Quant Mark Econ (2010) 8:207–239
DOI 10.1007/s11129-009-9073-0

Recording discrepancies in Nielsen Homescan data: Are they present and do they matter?

Liran Einav · Ephraim Leibtag · Aviv Nevo

Received: 18 September 2008 / Accepted: 17 July 2009 / Published online: 25 August 2009

Abstract We report results from a validation study of the Nielsen Homescan consumer panel data. We use data from a large grocery retailer to match transactions that were recorded by the retailer (at the store) and by the Homescan panelist (at home). The matched data allow us to identify and document discrepancies between the two data sets in reported shopping trips, products, prices, and quantities. We find that the discrepancies are largest for the price variable, and show that they are due to two effects: the first

We are grateful to two anonymous referees, to Peter Rossi (the Editor), and to participants at the Chicago-Northwestern IO-Marketing conference, the Hoover Economics Bag Lunch, the NBER Price Dynamics Conference, the NBER Productivity Potpourri, the Stanford Economics Junior Lunch, and the World Congress on National Accounts for many helpful comments. We thank Andrea Pozzi and Chris Taylor for outstanding research assistance. This research was funded by a cooperative agreement between the USDA/ERS and Northwestern University, but the views expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Department of Agriculture.

L. Einav
Department of Economics, Stanford University, Stanford, CA, USA
e-mail: leinav@stanford.edu

L. Einav · A. Nevo
NBER, Cambridge, MA, USA

E. Leibtag
USDA/ERS, Washington, DC, USA
e-mail: eleibtag@ers.usda.gov

A. Nevo (✉)
Department of Economics, Northwestern University, Evanston, IL, USA
e-mail: nevo@northwestern.edu

seems like standard recording errors (by Nielsen or the panelists), while the second is likely due to the way Nielsen imputes prices. We present two simple applications to illustrate the impact of recording differences, and we use one of the applications to illustrate how the validation study can be used to adjust estimates obtained from Nielsen Homescan data. The results suggest that while recording discrepancies are clearly present and potentially impact results, corrections, like the one we employ, can be adopted by users of Homescan to investigate the robustness of their results to such potential recording differences.

Keywords Measurement error · Validation study · Self-reported data

JEL Classification C81 · D12

1 Introduction

Nielsen Homescan (Homescan) is a large data set that tracks consumers' grocery purchases by asking consumers to scan barcodes of purchased products at home after each shopping trip. The Homescan data allow researchers, practitioners, and policymakers to study questions that cannot be addressed using other forms of data. For example, Homescan covers purchases at retailers that traditionally do not cooperate with scanner data collection companies, such as Wal-Mart and Whole Foods. Another advantage of the Homescan data is its national coverage, which provides wide variation in household location and demographics compared to other panel data sets in which most households are from a small number of markets with relatively limited variation in demographics. Indeed, there has been a recent surge in the use of Homescan in the academic literature (Dube 2004; Aguiar and Hurst 2007; Hausman and Leibtag 2007; Katz 2007; and Broda and Weinstein 2008 and forthcoming).

Questions have been raised regarding the credibility of the Homescan data since the data are self-recorded, and the recording process is time consuming. There are two common concerns. First, there are potential concerns about sample selection. Because of the time commitment, households who agree to participate in the sample might not be representative of the population of interest. Second, households who agree to participate in the sample might record their purchases incorrectly.

This paper reports results from a validation study of the Homescan data that allows us to examine the second concern—the accuracy of the recording. We use data from a single retailer to match records from Homescan with detailed transaction-level data from the retailer. Thus, we are able to observe the same transaction twice: as it was recorded by the retailer, and as recorded by the Homescan panelist. By comparing the two data sources we can document reporting discrepancies (or lack thereof) and propose ways to correct for any implication these differences will have on statistical analysis. In particular, we compare the data sets along two dimensions. First, we document differences in trip and product information. If the household reported a trip that cannot be found in the retailer data, we attribute this to mis-recorded trip information (store and/or date). Alternatively, if the household did not report a trip that appears in the retailer data, we will consider this as an unreported trip. As we discuss later, this is the one and only statistic that relies on loyalty card information and should be interpreted more cautiously as loyalty card information is known to be noisy. Within matched trips, we document if the household did not record or mis-recorded the product information. Second, for matched products, we document differences in the recording of the purchase price and quantity, and information on whether the product was purchased on promotion.

Our goal is to present the results of our comparison and let future users of Homescan decide whether they believe the discrepancies we report are potentially a problem in their application. Our analysis proceeds in three steps. First, we describe the magnitude of discrepancies between the two data sets in different dimensions. Second, we investigate whether recording differences are correlated with household or trip characteristics. Correlation of recording discrepancies with demographics may be suggestive of which types of research would be most sensitive to such differences. For example, we ask whether a correlation between the price paid and demographics, observed in the Homescan data, can be driven by systematic recording difference of price by demographic groups. Third, we show how to correct for the reporting errors, and we provide sufficient information form our validation study to allow future users of Homescan who wish to perform the proposed correction to do so.

We would like to clarify two important issues more related to terminology than to substance. First, through most of the paper we treat the retailer's data as the "truth," allowing us to attribute any differences between the data sets to "errors" or "mistakes" in the Homescan data. Of course, to the extent that there are recording errors in the retailer's data, these words should be interpreted accordingly. We discuss this further in the context of the results. Second, we often refer to "errors," "mistakes," or "mis-recording" in Homescan. These could be driven by various mechanisms that are discussed later in the paper: recording errors by the Homescan panelists themselves, misunderstanding of the Homescan instructions, or differences that are generated due to the way Nielsen puts together the data. We simply want to note that by using the words "errors," "mistakes," or "mis-recording" we mean *any* of these possible mechanisms.

In Section 2, we describe the study design and the data construction process. In Section 3, we document recording differences. For approximately 20% of trips recorded in the Homescan data we can say with a high degree of certainty that there is no corresponding transaction in the retailer's data. This suggests that either the store or date information was recorded with error. Using the retailer's loyalty card information, we find that there also seem to be many trips that are found in the retailer's data with no parallel in the Homescan data. Therefore, there seems to be evidence that households do not record all of their trips. For the trips we matched, we find that more than 10% of the items are not recorded. For those items recorded in both data sets, we find that quantity is reported fairly accurately: 94% of the quantity information matches in the two data sets, and conditional on a reported quantity of 1 in the Homescan data, this probability goes up to 99%.

The match for the price variable is worse. In about half of the cases the two data sets do not agree. However, the correlation between the Homescan price and the retailer's price is 0.88, and the recording error explains only about 22% of the variation in the reported price. We document two types of price errors. When the item is not associated with a loyalty card discount, the price recording errors are similar to classical errors, and are roughly normally distributed around the true price. In this case the correlation between the two prices is 0.96 and the error explains only 8.5% of the variation in the Homescan price. In contrast, when the item is associated with a loyalty card discount, prices in Homescan tend to over report the true price, sometimes by a large amount. It seems likely that much of this second case is driven by the way Nielsen imputes prices. When available, Nielsen uses the quantity weighted average store-level price instead of the actual price paid by the household. The store-level price will differ from the price the panelist paid for at least two reasons: loyalty card discounts and mid-week price changes. Both these reasons are likely to contribute to the price recording errors we document. We note that this type of error might not be present for data from other retailers that, for example, do not offer loyalty card discounts and do not change prices within a week, as defined by Nielsen.

We also investigate the heterogeneity across households in the quality of their data recording. We find that some households are extremely accurate, while others are much less so. We show that these latter households are more likely to be larger households in which the female head of household is fully employed. This points to opportunity cost of time as an important determinant of recording errors in Homescan. Since we find that recording errors are not mean zero and are correlated with different household attributes, using the Homescan data may result in biased estimates of coefficients of interest, and may lead to inaccurate conclusions. This motivates us to investigate how these recording errors may affect results, and to propose ways to correct for the impact of the recording errors. We present the impact of recording errors in the context of two examples that use both the retailer and Homescan data. We first study how the price paid varies with demographics, and we then estimate demand. We also use the first example to illustrate how the validation sample can be used to correct for recording errors. Indeed, we show that results using the true data and the Homescan data could be vastly different, and that our correction procedure makes them closer.

For our correction method to be applicable more broadly in the Homescan data, we rely on the assumption that the distribution of the recording errors is the same in our validation study as in the rest of the Homescan data. Because there is a reason to believe that our validation sample is not fully representative of the entire Homescan data, corrections using our validation sample should be done with caution and are probably best viewed as robustness checks.

This paper fits into a broader literature of validation studies. Responses to surveys and self-reported data are at the heart of many data sets used by researchers, executives, and policymakers. For example, the Panel Study of Income Dynamics (PSID), the Current Population Survey (CPS), and the Consumer Expenditure Survey (CEX) are used heavily by economists. One concern with self-reported data is that the data are recorded with error, and that the error is systematically related to the characteristics of the respondents or to the variables being recorded. Econometricians have developed theoretical models to examine the consequences of measurement error. To study the magnitude of the measurement error and to document the distribution of the error, an empirical literature has emerged that compares the self-reported sample to a validation sample. Bound et al. (2001) provide a detailed review of this literature. This paper adds to this literature by examining a different data set and using a different validation method. While most of the literature has focused on data sets that record labor market decisions and outcomes, we study the Nielsen Homescan data, which documents purchase decisions. We compare the recording errors we document to errors in commonly used economic data sets and find that errors in Homescan are of the same order of magnitude as errors in earnings and employment status data.

