Among CPS marketplace enrollees, public is the most common Non-Target coverage type reported (18.5%), while ESI and other/unspecified are roughly evenly split (9.0% and 8.4%). Finally, although the ACS generated higher levels of Target reporting for both non-group and marketplace enrollees, both these types of enrollees were more likely to be misreported as uninsured in the ACS than in the CPS. Among non-group enrollees, 0.6% and 1.8% are misreported as lacking coverage in the CPS and ACS, respectively, and among marketplace enrollees 1.1% and 2.1% are misreported as lacking coverage in the CPS and ACS.

We also explore over-reporting in more detail (see Table 5). Because we begin with the universe of respondents who reported non-group/marketplace, and the ACS does not distinguish between these two coverage types, we cannot split out results for non-group from marketplace. However, among those for whom non-group/marketplace coverage was reported, we can show how often non-group/marketplace coverage could be validated in the records, and how often non-group/marketplace could not be validated but a different type of coverage (ESI or public) could be validated instead. Note that results in the Target row are, again, simply a different expression of over-reporting results already shown in Table 2. For example, Table 2 shows the percentage of reports that could not be validated in the CPS is 44.5%, and Table 5 shows the percentage of reports that could be validated is 55.5%, and these two metrics sum to 100 (i.e., the universe of non-group/marketplace reports is accounted for when we sum under-reports and Target reports). What is new in Table 5 is the Non-Target results. These findings show that in the CPS Standard sample, among the non-group/marketplace reports that could not be validated to be the Target coverage type, most (36.2%) were validated to have ESI coverage and the remainder had public (8.3%). In the ACS, however, the Non-Target cases were roughly evenly split between ESI and public (20.5% and 19.4%, respectively). In the Augmented sample, the addition of the MinnesotaCare sample shifts these distributions, with both the CPS and ACS having more reports of non-group/marketplace being validated as public coverage.

Table 4. Target and under-reporting estimates and differences across CPS and ACS, non-group and marketplace enrollees[1].

Reported coverage type	Non-group enrollees				Marketplace enrollees			
	%		Difference		%		Difference	
	CPS	ACS[4]	CPS-ACS		CPS	ACS[4]	CPS-ACS	
Target (Non-group/marketplace)[2]	78.5	85.6	−7.1	***	62.9	83.6	−20.7	***
Non-target								
ESI only	15.3	6.4	8.9	***	9.0	6.0	3.0	***
Public only	3.2	5.4	−2.3	***	18.5	7.8	10.7	***
Other(s)/unspecified only[3]	2.4	1.4	1.0	***	8.4	4.8	3.7	***
Uninsured	0.6	1.8	−1.1	***	1.1	2.1	−1.0	**
TOTAL	100	100			100	104		

*** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$.

[1]Results for the Standard and Augmented samples are identical because no MinnesotaCare enrollees were among the non-group/marketplace enrollees.

[2]Target represents the percentage reported with the known coverage type – here non-group/marketplace alone or in combination with any other coverage type; non-target indicates the percentage reported as having a coverage type other than what is shown in enrollment records. ESI = employer sponsored insurance; Non-group/Marketplace = insurance purchased outside and within the marketplace; Uninsured = no coverage was reported. Public insurance includes Medicaid in the Standard sample; Public insurance includes Medicaid and MinnesotaCare in the Augmented sample.

[3]Coverage reported was Medicare, military and/or other/unspecified but NO ESI, non-group, marketplace or public coverage was reported.

[4]Because these are predicted values based on adjustments for varying demographics across treatments (see Section 2.4), the individual values may not sum to 100 percent.

Table 5. Target and Over-reporting estimates and differences across CPS and ACS, those for whom non-group/marketplace coverage was reported[1].

	Standard Sample				Augmented Sample			
	%		Difference		%		Difference	
Coverage type in enrollment records	CPS	ACS[3]	CPS-ACS		CPS	ACS[3]	CPS-ACS	
Target (non-group/marketplace)[2]	55.5	59.4	−3.9	***	45.2	45.8	−0.7	**
Non-target								
ESI only	36.2	20.5	15.8	***	29.5	16.3	13.2	***
Public only	8.3	19.4	−11.1	***	25.3	36.2	−10.9	***
ESI/Public	0	0	0		0.1	0	0.1	
TOTAL	100	99.2	n/a		100	98.2	n/a	

*** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$.

[1]Coverage type reported was non-group/marketplace alone, or in combination with any other coverage type.

[2]Enrollment records indicated non-group/marketplace (Target) alone, or in combination with any other coverage type; non-target indicates the percentage having other coverage type in enrollment records. ESI = employer sponsored insurance; Non-group/Marketplace = insurance purchased outside and within the marketplace; Public insurance includes Medicaid in the Standard sample; Public insurance includes Medicaid and MinnesotaCare in the Augmented sample.

[3]Because these are predicted values based on adjustments for varying demographics across treatments (see Subsection 2.4), the individual values may not sum to 100 percent.

Table 6. Target and Over-Reporting Estimates and Differences Across CPS and ACS, Those for Whom ONLY Non-Group/Marketplace Coverage was Reported[1].

Coverage Type in Enrollment Records	Standard Sample				Augmented Sample			
	%		Difference		%		Difference	
	CPS	ACS[3]	CPS-ACS		CPS	ACS[3]	CPS-ACS	
Target (Non-group/Marketplace)[2]	74.9	73.8	1.1	***	58.1	55.2	2.9	***
Non-Target								
ESI only	13.1	1.9	11.2	***	10.2	1.5	8.7	***
Public only	12.0	22.3	−10.3	***	31.6	42.3	−30.2	***
ESI/Public	0	0	0		0.1	0	−0.1	
TOTAL	100	98	n/a		100	99	n/a	

*** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$; n/a = not applicable.

[1]Coverage type reported was non-group/marketplace ONLY; no other type of coverage was reported.

[2]Enrollment records indicated non-group/marketplace (Target) alone, or in combination with any other coverage type non-target indicates the percentage having other coverage type in enrollment records. ESI = employer sponsored insurance; Non-group/Marketplace = insurance purchased outside and within the marketplace; Public insurance includes Medicaid in the Standard sample; Public insurance includes Medicaid and MinnesotaCare in the Augmented sample.

[3]Because these are predicted values based on adjustments for varying demographics across treatments (see Subsection 2.4), the individual values may not sum to 100 percent.

We take the non-group/marketplace results one step further to address the research suggesting that noncomprehensive non-group plans (e.g., dental and vision plans) may account for much of the observed over-reporting of non-group/marketplace coverage. Table 6 mimics Table 5 except that we limit the sample to those for whom ONLY non-group/marketplace coverage was reported (versus those for whom non-group/marketplace was reported in combination with one or more other types of coverage). Thus, for example, Table 6 EXCLUDES those who report having ESI and non-group/marketplace when they actually have ESI and a dental plan. Results show a fairly dramatic shift. Over-reporting (100 minus the validated reports of coverage shown in the Target row) drops by almost 20 percentage points in the CPS and by almost 15 percentage points in the ACS in the Standard sample. The same pattern is observed in the Augmented sample but the magnitude of the drop in over-reporting is somewhat lower. Further, when we limit the sample to those for whom only non-group/marketplace coverage was reported in the Standard sample (Table 6), those validated to have ESI drops by 23 percentage points in the CPS and by almost 19 percentage points in the ACS and those validated to have public increases by roughly 3–4 percentage points for both surveys compared to Table 5. The pattern is similar in the Augmented sample: those validated to have ESI drops by 19 percentage points in the CPS and by almost 15 percentage points in the ACS between Tables 5 and 6, and those validated to have public increases by roughly 6 percentage points for both surveys.

3.4. Uninsured

Finally, we examine how these patterns of over- and under-reporting by coverage type affect the measure of the uninsured, and how this varies across surveys (see Table 7). Columns indicate the coverage type according to the records, and rows indicate the reported coverage type – either Target, Non-Target or Uninsured. Public enrollees are more likely to be misreported as uninsured than private enrollees, across both survey treatments, by several fold. In the CPS, 5.0% of public versus only 0.8% of private enrollees are reported to have no insurance; in the ACS the rate is 6.6% for public and 2.2% for private. Across types of private coverage, results are fairly consistent; reports of no coverage are within a percentage point of each other for both surveys. For example in the CPS, uninsured rates for ESI, non-group/marketplace, non-group alone, and marketplace alone is 0.8%, 0.7%, 0.6% and 1.1%, respectively. In terms of differences across surveys, for public and private coverage overall, and for each component of private coverage, the uninsured rate is higher in the ACS than in the CPS, by 1.0 to 1.6 percentage points across coverage types.

4. Discussion

4.1. Moving Parts: The Inter-Relationship between Misreporting and Coverage Type Prevalence

There are several moving parts in a study like this. Among them are differences in the surveys' capacity to elicit true positives and avoid false positives, variation in over- and under-reporting across coverage types in both surveys, and the prevalence of the various coverage types in the population. While all of these are at play in the findings, one constant is the sheer dominance of ESI relative to other coverage types. For our particular insurer's

Table 7. Reporting of Target, Other and Uninsured Among those with Any Coverage According to Enrollment Records, By Coverage Type (Standard Sample).

	Private		Public		ESI		NongMkt		Non-Group		Market	
	%		%		%		%		%		%	
Reported	CPS	ACS[2]	CPS	ACS[2]	CPS	ACS[2]	CPS	ACS[2]	CPS	ACS[2]	CPS	ACS[2]
Target[1]	98.9	96.5	83.2	83.2	98.1	95.6	77.7	85.0	78.5	85.6	62.9	83.6
Non-Target	0.4	1.5	11.8	10.7	1.1	2.6	21.6	13.6	20.8	13.2	36.0	18.6
Uninsured	0.8	2.2	5.0	6.6	0.8	2.2	0.7	1.8	0.6	1.8	1.1	2.1
TOTAL	100	100	100	101	100	100	100	100	100	101	100	104

[1]Target represents the percentage reported with the known coverage type; non-target indicates the percentage reported as having a coverage type other than what is shown in enrollment records.

[2]Because these are predicted values based on adjustments for varying demographics across treatments (see Subsection 2.4), the individual values may not sum to 100 percent.

Private coverage is aggregate of employer sponsored insurance (ESI) and non-group insurance purchased outside (Nong) and within the marketplace (Market) with and without a premium subsidy; NongMkt is non-group and marketplace coverage combined; Public insurance includes Medicaid in the Standard Sample; Uninsured are the percentage for whom no insurance is reported.

population, the prevalence of ESI was 67%, public was 30% and non-group/marketplace was 3.4% (see Table 1). This distribution means that, consistent with Davern et al. (2008), reporting patterns of ESI enrollees have the greatest effect on the metrics for all the other coverage types. Another constant is simply the logic of aggregation. The private coverage metrics are a function of ESI and non-group/marketplace together, so respondents could report the wrong type of private coverage (e.g., ESI enrollees could misreport their coverage as non-group/marketplace, or vice versa), and metrics for the individual coverage types would be affected but the overall private metrics would not. Both of these factors played out in our results. For example, under-reporting of ESI was fairly low in both surveys (1.9% and 4.4% in CPS/ACS, Table 2), but for non-group/marketplace it was higher (22.3% and 15.0%). The low prevalence of non-group/marketplace coverage, the high prevalence of ESI combined with its low rate of under-reporting, and the fact that respondents could interchange ESI and non-group/marketplace coverage for the overall private coverage metric meant the impact of non-group/marketplace under-reporting, while high, had a negligible effect on private coverage metrics. Indeed, under-reporting for the aggregated private coverage type category (1.2% and 3.5% in CPS/ACS) was lower than for either of the two components of private coverage.

This kind of inter-play was also evident in the CPS-ACS differences. For example, on under-reporting, the ACS did better than the CPS for non-group/marketplace (by 7.2 percentage points, Table 2), but worse than the CPS for ESI (by 2.5 percentage points), and the surveys were identical on public coverage. For both aggregated categories of private coverage and the insured, the CPS did better than the ACS. Thus, the improved metric for non-group/marketplace in the ACS was not enough to compensate for its lower metric for ESI, given the low prevalence of non-group/marketplace relative to ESI. In other words, higher under-reporting of ESI in the ACS versus the CPS is the main driver of the differences between the two surveys in both the private and uninsured rate. A similar pattern was observed in over-reporting. Rates for ESI and public were lower in the CPS than the ACS, and higher for non-group/marketplace coverage. Due in large part to the weight of ESI relative to other coverage types, over-reporting of private coverage was 4.3 percentage points lower in the CPS than the ACS.

In terms of the point estimate, the difference between the survey estimate and the population prevalence is a function of not only the relative prevalence of different coverage types and levels of under- and over-reporting, but the nature of misreporting. For example, *under*-reporting of public coverage was identical in the CPS and ACS, but *over*-reporting of public coverage was a fair bit higher in the ACS than the CPS (by 6.5 percentage points, Table 2). However, the ACS under-estimated public coverage by 2.9 percentage points and the CPS under-estimated it by 4.3 percentage points. Thus, the lower over-reporting of public coverage in the CPS resulted in fewer false positives to make up for the false negatives, compared to the ACS. There is an additional nuance at work, which has to do with the difference in the Standard and Augmented samples. Recall that the Augmented sample includes MinnesotaCare enrollees, who contribute to the monthly premium based on a sliding scale. Where under-reporting of public coverage in the Standard sample was identical across surveys (Table 2), in the Augmented sample under-reporting in the CPS was lower than in the ACS (19.2% versus 22.0%, Table 3). Also, while over-reporting was still higher in the ACS versus CPS, the differential was

reduced (6.5 percentage points in the Standard sample versus 5.2 percentage points in the Augmented). The combination means that in the Augmented sample, the ACS picked up fewer legitimate reports of public coverage than it did in the Standard sample, AND it gained slightly fewer over-reports. Thus, the net estimate of public coverage in the ACS compared to the population prevalence was no longer as close as it was in the Standard sample; indeed in the Augmented sample it was on par with the CPS.

4.2. *Effects of Public Coverage that Requires a Premium Contribution*

More generally, when moving from the Standard to Augmented sample, under-reporting was equivalent for all coverage types except public, which increased by about 2.5 percentage points in the CPS and about 5 percentage points in the ACS. This suggests that in states where public programs require the enrollee to contribute to the premium, under-reporting goes up in both surveys, but more so in the ACS than in the CPS. In terms of over-reporting, the most pronounced difference between the Standard and Augmented samples was among those reporting non-group/marketplace coverage, where over-reporting increased by about 10 percentage points in the CPS (from 44.5% to 54.9%) and about 14 percentage points in the ACS. Because the only difference between the two samples is that the Augmented sample includes MinnesotaCare enrollees while the Standard sample does not, these results suggest that MinnesotaCare enrollees have a tendency to misreport their public coverage as non-group/marketplace coverage in both surveys, but more so in the ACS than in the CPS. With regard to overall prevalence, the gap between the survey estimate and population prevalence got slightly wider for private, public and non-group/marketplace coverage when moving from the Standard to the Augmented samples, and stayed about the same for ESI and uninsured. This pattern held for both surveys, but the size of the gap was slightly higher in the ACS than the CPS for private and non-group/marketplace and especially for public coverage. Again this suggests that in states where there is cost-sharing for public programs, measurement error will be slightly increased for private coverage (driven by non-group/marketplace) and public coverage compared to states where public programs have no premium cost-sharing, and that the ACS estimates will be somewhat more prone to measurement error than the CPS, particularly for public coverage.

4.3. *Non-Group/Marketplace Results*

Non-comprehensive plans – those that cover only a single service such as dental or vision – are common in the non-group market. Technically speaking, respondents should not report these non-comprehensive plans at all, because they are out of scope in the survey. Also because they are out of scope, they are not in the universe of plans that could be validated in the records. However, to the extent respondents are not paying attention to the "fine print" in the survey and report these non-comprehensive plans, they cannot be validated, and thus contribute to over-reporting. The large reduction in over-reporting when we eliminated those who reported non-group/marketplace in combination with another type of coverage (that is, the difference in the Target metrics when moving from Table 5 to 6) suggests that non-comprehensive plans could well be a major contributor to the over-reporting of non-group/marketplace coverage. In terms of misreporting, the

finding that non-group enrollees misreport their coverage as ESI more in the CPS than in the ACS is curious (15.3% versus 6.4%, Table 4). One possible explanation could reside with the self-employed who obtain coverage on the individual market but consider it a business expense. The CPS asks about coverage "through a job" while the ACS asks if coverage is through "a current or former employer or union." Some non-group enrollees may be inclined to select "job" in the CPS because the coverage is tangentially related to their self-employed status, which enables them to consider it a business expense. However, in the ACS when asked about coverage through an "employer or union" versus coverage "purchased directly from an insurance company," they may choose the latter. For these individuals the terms "employer/union" may signify more formal arrangements with a third party institution, which may not match the concept of their self-employment.

Marketplace coverage is relatively new in the landscape of health coverage options, and it is saddled with ambiguity with regard to self-reports in surveys. For instance, the very term "marketplace" can mean the portal through which coverage is obtained, and/or the marketplace coverage itself. There are also multiple other pathways to obtaining marketplace coverage, in addition to the portal (e.g., brokers). Furthermore, both public and private plans are available on the portal, some marketplace plans are fully subsidized, and some public plans charge a premium. Thus, any one question that could definitely establish marketplace coverage is elusive. The CPS and ACS surveys go about capturing marketplace coverage in very different ways. In the ACS, it is assumed that marketplace enrollees would report their coverage in response to the question asking about "health insurance purchased directly from an insurance company" (see Figure 1, second question). Indeed, 83.6% of marketplace enrollees did this (see Table 4). However, the ACS has not yet made any attempts to separate marketplace from non-group coverage. In the CPS, respondents are asked a series of questions about features of the coverage, such as general source (job, government/state), program name, portal, premiums and subsidies. For the reasons noted above, none of these individual questions alone determines coverage type. In a related research project, we used a supervised machine learning approach and enrollment records to guide an algorithm using these questions to classify coverage type in the CPS (Pascale et al. 2018a). There were multiple trade-offs and due to the high prevalence of public relative to marketplace coverage, for the small handful of ambiguous cases, we chose an algorithm that slightly favored public over marketplace classification. This choice could partially explain why, in Table 4, 18.5% of known marketplace enrollees are shown as reporting public coverage, which, in turn, contributes to the Target marketplace metrics being lower than non-group (62.9 versus 78.5).

4.4. The Uninsured Rate

As noted earlier, the few existing studies that linked enrollment records with survey reports of both public and private coverage were conducted under very different conditions than our study, and comparisons with regard to coverage type are of limited use. The most relevant metric from these earlier studies would be false negatives of insurance: what percent of those with any kind of coverage according to the records were misreported as uninsured. Our study found overall uninsured rates of 1.9% and 3.5% in the CPS and ACS, respectively. Nelson et al. (2000) and Marquis (1983) found rates of 2.2%

and 3%, respectively. Davern et al. (2008) found lower rates (0.3% to 0.6% across coverage types), which could be partly explained by their inclusion of those over 65 (where coverage is near-universal), and their exclusion of proxy reports. In terms of CPS-ACS differences, one reason the ACS uninsured rate was 1.4 percentage points higher than the CPS could be the fact that the CPS series begins with a global yes/no question on any coverage at all, while the ACS does not. Several qualitative and quantitative studies indicate that a single household respondent sometimes has only limited knowledge about the details of other household members' coverage, and when confronted with a series of questions about specific coverage type, some respondents simply fail to report any coverage at all (Pascale 2009). Another key difference in the surveys is the "verification question." After a battery of questions on different types of coverage is asked, if no coverage is reported the CPS (and several other surveys) ask if it is correct that the person is uninsured, and if not the survey allows for collecting detailed information on the coverage. The ACS does not include this verification question. A final compounding problem in the ACS could be household size. The eight-question "laundry list" series is repeated for each household member, which risks respondent fatigue once the series is administered for, say, the fifth or sixth person, particularly if those individuals listed later in the household roster are more socially distant from the household respondent (e.g., unrelated housemates). This kind of respondent fatigue can result in a failure to report any coverage (Blumberg et al. 2004).

5. Limitations

There are several limitations that could influence the results and their generalizability. First, Minnesota is an atypical state in terms of demographics; compared to the U.S. as a whole, the state has a higher proportion of whites and those with a high school diploma and college degree. The state population has high rates of health insurance, high income, low unemployment and very low poverty relative to other states (U.S. Census Bureau 2016a, 2016b). Second is the fact that the study represents coverage from a single health insurance provider which, on its face, limits generalizability of the results. More specifically, however, with regard to the marketplace, the insurer's market share is relatively low. In 2014, the insurer served four percent of the "MNsure" market (the name for private marketplace plans in Minnesota), compared to 59% and 25% by the dominant insurers in the marketplace (Minnesota Department of Health 2018). Furthermore, the insurer's marketplace plans had higher premiums than most MNsure plans. It is possible that CHIME participants in the marketplace strata are more educated and financially secure than those in the marketplace population overall, and that these characteristics affect reporting accuracy. To investigate this, in related analysis (Call et al. 2018) we examine socioeconomic and health status characteristics associated with reporting accuracy. Third, the study design does not allow us to determine with certainty whether apparent false positives were truly inaccurate. That is, a report of coverage that could not be validated in the insurer's records could actually be accurate if the person had insurance from a different carrier. Finally, due to the relatively low response rate we cannot ascertain, beyond our simple nonresponse analysis, whether our results are biased due to differential response; it could be that those well aware of their insurance status would be the most likely to respond.

6. Conclusions

The scant studies thus far that have examined reporting accuracy across a range of coverage types suggest private coverage is over-reported and public coverage is under-reported (Davern et al. 2008; Nelson et al. 2000). Our findings are generally consistent with these earlier studies, but for the first time we provide reporting accuracy metrics based on two major national survey instruments in a post-ACA era, and we compare the two surveys for both individual and aggregated coverage types. Because there is such an established literature on the role of the questionnaire in measurement error of health insurance estimates, and the ACA represents a major shift in the landscape of the U.S. health system, we offer these metrics as a baseline. That is, we reserve judgment on whether the metrics indicate high or low data quality from the surveys and simply offer these findings to inform researchers in their choice of datasets that are fit for purpose, adjustments for measurement error, and so on. We do suggest, however, that while many differences between the CPS and ACS are statistically significant, the magnitude of the difference is fairly small in most cases. In our opinion, this evidence suggests that data users can take data quality off the table as a factor in their decisions about which survey to when making estimates of coverage type. For the uninsured measure the question is debatable given the 1.6 percentage point gap between surveys.

With regard to Medicaid in particular, there is a substantial literature linking survey reports to enrollment records, and Medicaid under-reporting in the pre-redesigned CPS has been thoroughly documented. Therefore, for Medicaid we can go beyond baseline findings and offer results in the context of the CPS pre- and post-redesign. One recent study (Noon et al. 2019) examined results from the pre-redesigned (aka traditional) CPS from 2000–2010. The under-reporting rate ranged from 38.8–44.7%. In comparison, Table 2 shows the under-reporting rate for the redesigned CPS for public coverage to be 16.8%. In terms of over-reporting, the Noon et al. study of the traditional CPS ranged from 20.7–26.8%, while Table 2 shows over-reporting of public coverage to be 2.1%. The Noon et al. study also provides results in terms of the "Medicaid undercount" – the difference between enrollment records and the survey estimate as a percent of the population prevalence – which ranged from 22–39% in the traditional CPS. To produce parallel metrics from findings in Table 2, we take the net prevalence difference of 4.3% points and divide it by the 28.4 prevalence in the records to get an undercount of 15.1%. The same exercise in the Augmented sample yields a 17.8% undercount. While there are many conditions in each study that hinder direct comparisons (e.g., in contrast to CHIME data, Noon et al. (2019) use CPS production data that were fully edited and imputed, and use a calendar year measure of insurance), results are consistent across all three metrics and provide compelling evidence that measurement error in the CPS has been reduced post-redesign, perhaps by half or more. In other words, to the extent that the CHIME study conditions produce estimates that are comparable to the national CPS ASEC, there appear to be substantial improvements in Medicaid reporting accuracy in the CPS redesign.

In terms of next steps, we generally expect a survey with lower under- and over-reporting to produce a more accurate point estimate than a survey with higher levels of under- and over-reporting. However, as was demonstrated above, in some cases the point estimate is closer to the population prevalence even if both under- and over-reporting are higher, due the two types of measurement error netting out. We explore the impact of this empirically by examining the

demographic (e.g., age, household size, income) and health status characteristics of those reported to have a given coverage type in the survey and comparing that to the demographic profile of those with that coverage type according to the enrollment records for both survey treatments (Call et al. 2018). The trailing accuracy metrics for non-group/marketplace are also a subject for further investigation. Finally, future research will examine experimental questions embedded in the ACS about the marketplace, premiums and subsidies, which could be leveraged to separate public, non-group and marketplace coverage.

7. Appendix

7.1. Appendix A, Comparison of Matched Individuals by Survey Treatment Arm

Appendix A. Comparison of Matched Individuals by Survey Treatment Arm.

	CPS	ACS	p-value
Female	51%	54%	0.1284
Respondent	52%	52%	0.9311
Child of respondent	27%	27%	
Spouse of respondent	17%	18%	
Other person in household	3%	3%	
Resides in 1 person household	25%	24%	**0.0270**
Resides in 2–4 person household	57%	43%	
Resides in 5+ person household	18%	33%	
Family size unknown	0%	0%	
Non-Hispanic White	83%	81%	**0.0072**
Non-Hispanic Black	8%	7%	
Hispanic	4%	5%	
Other race, non-Hispanic	5%	7%	
Family income $<$138% FPL	23%	23%	**0.0547**
Family income 139–199% FPL	17%	20%	
Family income 200–400% FPL	32%	29%	
Family income $>$400% FPL	26%	26%	
Family income unknown	2%	2%	
Full-year Full-time employed	33%	31%	0.3850
Less than full-time employed	29%	30%	
Out of the labor force	15%	17%	
Under 15	21%	20%	
Employment status unknown	3%	3%	
Employer $<$10 employees	35%	35%	0.3884
Employer 10–50 employees	19%	20%	
Employer 51–99 employees	6%	5%	
Employer 100+ employees	32%	34%	
Unknown employer size	7%	6%	
Less than high school	8%	8%	0.5234
High school graduate	24%	27%	
Some college or Associate's degree	31%	30%	
Bachelor's degree or more	37%	35%	
Education is unknown	0%	0%	
Married	50%	50%	0.1826
Divorced/separated/widowed	15%	17%	
Never married	35%	33%	
Marital status is unknown	0%	0%	

Note: Chi-square tests were performed to test for differences across groups.

Appendix B. CPS versus Unadjusted ACS Estimates.
Standard Sample: Under-[1] and Over-Reporting,[2] Prevalence Estimates[3] and Differences Across CPS and ACS[4].

Coverage Type in Records[5]	Under-Reporting				Over-Reporting				Prevalence Estimates[5]							
	%		Difference		%		Difference		%		Difference		%		Difference	
	CPS	ACS	CPS-ACS		CPS	ACS	CPS-ACS		Recs	CPS	CPS-Recs		Recs	ACS	ACS-Recs	
Private	1.2	3.5	−2.3	***	2.3	4.4	−2.0	**	71.8	72.6	0.9	**	72.4	73.1	0.7	
ESI	1.9	4.9	−3.1	**	2.8	3.3	−0.6		67.9	68.5	0.6		68.6	67.4	−1.1	*
NongMkt	22.3	15.6	6.7	***	44.5	53.0	−8.4	*	3.8	5.4	1.6	***	3.8	6.8	3.0	***
Public	16.8	16.0	0.8		2.1	7.2	−5.1	***	28.4	24.1	−4.3	***	27.7	25.1	−2.6	***
Insured	1.9	3.0	−1.1	**	n/a	n/a	n/a	n/a	100.0	98.1	−1.9	***	100.0	97.0	−3.0	***

*** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$; n/a = not applicable.

[1]Under-reporting = false negatives or the % of those known to have Coverage Type X for whom Coverage Type X is not reported in the survey.

[2]Over-reporting = false positive or the % of those for whom Coverage Type X is reported, but who (a) cannot be validated in the enrollment records to have Coverage Type X and (b) can be validated in the enrollment records to have Coverage Type Y.

[3]Prevalence = survey estimates of Coverage Type X versus prevalence of Coverage Type X indicated in the enrollment records.

[4]Private coverage is the aggregate of employer sponsored insurance (ESI), non-group insurance purchased outside the marketplace (Nong) and within the marketplace (Mkt); NongMkt is non-group and marketplace coverage combined; Public insurance is Medicaid in the Standard Sample; Insured includes both private and public insurance, but does not include health insurance provided through the military, the Indian Health Service, or Medicare.

[5]The prevalence of public and private coverage indicated in the enrollment records may sum to more than 100 percent because some individuals may have both private and public coverage. The estimated prevalence of public and private coverage sums to less than the estimated insured prevalence because insurance provided through the military, the Indian Health Service, or Medicare may have been reported as the insurance type in the survey but was not categorized as public or private coverage in this analysis, but the estimated insured prevalence includes all individuals who were not reported as uninsured.

Augmented Sample: Under-[1] and Over-Reporting,[2] Prevalence Estimates[3] and Differences Across CPS and ACS[4].

Coverage Type in Records[5]	Under-Reporting				Over-Reporting				Prevalence Estimates[6]							
	%		Difference		%		Difference		%		Difference		%		Difference	
	CPS	ACS	CPS-ACS		CPS	ACS	CPS-ACS		Recs	CPS	CPS-Recs		Recs	ACS	ACS-Recs	
Private	1.2	3.9	−2.7	***	4.2	7.2	−3.0	***	68.1	70.3	2.1	***	68.5	70.9	2.4	***
ESI	1.9	5.4	−3.5	**	3.0	4.0	−1.0		65.5	65.2	−0.3	*	64.9	64.0	−0.9	
NongMkt	22.3	15.6	6.7	**	54.9	62.8	−8.0	**	3.7	6.3	2.6	***	3.6	8.2	4.6	***
Public	19.2	22.0	−2.8	***	1.8	6.4	−4.6	***	32.0	26.3	−5.7	***	31.9	26.6	−5.3	***
Insured	2.0	3.3	−1.3	***	n/a	n/a	n/a	n/a	100.0	98.0	−2.0	***	100.0	96.7	−3.3	***

*** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$; n/a = not applicable.

[1]Under-reporting = false negatives or the % of those known to have Coverage Type X for whom Coverage Type X is not reported in the survey.

[2]Over-reporting = false positive or the % of those for whom Coverage Type X is reported, but who (a) cannot be validated in the enrollment records to have Coverage Type X and (b) can be validated in the enrollment records to have Coverage Type Y.

[3]Prevalence = survey estimates of Coverage Type X versus prevalence of Coverage Type X indicated in the enrollment records.

[4]ACS estimates are predicted based on adjustments for varying demographics across treatments (see Subsection 2.4). Appendix B shows unadjusted estimates.

[5]Private coverage is the aggregate of employer sponsored insurance (ESI), non-group insurance purchased outside the marketplace (Nong) and within the marketplace (Mkt); NongMkt is non-group and marketplace coverage combined; Public insurance is Medicaid and MinnesotaCare in the Augmented Sample; Insured includes both private and public insurance.

[6]The prevalence of public and private coverage indicated in the enrollment records may sum to more than 100 percent because some individuals may have both private and public coverage. The estimated prevalence of public and private coverage sums to less than the estimated insured prevalence because insurance provided through the military, the Indian Health Service, or Medicare may have been reported as the insurance type in the survey but was not categorized as public or private coverage in this analysis, but the estimated insured prevalence includes all individuals who were not reported as uninsured.

7.3. Appendix C, CPS AND ACS Survey Modules

APPENDIX C. CPS AND ACS SURVEY MODULES

CPS Health Insurance Module

Section A: Coverage Status

HINTRO
These next questions are about health coverage between January 1, [CY-1] and now.
- Press 1 to continue → PINTRO

PINTRO
[First/Next] I'm going to ask about [your/NAME's] health coverage.
- Press 1 to continue → CK-MCARE1

CK-MCARE1
Is NAME either 65+?
- Yes → MCARE1
- No → ANYCOV

MCARE1
Medicare is health insurance for people 65 years and older and people under 65 with disabilities. [Are you/Is NAME] NOW covered by Medicare?
◆ Code Medicare Parts A, B and C and Medicare Advantage as "Yes".
1. Yes → BEFORAFT_LC1
2. No/DK/REF → ANYCOV

ANYCOV
[Do you/Does NAME] NOW have any type of health plan or health coverage?
1. Yes → SRCEGEN_LC1
2. No/DK/REF → MEDI

MEDI
[Are you/Is NAME] NOW covered by Medicaid, Medical Assistance [or] CHIP [if MCARE1 not yet asked: or Medicare]?
1. Yes → GOVTYPE_LC1
2. No/DK/REF → OTHGOVT

OTHGOVT
[Are you/Is NAME] NOW covered by a state or government assistance program that helps pay for healthcare, such as MinnesotaCare, Minnesota Comprehensive Health Association (MCHA), PMAP, MNsure or healthcare.gov?
[NOTE: Minnesota example is shown; question text fills all known state-specific program names for Medicaid and CHIP, all state-specific government program names, and all state-specific names for marketplace coverage]

◆ Stop reading the list if respondent says "YES."
1. Yes → GOVPLAN_LC1
2. No/DK/REF → If ever served in Armed Forces (AFEVER = 1) → VET; else → VERIFY

VET
[Are you/Is NAME] NOW covered by Veteran's Administration (VA) care?
1. Yes → BEFORAFT_LC1
2. No/DK/REF → VERIFY

VERIFY
I have recorded that [you are/NAME is] not currently covered by a health plan. Is that correct?
1. Yes, is NOT covered → ADDOTH1_L
2. No, is covered → SRCEGEN_LC1
3. DK/REF → ADDOTH1_L

Section B: Plan Type

SRCEGEN_LC1
ASK OR VERIFY
For the coverage you/NAME has/have NOW, [do you/does NAME] get it through a job, the government or state, or some other way?
◆ **JOB:** Former job/Retiree, Union, Spouse/parent's job, Job with the government, COBRA, TRICARE/TRICARE for Life
◆ **GOVERNMENT OR STATE:** Medical Assistance, Medicaid, Medicare (Parts A + B; Part C), Medicare Advantage, State-provided health coverage, VA Care/CHAMPVA/other military
◆ **OTHER:** Privately purchased, Parent or spouse, Medicare Supplements, Exchange plan/Marketplace, Group or association, School,
◆ **IF RESPONDENT CHOOSES MORE THAN ONE:** Ok let's talk about one plan at a time. Which would you like to tell me about first?
If VERIFY = 2 then fill: ◆ If respondent is not covered, go back to VERIFY and select "Yes"
1. Job (current or former) → MILPLAN_LC1
2. Government or State → JOBCOV_LC1
3. Other way → SRCEDEPDIR_LC1
- DK/REF → SRCEDEPDIR_LC1

SRCEDEPDIR_LC1
◆ ASK OR VERIFY
[Do you/Does NAME] get that coverage through a parent or spouse, [do you/does he/she] buy it [yourself/himself/herself], or [do you/does he/she] get it some other way?
PARENT/SPOUSE: Parent, Spouse
BUY IT DIRECTLY: Buy it, Parent or spouse buys it, Medicare Supplement

SOME OTHER WAY: Former employer, Group or association, Indian Health Service, School

1. Parent or spouse → POLHOLDER_LC1
2. Buy it → POLHOLDER_LC1
3. Other way → SRCEOTH_LC1

• DK/REF → SRCEOTH_LC1

SRCEOTH_LC1

◆ ASK OR VERIFY

[Do you/Does NAME] get it through a former employer, a union, a group or association, the Indian Health Service, a school, or some other way?

1. Former employer → POLHOLDER_LC1
2. Union → POLHOLDER_LC1
3. Group or association → POLHOLDER_LC1
4. Indian Health Service → BEFORAFT_LC1
5. School → POLHOLDER_LC1
6. Some other way → GOVPLAN_LC1

• DK/REF → GOVPLAN_LC1

JOBCOV_LC1

Is that coverage related to a JOB with the government or state?

◆ Include coverage through FORMER employers and unions, and COBRA plans.

1. Yes → MILPLAN_LC1
2. No → GOVTYPE_LC1

• DK/REF → GOVTYPE_LC1

Soft edit: If "yes" and no one in the household was reported to have a job (more than part time, seasonal or temp work), nor is anyone in the household a retiree, then ask soft edit: "Can I just check – I recorded that this coverage is related to a JOB. Is that correct?"

◆ If this is correct, continue to MILPLAN_LC1

◆ If this is not correct, go back to JOBCOV_LC1 and correct

MILPLAN_LC1

◆ ASK OR VERIFY

Is that plan related to military service in any way?

◆ Examples of military plans include:

- VA Care
- TRICARE
- TRICARE for Life
- CHAMPVA
- Other military care

1. Yes → MILTYPE_LC1
2. No → POLHOLDER_LC1

• DK/REF → POLHOLDER_LC1

GOVTYPE_LC1

◆ ASK OR VERIFY

Is that coverage Medicaid, CHIP, Medicare, a plan through the military, or some other program?