2 Data

2.1 Data sources

2.1.1 Homescan

The Homescan data consist of a panel of households who record their grocery purchases.[1] The purchases are from a wide variety of store types, including traditional food stores, non-traditional outlets such as supercenters and warehouse clubs, and online merchants. Consumers, who are at least 18 years old and interested in participating, register online and are asked to supply demographic information. Based on this information, Nielsen contacts a subset of these consumers. Consumers selected to become panel members are not paid for participating in the program. However, every week a panel member who scans at least one purchase receives a set amount of points. The points can be redeemed for merchandise. Panelists can earn additional points for answering surveys and by participating in sweepstakes that are open only to panel members.

Each participating household is provided with a scanner and instructed to scan all purchased items upon returning home after a given shopping trip. For

[1] See also http://www.nielsen.com/clients/index.html for additional information about the Homescan data.

each shopping trip the panelist is asked to identify the store from which items were purchased. They then scan the barcodes of the products they purchased, and enter the quantity of each item, whether the item was purchased at the regular or promotional ("deal") price, and the coupon amount (if used) associated with this purchase.

Nielsen then matches the barcode, or Universal Product Code (henceforth UPC), with detailed product characteristics. The recording of price is particularly important to understand some of the findings below. If the household purchased products at a store covered in Nielsen's store-level data ("Scantrack"), Nielsen does not require the household to enter the price paid for each item (as a way to make the scanning process less time-consuming for the household). Instead, the price is imputed from the store-level data. We believe (but could not verify) that this imputation is used for all products bought at the retailer's stores from which we obtained data. If the same item could be transacted at different prices within the same store during the same week, this imputation process can introduce errors into the price data. Different consumers will pay different prices for (at least) two reasons. First, in some weeks discounts are offered to loyalty card members and therefore consumers who use the loyalty card will pay a lower price than those who do not use a loyalty card. Second, the retailer typically changes the price at most once a week, but oftentimes the price changes do not align with the week definitions used for price imputations by Nielsen. Therefore, consumers may pay a different price depending on which part of the week, as defined by Nielsen, they visited the store. As we will see, this imputation process will lead to differences in price that are frequent and sometimes large.

In the analysis below we use data from 2004. We consider only households that are part of the "static" sample, which contains households who report purchases in at least 10 months of the year. These households are generally considered more reliable than those who report for fewer months, and these are the only data available, to date, to researchers outside of Nielsen. Overall, the data include purchases of almost 250 million different items by just under 40,000 households. We will focus on two metropolitan markets where the retailer has a significant presence. In these markets there are 1,249 households in Homescan who report over 900,000 items purchased.

2.1.2 The retailer's data

The second data set comes from a large national grocery chain, which we will refer to as "the retailer." This retailer operates hundreds of stores across the country and records all the transactions in all of its stores. For each transaction, the data record the exact time of the transaction, the cashier number, and the loyalty card number if one was used. The data list the UPCs purchased, the quantity purchased of each product, the price paid, and the loyalty card discount (if there was one). The retailer links loyalty cards that belong to members of the same household, primarily by matching the street addresses and telephone numbers individuals use when applying for a loyalty card. The retailer then assigns each household a unique identification number. Clearly, this definition of a household is more prone to errors than is Homescan's definition, in which a household is simply associated with the house at which the scanner resides. We return to this later.

In principle, we could try to match our Homescan data with all of the retailer's data in 2004. However, due to constraints on the size of the data we could obtain from the retailer, we had to limit our analysis to only a subsample of the retailer's data. We therefore obtained the retailer data in two steps as a way to maximize potential matches subject to the size constraint. In the first step we identified a set of consumers who claimed to go to the a retailer's store on a particular date. We then obtained complete transaction level data—including exactly what was bought and how much was paid—from the retailer for 1,603 distinct store-days. Our simple algorithm for matching shopping trips recorded in the Homescan data (Homescan trips) to shopping trips recorded in the retailer's data (retail trips) revealed 1,372 likely matches that are associated with 293 distinct households.

In a second step, we asked the retailer to use the loyalty card identified in these 1,372 retail trips and to provide us with all the retail trips available (during 2004) for the households associated with these cards. Throughout the paper we aggregate retail trips of a given household within a day. Figure 1 provides a schematic chart that sketches the key steps in the data construction process. The full process is described in more detail in the Appendix.

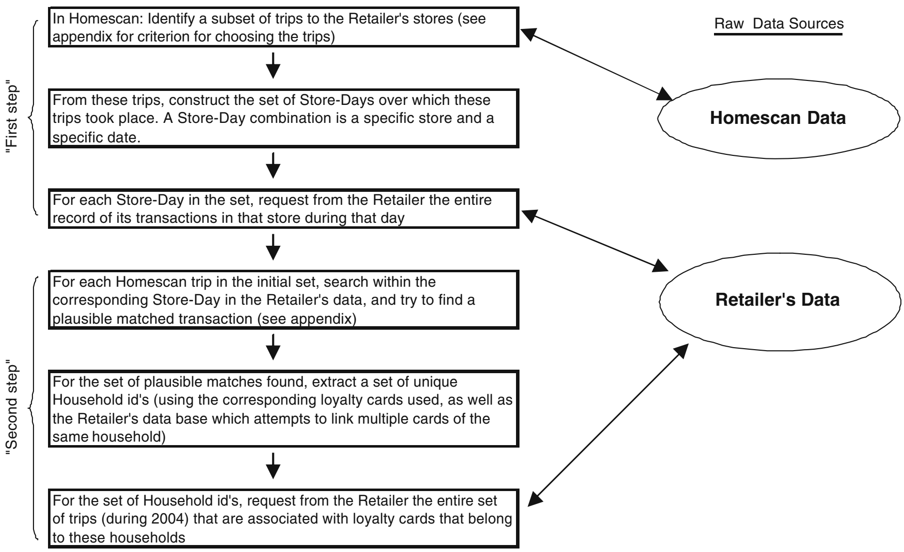

Fig. 1 Schematic sketch of the data construction process

Although the sample may seem a bit arbitrary, our sample selection method is impartial. Our data selection criteria was meant to generate the most matches given our data extraction size limitations. Additionally, the retailer was fully cooperative in the sense that the retailer provided all the data we requested. While we end up with a non-standard sample, we cannot think of any reason that our sample selection would be correlated with any (potential) recording errors. We also want to emphasize that we only used the loyalty card information to generate a data request. However, we did not use it for the record-matching strategy. Therefore, unless we explicitly note, the statistics we provide below will not be impacted by the way the retailer generates the loyalty card panel.

2.2 Record-matching strategy

We now describe our strategy for matching Homescan trips with retail trips.[2] We start by analyzing possible matches in the data obtained in the first step. Recall that a Homescan trip contains all products purchased by the household on a particular day in a particular store. The retailer's data contain the products purchased in each of the (more than 2,500 on average) retail trips at the same store and day. The goal is to match the Homescan trip to exactly one of the retail trips in the retailer's data, or to determine that none of the trips in the retailer's data is a good match. The latter case would be indicative of the household misrecording the date or the store information in Homescan.

Since this procedure relies on the coding of the items (UPCs), one may be concerned that certain items, especially non-packaged items, may have different codes at the retailer's stores and in Homescan. An additional concern is that the Homescan data we use only include the food items scanned by the household, while the retailer data also include non-food items. To deal with these concerns, we generated the universe of UPCs used by Homescan panelists in the entire 2004 Homescan panel and, separately, the universe of UPCs that are used by the retailer. We then restricted attention throughout the rest of the analysis to only the intersection of these two lists of UPCs (by eliminating from the analysis all data related to UPCs not in the intersection). Therefore, if a UPC is found in a retail trip but not in the corresponding Homescan trip, we can be sure that the UPC should have been recorded in the Homescan data but was not for some reason.

After reducing the data set as described above, we continue as follows. Our unit of observation is a Homescan trip. For each such Homescan trip,

[2]Earlier we mentioned a simple matching algorithm we used for the data construction. This was only used to speed up the data requesting process from the retailer, and we do not use its results further. In this section we describe a more systematic matching strategy that is used for the remainder of the paper.

for which we have the retailer's data for that store and that day, we count the number of distinct UPCs that overlap between the Homescan trip and each of the hundreds of retail trips on the same date and in the same store. We then keep the two retail trips with the largest number of UPC overlap, and define ratios between the UPC overlap in each retail trip and the number of distinct UPCs reported for the Homescan trip. The first, $r1$, is the ratio of the number of overlapping UPCs in the retail trip with the highest overlap to the total number of distinct UPCs reported in the Homescan trip. The higher this ratio, the higher the fraction of products matched, and the more likely that this trip is a correct match. The second ratio, $r2$, is similar, but is computed for the retail trip with the second-highest overlap. By construction, $r2$ will be less than or equal to $r1$. A higher $r2$ makes it more likely that the second retail trip is also a reasonable match. Since, in reality, there is, at most, a single retail trip that should be matched, this statistic tries to guard against a false positive. Our confidence in the match between the Homescan trip and the first retail trip increases the higher is $r1$ and the lower is $r2$. As will become clear below, in practice it turns out that false positives resulting from this algorithm do not seem to be a concern once the Homescan trip includes a large number of distinct UPCs.

Using these two statistics, $r1$ and $r2$, and the number of products purchased in the Homescan trip, we separate each Homescan trip into one of three categories: reliable matches, Homescan trips that with high probability do not have a match, and uncertain matches (i.e., we cannot classify these Homescan trips into either of the other groups with a reasonable level of certainty). The first group of transactions will be used to study recording errors of products, prices, and quantities. The second group will be used to document unrecorded trips or errors in recording trip information. We applied different criteria to define the three groups and verified that all our findings are robust to reasonable modifications of these criteria.