◆ Code Medicare Parts A, B and C and Medicare Advantage as "Medicare".

IF R CHOOSES MORE THAN ONE: Ok let's talk about one plan at a time. Which would you like to tell me about first?

1. Medicaid or Medical Assistance → GOVPLAN_LC1
2. CHIP → PORTAL_LC1
3. Medicare → soft edit then → BEFORAFT_LC1
4. Military → MILTYPE_LC1
5. Other → GOVPLAN_LC1

• DK/REF → GOVPLAN_LC1

Soft edit: if Medicare is selected and NAME is under 65 ask: "There are two programs that sound a lot alike. MediCARE is for people 65 years and older, or people under 65 with disabilities. MediCAID is a government-assistance plan for those with low-incomes or a disability. Just to be sure, which program are you/is NAME covered by?"

◆ If Medicare is correct, suppress and continue.

◆ If Medicare is not correct, go back to GOVTYPE_LC1 and correct.

MILTYPE_LC1

◆ ASK OR VERIFY

Is that plan through TRICARE, TRICARE for Life, CHAMPVA, VA care, military health care, or something else?

1. TRICARE
2. TRICARE for Life
3. CHAMPVA
4. Veterans Administration (VA) care
5. Military health care
6. Other

• DK/REF

[all] → POLHOLDER_LC1

POLHOLDER_LC1

◆ ASK OR VERIFY

Whose name is the policy in? (Who is the policyholder)?

1. *household member 1*
2. *household member 2*

. . . . . .

16. *household member 16*
17. Someone living outside the household

• DK/REF

[all] → CK-SRCEPTSP_LC1

CK-SRCEPTSP_LC1

- If SRCEDEPDIR_LC1 = "parent or spouse" then → SRCEPTSP_LC1
- Else if SRCEDEPDIR_LC1 = 2 = "buy it" then → PORTAL_LC1
- Else → CK-HIPAID_LC1

SRCEPTSP_LC1

◆ ASK OR VERIFY

Do they get that coverage through their job, do they buy it themselves, or do they get it some other way?

1. Job (current or former) → HIPAID_LC1
2. Buy it → PORTAL_LC1
3. Other way → GOVPLAN_LC1

- DK/REF → GOVPLAN_LC1

GOVPLAN_LC1

◆ ASK OR VERIFY

What do you call the program?

IF RESPONDENT ANSWERS WITH INSURANCE COMPANY NAME: OK, so that would be the plan name. What do you call the program? Some examples of programs in [STATE] are [read full list below].

NOTE: Some response categories are generic (regardless of state) and some are state-specific. The generic response categories are: 1, 2, 3, 13, 17 and 18. Response categories 4–12 fill up to nine state-specific names for Medicaid, CHIP and other state-sponsored government programs. If there are fewer than nine, only response categories with a program name are displayed. Response categories 14–16 display the state-specific names for the Marketplace and only response categories with Marketplace names are displayed.

1. Medicaid
2. Medical Assistance
3. Indian Health Service
4. MinnesotaCare
5. Minnesota Comprehensive Health Association (MCHA)
6. PMAP
13. Healthcare.gov
16. Plan through MNsure
17. Other government plan
18. Other (please specify)

- DK/REF

Skip Instructions

- if 3 (IHS) → BEFORAFT_LC1
- else if 17, 18 (non-specific other government plan or other/specify) then → MISC-SPEC_LC1
- else if 13–16 (marketplace plan) then → POLHOLDER2_LC1
- all others (Medicaid, CHIP, state-specific government plan, DK, REF) → PORTAL_LC1

MISCSPEC_LC1
[open text; 65 characters] → PORTAL_LC1

PORTAL_LC1
◆ ASK OR VERIFY
Is that coverage through MNsure, which may also be known as healthcare.gov?
1. Yes → EXCHTYPE_LC1
2. No → CK-POLHOLDER2_LC1
- DK/REF → CK-POLHOLDER2_LC1

EXCHTYPE_LC1
◆ ASK OR VERIFY
What do you call it – MNsure or healthcare.gov?
1. MNsure
2. Healthcare.gov
- DK/REF
[all] → CK-POLHOLDER2_LC1

CK-HIPAID_LC1
Is coverage related to employment?
- Yes → HIPAID_LC1
- No → BEFOREAFT_LC1

HIPAID_LC1
Does (name's/policyholder names's) employer or union pay for all, part, or none of the health insurance premium?
◆ Report here employer's contribution to employee's health insurance premiums, not the employee's medical bills.
1. All
2. Part
3. None
- DK/REF
[all] → BEFOREAFT_LC1

CK-POLHOLDER2_LC1
Was POLHOLDER_LC1 already asked?
- Yes → PREMYN_LC1
- No → POLHOLDER2_LC1

POLHOLDER2_LC1
◆ ASK OR VERIFY
Whose name is the policy in (Who is the policyholder)?
1. *household member 1*
2. *household member 2*
. . . . . .

16. *household member 16*
17. Someone living outside the household

- DK/REF

[all] → PREMYN_LC1

PREMYN_LC1

Is there a monthly premium for this plan?

◆ **READ IF NECESSARY:** A monthly premium is a fixed amount of money people pay each month to have health coverage. It does not include copays or other expenses such as prescription costs.

1. Yes → PREMSUBS_LC1
2. No → METAL_LC1

- DK/REF → METAL_LC1

PREMSUBS_LC1

Is the cost of the premium subsidized based on [if single-person hh and NAME is policyholder fill: your/else fill: family] income?

◆ **READ IF NECESSARY:** A monthly premium is a fixed amount of money people pay each month to have health coverage. It does not include copays or other expenses such as prescription costs.

◆ **READ IF NECESSARY:** Subsidized health coverage is insurance with a reduced premium. Low and middle income families are eligible to receive tax credits that allow them to pay lower premiums for insurance bought through healthcare exchanges or marketplaces.

1. Yes
2. No

- DK/REF

[all] → PREMCOST_LC1

PREMCOST_LC1

How much is the premium for this plan?

READ IF NECESSARY: A monthly premium is a fixed amount of money people pay each month to have health coverage. It does not include copays, deductibles, or other expenses such as prescription costs.

[open text] → PREMUNIT_LC1

- DK/REF → METAL_LC1

PREMUNIT_LC1

ASK OR VERIFY

Is that per month, quarter, year, or some other time period?

1. Every 2 weeks
2. Month
3. Quarter
4. Year
5. Other (please specify) → UNITSP_LC1 (open-text specify)

- DK/Ref
  ⇒ METAL_LC1

METAL_LC1
Some health plans are sold at different levels of coverage: bronze, silver, gold and platinum. And some people, including young people under 30, can purchase a catastrophic plan. Is this plan a. . .
[READ LIST; ENTER ONLY ONE].
NOTE: Catastrophic plans are only available for those under 30 years old or those with a "hardship exemption"
1. Bronze
2. Silver
3. Gold
4. Platinum or a
5. Catastrophic plan or
6. None of the above?
- DK/Ref
  ⇒ BEFORAFT_LC1

Section C: Months of Coverage

BEFORAFT_LC1
Did [your/NAME's] coverage from [PLANTYPE] start before January 1, [CY-1]?
◆ **READ IF NECESSARY:** Your best estimate is fine.
If PLANTYPE is job-related fill:
◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched employers or plans through [your/their] employer, consider it the same plan.
If PLANTYPE is directly-purchased fill:
◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched plans that you/he/she buys, consider it the same plan.
1. Yes → CNTCOV_LC1
2. No → MNTHBEG1_LC1
- DK/REF → ANYTHIS_LC1

MNTHBEG1_LC1
In which month did that coverage start?
◆ **READ IF NECESSARY:** Your best estimate is fine.
If PLANTYPE is job-related fill:
◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched employers or plans through [your/their] employer, consider it the same plan.
If PLANTYPE is directly-purchased fill:
◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched plans that you/he/she buys, consider it the same plan.
◆ This question refers to [PLANTYPE].

1. January
2. February

. . . . . .

12. December

- DK/REF

If MNTHBEG1_LC1 = current month or earlier → YEARBEG1_LC1
If MNTHBEG1_LC1 = later than current month → CNTCOV_LC1
If MNTHBEG1_LC1 = (D/R) → ANYTHIS_LC1

YEARBEG1_LC1

◆ ASK OR VERIFY

Which year was that?
If PLANTYPE is job-related fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched employers or plans through [your/their] employer, consider it the same plan.

If PLANTYPE is directly-purchased fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched plans that you/he/she buys, consider it the same plan.

◆ This question refers to [PLANTYPE].

1. CY-1 → CNTCOV_LC1
2. CY → CNTCOV_LC1

- DK/REF → ANYTHIS_LC1

CNTCOV_LC1

Has it been continuous since [January, CY-1/month and year from MNTH/YRBEG1]?
If PLANTYPE is job-related fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched employers or plans through [your/their] employer, consider it the same plan.

If PLANTYPE is directly-purchased fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched plans that you/he/she buys, consider it the same plan.

◆ **READ IF NECESSARY:** If the gap in coverage was less than three weeks, consider the coverage "continuous."

◆ This question refers to [PLANTYPE].

1. Yes → CK-OTHMEMB_LC1
2. No → MNTHBEG2_LC1

- DK → MNTHBEG2_LC1
- REF → ANYTHIS_LC1

MNTHBEG2_LC1

In which month did this most recent period of coverage start?

◆ **READ IF NECESSARY:** Your best estimate is fine.

If PLANTYPE is job-related fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched employers or plans through [your/their] employer, consider it the same plan.

If PLANTYPE is directly-purchased fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched plans that you/he/she buys, consider it the same plan.

◆ This question refers to [PLANTYPE].

1. January
2. February

. . . . . .

12. December

• DK/REF

If MNTHBEG2_LC1 = current month or earlier → YEARBEG2_LC1
If MNTHBEG2_LC1 = later than current month → SPELLADD_LC1
Else If MNTHBEG2_LC1 = (D/R) → if covered all months of CY => ANYLAST_LC1; else → ANYTHIS_LC1

YEARBEG2_LC1

◆ ASK OR VERIFY

Which year was that?

If PLANTYPE is job-related fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched employers or plans through [your/their] employer, consider it the same plan.

If PLANTYPE is directly-purchased fill:

◆ **READ IF NECESSARY:** If [you/POLICYHOLDER NAME] switched plans that you/he/she buys, consider it the same plan.

◆ This question refers to [PLANTYPE].

1. [CY-1] → SPELLADD_LC1
2. [CY] → SPELLADD_LC1

• DK → if covered all months of CY → ANYLAST_LC1; else → ANYTHIS_LC1
• REF → if covered all months of CY → ANYLAST_LC1; else → ANYTHIS_LC1

SPELLADD_LC1

I have recorded that [you were/NAME was] covered by [PLANTYPE] in [read months covered]. Were there any OTHER months between January [CY-1] and now that [you were/NAME was] also covered by [PLANTYPE]?

1. Yes → if covered all months of CY → ANYLAST_LC1; else → ANYTHIS_LC1
2. No → CK-OTHMEMB_LC1

• DK/REF → CK-OTHMEMB_LC1

ANYTHIS_LC1

Which months [were you/was NAME] covered by [PLANTYPE] THIS year – in [CY]?

◆ Choose all months that apply

1. January
2. February
3. March
4. April

20. All months of CY

21. No months of CY
- DK/REF

[all] → ANYLAST_LC1

ANYLAST_LC1

Which months [were you/was NAME] covered by [PLANTYPE] LAST year – in [CY-1]?

◆ Choose all months that apply

1. January
2. February

. . . . . .

12. December
20. All months of CY-1
21. No months of CY-1

- DK/REF

[all] → CK-OTHMEMB_LC1

CK-OTHMEMB_LC1

Does this household have 2 or more members?

- Yes → OTHMEMB_LC1
- No → CK-OTHOUT_LC1

Section D: Other Household Members Covered by Leader's Plan, and Months Covered

OTHMEMB_LC1

Between January 1, [CY-1] and now, was anyone in the household other than [you/NAME] ALSO covered by [PLANTYPE]?

1. Yes → COVWHO_LC1
2. No → CK-OTHOUT_LC1

- DK/REF → CK-OTHOUT_LC1

Hard edit: If NAME is a dependent on a job or direct-purchase plan and OTHMEMB_LC1 ne "yes" (that is, the respondent fails to report that the policyholder is also on the plan) store a "Yes"

COVWHO_LC1

Who else was covered? (Who else was covered by [PLANTYPE]?)

◆ **PROBE:** Anyone else?

0. *household member 1*
1. *household member 2*

. . . . . .

16. *household member 16*
96. all persons listed
97. DK/REF

- Any household member → CK-SAMEMNTHS_LC1
- DK/REF => CK-OTHOUT_LC1

Hard edit: If NAME is a dependent on a job or direct-purchase plan and the policyholder is not selected, store policyholder's name in COVWHO_LC1

CK-SAMEMNTHS_LC1

- If leader was covered all months → SAMEMNTHS_LC1
- If leader was NOT covered all months → MNTHS_LC1

SAMEMNTHS_LC1

[Was/Were] [NAME/NAMEs] also covered from January 1, CY-1 until now?

◆ This question refers to [PLANTYPE].

1. Yes (all also covered from January CY-1 until now) → CK-OTHOUT_LC1
2. No (at least one person not covered from January, CY-1 until now)

- DK/REF → MNTHS_LC1

MNTHS_LC1

[First person] Which months between January [CY-1] and now was [NAME from COVWHO_LC1] covered?

[Second + person] How about NAME? (Which months between January [CY-1] and now was [NAME] covered?)

◆ Choose all months that apply

◆ This question refers to [PLANTYPE].

1. January CY-1
2. February CY-1

. . . . . .

12. December CY-1
13. January CY
14. February CY
15. March CY
16. April CY
17. DK/REF
20. All months from January 2013 until now
21. No months from January 2013 until now

[all] → Loop through all persons reported in COVWHO_LC1; then => CK-OTHOUT_LC1

CK-OTHOUT_LC1

- If PLANTYPE is private → OTHOUT_LC1
- Else → CK-ADDGAP1_L

OTHOUT_LC1

Does that plan cover anyone living outside this household?

1. Yes → OTHWHO_LC1
2. No → CK- ADDGAP1_L

- DK/REF → CK- ADDGAP1_L

OTHWHO_LC1

How old are they – under 19, 19–25 or older than 25? [MARK ALL THAT APPLY]?

1. Under 19
2. 19–25 years old

3. Older than 25
- DK/REF

[all] → CK-ADDGAP1_L

Additional Plans

CK-ADDGAP1_L

Are there any gaps in coverage for NAME?
- Yes (gaps in coverage) → ADDGAP1_L
- No (no gaps in coverage) → ADDOTH1_L

ADDGAP1_L

So far, I have recorded that [you were/NAME was] NOT covered in [months not covered]. [Were you/Was NAME] covered by any type of health plan or health coverage in [that/those] month(s)?

◆ **READ IF NECESSARY:** Do not include plans that cover only one type of care, such as dental or vision plans.

1. Yes → SRCEGEN_LP1
2. No → ADDOTH1_L
- DK/REF → ADDOTH1_L

Past Loop

The Past Loop is designed to capture plan type, months of coverage, other household members covered by the same plan, and the months they were covered. As such, the Past Loop consists of all items in Sections B through D above, but with the following exceptions. First, all items in the Past Loop are worded in the past tense. Second, for Section C of the past loop, there is only a single item asking about months of coverage. This is because for current coverage the questionnaire anchors the respondent in their day-of coverage and then establishes the start month of the spell. For coverage that is not held on the day of the interview it is not possible to employ this same technique so we simply ask what months throughout the 16-month reference period the coverage was held, as follows:

WMNTHS_LP1

Which months between January [CY-1] and now [were you/was NAME] covered by [PLANTYPE]?

◆ Choose all months that apply

1. January CY-1
2. February CY-1

. . . . . .

12. December CY-1
13. January CY
14. February CY
15. March CY
16. April CY
17. DK/REF

20. All months from January 2013 until now
21. No months from January 2013 until now

[all] → CK-OTHMEMB_LP1

Once months of coverage are established for the leader, the respondent skips to Section D to determine whether other household members were also covered by the same plan.

SRCEGEN_LP1 thru OTHWHO_LP1

- *Copy all items in Sections B through D in the Current Loop (with the exception above for Section C) and replace "_LC1" with "__LP1."*
- *All answer choices at end of Section D => ADDOTH1_L*

ADDOTH1_L

[Other than [PLANTYPEs],] [W/were you/W/was NAME] covered by any [other] health plan or health coverage AT ANY TIME between January 1, CY-1 and now?

◆ **READ IF NECESSARY:** Do not include plans that cover only one type of care, such as dental or vision plans.

1. Yes → SRCEGEN_LP2
2. No → CK-NEXTMEMB

- DK/REF → CK-NEXTMEMB

If ADDOTH1_L is answered for Person 1 then set MARKTWO = 2 (sufficient partial)

SRCEGEN_LP2 thru OTHWHO_LP2

- *Copy all items in Past Loop and replace "_LP1" with "__LP2."*
- *All answer choices at end of Section D => ADDOTH2_L*

ADDOTH2_L

[Other than [PLANTYPEs],] [W/were you/W/was NAME] covered by any [other] health plan or health coverage AT ANY TIME between January 1, CY-1 and now?

◆ **READ IF NECESSARY:** Do not include plans that cover only one type of care, such as dental or vision plans.

1. Yes → SRCEGEN_LP3
2. No → CK-NEXTMEMB

- DK/REF → CK-NEXTMEMB

SRCEGEN_LP3 thru OTHWHO_LP3

- *copy all items in Past Loop and replace "_LP1" with "__LP3."*
- *All answer choices at end of Section D => CK-NEXTMEMB*

CK-NEXTMEMB

Have all household members been asked about explicitly?

- Yes → HEALTHSTATUS_INTRO
- No → FINTRO

Additional Plans for Follower

FHINTRO
Next I'm going to ask you about NAME's health coverage.
◆ Press 1 to Continue

CK-ADDGAP1_F
Are there any gaps in coverage for NAME?
- Yes (gaps in coverage) → ADDGAP1_F
- No (no gaps in coverage) → ADDOTH1_F

ADDGAP1_F
So far, I have recorded that [you were/NAME was] NOT covered in [months not covered]. [Were you/Was NAME] covered by any type of health plan or health coverage in [that/those] month(s)?
◆ **READ IF NECESSARY:** Do not include plans that cover only one type of care, such as dental or vision plans.
1. Yes → SRCEGEN_FP1
2. No → ADDOTH1_F
- DK/REF → ADDOTH1_F

SRCEGEN_FP1 thru OTHWHO_FP1
- *copy all items in Past Loop and replace "_LP1" with "__FP1."*
- *All answer choices at end of Section D => ADDOTH1_F*

ADDOTH1_F
[Other than [PLANTYPEs],] [W/were you/W/was NAME] covered by any [other] health plan or health coverage AT ANY TIME between January 1, CY-1 and now?
◆ **READ IF NECESSARY:** Do not include plans that cover only one type of care, such as dental or vision plans.
1. Yes → SRCEGEN_FP2
2. No → CK-NEXTMEMB2
- DK/REF → CK-NEXTMEMB2

SRCEGEN_FP2 thru OTHWHO_FP2
- *copy all items in Past Loop and replace "_LP1" with "__FP2."*
- *All answer choices at end of Section D => ADDOTH2_F*

ADDOTH2_F
[Other than [PLANTYPEs],] [W/were you/W/was NAME] covered by any [other] health plan or health coverage AT ANY TIME between January 1, CY-1 and now?
◆ **READ IF NECESSARY:** Do not include plans that cover only one type of care, such as dental or vision plans.
1. Yes → SRCEGEN_FP3
2. No → CK-NEXTMEMB2
- DK/REF → CK-NEXTMEMB2

SRCEGEN_FP3 thru OTHWHO_FP3

- *copy all items in Past Loop and replace "_LP1" with "__FP3."*
- *All answer choices at end of Section D => HEALTHSTATUS_INTRO*

CK-NEXTMEMB2

Have all household members been asked about explicitly?

- Yes → HEALTHSTATUS_INTRO
- No → FINTRO for next person

ACS Health Insurance Module

ACSJOB

I am now going to ask you some questions about [your/NAME's] health insurance and health coverage. [Are you/Is NAME] currently covered by health insurance through a current or former employer or union of [yours/yours or another family member/ <him/her> or another family member]?

◆ NOTE: If the respondent says this person has health coverage through the military, mark "2" and tell them that military health insurance/coverage will be discussed later.

1. Yes
2. No

- DK/Ref
  ⇒ ACSDIR

ACSDIR

[Are you/Is NAME] currently covered by health insurance purchased directly from an insurance company by [you/you or another family member/ <him/her> or another family member]?

1. Yes
2. No

- DK/Ref
  ⇒ ACSMCARE

Soft Edit: if ACSJOB = 1 and ACSDIR = 1 ask: "I recorded that (Fill 1: you/ <NAME>) (have/has) both insurance through an employer or union AND insurance directly purchased through an insurance company. These are two different plans, is that correct?"

◆ If correct, suppress and continue.

◆ If not, determine which is the primary plan and go back to and change the "yes" to a "no" for the other plan

ACSMCARE

[Are you/Is NAME] currently covered by Medicare, for people age 65 or older or people with certain disabilities?

1. Yes
2. No

- DK/Ref
  ⇒ ACSMCAID

ACSMCAID
[Are you/Is NAME] currently covered by Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability?
1. Yes
2. No
- DK/Ref
  ⇒ ACSMIL

ACSMIL
[Are you/Is NAME] currently covered by TRICARE or other military health care?
1. Yes
2. No
- DK/Ref
  ⇒ ACSVA

ACSVA
[Are you/Is NAME] currently covered through the Veteran's Administration or [have you/has NAME] ever used or enrolled for VA health care)?
1. Yes
2. No
- DK/Ref
  ⇒ ACSIHS

ACSIHS
[Are you/Is NAME] currently covered through the Indian Health Service?
1. Yes
2. No
- DK/Ref
  ⇒ ACSOTHER

ACSOTHER
[Are you/Is NAME] currently covered by any other health insurance or health coverage plan?
1. Yes → ACSOTHERS
2. No → CK-ACSLAST
- DK/Ref → CK-ACSLAST

ACSOTHERS
What is the name of the health care plan?
[open text; allow 30 characters]
  ⇒ CK-ACSLAST

CK-ACSLAST

- If there is another person on the roster (regardless of age) → ACSJOB
- Else if at least one plan was reported → ACS_MKT
- Else → HEALTHSTAT

ACS_MKT

Was this plan obtained through a State or Federal Marketplace, Healthcare.gov, or a similar state website?

1. Yes
2. No
- DK/REF
  ⇒ ACS_PREM

ACS_PREM

Do you or another family member pay a premium for this health insurance plan? A premium is a fixed amount of money paid on a regular basis for health coverage. It does not include copays, deductibles, or other expenses such as prescription costs.

1. Yes → ACS_SUBS
2. No → ACS_METAL
- DK/REF → ACS_METAL

ACS_SUBS

Based on family income, do you or another family member receive financial assistance through a subsidy or tax credit to help pay part or all of the cost of the premium for this plan?

1. Yes
2. No
- DK/REF
  ⇒ ACS_PREMCOST

ACS_PREMCOST

How much is the premium for this plan?

READ IF NECESSARY: A premium is a fixed amount of money paid on a regular basis for health coverage. It does not include copays, deductibles, or other expenses such as prescription costs.

[open text] → ACS_PREMUNIT

- DK/REF → ACS_METAL

ACS_PREMUNIT

ASK OR VERIFY

Is that per month, quarter, year, or some other time period?

1. Every 2 weeks
2. Month
3. Quarter
4. Year

5. Other (please specify) → ACS_UNITSP (open text specify)
- DK/Ref
  ⇒ ACS_METAL

ACS_METAL
Some health plans are sold at different levels of coverage: bronze, silver, gold and platinum. And some people, including young people under 30, can purchase a catastrophic plan. Is this plan a. . .
[READ LIST; ENTER ONLY ONE].
NOTE: Catastrophic plans are only available for those under 30 years old or those with a "hardship exemption"
1. Bronze
2. Silver
3. Gold
4. Platinum or a
5. Catastrophic plan or
6. None of the above?
- DK/Ref
  ⇒ ACS_PATHWAY

ACS_PATHWAY
There are many different ways to obtain information on the health insurance plans in the marketplace. Which of the following sources of information did you use or try to use to obtain information?
MARK ALL THAT APPLY
1. Website, including online chat option
2. Newspaper, radio, or television
3. Call center
4. Assistance from navigators, application assisters, certified application counselors, or community health workers
5. Assistance from an insurance agent or broker
6. Assistance from family or friends
7. Assistance from an employer
8. Assistance from a tax preparer
9. Assistance from Medicaid or another program agency such as TANF, SNAP, or WIC
10. Assistance from a hospital, doctor's office, or clinic
11. Other (please specify) → ACS_PATHSP (open text specify)
- DK/Ref
  ⇒ HEALTHSTAT

HELP SCREENS

For ACSMCAID:
Medicaid, medical assistance, or government assistance plans for those with low incomes or a disability may be known by different names in different states. Below is a list of

program names by state. This list is not comprehensive, but provides guidance for those not familiar with the term Medicaid and may only know their specific state program name. [fill state-specific program name(s) based on the attachment]

For all items except ACSMCAID:
DATA USES

- Used to allocate funds to states and local areas for governmentprovided health care.
- Used by federal agencies, such as the Department of Health and Human Services, to evaluate the effectiveness of government health care programs.
- Used by federal and local agencies to examine the adequacy of existing health care facilities in meeting current and future health care needs.

WHY WE ASK IT THIS WAY

- These questions ask about each type of insurance a respondent may have.
- Insurance can include both private coverage (provided by an employer or purchased) as well as public coverage (from government programs such as Medicare, Medicaid, and VA).
- The reason the question specifies (health insurance or health coverage plans is because many types of public (government) coverage are not technically health insurance plans. The goal of the item is to obtain information on whether an individual has health insurance coverage and if so, what kind of coverage he/she has.

7.4. Appendix D, Sample Distribution by Strata for Standard and Augmented Samples

Appendix D. Sample Distribution by Strata for Standard and Augmented Samples[1].

	Standard Sample					Augmented Sample				
	CPS		ACS		Total	CPS		ACS		Total
Strata[2]	n	%	n	%	n	n	%	n	%	n
ESI	280	17.6	225	15.6	505	280	14.6	227	12.8	507
Non-group	629	39.6	528	36.5	1,157	629	32.8	528	29.8	1,157
Marketplace	178	11.2	152	10.5	330	178	9.3	152	8.6	330
MinnesotaCare	46[3]	2.9	24[3]	1.7	70[3]	332	17.3	292	16.5	624
Medicaid	390	24.5	431	29.8	821	414	21.6	457	25.8	871
Transition	66	4.2	87	6.0	153	86	4.5	118	6.7	204
TOTAL	1,589	100	1,447	100	3,036	1,589	100	1,447	100	3,693

[1]Standard Sample excludes enrollees in the MinnesotaCare strata; Augmented sample includes enrollees in the Medicaid and MinnesotaCare strata.

[2]ESI refers to employer sponsored insurance; Non-group is insurance that is purchased directly, not through an employer group or association and not on the portal; Marketplace is non-group/direct-purchase coverage available on the portal for which many enrollees receive a subsidy for the monthly premium; MinnesotaCare is a state-specific program for low-income families that charges a sliding-fee premium; the Transition strata is comprised of policyholders who switched from ESI to public or vice versa in 2014.

[3]Includes cases enrolled in both MinnesotaCare and another type of coverage.

8. References

American Association of Public Opinion Research. 2016. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed March 2019).

Bennefield, R.L. 1996. Dynamics of Economic Well-Being: Health Insurance, 1992 to 1993, Who Loses Coverage and for How Long? Available at: https://www.census.gov/prod/1/pop/p70-54.pdf (accessed March 2019).

Blewett, L.A. and M. Davern. 2006. "Meeting the Need for State-Level Estimates of Health Insurance Coverage: Use of State and Federal Survey Data." *Health Services Research* 41(3p1): 946–975. Doi: https://doi.org/10.1111/j.1475-6773.2006.00543.x.

Blumberg, S.J. and M.L. Cynamon. 1999. Misreporting Medicaid Enrollment: Results of three studies linking telephone surveys to state administrative records. Available at: https://www.cdc.gov/nchs/data/hsrmc/hsrmc_7th_proceedings_1999.pdf (accessed March 2019).

Blumberg, S.J., L. Osborn, J.V. Luke, L. Olson, and M.R. Frankel. 2004. "Estimating the prevalence of uninsured children: an evaluation of data from the National Survey of Children with Special Health Care Needs, 2001." *Vital and Health Statistics. Series 2,* (136): i–38. Available at: https://www.cdc.gov/nchs/data/series/sr_02/sr02_136.pdf (accessed March 2019).

Call, K.T., M.E. Davern, J.A. Klerman, and V. Lynch. 2013. "Comparing errors in Medicaid reporting across surveys: Evidence to date." *Health Services Research* 48(2 PART1): 652–664. Doi: https://doi.org/10.1111/j.1475-6773.2012.01446.x.

Call, K.T., A.R. Fertig, J. Pascale, and D. Oellerich. 2018. Who gets it right? Characteristics associated with accurate reporting of health insurance coverage. In Paper presented at the Academy Health Annual Research Meeting, Seattle, WA. U.S.A. June 25, 2018. Available at: https://www.academyhealth.org/events/2018-06/2018-annual-research-meeting (accessed March 2019).

Cantor, J.C., A.C. Monheit, S. Brownlee, and C. Schneider. 2007. "The adequacy of household survey data for evaluating the nongroup health insurance market." *Health Services Research* 42(4): 1739–1757. Doi: https://doi.org/10.1111/j.1475-6773.2006.00662.x.

Czajka, J.L. and K. Lewis. 1999. Using National Survey Data to Analyze Children's Health Insurance Coverage: An Assessment of Issues. Available at: https://www.mathematica-mpr.com/our-publications-and-findings/publications/using-national-survey-data-to-analyze-childrens-health-insurance-coverage-an-assessment-of-issues (accessed March 2019).

Davern, M., K.T. Call, J. Ziegenfuss, G. Davidson, T.J. Beebe, and L. Blewett. 2008. "Validating health insurance coverage survey estimates: A comparison of self-reported coverage and administrative data records." *Public Opinion Quarterly* 72(2): 241–259. Doi: https://doi.org/10.1093/poq/nfn013.

Eberly, T., M.B. Pohl, and S. Davis. 2009. "Undercounting Medicaid enrollment in Maryland: Testing the accuracy of the current population survey." *Population Research and Policy Review* 28(2): 221–236. Doi: https://doi.org/10.1007/s11113-008-9078-5.

Farley-Short, P. 2001. Counting and Characterizing the Uninsured. Available at: http://rwjf-eriu.org/pdf/farleyshort-final.pdf (accessed March 2019).

Fertig, A.R., J. Pascale, K.T. Call, and D. Oellerich. 2018. Design and Sampling Strategy for a Validation Study Linking Enrollment Records to Survey Reports: the CHIME Study. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/rsm2018-10.pdf (accessed December 2018).

Hill, S.C. 2007. "The Accuracy of Reported Insurance Status in the MEPS." *Inquiry* 44(4): 443–468. Doi: https://doi.org/10.5034/inquiryjrnl_44.4.443.

Klerman, J.A., M. Davern, K.T. Call, V. Lynch, and J.D. Ringel. 2009. "Understanding The Current Population Survey's Insurance Estimates And The Medicaid 'Under-count'." *Health Affairs* 28(6): w991–w1001. Doi: https://doi.org/10.1377/hlthaff.28.6.w991.

Lewis, K., M.R. Ellwood, and J.L. Czajka. 1998. Counting the Uninsured: A Review of the Literature. Available at: https://www.urban.org/research/publication/counting-uninsured (accessed March 2019).

Lurie, I.Z. and J. Pearce. 2018. Health Insurance Coverage from Administrative Tax Data. Available at: https://www.treasury.gov/resource-center/tax-policy/Pages/tax_analysis_paper.aspx (accessed May 2019).

Mach, A. and B. O'Hara. 2011. Do people really have multiple health insurance plans? Estimates of Nongroup Health Insurance in the American Community Survey. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2011/demo/SEHSD-WP2011-28.pdf (accessed March 2019).

Marquis, M.S. 1983. "Consumers' Knowledge about their Health Insurance Coverage." *Health Care Financ Rev* 5(1): 65–80. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191335/ (accessed May 2019).

Minnesota Department of Health. 2018. Section 4: Individual and Small Group Health Insurance Markets – Chart Summaries. Available at: https://www.health.state.mn.us/data/economics/chartbook/summaries/section4summaries.html (accessed March 2019).

Nelson, D.E., B.L. Thompson, N.J. Davenport, and L.J. Penaloza. 2000. "What people really know about their health insurance: A comparison of information obtained from individuals and their insurers." *American Journal of Public Health* 90(6): 924–928. Doi: https://doi.org/10.2105/AJPH.90.6.924.

Noon, J.M., L.E. Fernandez, and S.R. Porter. 2019. "Response error and the Medicaid undercount in the current population survey." *Health Services Research*, 54(1): 34–43. Doi: https://doi.org/10.1111/1475-6773.13058.

Pascale, J. 2008. "Measurement Error in Health Insurance Reporting." *Inquiry* 45(4): 422–437. Doi: https://doi.org/10.5034/inquiryjrnl_45.04.422.

Pascale, J. 2009. Health Insurance Measurement A Synthesis of Cognitive Testing Findings. In *Questionnaire Evaluation Standards (QUEST) Meeting, May 18–20*. Bergen, Norway. Available at: https://wwwn.cdc.gov/qbank/QUest/2009/pres10.pdf (accessed March 2019).

Pascale, J. 2016. "Modernizing a major federal government survey: A Review of the redesign of the current population survey health insurance questions." *Journal of Official Statistics* 32(2): 461–486. Doi: https://doi.org/10.1515/JOS-2016-0024.

Pascale, J., M. Boudreaux, and R. King. 2016. "Understanding the New Current Population Survey Health Insurance Questions." *Health Services Research* 51(1): 240–261. Doi: https://doi.org/10.1111/1475-6773.12312.

Pascale, J., K.T. Call, and A.R. Fertig. 2018a. Using a Machine Learning Approach to Classify Health Insurance Type from Survey Responses Using Enrollment Data. In Paper presented at the AcademyHealth Annual Research Meeting, Seattle, WA.U.S.A. June 25, 2018. Available at: https://www.academyhealth.org/events/2018-06/2018-annual-research-meeting (accessed March 2019).

Pascale, J., K.T. Call, and A.R. Fertig. 2018b. Using Enrollment Records to Guide Categorization of Health Insurance Coverage Type Post-ACA. In Paper presented at the AcademyHealth Annual Research Meeting, Seattle, WA.U.S.A. June 25, 2018. Available at: https://www.academyhealth.org/events/2018-06/2018-annual-research-meeting (accessed May 2019).

Pascale, J., J. Rodean, L. Leeman, C. Cosenza, and A. Schoua-Glusberg. 2013. "Preparing to Measure Health Coverage in Federal Surveys Post-Reform." *Inquiry* 50(2): 106–123. Doi: https://doi.org/10.1177/0046958013513679.

Pascale, J., M.I. Roemer, and D.M. Resnick. 2009. Medicaid underreporting in the CPS: Results from a Record Check Study. *Public Opinion Quarterly* 73(3): 497–520. Doi. https://doi.org/10.1093/poq/nfp028.

Rosenbach, M. and K. Lewis. 1998. Estimates of Health Insurance Coverage in the Community Tracking Study and the Current Population Survey. Available at: https://www.researchgate.net/publication/267196753_Estimates_of_Health_Insurance_Coverage_in_the_Community_Tracking_Study_and_the_Current_Population_Survey (accessed March 2019).

Swartz, K. 1986. "Interpreting the Estimates from Four National Surveys of the Number of People Without Health Insurance." *Journal of Economic and Social Measurement* 14(3): 233–242. Doi: https://doi.org/10.3233/JEM-1986-14306.

U.S. Census Bureau. 2016a. Table DP05 ACS Demographic and Housing Estimates. Available at: http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF (accessed June 2016).

U.S. Census Bureau. 2016b. Table S1201 Educational Attainment. Available at: http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF (accessed June 2016).

Received April 2018
Revised February 2019
Accepted March 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 461–486, http://dx.doi.org/10.2478/JOS-2019-0020

Decomposing Multilateral Price Indexes into the Contributions of Individual Commodities

Michael Webster[1] *and Rory C. Tarnow-Mordi*[2]

This article describes methods for decomposing price indexes into contributions from individual commodities, to help understand the influence of each commodity on aggregate price index movements.