Matching Homescan trips to retail trips from the second step of the data construction process is a slightly different task. Recall that here we are not supplied with a list of all retail trips for the day and store. Instead, we are given a single retail trip that the retailer believes represents the household's purchases on that day. Thus, the matching problem here is not which retail trip matches the Homescan trip, but rather whether a given retail trip is a good match or not. We match trips by computing the ratio $r1$, the number of distinct UPCs that overlap divided by the number of items in the Homescan trip. Using the statistic $r1$ and the total number of distinct items purchased, we classify the Homescan trips into three categories, as we do with the first step data. In principle, in this step the thresholds for $r1$ used to classify the trips can be different from the thresholds used in the first step. It turns out, however, that the vast majority of $r1$s we compute are either close to one or close to zero, making the choice of a threshold largely irrelevant. As an additional guard against false positives, we also report some of the results when eliminating from the data certain households that seem to be inconsistent in the way they use their loyalty cards.

3 Documenting recording differences

We now summarize our main findings of recording errors in the Homescan data. We organize the discussion around the three dimensions of potential errors: trip information, product (UPC) information, and price/quantity information. As mentioned earlier, for most of what follows we treat the retailer's data as the "truth" and ask if, or how well, the Homescan trip matches it. In that sense, Homescan recording "errors" are defined as records that do not match the retailer's data. Although it could be the case that the retailer's cashier is the one making the error, rather than the Homescan panelist, we think that this is less likely, especially for analysis at the product, price, and quantity level. At the trip level, when we sometimes rely on loyalty card information, it is not clear that the retailer's data are necessarily more accurate. For example, if a household borrows a loyalty card once, then all the retail trips associated with that card will be linked to the household's record. We discuss this further below.

3.1 Trip and product information

We separate Homescan trips according to the number of distinct UPCs in the Homescan data. A small trip is defined as one with 4 or fewer UPCs, a medium trip has 5–9 UPCs, and a large trip is a trip with 10 UPCs or more. A potential concern is that we have false positives, i.e., that we match trips incorrectly. Our preliminary analysis (summarized in Einav et al. 2008) found that for the medium and large trips mis-classification of a match is not a concern. The real issue is whether a match exists at all, which can be diagnosed by focusing on $r1$.

The distribution of $r1$ is displayed in Fig. 2. The information in the top panel helps us address the question of how many of the Homescan trips that are associated with the first step of the data construction process seem to have mis-recorded store and date information. Focusing on large trips, we find that there are 150 trips with $r1$ less than 0.2, 175 with $r1$ less than 0.3, and 180 with $r1$ less than 0.4 (corresponding to 18.5, 21.6, and 22.4%, respectively). For medium trips the corresponding numbers are 113, 155, and 223 (or 9.5, 13.0, and 18.7%). Taken together, these numbers suggest that for about 20% of the medium and large Homescan trips we can say with a high degree of certainty that they do not match any retail trip. Therefore, we conclude that approximately 20% of the Homescan trips have mis-recorded date or store information.[3] The bottom panel of Fig. 2 shows a similar pattern for the second

[3] A natural speculation is that some of these mis-recorded trips simply mis-record the date by a day (e.g., because the household did not get around to actually scanning the purchased products at home until the next day). Using the retailer's data from the second step we found that while such cases occur, they do not account for a large fraction of the 20% mis-recorded trips reported here.

step data, with the distribution of $r1$ being even more bimodal. Here, again, we cannot find a match for about 20% of the Homescan trips, likely due to misrecorded trip information. In the bottom panel this could also be due to misuse of loyalty cards (for example, if the household forgot the card at home and did

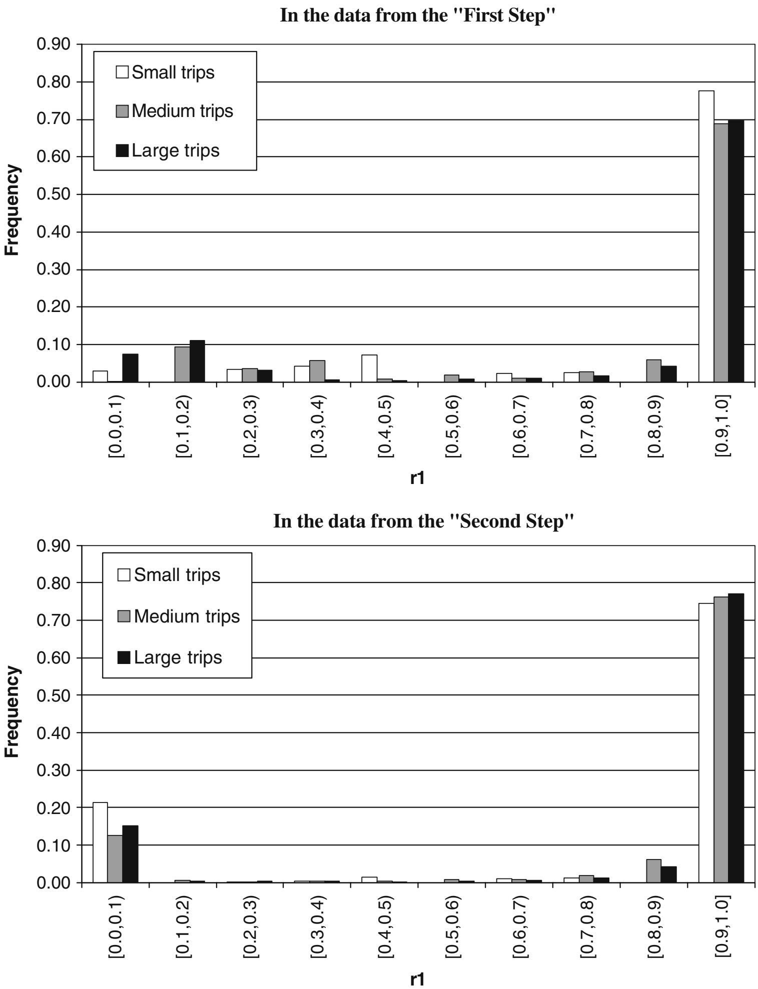

Fig. 2 The bimodal distribution of $r1$. UPC counts (which are used to classify trips as small, medium, or large) are based on the number of distinct UPCs in a trip as reported in the Homescan data. Each histogram plots the distribution of the $r1$ statistic. In the "first step," this is the transaction with the highest UPC overlap in the same store and day. In the "second step," this is the specific transaction in the same store and day by the matched household. Both histograms show a very clear bimodal pattern, where $r1$ is either very close to one or very close to zero, and especially so for large trips. This makes it clear why the results remain essentially unchanged when we change the cutoff value of $r1$ above which we define a match to be successful (throughout the paper we report results that use 0.7 for this cutoff value)

not use it, the trip would be reported in Homescan but will not show up in the second step retailer's data.) However, given that the fraction of unmatched trips in the top panel (where misuse of loyalty cards is not an issue) is very similar, we suspect that much of these unmatched trips are due to mis-recorded trip information.

So far, we have looked for Homescan trips that cannot be matched to retail trips. The data from the second step allow us to also look for the opposite case: retail trips that cannot be found in Homescan. Recall that the data obtained in the second step include all the retail trips associated with certain households. We find that only 40% of these retail trips appear in Homescan, but we suspect that this number is over estimating the fraction of missed trips, and that at least part of it is driven by the retailer classifying multiple loyalty cards as belonging to the same household, or by multiple households sharing the same card. To address this concern, we focus on 273 households that seem to have more reliable loyalty card use.[4] On average, across these households, 53% of the retail trips are not reported in Homescan.

There is heterogeneity across households in their accuracy of reporting. In Fig. 3 we plot, for these 273 households, the ratio of the number of Homescan trips to the number of retail trips, and the fraction of Homescan trips that are matched well on their UPCs. We consider a match as good when the $r1$ is greater than 0.7. Given that these are trips of the same household to the same store on the same day, even trips that do not match well on UPCs are very likely to be the same trips, only with significant mis-recording of items.

Figure 3 suggests that there are two types of households, as the correlation between the two ratios is highly positive (0.47). The first group includes those in the upper right corner, who do not miss many trips and also record the trip information fairly accurately. Households in the second group are those that do poorly on both counts: they fail to report a large fraction of their trips and even when they do report a trip, they do not record its items accurately. Using this rough classification, we use the metric of Fig. 3 to classify households as more or less reliable, depending on how far they are from perfection, which is the point (1,1) in the figure. Table 1 then summarizes the key characteristics for each group and highlights those demographics that are significantly different between the groups. The quality of recording is associated with household composition, as well as with whether the female head of household is fully

[4] For each of the 291 Homescan households for which we obtained data in the second step, we compute the fraction of their retail trips that produced a match, where a match is defined as a trip, of any size, with $r1$ greater than 0.7. A higher fraction implies that this household made fewer errors in recording the store and date. The distribution of this fraction is bimodal. We define a poor match household as one in which the fraction is less than 0.3. This procedure eliminated 18 households and left us with 273 households, who used the same loyalty cards (or matched cards, as linked by the retailer) consistently. We then applied a similar procedure to specific cards of these households, which made us drop a small number of cards.