Previous authors have addressed the decomposition of bilateral price indexes, which aggregate changes in commodity prices from one time period to another. Our focus is the decomposition of multilateral price indexes, which aggregate commodity prices across more than two time periods or countries at once. Multilateral indexes have historically been used for spatial comparisons, and have recently received attention from statistical agencies looking to produce temporal price indexes from large and high frequency price data sets, such as scanner data. Methods for decomposing these indexes are of practical relevance.

We present decompositions of three multilateral price indexes. We also review methods proposed by other researchers for extending multilateral indexes without revising previously published index levels, and show how to decompose the extended indexes they produce. Finally, we use a data set of seasonal prices and quantities to illustrate how these decomposition methods can be used to understand the influence of individual commodities on multilateral price index movements, and to shed light on the relationships between various multilateral and extension methods.

Key words: Scanner data; time product dummy; GEKS; Geary-Khamis; linking indexes.

1. Decomposition of Bilateral Price Indexes

Price indexes are used to combine the price changes of individual commodities into an aggregate measure of price change. Statistical agencies also find it useful to work in the opposite direction: to decompose a price index into the contributions of individual commodities. This facilitates the identification of the commodities with the greatest contributions to change, which is helpful for validating the inputs and explaining the index (ILO et al. 2004, chap. 9).

It is useful to start with a few straightforward examples. A price index that takes the form of an arithmetic mean of commodity price ratios or relatives has an *additive*

[1] Australian Bureau of Statistics, 45 Benjamin Way, Belconnen, ACT 2617, Australia. Email: michael.webster@abs.gov.au

[2] Exposé: Data Exposed, Margaret Graham Building, Lot 14, Frome Road, Adelaide, SA 5000 Australia. Email: rory.tarnow-mordi@exposedata.com.au

Acknowledgments: This research was performed while Rory C. Tarnow-Mordi was employed by the Australian Bureau of Statistics. Views expressed are those of the authors and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted or used, they should be attributed clearly to the authors. The authors would like to thank an Associate Editor, two anonymous referees, Frances Krsinich, Jan de Haan, Lyndon Ang, Marcel van Kints, Justin Farrow, Siu-Ming Tam and Daniel Melser for helpful comments on earlier drafts. We take full responsibility for any errors in the manuscript that remain.

decomposition. In other words, it can be decomposed into a sum of contributions, each depending on prices (or price changes) of an individual commodity:

$$P^{0,1} = \sum_i c_i(\mathbf{p}_i, \mathbf{w}_i) \tag{1}$$

where $\mathbf{p}_i$ is a vector of prices for commodity i, $\mathbf{w}_i$ is a weight (vector) used to aggregate its prices with the prices of other commodities, and c_i is some unspecified function that depends only on prices and weights of commodity i. Note that the subscript in c_i is not strictly necessary but is included to simplify references to summation terms.

For instance, the Laspeyres index between two periods (0 and 1) can be expressed as

$$P_{\mathrm{L}}^{0,1} = \frac{\sum_i p_i^1 q_i^0}{\sum_i p_i^0 q_i^0} = \sum_i s_i^0 \frac{p_i^1}{p_i^0} \tag{2}$$

where p_i^0 and p_i^1 are the prices of commodity i in periods 0 and 1, q_i^0 is the quantity of commodity i in period 0, and $s_i^0 = p_i^0 q_i^0 / \sum_{j=1}^N p_j^0 q_j^0$ is the expenditure share of commodity i in period *0*.

Similarly, an index that can be expressed as a geometric mean of price relatives has a simple *multiplicative decomposition*:

$$P^{0,1} = \prod_i c_i(\mathbf{p}_i, \mathbf{w}_i) \tag{3}$$

For instance, the Törnqvist index between 0 and 1 can be expressed as

$$P_{\mathrm{T}}^{0,1} = \prod_i \left(\frac{p_i^1}{p_i^0}\right)^{\frac{1}{2}(s_i^0 + s_i^1)} \tag{4}$$

where s_i^1 is the expenditure share of commodity i in period 1.

Several authors have written about the decomposition of common bilateral price indexes. Balk (2008, chap. 4) provides a good overview of the topic. As well as presenting decompositions of the straightforward type above — additive decompositions of arithmetic mean indexes, and multiplicative decompositions of geometric mean indexes — Balk also presents additive decompositions of geometric mean indexes, multiplicative decompositions of arithmetic mean indexes, and both arithmetic and multiplicative decompositions of Fisher and Walsh indexes, referencing earlier publications by Van IJzeren (1952, 1983), Vartia (1974, 1976), Diewert (2002) and Reinsdorf et al. (2002).

Many of these decompositions feature a logarithmic mean involving the price index that is being decomposed. For example, Balk shows that a general arithmetic mean index $P^{0,1} = \sum_i w_i \frac{p_i^1}{p_i^0}$ can also be written as

$$P^{0,1} = \prod_i \left(\frac{p_i^1}{p_i^0}\right)^{\sigma_i} \tag{5}$$

where $\left(\frac{p_i^1}{p_i^0}\right)^{\sigma_i}$ is the contribution of commodity i to the arithmetic mean index:

$$\sigma_i = \frac{w_i \times L\left(P^{0,1}, p_i^1/p_i^0\right)}{\sum_j w_j \times L\left(P^{0,1}, p_j^1/p_j^0\right)}$$

and L is the logarithmic mean function, defined as $L(x, y) = \begin{cases} (y - x)/(\ln y - \ln x) & x \neq y \\ x & x = y \end{cases}$ for positive arguments x and y.

There are several possible decompositions of a single price index: for example, the Laspeyres index has an additive decomposition given by Equation 2, as well as a multiplicative decomposition given by Equation 5 where $w_i = s_i^0$. The commodity contributions from some decompositions, such as Equation 5, depend on the aggregate price changes (or levels): in the remainder of this article, we refer to such decompositions as *reflexive*. We refer to decompositions with commodity contributions that depend only on the prices of the relevant commodity and the expenditures (or quantities) of any or all commodities, such as Equations 2 and 4, as *simple*. This distinction has not previously been named in any source that we are aware.

Different decompositions may be useful in different scenarios. For instance, when we are comparing the properties of two price indexes, it is useful to decompose them in similar ways; when we are combining index movements additively or multiplicatively, a corresponding (additive or multiplicative) decomposition facilitates the calculation of contributions to the combined index.

Fundamentally, however, if we are interested in separating out the contributions of individual commodities to a price index, a simple decomposition seems preferable to a reflexive decomposition. This is because the aggregate price change, which the reflexive decomposition explicitly references, necessarily depends on the prices of all commodities.

It seems unavoidable for the contributions to depend on expenditures (or quantities) as these reflect measures of economic importance that are used to aggregate the price index. For a simple decomposition, what is important is that the expenditures (or quantities) do not depend on the price index.

Note that Equation 4 also yields a simple decomposition into the contributions of individual price observations:

$$P_{\mathrm{T}}^{0,1} = \prod_i \prod_t \left(p_i^t\right)^{f_i^t(\mathbf{s})} \tag{6}$$

$$\text{where } f_i^t(\mathbf{s}) = \begin{cases} -1/2\left(s_i^0 + s_i^1\right) & t = 0 \\ 1/2\left(s_i^0 + s_i^1\right) & t = 1 \end{cases}$$

It can be shown that decompositions of this general form are unique: if, for a given price index formula, there exist functions of expenditure shares $f_i^t(\mathbf{s})$ and $g_i^t(\mathbf{s})$ satisfying $P^{0,1} = \prod_i \prod_t \left(p_i^t\right)^{f_i^t(\mathbf{s})} = \prod_i \prod_t \left(p_i^t\right)^{g_i^t(\mathbf{s})}$ for any sets of prices p_i^t and expenditure shares s_i^t, then $f_i^t(\mathbf{s}) = g_i^t(\mathbf{s})$ for all i and t. We meet other decompositions of this form later in the article.

2. Decomposition of Multilateral Price Indexes

The price indexes mentioned in the previous section are *bilateral*, in the sense that they measure price change between two time periods 0 and 1. Suppose we are interested in measuring price change over a *window* of adjacent time periods between 0 and T, with $T > 1$. Traditional practice involves either calculating a sequence of bilateral indexes between 0 and each subsequent period $\{P^{0,1}, P^{0,2}, \cdots, P^{0,T}\}$ or a sequence of bilateral indexes between consecutive periods $\{P^{0,1}, P^{1,2}, \cdots, P^{T-1,T}\}$. The former sequence yields a *direct* bilateral index and the latter sequence yields a *chained* bilateral index. Alternatively, we can use a *multilateral* index method to simultaneously estimate a system of price comparisons $\{P^0, \cdots, P^T\}$.

Ivancic et al. (2011) proposed using multilateral methods to produce price indexes from data sets of retail transactions, finding they gave more satisfactory results than either direct or chained bilateral indexes. This has inspired further studies at several statistical agencies with access to scanner data: see, for instance, De Haan and Krsinich (2014), De Haan (2015), Howard et al. (2015), Chessa (2015), Krsinich (2016), Australian Bureau of Statistics (2016, 2017).

A feature of multilateral indexes is that the price comparison between any pair of time periods a and b may depend on prices in other periods, and on commodities that are sold in a and not b or vice versa. This makes it important to be able to decompose multilateral index movements: without this, it is challenging to interpret which commodities' price changes have the greatest influence on price comparisons.

The decomposition of multilateral price indexes is the focus of the remainder of this paper. We decompose three multilateral methods considered in the studies cited above:

1. The Time Product Dummy (TPD) method advocated by Krsinich (2016), which is a temporal analogue of the Country Product Dummy method introduced by Summers (1973),
2. The GEKS method proposed by Gini (1931), Eltetö and Köves (1964) and Szulc (1964), especially the GEKS-Törnqvist or CCD variant proposed by Caves et al. (1982),
3. The Geary-Khamis (GK) method proposed by Geary (1958) and Khamis (1972).

We focus on these specific multilateral methods because a number of statistical agencies are either researching them or starting to use them for the production of official price indexes.

2.1. *Decomposition of the TPD Method*

Suppose we have a set of price observations p_i^t pertaining to periods $t \in \{0, \ldots, T\}$ and commodities $i \in \{1, \ldots, N\}$, possibly with some missingness (combinations of i and t for which p_i^t is not observed or does not exist).

The TPD method involves calculating a system of price comparisons by fitting the model

$$\ln p_i^t = \alpha + \delta^t + \gamma_i + \varepsilon_i^t \tag{7}$$

where α is the intercept, δ^t is the *time effect* parameter for period t, γ_i is the *product (commodity) effect* parameter for commodity i and ε_i^t is an error term. In estimating

the model, we choose an arbitrary time period and commodity to treat as reference categories, and set their effects to zero: for notational convenience, we select period 0 and commodity N.

The remaining parameters in the model are estimated by minimising the sum of squared residuals. Where expenditure information is available, a common approach is to minimise the sum of weighted squared residuals using the expenditure shares s_i^t as weights (see Rao 2005; De Haan and Krsinich 2014; Chessa 2015; Krsinich 2016; Australian Bureau of Statistics 2016).

The time effect parameter estimates reflect the natural logarithm of the price level in each period, relative to period 0, so it is natural to estimate the price level in each period by taking the exponential of the time effect estimates. The TPD price comparison between periods a and b is thus the ratio of price levels

$$P_{\text{TPD}}^{a,b} = \frac{\exp\left(\hat{\delta}^b\right)}{\exp\left(\hat{\delta}^a\right)} = \exp\left(\hat{\delta}^b - \hat{\delta}^a\right) \tag{8}$$

Strictly the exponential transformation introduces a model bias, which in this context is usually implicitly or explicitly treated as small enough to ignore (see, for instance, De Haan et al. 2016).

2.1.1. Simple TPD Decompositions

We can decompose TPD price comparisons by following the weighted least squares process used to derive the parameters. In general, regression model parameter estimates under the weighted least squares process are given by the product of matrices

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{p} = \mathbf{A}\mathbf{p} \tag{9}$$

In cases where the design matrix $\mathbf{X}$ and the weight matrix $\mathbf{W}$ are considered fixed and known, this equation demonstrates that each parameter is a linear combination (represented by matrix $\mathbf{A}$) of the observed variables $\mathbf{p}$; that is $\beta_i = \sum_j A_{i,j}p_j$. This fact, combined with the exponential transformation in Equation 8, gives a natural multiplicative decomposition of the price change between two periods.

Specifically, the weighted least squares equation in our case is composed of

- The parameter estimate vector $\hat{\boldsymbol{\beta}}$ which is $\begin{bmatrix} \hat{\alpha} & \hat{\delta}^1 & \cdots & \hat{\delta}^T & \hat{\gamma}_1 & \cdots & \hat{\gamma}_{N-1} \end{bmatrix}^T$
- The design matrix $\mathbf{X}$ corresponding to the parameter vector and price vector, with a simple structure:

$$\left[\begin{array}{cccc|ccc} 1 & D^1(1) & \cdots & D^T(1) & D_1(1) & \cdots & D_{N-1}(1) \\ & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & D^1(K) & \cdots & D^T(K) & D_1(K) & \cdots & D_{N-1}(K) \end{array}\right] = [\mathbf{X}_\tau|\mathbf{X}_\pi]$$

where $D^t(k)$ and $D_i(k)$ are dummy variables with values of 1 if the k-th price observation pertains to period t and commodity i respectively and zero otherwise. K is the total number of price observations.

- The weight matrix $\mathbf{W}$, a diagonal matrix of expenditure shares: $diag\left(s_i^t\right)$
- The price vector $\mathbf{p}$, which contains the log price observations for each commodity-time, $\begin{bmatrix} \ln p_1 & \ln p_2 & \cdots & \ln p_k & \cdots & \ln p_K \end{bmatrix}^T$

Note that for simplicity, the price vector is indexed with a single variable k instead of the separate time variable t and commodity variable i shown in Equation 7. This difference is superficial: both representations are equivalent, but a single index variable makes the linear algebra simpler.

As only the time effect parameters are needed to estimate TPD comparisons, the weighted least squares solution of Equation 9 can be simplified using the Banachiewicz formula for block matrix inversion (see Puntanen and Styan 2006):

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}$$

where the block matrices are of appropriate dimensions and $\mathbf{A}$, $\mathbf{D}$, and $(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})$ are invertible. Applying this result, we obtain

$$\hat{\delta}^a = \left[\left(\mathbf{W}_\tau - \mathbf{W}_{\tau\pi}\mathbf{W}_\pi^{-1}\mathbf{W}_{\tau\pi}^T\right)^{-1}\left(\mathbf{X}_\tau^T - \mathbf{W}_{\tau\pi}\mathbf{W}_\pi^{-1}\mathbf{X}_\pi^T\right)\mathbf{W}\right]_{a+1}\mathbf{p} = \sum_{k=1}^{K} w_{a,k}\ln(p_k) \quad (10)$$

where $\mathbf{W}_\tau = \mathbf{X}_\tau^T\mathbf{W}\mathbf{X}_\tau$, $\mathbf{W}_\pi = \mathbf{X}_\pi^T\mathbf{W}\mathbf{X}_\pi$, $\mathbf{W}_{\tau\pi} = \mathbf{X}_\tau^T\mathbf{W}\mathbf{X}_\pi$, and the $a+1$ subscript indicates we take row $a+1$ of the matrix in the square brackets. Note that when in the proceeding paragraphs, variable a may be replaced with variable b, but the analogical formulas apply.

Equation (10) defines the weights $w_{a,k}$ for $a > 0$; for $a = 0$ (the reference period) we set $w_{0,k} = 0$ for every k, as the corresponding weights from (10) would yield the parameter estimate $\hat{\alpha}$.

This simplification is useful for computation, as it limits the size of the matrix required to be inverted to the number of time periods included in the model. This is a particular advantage for TPD methods that aggregate the prices of an arbitrary number of commodities over a window of a fixed size, as it protects the performance of any implementation.

Combining Equations 8 and 10, it follows that a decomposition of the TPD price index in terms of commodity price observations is

$$\begin{aligned} P_{\text{TPD}}^{a,b} &= \prod_{k=1}^{K} p_k^{w_{b,k}-w_{a,k}} \\ &= \prod_i \prod_t \left(p_i^t\right)^{w_{b,k(i,t)}-w_{a,k(i,t)}} \\ &= \prod_i \prod_t c_{\text{TPD},i}^t(a,b) \end{aligned} \quad (11)$$

where $k(i,\ t)$ is the observation corresponding to commodity i and period t, and $c_{\text{TPD},i}^t(a,b) = \left(p_i^t\right)^{w_{b,k(i,t)}-w_{a,k(i,t)}}$ is the contribution of price observation p_i^t to the TPD price comparison between periods a and b.

We can alternatively decompose TPD price comparisons by deriving the parameters in a manner similar to Diewert and Fox (2017). The weighted sum of squared errors can be written as

$$E = \sum_i \sum_t s_i^t \left(\log p_i^t - \alpha - \delta^t - \gamma_i \right)^2 \tag{12}$$

The time and commodity effect parameters that minimize E satisfy $\frac{\partial E}{\partial \delta^t} = 0$ and $\frac{\partial E}{\partial \gamma_i} = 0$. This yields a pair of equations

$$\hat{\delta}^t = \sum_i s_i^t \left(\log \mathrm{p}_i^t - \left(\hat{\alpha} + \hat{\gamma}_i \right) \right) \tag{13}$$

$$\hat{\alpha} + \hat{\gamma}_i = \frac{\sum_t s_i^t \left(\log \mathrm{p}_i^t - \hat{\delta}^t \right)}{\sum_t s_i^t} \tag{14}$$

Substituting Equation 14 into Equation 13 to eliminate $\hat{\alpha}$ and $\hat{\gamma}_i$ yields

$$\hat{\delta}^t = \sum_u \left(\sum_i \frac{s_i^t s_i^u}{s_i^\Sigma} \right) \hat{\delta}^u + \sum_i s_i^t \left(\log \mathrm{p}_i^t - \frac{\sum_u s_i^u \log \mathrm{p}_i^u}{s_i^\Sigma} \right) \tag{15}$$

where $s_i^\Sigma = \sum_t s_i^t$

Equation 15 can be written in vector-matrix form as $\mathbf{Id} = \mathbf{Md} + \mathbf{b}$, where $\mathbf{I}$ is an identity matrix of size $T + 1$, $\mathbf{d}$ is a vector of the time effect estimates, $\mathbf{M}$ is a matrix with the element in the t-th row and u-th column equal to

$$\sum_i \frac{s_i^t s_i^u}{s_i^\Sigma},$$

and $\mathbf{b}$ is a vector with the t-th element equal to

$$\sum_i s_i^t \left(\log \mathrm{p}_i^t - \frac{\sum_u s_i^u \log \mathrm{p}_i^u}{s_i^\Sigma} \right)$$

The solution to this equation satisfies

$$(\mathbf{M} - \mathbf{I})\mathbf{d} = -\mathbf{b} \tag{16}$$

The matrix $\mathbf{M} - \mathbf{I}$ is singular so we cannot invert it to solve Equation (16). However, we usually constrain the time effects by setting $\delta^0 = 0$. This constraint can be expressed in matrix form as

$$\mathbf{Cd} = \mathbf{0} \tag{17}$$

where $\mathbf{C}$ is a matrix with all entries in the first column equal to 1 and 0 elsewhere and $\mathbf{0}$ is a vector of zeroes. Collier (1999) uses a similar technique in a different context (deriving Geary-Khamis indexes). Adding Equations 16 and 17 yields

$$(\mathbf{M} - \mathbf{I} + \mathbf{C})\mathbf{d} = -\mathbf{b} \tag{18}$$

$$\mathbf{d} = -(\mathbf{M} - \mathbf{I} + \mathbf{C})^{-1}\mathbf{b} \tag{19}$$

Let us denote the element in the r-th row and c-th column of $(\mathbf{M} - \mathbf{I} + \mathbf{C})^{-1}$ as m_{rc}. Then from Equation 19 we can write an arbitrary time effect estimate $\hat{\delta}^a$ as

$$\hat{\delta}^a = -\sum_t m_{at} \sum_i s_i^t \left(\log p_i^t - \frac{\sum_u s_i^u \log p_i^u}{s_i^\Sigma} \right) = -\sum_i \sum_t s_i^t \log p_i^t \left(m_{at} - \frac{\sum_u s_i^u m_{au}}{s_i^\Sigma} \right) \tag{20}$$

It follows that

$$\begin{aligned} P_{\text{TPD}}^{a,b} &= \exp\left(\hat{\delta}^b - \hat{\delta}^a\right) \\ &= \prod_i \prod_t \left(p_i^t\right)^{-s_i^t \left([m_{bt} - m_{at}] - \sum_u s_i^u [m_{bu} - m_{au}]/s_i^\Sigma \right)} \end{aligned} \tag{21}$$

Equations 11 and 21 give apparently distinct, but actually equivalent, formulations of a simple TPD decomposition. This fact is a consequence of the uniqueness of the solution to a full rank weighted least squares problem. It also follows from our earlier observation (from Section 1) that decompositions of this form are unique. Both formulations can be used to explain the impact of individual commodities by combining the relevant terms, that is, $\prod_t c_{TPD,i}^t(a,b)$ gives the contribution of commodity i to the price comparison between periods a and b.

Despite the equivalence of the two formulations of this decomposition, the former formulation focuses on simplicity in linear algebra, but loses the explicit separation of time and commodity terms in the decomposition, requiring these to be recovered after the decomposition is derived. The latter formulation carefully maintains the separate time and commodity terms, but is more difficult to express in terms of matrix operations. Both of these decompositions will be referred to henceforth as the Simple TPD Decomposition.

2.1.2. Reflexive TPD Decomposition

A third decomposition of the TPD can be derived from a multilateral method proposed by Rao (1990), which involves solving a set of equations

$$P_{\text{Rao}}^t = \prod_i \left(\frac{p_i^t}{\pi_i} \right)^{s_i^t} \tag{22}$$

$$\pi_i = \prod_t \left(\frac{p_i^t}{P_{\text{Rao}}^t} \right)^{\frac{s_i^t}{\sum_u s_i^u}} \tag{23}$$

simultaneously for the unknown parameters π_i and P_{Rao}^t. π_i can be interpreted as a *reference price* for commodity i, and we typically impose the condition $P_{\text{Rao}}^0 = 1$ to obtain a unique solution.

Rao (2005) demonstrates that the system of price comparisons obtained by solving Equations 22 and 23 simultaneously is equivalent to the (weighted) TPD system.

From Equation 22, the TPD price change between two periods a and b can be expressed as

$$P_{\text{TPD}}^{a,b} = \prod_i \frac{\left(p_i^b\right)^{s_i^b}}{\left(p_i^a\right)^{s_i^a}} (\pi_i)^{s_i^a - s_i^b} \tag{24}$$

We need to be careful expressing the price change in this way, because some commodities may only have a price in one of the two periods a and b. Where, for instance, a commodity is not sold in period a, we simply replace the expenditure share s_i^a with a 0, and consequently replace the exponentiated missing price $\left(p_i^a\right)^{s_i^a}$ with a 1.

Equation 24 has each commodity's contribution to the price change expressed in terms of its weighted prices, as well as the reference price π_i. From Equation 23, we can see that the reference prices depend on the aggregate price levels, which makes the decomposition reflexive. It also has the interesting property that choosing a period other than 0 as the reference would not alter the price comparisons, but could alter the reference prices π_i, and consequently the commodity contributions to those price comparisons. We will continue to refer to this decomposition as the Reflexive TPD Decomposition.

2.2. *Decomposition of the CCD and GEKS Methods*

The GEKS method involves calculating multilateral price comparisons by combining bilateral Fisher indexes:

$$P_{\text{GEKS}}^{a,b} = \prod_t \left(\frac{P_{\text{F}}^{t,b}}{P_{\text{F}}^{t,a}} \right)^{\frac{1}{T+1}} \tag{25}$$

where P_F is a Fisher price index:

$$P_{\text{F}}^{0,1} = \left[\left(\frac{\sum_i p_i^1 q_i^0}{\sum_i p_i^0 q_i^0} \right) \left(\frac{\sum_i p_i^1 q_i^1}{\sum_i p_i^0 q_i^1} \right) \right]^{1/2} \tag{26}$$

GEKS-Törnqvist or CCD price comparisons are obtained by replacing the Fisher indexes in Equation 25 with Törnqvist indexes:

$$P_{\text{CCD}}^{a,b} = \prod_t \left(\frac{P_T^{t,b}}{P_T^{t,a}} \right)^{\frac{1}{T+1}} \tag{27}$$

We can easily derive multiplicative decompositions of the CCD index using a multiplicative decomposition of the Törnqvist index. From Equation 4, we know that the Törnqvist index can be written as a product of commodity contributions $c_{\text{T},i}(0,1) = \left(\frac{p_i^1}{p_i^0}\right)^{\frac{1}{2}\left(s_i^0 + s_i^1\right)}$. Substituting this into Equation 27, we obtain

$$
\begin{aligned}
P_{\mathrm{CCD}}^{a,b} &= \prod_i \prod_t \left(\frac{c_{\mathrm{T},i}(t,b)}{c_{\mathrm{T},i}(t,a)} \right)^{\frac{1}{T+1}} \\
&= \prod_i \frac{\left(p_i^b\right)^{w_i(\bullet,b)}}{\left(p_i^a\right)^{w_i(\bullet,a)}} \left[\prod_t \left(p_i^t\right)^{\frac{w_i(t,a)-w_i(t,b)}{T+1}} \right] \\
&= \prod_i c_{\mathrm{CCD},i}(a,b)
\end{aligned}
\tag{28}
$$

where

$w_i(t,a) = \frac{1}{2}\left(\frac{s_i^t}{\sum_{i \in (t \cap a)} s_i^t} + \frac{s_i^a}{\sum_{i \in (t \cap a)} s_i^a} \right)$ is the weight of commodity i in the Törnqvist price comparison between periods t and a (represented by the notation $i \in (t \cap a)$)

$w_i(\bullet,a) = \frac{1}{T+1} \sum_t w_i(t,a)$ is the average weight across comparisons involving period a

and $c_{\mathrm{CCD},i}(a,b) = \frac{\left(p_i^b\right)^{w_i(\bullet,b)}}{\left(p_i^a\right)^{w_i(\bullet,a)}} \left[\prod_t \left(p_i^t\right)^{\frac{w_i(t,a)-w_i(t,b)}{T+1}} \right]$ is the contribution of commodity i to the CCD price comparison between periods a and b.

Note that this CCD decomposition is simple: it inherits this property from the Törnqvist decomposition.

If there are any missing prices p_i^t, we replace the corresponding term(s) with a 1. If there are no missing prices (i.e., the same set of commodities is sold every period), Equation 28 can be simplified to an expression based on the CCDI index presented by Diewert and Fox (2017):

$$
P_{\mathrm{CCD}}^{a,b} = \prod_i \frac{\left(p_i^b\right)^{\frac{1}{2}\left(s_i^\bullet + s_i^b\right)}}{\left(p_i^a\right)^{\frac{1}{2}\left(s_i^\bullet + s_i^a\right)}} \left(p_i^\bullet\right)^{\frac{1}{2}\left(s_i^a - s_i^b\right)}
\tag{29}
$$

where $s_i^\bullet = \frac{1}{T+1} \sum_t s_i^t$ and $p_i^\bullet = \prod_t \left(p_i^t\right)^{\frac{1}{T+1}}$

As observed by Chessa et al. (2017), Equation 29 can be expressed as a geometric average of two factors. The second factor is very similar to Equation 24, revealing that the TPD and CCD indexes are closely related. However, the first factor reveals that the CCD gives more influence to local price changes between periods a and b.

$$
P_{\mathrm{CCD}}^{a,b} = \prod_i \left[\left(\frac{p_i^b}{p_i^a} \right)^{s_i^\bullet} \right]^{\frac{1}{2}} \left[\frac{\left(p_i^b\right)^{s_i^b}}{\left(p_i^a\right)^{s_i^a}} \left(p_i^\bullet\right)^{\left(s_i^a - s_i^b\right)} \right]^{\frac{1}{2}}
\tag{30}
$$

We could obtain a multiplicative decomposition of a GEKS price comparison in a similar way, by substituting a multiplicative Fisher decomposition into Equation 25. The results are not presented here. We note, however, that a GEKS decomposition will inherit the simple/reflexive property of the corresponding Fisher decomposition. The multiplicative Fisher decompositions presented by Balk (2008, chap. 4) are both reflexive.

2.3. Decomposition of the GK Method

The GK method involves solving a set of simultaneous equations, similar to those from the Rao method:

$$P^t_{\mathrm{GK}} = \frac{\sum_i p^t_i q^t_i}{\sum_i \pi_i q^t_i} \tag{31}$$

$$\pi_i = \frac{\sum_t p^t_i q^t_i / P^t_{\mathrm{GK}}}{\sum_t q^t_i} \tag{32}$$

where again, π_i can be interpreted as a reference price for commodity i, and we typically impose the condition $P^0_{GK} = 1$ to obtain a unique solution.

To decompose GK index movements, it is helpful to first rewrite Equation 31 as

$$P^t_{\mathrm{GK}} = \sum_i \sigma^t_i \frac{p^t_i}{\pi_i} \tag{33}$$

where $\sigma^t_i = \frac{\pi_i q^t_i}{\sum_j \pi_j q^t_j}$ can be interpreted as an expenditure share of commodity i in period t, if all commodities were sold at reference prices.

Instinctively, one might seek an additive decomposition of the GK price change between two periods using Equation 33: some algebraic manipulation yields

$$\begin{aligned} P^{a,b}_{\mathrm{GK}} &= \frac{\sum_i \sigma^b_i \frac{p^b_i}{\pi_i}}{\sum_i \sigma^a_i \frac{p^a_i}{\pi_i}} \\ &= \left[\sum_{q^b_i > 0} s^b_i \frac{\sigma^a_i}{\sigma^b_i} \frac{p^a_i}{p^b_i} + \left(P^b_{\mathrm{GK}}\right)^{-1} \sum_{q^b_i = 0} \sigma^a_i \frac{p^a_i}{\pi_i} \right]^{-1} \\ &= \sum_{q^a_i > 0} s^a_i \frac{\sigma^b_i}{\sigma^a_i} \frac{p^b_i}{p^a_i} + \left(P^a_{\mathrm{GK}}\right)^{-1} \sum_{q^a_i = 0} \sigma^b_i \frac{p^b_i}{\pi_i} \end{aligned} \tag{34}$$

where the first sum includes commodities sold in both a and b, and the second sum includes commodities sold in b and not a (last line of Equation 34) or vice versa (second last line). The last line of Equation 34 is an additive decomposition that is reflexive through its inclusion of the aggregate price level P^a_{GK}, and also indirectly through the shares σ^t_i and the reference prices π_i.

When an identical set of commodities is sold in a and b, the second term of this decomposition disappears and the first term seems quite appealing as an additive decomposition. However, in general, the asymmetric manner in which it handles prices

missing from only one of a or b seems unsatisfactory. Taking the mean of the second last and last lines of Equation 34 would address the asymmetry but the result is no longer an additive decomposition.

We can obtain a more symmetric GK decomposition by first using Equation 5 to convert Equation 33 to a multiplicative form:

$$P^t_{\mathrm{GK}} = \prod_i \left(\frac{p^t_i}{\pi_i}\right)^{\theta^t_i} \tag{35}$$

where

$$\begin{aligned}
\theta^t_i &= \frac{\sigma^t_i \times L\left(P^t_{\mathrm{GK}}, p^t_i/\pi_i\right)}{\sum_j \sigma^t_j \times L\left(P^t_{\mathrm{GK}}, p^t_j/\pi_j\right)} \\
&= \frac{q^t_i \times L\left(\pi_i P^t_{\mathrm{GK}}, p^t_i\right)}{\sum_j q^t_j \times L\left(\pi_j P^t_{\mathrm{GK}}, p^t_j\right)} \\
&\approx \frac{q^t_i \times p^t_i}{\sum_j q^t_j \times p^t_j} \\
&= s^t_i
\end{aligned}$$

where the second equality follows from the definition of σ^t_i and the homogeneity of the logarithmic mean, and the approximation follows from the $p^t_i \approx \pi_i P^t_{\mathrm{GK}}$ relationship implicit in the GK method.

It follows that

$$\begin{aligned}
P^{a,b}_{\mathrm{GK}} &= \frac{\prod_i \left(\frac{p^b_i}{\pi_i}\right)^{\theta^b_i}}{\prod_i \left(\frac{p^a_i}{\pi_i}\right)^{\theta^a_i}} \\
&= \prod_i \frac{\left(p^b_i\right)^{\theta^b_i}}{\left(p^a_i\right)^{\theta^a_i}} (\pi_i)^{\theta^a_i - \theta^b_i}
\end{aligned} \tag{36}$$

Equation 36 is a multiplicative GK decomposition that is reflexive through both the reference prices π_i and the exponents θ^t_i. Note the similarity to Equation 24.

3. Decomposition of Extended Multilateral Indexes

Statistical agencies compute and publish price indexes as new periods of price data become available. The published index series is extended by linking or "splicing" price comparisons involving the latest period onto published index levels for previous periods. This section focusses on how to extend the index series when multilateral methods are

used to generate price comparisons, and how we can use the results from the previous section to decompose published index movements.

It is relatively straightforward to extend a bilateral index. For instance, when data from period t becomes available, we would extend a direct bilateral index to period t by first calculating the price comparison $P^{0,t}$ between the reference period (0) and t, and then multiplying this by the index level in the reference period: $P^t = P^0 \times P^{0,t}$. Similarly, we would extend a chained bilateral index by multiplying the previous index level P^{t-1} by the price comparison between the previous and current periods $P^{t-1,t}$.

How best to extend a multilateral price index is more ambiguous. In period t, we simultaneously estimate price comparisons between t and several historical periods. Using the price comparison from one historical period to extend the index may yield a different result to using the price comparison from another.

Several authors have proposed splicing methods for extending multilateral indexes. In this article, we focus on decomposing the methods considered in Australian Bureau of Statistics (2017):

- *Rolling window* methods, including the *movement splice* proposed by Ivancic et al. (2011), the *window splice* proposed by Krsinich (2016), the *half (window) splice* proposed by De Haan (2015), and the *mean splice* proposed by Diewert and Fox (2017). These methods involve selecting a fixed window length ($T + 1$ periods) for multilateral comparisons. As each new period of data becomes available, we calculate a new system of comparisons over the window spanning from $t - T$ to t and splice it together with the previous system of comparisons (using a window spanning from $t - T - 1$ to $t - 1$) to estimate the index movement from $t - 1$ to t,
- The *direct* method proposed by Chessa (2015). This method involves selecting a fixed base period b (say, December) as the start of the multilateral comparison window. As each new period of data becomes available, we calculate a system of comparisons spanning from b to t and use the direct price comparison between b and t to estimate the price change between these periods. The base period can be updated regularly (e.g., annually).

Table 1 expresses the extended price movements between consecutive periods ($t - 1$ and t) in terms of multilateral price movements from the current window (ending in t) and the previous window (ending in $t - 1$). The methods are algebraically similar, though in practice the indexes may yield different trends. The next section presents empirical results.

Of most relevance here is that they all combine multilateral price movements in a multiplicative manner (through division or geometric averaging). This means that we can substitute a multiplicative decomposition for each of the multilateral price movements that feature in the extended price movement, and collect like terms to obtain a multiplicative decomposition of the extended price movement. Importantly, if the multilateral decomposition is simple, then the decomposition of the extended price movement is also simple.

In Table 1, $P^{x,y}(z)$ denotes the aggregate price comparison between periods x and y derived from a multilateral window ending in period z, and $c_i(x, y; z)$ denotes the contribution of commodity i to that aggregate comparison.

Table 1. Comparison of extension methods.

Extension method	Price movement between consecutive periods	Decomposition of consecutive movement
Movement splice	$P^{t-1,t} = P^{t-1,t}(t)$	$P^{t-1,t} = \prod_i c_i(t-1,t;t)$
Window splice	$P^{t-1,t} = \frac{P^{t-T,t}(t)}{P^{t-T,t-1}(t-1)}$	$P^{t-1,t} = \prod_i \frac{c_i(t-T,t;t)}{c_i(t-T,t-1;t-1)}$
Half splice (assuming T is even)	$P^{t-1,t} = \frac{P^{t-T/2,t}(t)}{P^{t-T/2,t-1}(t-1)}$	$P^{t-1,t} = \prod_i \frac{c_i(t-T/2,t;t)}{c_i(t-T/2,t-1;t-1)}$
Mean splice	$P^{t-1,t} = \prod_{s=1}^{T} \left[\frac{P^{t-s,t}(t)}{P^{t-s,t-1}(t-1)}\right]^{\frac{1}{T}}$	$P^{t-1,t} = \prod_i \left[\prod_{s=1}^{T} \frac{c_i(t-s,t;t)}{c_i(t-s,t-1;t-1)}\right]^{\frac{1}{T}}$
Direct	$P^{t-1,t} = \frac{P^{b,t}(t)}{P^{b,t-1}(t-1)}$	$P^{t-1,t} = \prod_i \frac{c_i(b,t;t)}{c_i(b,t-1;t-1)}$

It may be of interest to decompose longer term (e.g., annual) price comparisons of an extended price index. These longer term movements can be expressed as a product of consecutive price movements:

$$P^{a,b} = \prod_{t=a+1}^{b} P^{t-1,t} \tag{37}$$

As above, we can substitute a multiplicative decomposition for each element of Equation 37 and collect like terms to obtain a multiplicative decomposition into commodity contributions

$$P^{a,b} = \prod_i \prod_{t=a+1}^{b} c_i(t-1,t)$$

where $c_i(t-1,t)$ is the contribution of commodity i to the extended movement between $t-1$ and t (as given in the third column of Table 1). Once again, if the underlying multilateral decomposition is simple, this will be preserved.