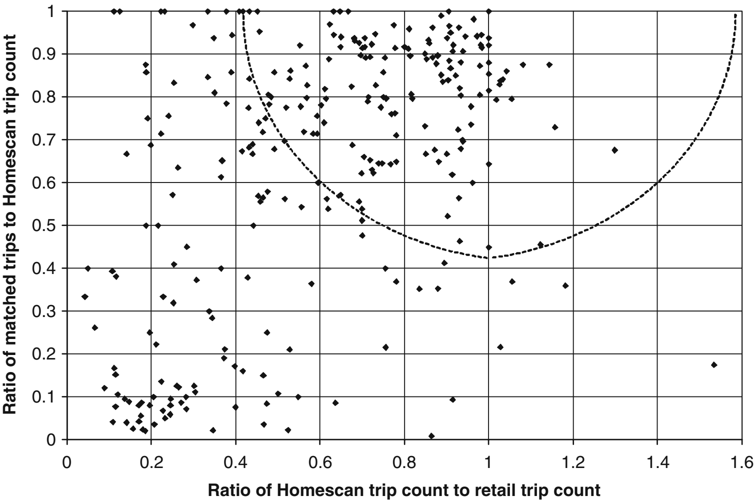

Fig. 3 Household heterogeneity. A data point in this figure is a household whose transactions match consistently (see text for exact definition). There are 273 such households. The horizontal axis reports the ratio between the number of reported Homescan trips and the number of reported retail trips (based on the retailer's loyalty card use). Ratios below one suggest unreported trips (in Homescan). Ratios above one suggest trips to the Retailer's store without using the loyalty card (or using a card that the retailer did not link to the household). The vertical axis reports the fraction of Homescan trips for which we could match a significant number of the UPCs (at least 70%). The figure shows a clear distinction between two types of households. Those with both ratios close to 1 report most of their trips, and report the UPCs in each trip relatively well. In contrast, those households that are close to the origin are households that do not report a large fraction of their trips, and do not report (or report incorrectly) many of the UPCs in those trips they do report. The dashed half circle is the cut point we use to define households as more or less reliable for the statistics reported in Table 1

employed. The pattern is likely driven by the opportunity cost of time of carefully recording purchases. This cost is likely higher for fully employed females and for larger households. In contrast, other demographic variables, such as income and race, are not systematically correlated with the quality of recording.

We now turn our attention to mistakes in recording items (UPCs) conditional on a trip being matched. Since we do not want matching errors to drive our findings we focus on reliable matches only. We use two different criteria for defining a reliable match. First, we look at large Homescan trips, involving 10 or more products, with $r1$ greater than 0.7. There are 2,477 such trips. Second, we examine medium size Homescan trips, with at least 5 but no more than 9 distinct UPCs, and with $r1$ greater than 0.7. There are 3,168 such trips. We do not use the remaining, small trips for the rest of the analysis.

We find that for the typical Homescan trip almost all the products (98% in both large and medium transactions) scanned by the Homescan panelist exist in the matched retail trip. Selection into the sample was conditional on this fraction being at least 70% ($r1 > 0.7$). Nevertheless, we still view this as

Table 1 Comparison of more and less reliable households

	Households who are reporting more accurately	Households who are reporting less accurately	p-value
Number of households	144	129	
Household size	1.96	2.50	0.000
Household income ($000)	48.89	53.82	0.182
No female head of household	0.16	0.05	0.005
Age female	47.90	51.63	0.135
No male head of household	0.28	0.21	0.191
Age male	41.08	44.90	0.232
Number of kids	0.13	0.22	0.029
Number of little kids	0.02	0.05	0.143
Male employed	0.47	0.49	0.704
Male fully employed	0.42	0.45	0.585
Female employed	0.42	0.50	0.189
Female fully employed	0.26	0.38	0.040
Male education (category)	3.04	3.30	0.302
Female education (category)	3.46	3.92	0.017
Married (or widower)	0.58	0.78	0.000
Non-white	0.10	0.13	0.481
“15K” homescan household	0.07	0.08	0.799

This table compares demographics of more and less reliable households, as defined by their recording behavior (see Fig. 3). The reported p-value is for the test of the null hypothesis that the means are equal in both columns. The highlighted demographics are those for which this null hypothesis can be rejected (using a 5% confidence level)

a remarkably high number. This may not be surprising, as the products are scanned, so it is, in fact, hard to imagine how mis-recording at this level could take place.

On the other hand, on average about 10% (14% for medium transaction) of the items that show up in the retail trip are not recorded by the Homescan panelist. Recall that we eliminated from the analysis products with UPCs that only show up in one of the data sets. Thus, these missing items cannot be attributed to categories that the Homescan panelist was not supposed to scan. We qualitatively tried to analyze which items are more likely to be missing in the Homescan trip, by grouping the missed items into product categories, and investigating whether particular categories stand out. While there are many items that belong to various categories that are occasionally missing and no single category accounts for a large fraction of the missing items, two specific types of items are common. The first group includes consumable products: small bottles of drinks, snacks, etc. that are likely often consumed on the way home, before the purchase is scanned. The second group includes items that belong to product categories that include many distinct, yet similar UPCs. Yogurts of different flavors and baby food of different flavors are typical examples. In such cases, it seems likely that individuals simply scan one of the flavors and enter a large quantity instead of scanning each of the flavors (which has a distinct UPC) separately. These will appear as missed products, but in reality might not matter, unless we care about the exact flavor bought.

To measure this we examine the total number of items bought in the trip. In this example, the total quantity would match even if the distinct UPC count does not. This slightly reduces the differences, but not by much, implying that mis-recorded quantity cannot fully explain the difference in the number of products.

In order to check if the mistakes in recording products are systematic, we regress the missed expenditure on the total trip expenditure and find that a larger fraction of the expenditure is missed on larger trips. On large trips the household is more likely to forget to scan, not go through the trouble of doing so, or consume items on the way home.

3.2 Price and quantity information

We now focus on errors in the recording of price and quantity variables. For this purpose we look at the products that appeared in both data sets from the reliably matched trips using the two definitions of reliable trips. It turns out that the statistics we present below hardly vary across the groups, so match reliability does not seem to be a concern. For the rest of this section we will refer to the first set of matched products, those from Homescan trips with at least 10 products and $r1$ greater than 0.7, as "matched large trips"; similarly, we will refer to the products matched from medium Homescan trips as "matched medium trips." For matched large trips we have 41,158 products purchased, an average of 16.6 products per trip. For matched medium trips we have 21,386 matched items, for an average of 6.8 products per trip (recall that these are trips with 5–9 products).

We present summary statistics for the key variables first and then discuss in more detail additional patterns. Table 2 presents the fraction of observations of quantity, expenditure, price, and deal indicator, that match between the reports in the Homescan data and in the retailer's data. We present results separately for large and medium trips to illustrate the robustness of the patterns, but given that the summary statistics are so similar across these two types of trips, we focus the discussion and the subsequent analysis on large trips alone. We find that 94% of the time the two data sources report the same quantity. The total expenditure on the item is the same in both data sets much less frequently, and only 48% of the time do the two data sets report identical expenditure. On average, the expenditure reported in Homescan is about 10% higher than the expenditure recorded by the retailer, although there is wide dispersion around this average (see Table 2). The pattern for price is similar to that of expenditure. It is slightly better matched (50% match rate and 7% higher prices in Homescan on average), possibly because the expenditure variable (price times quantity) is further prone to errors due to mis-recorded quantities. Finally, we examine the deal indicator. In the retailer's data the deal variable equals one if the gross and net price differ. In the Homescan data this is a self reported variable. Overall, this indicator matches in 80% of the observations, a worse match than the quantity data, but better than price.

Table 2 Summary match statistics

	Matched large trips				Matched medium trips			
	Mean	Std.	5%	95%	Mean	Std.	5%	95%
Quantity								
Homescan	1.44	1.16	1	3	1.51	1.36	1	4
Retailer	1.35	0.87	1	3	1.38	0.99	1	3
Fraction same	0.938				0.924			
Expenditure								
Homescan	3.14	2.44	0.69	7.38	3.23	2.74	0.69	7.58
Retailer	2.76	2.03	0.65	6.00	2.82	2.15	0.66	6.29
Fraction same	0.479				0.486			
Log(homescan/retailer)	0.10	0.41	−0.38	0.69	0.10	0.44	−0.42	0.70
Price								
Homescan	2.44	1.63	0.50	4.99	2.44	1.67	0.50	4.99
Retailer	2.25	1.53	0.50	4.89	2.27	1.55	0.50	4.99
Fraction same	0.503				0.512			
Log(homescan/retailer)	0.07	0.37	−0.37	0.61	0.05	0.39	−0.42	0.60
Deal indicator								
Homescan	0.520				0.534			
Retailer	0.554				0.549			
Fraction same	0.795				0.820			
Number of obs. (UPCs)	41,158				21,386			
Distinct shopping trips	2,477				3,168			
Distinct households	263				318			

Large and medium trips are defined using the count of distinct UPCs as reported in Homescan (Medium: 5–9, Large: 10+). An observations in this table is a distinct item (UPC) in a given trip

We now explore in more detail the patterns we found for each of the variables. We start with quantity. The overall match rate is reasonable. However, for 73% of the Homescan data and 76% of the retailer's data (in matched large trips), reported quantities are 1, so a high number of cases in which the two quantities are the same might not be surprising. Indeed, conditional on the Homescan data reporting a quantity of 1, the probability of this report matching the retailer's data is 0.99, while conditional of the Homescan data reporting a quantity larger than 1 the probability of a match is only 0.86. So a reported quantity of 1 seems to be very reliable, while a quantity greater than 1 might be somewhat more prone to mistakes, but still reasonable.