4. Empirical Results

In this section, we illustrate how the decomposition methods described in the previous sections can be used to quantify the contributions of individual commodities to multilateral price comparisons. In Subsection 4.1 we introduce the data used for this analysis. In Subsection 4.2, we decompose indexes calculated using a range of multilateral methods, and in Subsection 4.3 we decompose indexes calculated using a range of extension methods. This allows us to compare and contrast the methods considered. However, in practice, a statistical agency may prefer a single combination of multilateral and extension methods for various reasons. In this context, the comparison between methods is less important than the illustration that we can decompose an index calculated using any combination of the multilateral and extension methods described above.

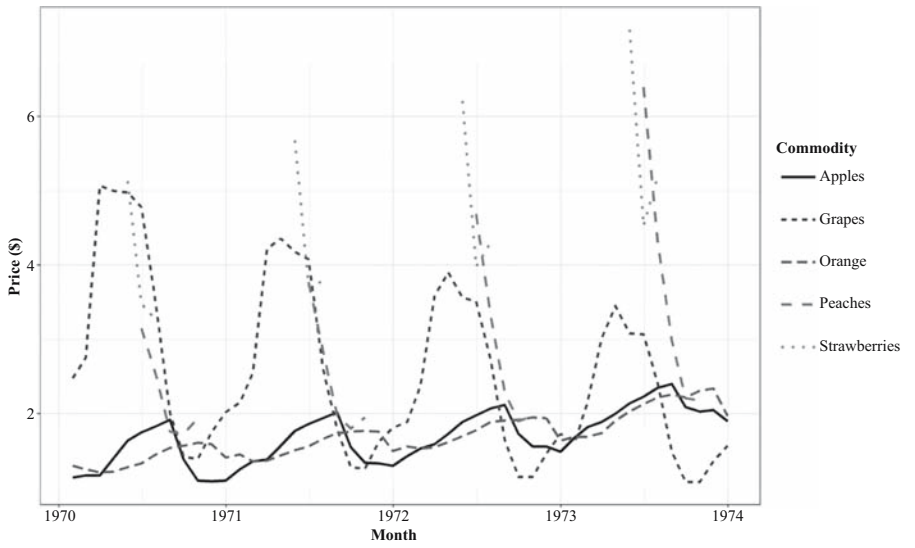

Fig. 1. Price of fruit commodities.

4.1. Data

The main data set we use for this illustration contains monthly price and quantity information relating to five fruit commodities over a period of four years. It is taken from the IWGPS Consumer Price Index Manual (ILO et al. 2004, chap. 22) and is a modified version of a data set from Turvey (1979). Three of the commodities (Apples, Grapes and Oranges) are sold every month whereas the remaining two (Peaches and Strawberries) are sold only for a few months each year. Figures 1 and 2 plot the prices and quantities of each commodity.

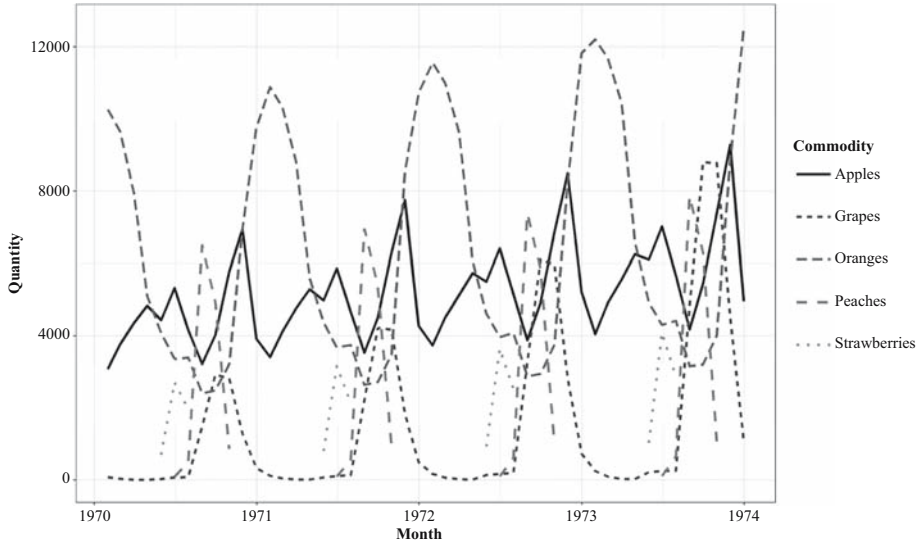

Fig. 2. Quantity of fruit commodities.

Table 2. Features of fruit and scanner data sets.

Commodity class	Fruit	Cookies	Oatmeal	Toothbrushes
Time span used	January 1970 to December 1973	October 1989 to September 1993	July 1991 to June 1995	October 1989 to September 1993
Number of monthly observations	176	18,403	2,617	8,027
Number of commodities	5	763	87	362
Proportion of commodities sold in every month	60%	19%	41%	11%

This Fruit data is useful for our illustration because it contains a small number of products, some of which are not sold every month. However, it does not share all the features of the data to which these methods are applied, such as truly new or disappearing products. For evidence that some of our findings are applicable in practice, we use scanner data for sales of Cookies, Oatmeal and Toothbrushes from Dominick's stores in Chicago, obtained from the James M. Kilts Center, University of Chicago Booth School of Business. For comparability with the Fruit data, we convert this (weekly) scanner data to monthly frequency by assigning each week to the month in which the majority of its sales fall and subset to 48 months of data; we also remove observations that are flagged as suspect (University of Chicago 2018). Table 2 summarizes a few features of the Fruit data and the three scanner data sets.

4.2. Multilateral Indexes Calculated on the Entire Fruit Data Set

Figure 3 compares the TPD, CCD and GK price indexes calculated on the entire Fruit data set, with January 1970 as the base period. For this data set, we observe that the three methods produce numerically similar indexes, with the TPD and GK particularly close. Corresponding figures for the other data sets are included as Supplemental Data (Figures A1, A2 and A3). Figure 4 compares the month-on-month price changes corresponding to these indexes.

The multilateral indexes show steep price increases every May. We can use the decomposition methods to understand which commodities are driving these price changes. Table 3 presents the contributions of each commodity to the price change between April and May 1973, using the Simple TPD, Reflexive TPD, CCD and GK decomposition methods presented in this article (recall that the Simple TPD decompositions presented in Subsubsection 2.1.1 are mathematically equivalent). The prices and expenditure shares of each commodity are included for reference. Note that Peaches are not sold in either month, and Strawberries are sold in May but not April.

Overall, the commodity contributions obtained from the Reflexive TPD and GK decompositions are very similar, as would be expected given their mathematical

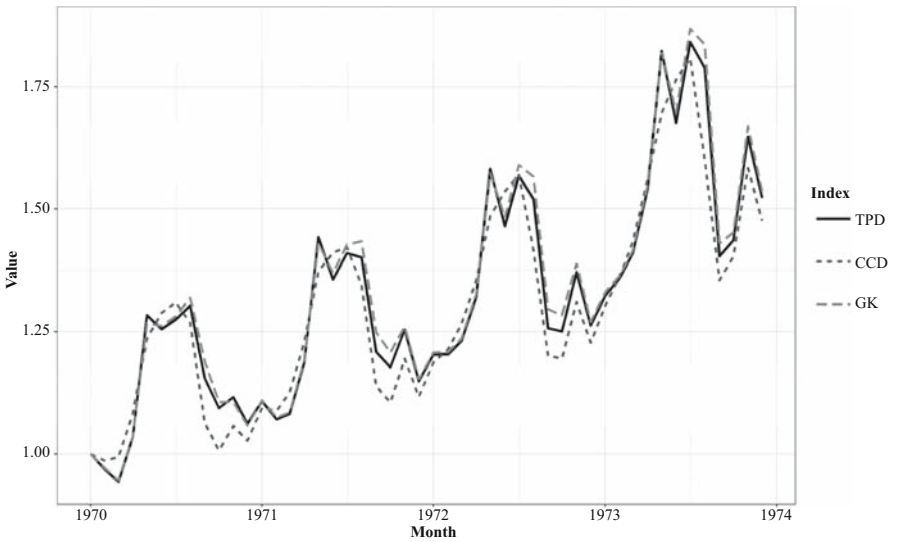

Fig. 3. Multilateral index values.

similarity. It is difficult to draw general conclusions about the numerical similarity of simple and reflexive decompositions of the TPD index, given the latter are not unique—as mentioned in Subsubsection 2.1.2, changing the reference period would yield a different reflexive decomposition.

These contributions reveal a few interesting features of the methods examined.

First, the reappearance of Strawberries, a strongly seasonal commodity with an intermittent sales pattern, contributes to an aggregate price increase in May 1973 (it has a contribution greater than one). As this commodity is not sold in the previous month, it

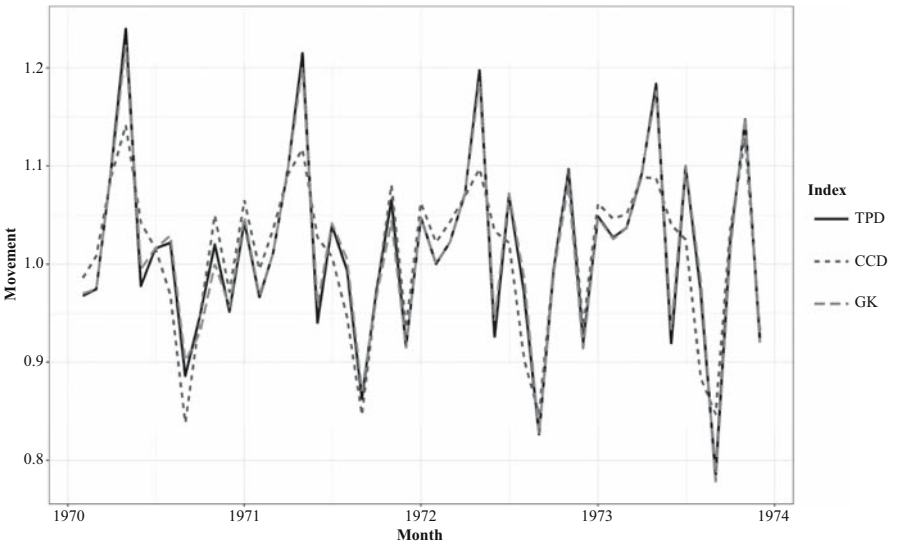

Fig. 4. Month on month multilateral index movements.

Table 3. Decomposition of multilateral index movement between April and May 1973.

Commodity	Contribution to multilateral index movement between April and May 1973				Price		Expenditure share	
	Simple TPD	Reflexive TPD	CCD	GK	April 1973	May 1973	April 1973	May 1973
Apples	1.039	0.993	1.035	1.004	2.00	2.14	0.50	0.42
Grapes	1.017	1.018	0.998	1.016	3.45	3.08	<0.01	0.02
Oranges	0.997	0.949	1.017	0.956	1.91	2.03	0.50	0.32
Peaches	1.004	1.000	1.000	1.000	NA	NA	NA	NA
Strawberries	1.118	1.233	1.034	1.206	NA	7.17	NA	0.24
Aggregate	1.184	1.184	1.087	1.176			1.00	1.00

would not contribute to a chained bilateral index unless explicit imputation was used. However, the multilateral indexes take into account the prices of this commodity in other periods, compared to which the May 1973 price is relatively high. This capacity to capture the price changes of commodities with intermittent sales is an advantage of using multilateral methods with scanner data: as seen from Table 2, such commodities are common in scanner data sets.

Second, the TPD and GK decompositions show some commodities have price increases between April and May 1973, but contribute to an aggregate price decrease between those periods (contribution less than one) or vice versa. This can occur because the contributions depend on changes in weights, as well as changes in prices. Moreover, the simple TPD decomposition suggests Peaches have a non-trivial contribution to change despite being absent from both periods. These are unintuitive observations, but not disqualifying – by their very definition, multilateral comparisons between two periods take the prices in other periods into account, which may help to mitigate drift (Ivancic et al. 2011) including in the presence of seasonal patterns (Ribe 2012).

Figure 5 shows the relationship between price change and contribution to aggregate price change, for every instance in the Fruit data set where a commodity is sold in consecutive months. They reveal that there is a correlation between commodities' month-on-month price changes and their contribution to change, but also that it is not uncommon for the price changes and contributions to be in opposite directions (observations in the upper left and lower right quadrants). Table 4 illustrates that this phenomenon occurs in scanner data as well. It is consistently less pronounced for the CCD than the TPD index, reflecting that local price changes have greater influence on the CCD index than the TPD index, as observed in Subsection 2.2.

A feature of simple decompositions is that changing the price of one commodity without changing the weights does not affect the contributions of other commodities. Suppose we adjust the price of Oranges in April 1973 to be five times its original value (9.55 instead of 1.91), while leaving the expenditure share unchanged. This price spike might result from adverse production conditions (e.g., a natural disaster), with consumers responding by allocating a fixed expenditure to each commodity and reacting to price changes with reciprocal quantity changes. Observe that this leaves the TPD and CCD weights

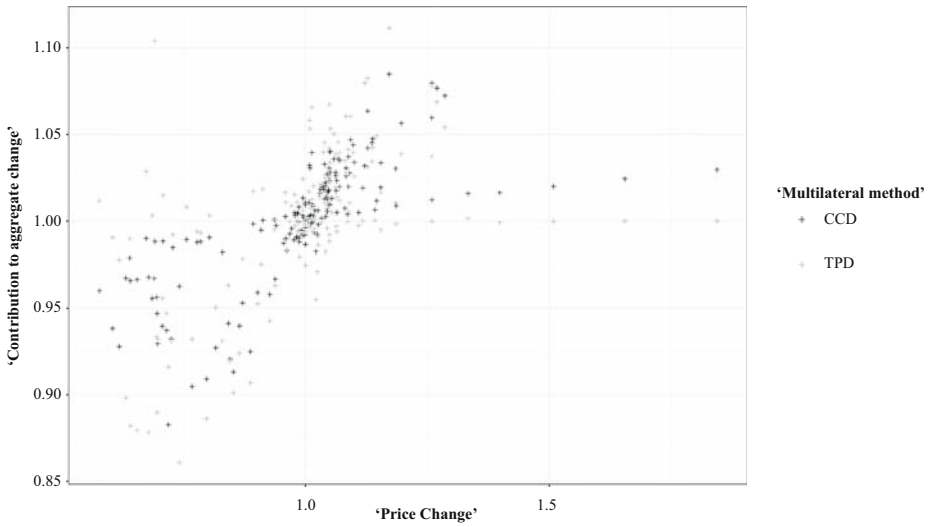

Fig. 5. Commodity price change verus index contribution.

unchanged, but alters the GK weights, so we exclude the latter method from the analysis that follows.

We can recalculate the TPD and CCD indexes using the adjusted data set and derive the contributions of each commodity to the index movement between April and May 1973. Table 4 presents the contributions of each commodity to the index movements from the adjusted data set and their relationship between the commodity contributions from the original data set (in Table 2). For the simple (Simple TPD and CCD) decompositions, only the contribution of Oranges is affected; however, the Reflexive TPD contributions for Apples and Strawberries are slightly altered by the price change of Oranges. This illustrates an advantage of simple decomposition methods.

4.3. Extended Multilateral Indexes

In practice, we would not calculate a multilateral index using the entire data set, but instead use one of the methods described in Section 3 to extend the series one period at a

Table 4. Price changes and contributions in opposite directions.

Commodity class	Fruit	Cookies	Oatmeal	Toothbrushes
Instances where a commodity is sold in consecutive months	165	17,385	2,520	7,467
Instances where contribution is in the opposite direction to price change (Simple TPD)	26%	25%	19%	22%
Instances where contribution is in the opposite direction to price change (CCD)	10%	22%	14%	17%

Table 5. Impact of changing one commodity's price on commodity contributions.

Commodity	Contribution based on adjusted data set			Ratio of adjusted/ original contribution		
	Simple TPD	Reflexive TPD	CCD	Simple TPD	Reflexive TPD	CCD
Apples	1.039	0.990	1.035	1.000	0.997	1.000
Grapes	1.017	1.019	0.998	1.000	1.000	1.000
Oranges	0.448	0.425	0.461	0.449	0.448	0.453
Peaches	1.004	1.000	1.000	1.000	1.000	1.000
Strawberries	1.118	1.239	1.034	1.000	1.005	1.000
Aggregate	0.531	0.531	0.492	0.449	0.449	0.453

time. Figure 6 presents TPD indexes that are extended using the movement splice, window splice, half splice, mean splice, and direct methods. The TPD index based on the entire data set is included for comparison. Corresponding figures for the CCD and GK methods are included as Supplemental Data (Figures A4 and A5). We use a window of length 13 months to calculate the rolling window methods and use a base month of January for the direct method. As the rolling window methods cannot be used to extend the index until a full window of historical data is available, we start the extended indexes in January 1971.

The indexes in Figure 6 are more dispersed than the indexes in Figure 3, indicating that the choice of extension method makes a greater difference to the series than the choice of multilateral method in this example. The mean splice index tracks the index with no extension closely. The direct multilateral index is typically lower than the mean splice in the middle of each calendar year, but similar at the end of the year. The half and movement splice indexes drift a little higher and lower than the mean splice respectively. The window splice index diverges substantially.

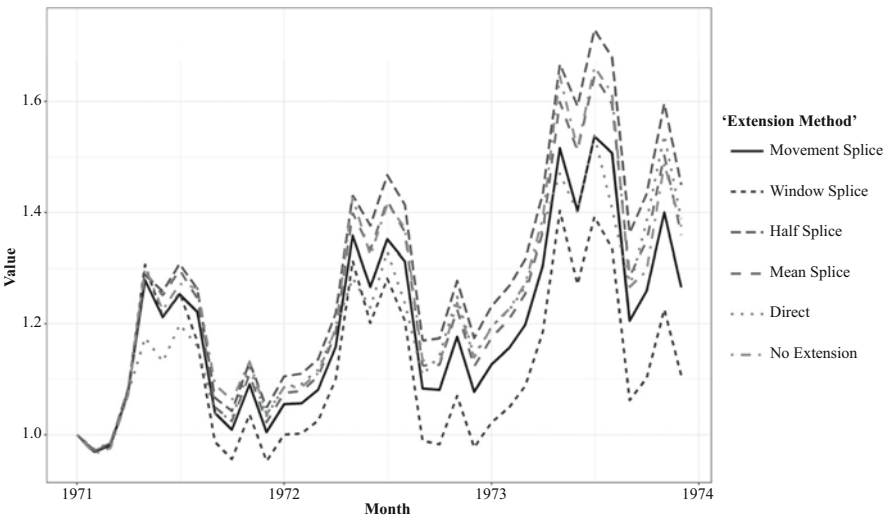

Fig. 6. Extended TPD index values.

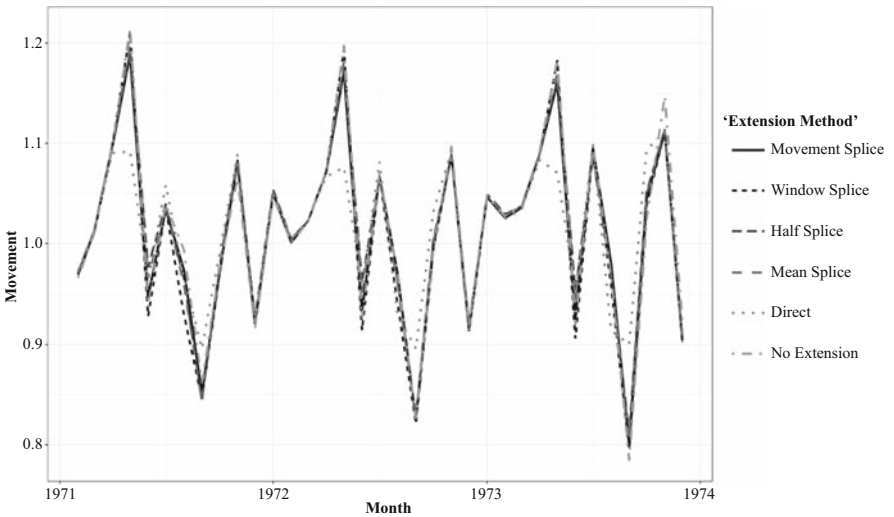

Fig. 7. Month on month extended TPD index movements.

Figures 7 and 8 compare the month-on-month and annual index movements using the various extension methods. The clearest difference in the month-on-month movements is that the direct index movements have a less extreme peak each May and a less extreme trough each September. Otherwise the month-on-month movements appear similar. However, there are systematic differences between the annual movements of the various rolling window methods, implying that these indexes diverge gradually. The decompositions can help to explain these differences.

Table 6 compares the Simple TPD contributions of each commodity to the extended TPD price movements between April and May 1973. The main difference between

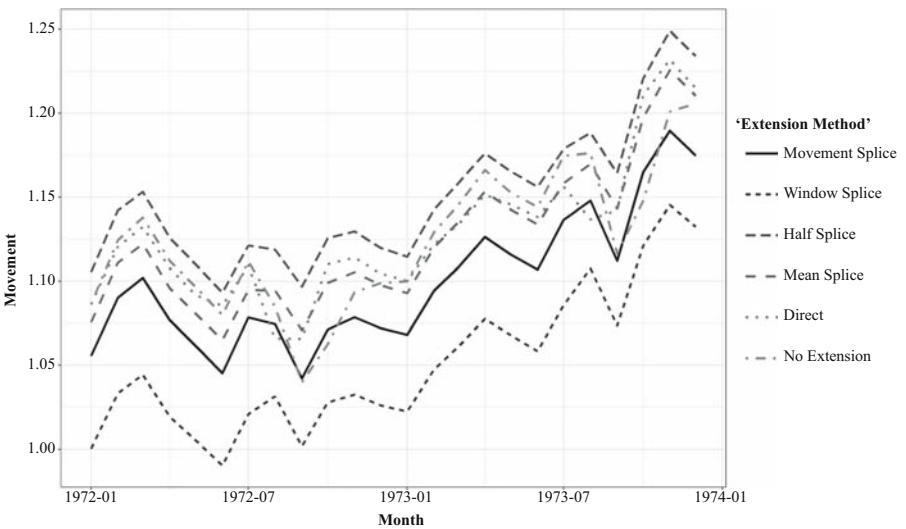

Fig. 8. Annual extended TPD index movements.

Table 6. Decomposition of extended index movement between April and May 1973.

Commodity	Contribution to movement of extended TPD index between April and May 1973					Price		Expenditure share	
	Movement splice	Window splice	Half splice	Mean splice	Direct	April 1973	May 1973	April 1973	May 1973
Apples	1.035	1.034	1.034	1.033	1.034	2.00	2.14	0.50	0.42
Grapes	1.018	1.018	1.020	1.020	1.008	3.45	3.08	<0.01	0.02
Oranges	1.013	1.008	1.013	1.012	1.027	1.91	2.03	0.50	0.32
Peaches	1.003	1.004	1.003	1.004	1.000	NA	NA	NA	NA
Strawberries	1.085	1.109	1.085	1.094	1.000	NA	7.17	NA	0.24
Aggregate	1.161	1.183	1.163	1.170	1.071				

methods is in the contribution of Strawberries. Strawberries do not contribute to the direct index movement between April and May, as the expanding window (starting in January 1973) does not contain any observations for Strawberries until May, and we need two observations for a commodity to contribute to price comparisons.

To understand the differing contributions of Strawberries to the three rolling window methods, note that the high price of Strawberries in May 1973 makes the previous year's prices in the current window (May 1972 to May 1973) appear lower than they did in the previous window (April 1972 to April 1973). In consequence, the contribution of Strawberries to the price movement between the start of the current window (May 1972) and the previous period (April 1973) is more positive in the current window than in the previous window. As Krsinich (2016) argues, the window splice implicitly revises this price movement in extending the index series from the previous to the current period, whereas the movement splice makes no such revision, and the mean splice makes a partial revision (Australian Bureau of Statistics 2017). The half splice implicitly revises the movement over part of the previous window (November 1972 to April 1973), but as Strawberries are not sold over this period their contribution to the half splice is the same as their contribution to the movement splice.

Table 7 decomposes the annual movement of each extended TPD index between May 1972 and May 1973. Again, we can see that the commodities with strong seasonality have neutral contributions to the direct index movement because they are not sold between

Table 7. Decomposition of extended index movement between May 1972 and May 1973.

Commodity	Contribution to movement of extended TPD index between May 1972 and May 1973					Price		Expenditure share	
	Movement splice	Window splice	Half splice	Mean splice	Direct	May 1972	May 1973	May 1972	May 1973
Apples	1.026	1.021	1.060	1.040	1.020	1.89	2.14	0.42	0.42
Grapes	0.994	0.984	1.002	0.995	0.999	3.56	3.08	0.02	0.02
Oranges	1.057	1.051	1.086	1.075	1.124	1.70	2.03	0.32	0.32
Peaches	1.010	1.001	1.001	1.009	1.000	NA	NA	NA	NA
Strawberries	1.025	1.011	1.010	1.019	1.000	6.21	7.17	0.24	0.24
Aggregate	1.116	1.068	1.165	1.143	1.146				

January and April of either year. On the other hand, Oranges have a relatively positive contribution to the direct index movement, which likely relates to their high expenditure shares (about 0.75) in January 1972 and January 1973, the base months of the expanding windows used to estimate price changes within those years.

As observed in Figure 8, the half splice has the largest annual movements of the rolling window methods, followed by the mean splice, the movement splice, and the window splice. From Table 5, we can see that the contributions of the commodities that are sold all year round (Apples, Grapes and Oranges) follow the same ordering. Note that Peaches make a non-trivial contribution to the rolling window index movements — even though they are not sold in May, their prices in intervening months contribute to the month-on-month movements of the extended index, and ultimately to the annual movement.

5. Conclusions

Index decomposition is useful in practice for interpreting price indexes: it allows one to break down aggregate price movements into contributions from individual or groups of commodities. Decomposition is particularly important for understanding multilateral indexes, which combine many different time comparisons yielding complex dependencies on any individual commodity's price observations. We defined reflexive or simple decompositions based on whether the contribution for each commodity depends on an aggregate price level. Simple decompositions ensure contributions for commodities are invariant under changes in other commodity prices.

We introduced a simple decomposition for the TPD index and a reflexive decomposition for the GK index, and showed how these compare to the reflexive decomposition for the TPD index and the simple decomposition for the CCD index. These decompositions demonstrate that movements can be attributed to the price observations for each commodity. The theoretical and empirical results provide evidence of similarities between these three indexes and subtle differences between the CCD and the other methods. They also show how commodities sold in only one of two time periods can influence the price comparison between those periods, and reveal that is not uncommon for a commodity's contribution to aggregate price change to be in the opposite direction to its individual price change.

The comparison between decompositions raises questions for price index implementations. Where decompositions disagree on the direction of the effect of particular commodities, how should this be interpreted? Under what conditions should commodity contributions remain invariant?

We do not fully address these questions here. Where several decomposition methods are available, each may yield additional information about price movements. The circumstances in which the price index is applied may dictate which decomposition is most useful, such as the choice between an additive and a multiplicative method.

However, we have touched on several properties that it seems advantageous for a decomposition to possess, including that one commodity's contribution should be invariant to changes in the prices of other commodities (conditional on the expenditure shares), and that a commodity not sold in either of two periods should have a trivial contribution to the price change between those periods. Other desirable properties might

include invariance to the ordering of commodities and time periods, or invariance to the price changes of other commodities under different conditions. Development of a more complete set of desirable properties for index decomposition functions would be an interesting area for further research.

6. References

Australian Bureau of Statistics (ABS). 2016. "Making Greater Use of Transactions Data to Compile the Consumer Price Index, Australia." Cat. no. 6401.0.60.003. Canberra: ABS. Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/6401.0.60.003 (accessed May 2019).

Australian Bureau of Statistics (ABS). 2017. "An Implementation Plan to Maximise the Use of Transactions Data in the CPI." Cat. no. 6401.0.60.004. Canberra: ABS. Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/6401.0.60.004 (accessed May 2019).

Balk, B.M. 2008. *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*. New York: Cambridge University Press. Doi: https://doi.org/10.1017/cbo9780511720758.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. "Multilateral Comparisons of Output, Input, and Productivity Using Superlative Index Numbers." *The Economic Journal* 92(365): 73–86. Doi: http://dx.doi.org/10.2307/2232257.

Chessa, A. 2015. "Towards a Generic Price Index Method for Scanner Data in the Dutch CPI." Paper presented at the 14th meeting of the Ottawa Group, Tokyo, 20–22 May 2015. Available at: http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s1room2.pdf (accessed August 2017).

Chessa, A., J. Verburg, and L. Willenborg. 2017. "A Comparison of Price Index Methods for Scanner Data." Paper presented at the 15th meeting of the Ottawa Group, Eltville, 10–12 May 2017. Available at: https://www.bundesbank.de/Redaktion/EN/Downloads/Bundesbank/Research_Centre/Conferences/2017/2017_05_10_ottawa_group_07_1_paper.pdf (accessed February 2018).

Collier, I.L. 1999. "Comment." *International and Interarea Comparisons of Income, Output, and Prices*, edited by A. Heston and R. E. Lipsey, 87–107. Chicago: University of Chicago Press. Doi: https://doi.org/10.7208/chicago/9780226331126.001.0001.

De Haan, J. 2015. "A Framework for Large Scale Use of Scanner Data in the Dutch CPI." Paper presented at the 14th meeting of the Ottawa Group, Tokyo, 20–22 May 2015. Available at: http://www.stat.go.jp/english/info/meetings/og2015/pdf/t6s11p33_pap.pdf (accessed August 2017).

De Haan, J., R. Hendriks, and M. Scholz. 2016. "A Comparison of Weighted Time-Product Dummy and Time Dummy Hedonic Indexes." Paper presented at the 15th meeting of the Ottawa Group, Eltville, 10–12 May 2017. Available at: https://www.bundesbank.de/resource/blob/636054/f20b4679b121998a5ae5f106a11ee5a1/mL/2017-05-10-ottawa-group-07-4-paper-data.pdf (accessed February 2019).

De Haan, J. and F. Krsinich. 2014. "Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes." *Journal of Business & Economic Statistics* 32(3): 341–358. Doi: http://dx.doi.org/10.1080/07350015.2014.880059.

Diewert, W.E. 2002. "The Quadratic Approximation Lemma and Decompositions of Superlative Indexes." *Journal of Economic and Social Measurement* 28(1,2): 63–88. Doi: https://doi.org/10.3233/JEM-2003-0200.

Diewert, W.E. and K.J. Fox. 2017. "Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data." Paper presented at the 15th meeting of the Ottawa Group, Eltville, 10–12 May 2017. Available at: https://www.bundesbank.de/resource/blob/635970/864de61dee8f2b67dcb3ed2a2ab479c5/mL/2017-05-10-ottawa-group-07-2-paper-data.pdf (accessed February 2019).

Éltetö, Ö., and P. Köves. 1964. "On an Index Number Computation Problem in International Comparison" (in Hungarian). In *Statisztikai Szemle*, 42: 507–18. Available at: https://www.ksh.hu/statszemle_archivum#year=1964/issue=05 (accessed May 2019).

Geary, R.C. 1958. "A Note on the Comparison of Exchange Rates and Purchasing Power Between Countries." *Journal of the Royal Statistical Society. Series A (General)* 121(1): 97–99. Doi: http://dx.doi.org/10.2307/2342991.

Gini, C. 1931. "On the Circular Test of Index Numbers." *Metron* 9(9): 3–24.

Howard, A., K. Dunford, J. Jones, M. van Kints, K. Naylor, and R. Tarnow-Mordi. 2015. "Using Transactions Data to Enhance the Australian CPI." Paper presented at the 14th meeting of the Ottawa Group, Tokyo, 20–22 May 2015. Available at: http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/d012f001b8a1cf6cca257eed008074c9/$FILE/Australian_Bureau_of_Statistics-Using_transactions_data_to_enhance_the_Australian_CPI.pdf (accessed August 2017).

ILO, IMF, OECD, UNECE, Eurostat, The World Bank. 2004. *Consumer Price Index Manual: Theory and Practice*. Geneva: International Labour Organization. Doi: https://doi.org/10.5089/9789221136996.069.

Ivancic, L., W. Erwin Diewert, and K.J. Fox. 2011. "Scanner Data, Time Aggregation and the Construction of Price Indexes." *Journal of Econometrics* 161(1): 24–35. Doi: http://dx.doi.org/10.1016/j.jeconom.2010.09.003.

Khamis, S.H. 1972. "A New System of Index Numbers for National and International Purposes." *Journal of the Royal Statistical Society. Series A (General)* 135(1): 96–121. Doi: http://dx.doi.org/10.2307/2345041.

Krsinich, F. 2016. "The FEWS Index: Fixed Effects with a Window Splice." *Journal of Official Statistics* 32(2): 375. Doi: http://dx.doi.org/10.1515/jos-2016-0021.

Puntanen, S., and G.P.H. Styan. 2006. "Historical Introduction: Issai Schur and the Early Development of the Schur Complement." In *The Schur Complement and its Applications*, edited by F. Zhang. 1–16. New York: Springer. Doi: https://doi.org/10.1007/0-387-24273-2_1.

Rao, D.S.P. 1990. "A System of Log-Change Index Numbers for Multilateral Comparisons." In *Contributions to Economic Analysis* 194: 127–139. Doi: https://doi.org/10.1016/b978-0-444-88409-1.50013-0.

Rao, D.S.P. 2005. "On the Equivalence of Weighted Country-Product-Dummy (CPD) Method and the Rao-System For Multilateral Price Comparisons." *Review of Income and Wealth* 51(4): 571–80. Doi: http://dx.doi.org/10.1111/j.1475-4991.2005.00169.x.

Reinsdorf, M.B., W.E. Diewert, and C. Ehemann. 2002. "Additive Decompositions for Fisher, Törnqvist and Geometric Mean Indexes." *Journal of Economic and Social Measurement* 28(1,2): 51–61. Doi: https://doi.org/10.3233/jem-2003-0194.

Ribe, M. 2012. "Some Properties of the RGEKS Index for Scanner Data." Paper presented at Statistics Sweden's scanner data workshop, 7–8 June 2012, Stockholm, Sweden. Available at: https://www.scb.se/Statistik/PR/PR0101/_dokument/Some%20properties%20of%20the%20RGEKS%20index%20for%20scanner%20data.pdf (accessed February 2019).

Summers, R. 1973. "International Price Comparisons based upon Incomplete Data." *Review of Income and Wealth* 19(1): 1–16. Doi: http://dx.doi.org/10.1111/j.1475-4991.1973.tb00870.x.

Szulc, B. 1964. "Index Numbers of Multilateral Regional Comparisons" (in Polish). In *Przeglad Statysticzny*, 3: 239–54.

Turvey, R. 1979. "Treatment of Seasonal Items in Consumer Price Indices." In *Bulletin of Labor Statistics,* 4, 13–23. Geneva: ILO. Available at: https://www.ilo.org/public/libdoc/ilo/P/09606/09606(1979-4)XIII-XXIII.pdf (accessed May 2019).

University of Chicago. 2018. "Kilts Center for Marketing: Dominick's Data Manual." Chicago: University of Chicago Booth School of Business. Available at: https://www.chicagobooth.edu/-/media/enterprise/centers/kilts/datasets/dominicks-dataset/dominicks-manual-and-codebook_kiltscenter.aspx (accessed November 2018).

Van IJzeren, J. 1952. "On the Plausibility of Fisher's Ideal Indices" (in Dutch). In *Statistische en Econometrische Onderzoekingen, Nieuwe Reeks*, 7, 104–15. The Hague: CBS.

Van IJzeren, J. 1983. "Index Numbers for Binary and Multilateral Comparison: Algebraical and Numerical Aspects." In *Statistical Studies,* 34. The Hague: CBS.

Vartia, Y.O. 1974. "Relative Changes and Economic Indices." Doctoral dissertation, University of Helsinki.

Vartia, Y.O. 1976. "Ideal Log-Change Index Numbers." *Scandinavian Journal of Statistics* 3(3): 121–126. Available at: http://www.jstor.org/stable/4615624 (accessed May 2019).

Received June 2018
Revised December 2018
Accepted February 2019