Using the data from the matched large trips, conditional on quantities not matching, 82% of the time the quantity reported in Homescan is higher. Recording errors seem to be of various types, including six-packs that are recorded as quantities of 6 (the fraction of mistakes for reported quantities of 6, 12, 18 and 24 are 0.60, 0.85, 1.00 and 0.78, respectively), typing errors (e.g., 11 instead of 1), and occasional "double scanning" (quantity of 2 instead of 1). Together, this suggests that the Homescan data might be problematic for studying the quantity purchased. It seems to be better suited to measure whether or not a purchase occurred. Overall, the variance of error in the quantity variable constitutes 48.7% of the variance in the Homescan reported quantity. The correlation coefficient between the two quantity variables is 0.72.

While in the case of quantity recording errors are likely driven by the panelist's recording error, the case of price is somewhat different, given our understanding of how the Homescan prices are generated. As described in Section 2, if the consumer purchased the product at a store for which Nielsen has store-level data, then the panelist is not asked to record the price paid. Instead, Nielsen imputes a price from the store-level data. If some of the shoppers in a store during a given week paid the full price while some got a discount then the imputed price will be between the discounted price and non-discount price, and the Homescan data will over report or under report the actual transaction price. There are at least two reasons for heterogeneity in the price paid in a particular week. First, some discounts are offered only to loyalty card holders. Analysis of the retailer data suggests that loyalty cards are used in about 75–80% of the transactions,[5] and that about 60% of the transacted items are associated with loyalty card discounts, so errors due to this data construction process could be important. Second, even though typically the retailer changes the price at most once a week, the day when the price changes is in the middle of the week as defined by Nielsen. Thus, consumers who purchase on different days might pay different prices.

This suggests that the recording errors in price may be either due to the panelist's recording error or due to the price imputation, and the statistical properties of these errors are likely different. Other retailers might not offer loyalty card discounts or set prices exactly in alignment with Nielsen's definition of a week; thus, the price imputation error might not be present in data from these retailers.

To examine this issue, we present in Fig. 4 the distribution of the logarithm of the ratio of price in the Homescan data to the price in the retailer's data. The overall distribution is presented for comparison in both panels (dashed grey line). The solid black line in the top panel presents this distribution for transactions for which the store did not have a loyalty card discount for that item on that day, while the bottom panel repeats the same exercise for the cases in which a loyalty card discount was present. The overall pattern largely follows our discussion above. That is, the solid black line in the top panel is close to a standard "classical" error, with mean close to zero and most of the mass around zero. In other words, in these cases, even when the price does not match, the differences are small. In contrast, in the bottom panel there is a very fat right tail of the distribution, and the average difference of log prices is more

[5] While we do not have direct store-level data on card usage, we can get a rough idea of this. Specifically, for each observation in the transaction-level data that is associated with a loyalty card discount that reduced the price from p to $p - d$, we ask what is the corresponding store-level (average) price $\overline{p}$ at that store and week. Our estimate of loyalty card use (for a given item at a given store and week) is then given by $u = (p - \overline{p})/d$. Of course, this may vary due to sampling variation, but across items, stores, and weeks, the distribution of u is centered around 75–80%.

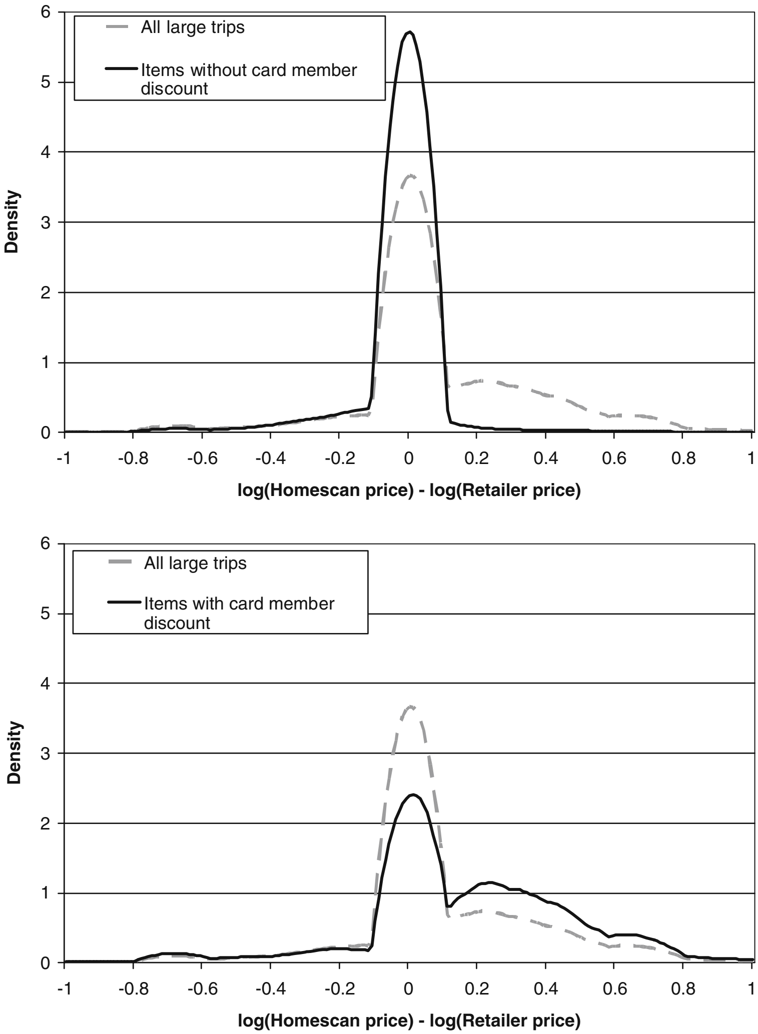

Fig. 4 Two sources of price recording errors. This figure presents kernel densities of the difference between the log Homescan price and the log Retailer price. The gray dashed line is common to both panels and uses all matched large trips, which parallels the corresponding row in the left panel of Table 2. The black solid line in each panel uses a subset of these data. The top panel uses the observations for which the item was not associated with any store discount at the time of purchase (about 34% of the cases), while the bottom panel uses the observations for which the item was associated with a card member store discount at the time of purchase (about 54% of the cases). In 12% of the cases we could not determine whether the item was associated with a discount (8% of the cases did not have a matched item in the store level data and in 4% of the cases it was difficult to determine whether a discount was available to all customers). Summary statistics for the solid lines in the top (bottom) panel are: mean −0.040 (0.143), standard deviation 0.305 (0.391) and 5%−95% range −0.288 to 0 (−0.434 to 0.693)

than 14%. This is consistent with the fact that all of our data is associated with users of loyalty cards (this is how we matched them), while the price imputed for them is aggregated over a population of which some do not use the loyalty card. Therefore, imputed Homescan prices are higher in such cases.

Overall, the variance of the error in the price variable constitutes 21.8% of the variance in the Homescan price (8.5% if no loyalty card discount is offered). The correlation coefficient between the two price variables is 0.88. The correlation increases to 0.96 if we condition on the Homescan deal indicator equal to 0 (and to 0.96 if we look at observations where no loyalty card discount was offered), and it decreases to 0.83 if the Homescan deal indicator is equal to 1 (0.84 if a loyalty card discount was offered). The variance in the error of the expenditure data explains 37.1% of the variation in the per item expenditure of the Homescan data. The correlation coefficient is 0.79.

In summary, we find that for the matched products, quantity is reported fairly accurately, although, when quantity reported is higher than 1, the reported data are less accurate and therefore the correlation between the two quantity variables is quite low. Prices and expenditures are reported with less accuracy. We suspect that this is due mostly to the Nielsen matching procedure that imputes store-level prices when possible.

3.3 Comparison to measurement errors in other data sets

It may be useful to compare the magnitude and frequency of recording errors in Homescan to those reported in other validation studies. To do so, we use Bound et al. (2001, Section 6) who summarize the evidence on measurement errors in data sets often used by labor economists. They report errors in earnings, transfer program income, assets, hours worked, unemployment status, industry and occupation, education, and health related variables. While it is hard to compare across contexts and over a large set of variables, our overall impression is that the magnitude of recording errors we document for Homescan are on the lower end of the range of recording errors reported by Bound et al. (2001). For example, Bound and Krueger (1991) compare the annual earnings reported in the CPS with Social Security administrative records. They find that the variance of the log of the ratio of earnings reported in the two data sets is 0.114 for men and 0.051 for women. The correlation coefficient between the two variables is 0.884 for men and 0.961 for women. Ashenfelter and Krueger (1994) study the years of schooling reported by twins: they compare the own report to the report of the twin. They find a correlation coefficient of 0.9. We, on the other hand, find that the overall variance in the log of the ratio of the Homescan and retailer price is 0.139. The variance is as low as 0.046 when the Homescan deal indicator is equal to zero, and 0.092 if no loyalty card discount is offered. So overall it seems like the errors we document in Homescan are comparable to what is found in other commonly used data sets.

4 Correcting for recording errors

Up to this point we used the validation sample to document recording errors. In this section, we discuss how the validation sample can be used to control for recording errors. Our discussion follows Chen et al. (2005), who provide more details and additional references. The basic idea is to use the validation sample to learn the distribution of the error, conditional on variables observed in the primary data. One can then use this distribution and "integrate over" it in the primary data. Of course, a key assumption is that the (conditional) distribution of the error is the same in both the validation data and in the primary data. This assumption can be evaluated on a case-by-case basis, and we revisit it below in the context of our application.

Formally, suppose the model we want to estimate implies a moment condition:

$$E\left[m\left(X^{*},\beta_{0}\right)\right]=\int m(x^{*},\beta_{0})\,f_{X^{*}}(x^{*})dx^{*}=0 \tag{1}$$

where $m(\cdot)$ is an $r\times 1$ vector of known functions, X^{*} are variables, which might not be fully observed, and $\beta_0\in B$, a compact subset of $\Re^{q}$ with $1\le q\le r$, is a vector of the true value of unknown parameters that uniquely sets the moment condition to zero. We observe two data sets. In the first, "primary" data set $\{X_{pi}:i=1...N_p\}$, we do not observe X^{*}, rather we only observe X, which is measured with error of unknown form. In our context, Homescan is the primary data set. In the second, "validation" data set we observe $\{(X^{*}_{vj},X_{vj}):j=1...N_v\}$, i.e., both the variable that is measured with error and its true value. The matched Homescan-retailer data is the validation sample in our case. We denote by $f_{X^{*}_{p}}$, f_{X_p}, $f_{X^{*}_{v}}$, and f_{X_v}, as the marginal densities of the latent variable and the mis-measured variable in the primary and validation data sets. We also denote by $f_{X^{*}_{p}|X_p}$ and $f_{X^{*}_{v}|X_v}$ the conditional densities of the latent variable given the mis-measured variable in the primary and validation data sets, respectively.

The key assumption is that

$$f_{X^{*}_{v}|X_v=x}=f_{X^{*}_{p}|X_p=x}\ \text{ for all } x. \tag{2}$$

That is, that the distribution of the true variables, conditional on the observed variables, is the same in both the primary and the validation samples. This is not a trivial assumption. For example, to use our validation sample for the entire Homescan data, it would require that the recording error is the same for the retailer we observe and for all other retailers. Even though we assume that $f_{X^{*}_{v}|X_v=x}=f_{X^{*}_{p}|X_p=x}$, we note the marginal density f_{X_v} might be different than f_{X_p} and therefore $f_{X^{*}_{v}}$ might be different than $f_{X^{*}_{p}}$.

We do not observe X^{*} in the primary data set and therefore cannot directly use the moment condition in Eq. 1 to estimate β. However, we could use the validation sample to estimate $f_{X_v^*|X_v}$, and the primary data set to estimate f_{X_p}. Thus,

$$f_{X_p^*}(x^*) = \int f_{X_p^*|X_p=x}(x^*) f_{X_p}(x)dx = \int f_{X_v^*|X_v=x}(x^*) f_{X_p}(x)dx \qquad (3)$$

where the second equality uses the key assumption that $f_{X_v^*|X_v=x} = f_{X_p^*|X_p=x}$. We can estimate this density by $\widehat{f_{X_p^*}}(x^*) = \int \widehat{f_{X_v^*|X_v=x}}(x^*) \widehat{f_{X_p}}(x)dx$ where $\widehat{f_{X_v^*|X_v=x}}(x^*)$ is the estimate of the density of X_v^* conditional on $X_v = x$, and $\widehat{f_{X_p}}(x)$ is the estimated density of X_p in the primary data. Now, we can use the moment condition to estimate the parameters of interest by

$$\widehat{\beta} = \arg\min_{\beta} \left(\int m(x^*, \beta) \widehat{f_{X_p^*}}(x^*)dx^* \right)' \widehat{W} \left(\int m(x^*, \beta) \widehat{f_{X_p^*}}(x^*)dx^* \right), \qquad (4)$$

where $\widehat{W}$ is a positive definite symmetric weight matrix.

While intuitive, this estimator involves estimating two distributions, $\widehat{f_{X_v^*|X_v=x}}(x^*)$ and $\widehat{f_{X_p}}(x)$, potentially non-parametrically, and then using them in a non-linear moment condition. Instead, Chen et al. (2005) propose to define

$$g(x, \beta) \equiv E\left[m\left(X_p^*, \beta\right) | X_p = x\right] = \int m\left(x^*, \beta\right) f_{X_p^*|X_p=x}(x^*)dx^*. \qquad (5)$$

Note, that $g(\cdot)$ is a function of the variable measured with error X_p, that is observed in the primary data set, rather than with respect to the true (latent) variable X_p^*. We can now apply the law of iterated expectations, so that

$$\begin{aligned} E_p\left[g(X, \beta_0)\right] &= E_p\left[E\left[m\left(X_p^*, \beta_0\right) | X_p = x\right]\right] \\ &= E_p\left[E\left[m\left(X_p^*, \beta_0\right)\right] | X_p = x\right] = E\left[0 | X_p = x\right] = 0. \qquad (6) \end{aligned}$$

Thus, the original moment condition implies that

$$E_p\left[g(X, \beta_0)\right] = \int g\left(x, \beta_0\right) f_{X_p}(x)dx = 0, \qquad (7)$$

and we can estimate the parameters of interest by

$$\widehat{\beta} = \arg\min_{\beta} \left(\frac{1}{N_p} \sum_{i=1}^{N_p} \widehat{g}(X_{pi}, \beta) \right)' \widehat{W} \left(\frac{1}{N_p} \sum_{i=1}^{N_p} \widehat{g}(X_{pi}, \beta) \right) \qquad (8)$$

where $\widehat{W}$ is a positive definite symmetric weight matrix, and $\widehat{g}(X_{pi}, \beta)$ is a non-parametric estimate of $g(X_{pi}, \beta)$, estimated using the validation sample. Using the validation sample to estimate $\widehat{g}(X_{pi}, \beta)$ yields a consistent estimate because of the key assumption (Eq. 2). Chen et al. (2005) propose using a series (sieve) estimator of $g(x, \beta)$:

$$\widehat{g}(x, \beta) = \sum_{j=1}^{N_v} m\left(X_{vj}^*, \beta\right) p^k\left(X_{vj}\right)' \left(P_v' P_v\right)^{-1} p^k(x), \qquad (9)$$

where $\{p_l(x), l = 1, 2, ...\}$ denotes a sequence of known basis functions, $p^{k_{nv}}(x) = (p_1(x), ...p_k(x))'$ and $P_v = (p^k(X_{v1})...p^k(X_{vN_v}))'$ for an integer k that increase with the sample size N_v, such that $k \to \infty$, and $k/N_v \to 0$ as $N_v \to \infty$. In words, $\widehat{g}(x, \beta)$ is estimated by projecting it onto the basis functions. In general, the optimization in Eq. 8 is non-linear, but not more complex than the optimization implied by using the moment condition given in Eq. 1.

In a linear model this simplifies to a fairly simple procedure. For example, suppose we want to estimate a regression of price paid p^* on demographics D, as we do in the next section. In the primary data (Homescan) we observe p and D, but we are concerned about possible recording errors in p. In the validation sample, the matched Homescan-retailer data, we observe p, p^*, and D, where p is the Homescan-reported price and p^* is the retailer-reported price. We then first use the validation data to regress the retailer's price p^* on p and D, to obtain, for example,

$$E(p^*|p, D) = D'\widehat{\beta} + \widehat{\alpha}p. \tag{10}$$

We then use the estimated coefficients, $\widehat{\beta}$ and $\widehat{\alpha}$, to construct $E(\widehat{p^*|p, D})$ in the Homescan data. Using Homescan, we then regress this predicted price on D to obtain the error-adjusted estimates.

It may be instructive to present the simplest case, in which both the prediction and estimating equations are linear, and the set of covariates D is identical. In this case, the "naive" regression in Homescan would be to regress p on D, while the corrected regression will be to regress $D'\widehat{\beta} + \widehat{\alpha}p$ on D. If the true coefficient on D is γ, then the "naive" coefficient will be (roughly) $\frac{\gamma-\beta}{\alpha}$. That is, with no measurement errors or with classical measurement error, we would have $\alpha = 1$ and $\beta = 0$, and no bias. However, either $\alpha \neq 1$ (the case where the mean of measurement error is not zero) or $\beta \neq 0$ (which would arise if the measurement error is correlated with D) will result in a bias.

5 Applications

In this section we illustrate the impact of measurement error and the proposed correction in the context of two applications. In the first application the price is the dependent variable, while in the second example we explore an illustrative demand equation in which price is a regressor.

5.1 An illustrative application I: price regressions

Recently, researchers have used Homescan to study how the prices paid vary with household demographics (e.g., Aguiar and Hurst 2007). We perform a simple version of such a study in order to evaluate the impact of recording errors. Our goal is twofold. First, the application provides a more meaningful way to evaluate the importance of recording errors. That is, while describing the recording errors is potentially interesting, it is not sufficient for determining whether the recording errors should be a serious concern. In this section we ask if the recording errors matter for conclusions drawn from the analysis. Second, we use this application to demonstrate how one could use our validation study to correct for recording errors in Homescan, and we hope that future users of Homescan will do so too, at least as a robustness check.

We note that our goal here is not to replicate any particular study, just to demonstrate that the errors could have important implications for certain conclusions, and to show how the validation study can be used to address these errors. We chose this particular application for two reasons. First, it is important and active line of work, making it more likely that researchers who use Homescan will perform a similar analysis. Second, it is simple. The key regression here is linear, and has price on the left hand side. This makes this analysis robust to classical recording errors in the Homescan price, and it is only non-classical recording errors that would lead to bias. Other settings, in which the model is non-linear or the variable of interest is on the right hand side will make the analysis more sensitive to errors, and the correction slightly more complex.

The regression of interest in this application is

$$p_{ik} = \alpha_k + \beta' D_i + \varepsilon_{ik} \tag{11}$$

where i is a household, k is a specific product (distinct UPC), p is the unit price paid for this product,[6] and D_i is a vector of demographic characteristics. The α_k's are a set of UPC fixed effects, and β is a vector of coefficients of interest. The economic question is whether certain demographic groups pay more or less for the same product, relative to the rest of the population. Aguiar and Hurst (2007), for example, focus on the price paid over the life cycle and emphasize their finding that the elderly pay lower prices for the same item, compared to other age groups. One could analyze the corresponding effect of other demographic groups, such as gender and race.

We start with Table 3, which presents results from estimating the above regression. An observation is a product (UPC) in a matched large trip, i.e., in a large trip with $r1$ greater than 0.7. The regression reported in the first column uses as the dependent variable the price, in cents, as recorded in Homescan, while the regression reported in the second column uses the price in the retailer's data. The covariates are identical in both cases as is the sample, only the dependent variable is different. The results in the two columns are somewhat different. Out of the twenty regression coefficients, five have different signs, nine do not agree on their statistical significance, and the point estimate (when they have the same sign) are off by an average of more than 40%.

[6]The reported results do not account for coupons. Results that use prices net of coupons are qualitatively similar, and are available from the authors upon request.

Table 3 Comparison of simple price regressions

Sample	All matched items		Items with card member discount		Items w/o card member discount	
Dependet variables (in cents)	Homescan price	Retailer price	Homescan price	Retailer price	Homescan price	Retailer price
Household size	−1.321 (0.585)	−3.110 (0.558)	−1.195 (0.874)	−3.231 (0.756)	−0.525 (0.910)	−1.226 (0.599)
Household income ($000)	0.014 (0.016)	0.094 (0.015)	−0.006 (0.024)	0.077 (0.021)	0.002 (0.023)	0.044 (0.015)
No female head of household	−41.118 (9.433)	−32.854 (8.987)	−43.188 (14.805)	−38.610 (12.812)	−36.211 (13.938)	−28.823 (9.177)
Age female	−1.247 (0.361)	−1.713 (0.344)	−1.172 (0.569)	−1.794 (0.493)	−0.972 (0.531)	−1.059 (0.350)
Age female squared	0.010 (0.003)	0.020 (0.003)	0.008 (0.005)	0.020 (0.005)	0.009 (0.005)	0.011 (0.003)
No male head of household	11.512 (9.730)	−33.063 (9.270)	30.465 (15.157)	−5.164 (13.117)	17.157 (14.169)	−3.175 (9.328)
Age male	−0.395 (0.382)	−1.342 (0.364)	0.092 (0.597)	−0.158 (0.517)	−0.032 (0.545)	−0.204 (0.359)
Age male squared	0.005 (0.004)	0.012 (0.004)	0.001 (0.006)	0.001 (0.005)	0.001 (0.005)	0.001 (0.003)
Number of kids	3.423 (1.409)	1.835 (1.343)	4.898 (2.097)	3.325 (1.815)	−0.633 (2.164)	−0.847 (1.425)
Number of little kids	−0.808 (2.060)	3.609 (1.962)	−0.119 (3.130)	5.046 (2.709)	3.423 (3.172)	2.112 (2.089)
Male employed	−0.585 (2.180)	−11.024 (2.077)	4.459 (3.157)	−6.299 (2.732)	−5.284 (3.822)	−2.155 (2.517)
Male fully employed	5.478 (2.078)	17.662 (1.980)	2.856 (3.026)	13.405 (2.619)	10.251 (3.70)	4.042 (2.436)
Female employed	5.256 (1.228)	1.014 (1.170)	6.485 (1.895)	1.257 (1.640)	0.588 (1.779)	−0.921 (1.172)
Female fully employed	−4.082 (1.213)	−3.285 (1.155)	−4.228 (1.881)	−2.490 (1.628)	−4.088 (1.733)	−2.149 (1.141)
Male education (category)	1.194 (0.443)	−1.318 (0.422)	1.558 (0.686)	−1.444 (0.593)	1.978 (0.636)	0.088 (0.419)
Female education (category)	−1.335 (0.487)	1.249 (0.464)	−1.520 (0.763)	0.749 (0.660)	−2.404 (0.688)	−0.498 (0.453)
Married (or widower)	4.787 (1.210)	1.896 (1.153)	3.830 (1.874)	−0.361 (1.621)	5.236 (1.723)	1.558 (1.134)
Nonwhite	−3.627 (1.540)	1.303 (1.468)	−8.499 (2.382)	−0.916 (2.062)	3.146 (2.238)	0.836 (1.473)
Hispanic	−3.453 (1.841)	−2.990 (1.754)	−4.949 (2.743)	−2.198 (2.374)	1.887 (2.868)	1.314 (1.888)
"15K" homescan household	−1.142 (1.377)	−2.475 (1.312)	−0.031 (2.046)	−1.083 (1.771)	−5.500 (2.144)	−3.328 (1.412)
Constant	286.150 (10.447)	295.098 (9.954)	274.680 (16.182)	254.138 (14.004)	268.066 (15.268)	287.018 (10.052)
R-squared	0.912	0.910	0.896	0.896	0.966	0.986
UPC fixed effects (number of UPCs)	yes (10,470)		yes (6,793)		yes (5,764)	
Number of observations	41,158		22,291		13,818	

The table reports results from regressions where price is the dependent variable. Each column reports results from a different regression, with standard errors in parentheses. The first two columns report results from regressions that use all the matched items in the matched large trips, with the first column using the Homescan price as the dependent variable and the second column using the Retailer price as the dependent variable. The two other pairs of columns repeat the same exercise for items that were sold on card member discount and those who were not (at the time of purchase). All regressions include UPC fixed effects

It is interesting to note that in almost all the cases when a coefficient is significant in one regression but not in the other, the retailer's data generate the significant estimate, while the Homescan data do not. In many cases the difference is also economically meaningful. For example, in the Homescan data the coefficient on race dummy variable is negative and significant, which implies that non-white consumers pay a lower price. On the other hand, in the retailer's data the coefficient is positive but not significant. A researcher using the Homescan data to study discrimination would probably reach different conclusions than one using the retailer's data to study the same question, using the very same set of shopping trips. Another example is in the impact of age on price paid. The Homescan data suggest a flatter impact of age, especially for males, than the retailer's data. Once again researchers using the data to study life cycle consumption might reach incorrect conclusions using the Homescan data.

As already noted in the previous section, there are two effects that cause the difference in the results. One is of pure recording errors, while the other arises from the way Nielsen imputes prices. We separate items that were purchased with a loyalty card discount and those without a discount to identify cases in which one type of recording error is more likely than the other (just as we did in generating Fig. 4). We then repeat the same regressions for these two different cases, separately. In general, the regression results are quite different for each of the regression pairs, but the differences are much more subtle in the case in which price imputation is likely to be an issue (when the item was associated with a loyalty card discount). For example, in the results for products with loyalty card discounts, the coefficient signs do not agree for eight out of the twenty coefficients. When a loyalty card discount is not available for the item and price imputation is unlikely to introduce errors, only two of the point estimates do not agree in sign.

The channels by which the coefficients are biased are quite different depending on the nature of the recording errors. Consider the case in which the recording errors are driven by Nielsen's price imputation, and focus on the race dummy variable. In this case, the regression using the Homescan data tells us that non-white households tend to buy at cheaper stores, i.e., stores where the average consumer in the store pays less for the same item. The regression using the retailer's data tells us that despite going to cheaper stores non-white panelists do not pay less on average. In contrast, the channel is different if none of the prices are imputed and the only difference is due to recording mistakes made by the panelist. Once again we use the race dummy variables as an example. The regression using the Homescan data tells us that non-white consumers report a lower price. On the other hand, the regression using the retailer's data suggests that they do not actually pay less, maybe even slightly more. Together these suggest that white consumers tend to over report prices relative to non-white consumers, not that they are likely to pay more.

In order to further study the effect of recording errors and to illustrate how the validation study can be used to fix them, Table 4 presents Homescan regressions in which we only focus on the age effect. That is, we use the

Table 4 Correcting for recording errors

Dependet variable Corrected (Sample)	Homescan price Not corrected (all) (1)	Retailer price NA (matched items) (2)	Homescan price Not corrected (matched items) (3)	Homescan price Corrected (all) (4)
Constant	282.089	237.215	249.356	274.398
	(0.487)	(1.108)	(1.026)	(0.592)
Female age 29	21.887	–2.665	7.127	7.436
or younger	(1.668)	(2.818)	(2.612)	(2.031)
Female age 30–34	14.095	–4.332	6.024	1.767
	(1.215)	(2.211)	(2.049)	(1.479)
Female age 35–39	12.617	–9.762	0.728	–3.625
	(0.927)	(2.619)	(2.427)	(1.128)
Female age 40–44	10.956	–13.430	–2.800	–6.309
	(0.802)	(1.596)	(1.479)	(0.977)
Female age 45–49	5.913	–10.483	0.705	–2.586
	(0.713)	(1.552)	(1.438)	(0.868)
Female age 50–54	15.873	–10.805	–0.700	–3.545
	(0.766)	(1.686)	(1.563)	(0.933)
Female age 55–64	12.123	–6.731	–1.588	0.996
	(0.669)	(1.425)	(1.320)	(0.815)
Female age 65 or older		——— Omitted	category ———	
Number of obs.	790,526	27,511	27,516	790,526

All columns report regression results where price is the dependent variable, with standard errors in parentheses. Column (1) reports results from regressions for the entire Homescan transactions in this market, columns (2) and (3) report the results for the matched transactions using the retailer and Homescan price, respectively, and column (4) reports results using the correction method (see text for details) to correct for the recording errors. All regressions use UPC fixed effects

regression in Eq. 11 with only the age variable (of the female head of the household). Column (1) uses all the Homescan observations in one market.[7] Note that these are all the observations in the 2004 data, not just the ones we matched. The excluded category are the elderly, 65 years or older. The results are qualitatively similar to the main finding of Aguiar and Hurst (2007): older consumers tend to pay less for the same products.

Columns (2) and (3) replicate the analysis using the matched sample. Column (2) presents the results using the retailer data and column (3) uses the matched Homescan transactions for the larger market. An important observation here is that the results of *either* column (2) or (3) are quite different from the results using the full Homescan sample in column (1). For this selected set of transactions, the pattern across ages is much flatter, and often reversed (younger individuals pay less, rather than more, than the elderly). It seems likely that the different results arise due to non-random selection into

[7] Our analysis so far used data from two metropolitan areas (see Appendix). Here we only use data from the larger metropolitan area, as a way to minimize confounding the results due to pricing differences between the two areas. Coincidentally, this is also the metropolitan area covered by the Homescan data used in Dube (2004) and Aguiar and Hurst (2007).

our matched sample. For example, in the larger Homescan sample the elderly are *more* likely to use coupons, while in our (matched) sample they are *less* likely to use them (not reported). It may not be totally surprising that the validation data is not representative of the larger Homescan data. We select on matched trips, which are associated with more "careful" individuals. The change in the results between the matched and overall Homescan sample may indicate that this selection is differential across age groups. For this reason we should probably be careful in drawing conclusions based on either columns (2) or (3) of Table 4. On the other hand, the difference between the results in columns (2) and (3) also suggests that we should be careful in drawing conclusion based on column (1) because the results can potentially be driven by recording errors.

In column (4) we present results that use the validation sample to correct for the recording error. We follow the procedure described in the previous section. We first use the validation sample to predict the "true" retail price as a function of the demographics (age dummies, in this case) and the Homescan reported price. We then use the full sample to impute predicted prices for each of the observations and regress this predicted price on age. The resulting coefficients significantly change compared to column (1), and the age pattern is different and economically less important. Loosely, the correction makes the estimated coefficients somewhere in between the original estimates (column (1)) and the true regression on the matched sample (column (2)).

We again note that an important assumption that makes this correction valid is that the conditional distribution of recording errors is the same in the validation sample and in the overall Homescan data (in that market). We note that although it seems likely that the matched sample is non-randomly selected, this by itself does not violate the assumption; the conditional distribution of the recording errors may be the same even if the unconditional distribution of prices is not. For this reason, we view this correction as a useful robustness test of existing estimates, rather than a recipe that should always be followed.

5.2 An illustrative application II: demand estimation

As a second application, we now illustrate the effect of measurement errors in the context of demand estimation. Here price is a right hand side variable, so even classical measurement errors would lead to bias. Because we view this exercise as illustrative, we favor making the comparison in the context of a simple and arguably more transparent model, rather than in a context of a full-blown demand system.[8]

[8]We note that our exercise is somewhat similar in spirit to the exercise reported by Gupta et al. (1996) who compare demand elasticities estimated from consumer-level data to those estimated from store-level data. Unlike them, however, we use the same set of transactions, so we can focus on the measurement error; their results are likely driven by selection issues: consumers in the panel might not represent the population of shoppers in the store.

The basic regression we estimate in this exercise is

$$q_{ik} = \alpha_k + \beta_k \log p_{ik} + u_{ik} \tag{12}$$

where i is a household, k is a specific product (distinct UPC), p is the unit price paid for this product, and q is the number of units bought. The α_k's are a set of UPC fixed effects, and β_k's are a set of product-specific semi-elasticities of demand.

We first estimate Eq. 12, using OLS, assuming the same semi-elasticities across all products (that is, assuming $\beta_k = \beta$ for all k). The sample consists of the entire set of matched transactions ($N = 41{,}158$ as in the first column of Table 3). Although the majority of transactions in our data have quantity (of a given product in a given trip) equal to 1, in 25% of the transactions multiple units of the same product are bought. Using the Homescan data, we obtain an estimate of the price semi-elasticity of −0.217 with a standard error of 0.018. In contrast, just as in the previous application, when we use the same set of transactions, but use the price and quantity data as recorded by the retailer, we obtain a much greater (in absolute value) estimate of −0.503 with a standard error of 0.020. The direction of the difference between the two estimates is consistent with measurement errors in the Homescan price variable, which would bias the coefficient estimate toward zero. We note that repeating the same exercise using the Homescan reported price and the retailer's reported quantity results in estimates that are almost the same as those that use both price and quantity from Homescan. This is consistent with our earlier report that the quantity variable is much less likely to significantly affect results, and that the measurement errors in price are those driving the differences.

To investigate further, and to address a reasonable concern that the average elasticity may be of less interest compared to the product-specific elasticity, we estimate Eq. 12 for each product separately. We focus on those products that are most frequently transacted. Specifically, we select the 28 products (UPCs) that are transacted at least 50 times in our matched sample. Twelve of these UPCs are different sizes and brands of milk, 6 of them are pre-packaged produce (e.g., carrots and strawberries), and 10 are some other commonly purchased items including carbonated soft drinks, sugar, and hotdog buns. Overall, this sample includes 2,570 matched transactions.[9] We then report the results of the product-specific semi elasticities in Fig. 5. Each point in the figure represents the semi-elasticity estimates for a given product; the estimates obtained from Homescan data are on the horizontal axis, and those obtained from the retailer data (using the same set of matched transactions) are on the vertical axis. The dashed line is the 45-degree line, so a point below it represents a case in which the retailer data give rise to estimates of greater price sensitivity than the Homescan data. Overall, 23 out of the 28

[9]The average semi-elasticity in this selected sample is lower than that reported earlier, but the difference between the data sets is similar. In this selected sample, we estimate semi-elasticities of −0.128 (0.022) and −0.378 (0.039) using the Homescan and the retailer data, respectively.

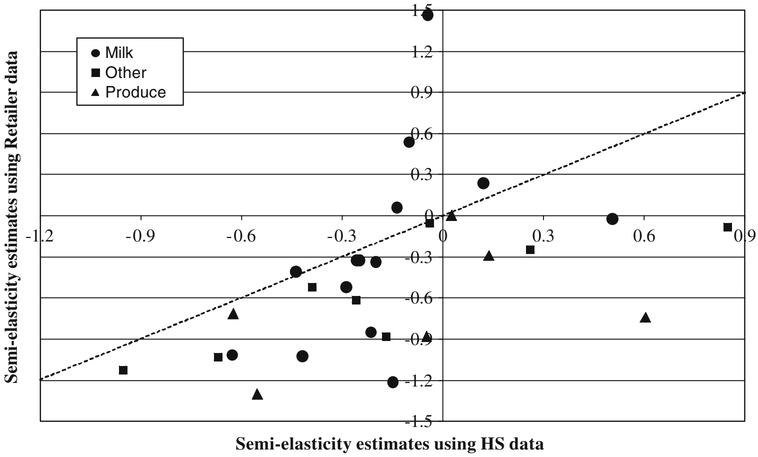

Fig. 5 Comparison of estimates of demand semi-elasticities. This figure presents point estimates from Eq. 12 for 28 distinct products. Each point in the graph represents a pair of estimates for the same product, one from Homescan data (horizontal axis) and the other from the Retailer data (vertical axis). The dashed line is the 45-degree line, and if the two data sets were identical all points should have been on this line. The fact that 23 out of the 28 points are below the line suggests that the Homescan estimates imply, overall, lower price sensitivity vis-a-vis the estimates obtained from the Retailer data

estimates lie below the line, and many of them are significantly below the line. This is consistent with the more aggregate results, and it is suggestive that measurement errors of price lead to lower estimated elasticities in the Homescan data.

6 Concluding comments

This paper presents a validation study of the Nielsen Homescan data. We were able to observe the same transaction twice: as it was recorded by the retailer, and as recorded by the Homescan panelist. By comparing the two data sources we can document reporting discrepancies. While through most of the paper we treat the retailer's data as the "truth" and attribute any differences between the data sets to "errors" or "mistakes" in the Homescan data, we should emphasize that to the extent that there are recording errors in the retailer's data, these words should be interpreted accordingly.

We described the magnitude of recording errors along several dimensions (trips, items, price, and quantity), and then demonstrated how the validation study could be used to correct bias in estimates caused by these recording errors. We think that our work has two distinct implications. First, it may provide guidance to Nielsen as to where and how to improve data collection and reporting. Second, it provides guidance to users of Homescan as to how to correct estimates for possible recording errors. We discuss each in turn.

We find that the price variable is the variable most poorly recorded in the Homescan data. This is due, at least in part, to the way Nielsen imputes store-level prices when available. There are several good reasons to provide researchers the store level prices instead of the prices reported by the panelists. However, given our findings, it seems important to also provide an indicator that an imputed price is used. Ideally, users of Homescan would have information on both the imputed store-level price and the price reported by the household. This information is not currently collected by Nielsen, but collecting this information, at least on an experimental basis, would allow for additional analysis of the magnitude of this discrepancy. One situation in which the store-level data could be very useful would be to help identify purchases on sale. A sale could then be defined as any situation in which the price reported by the consumer is less than the non-sale price reported by the store.

Nielsen could also improve the quality of the data by requiring the panelists to send in their receipts. The reported data could then be compared to these receipts. We are aware of at least one other consumer panel level data set that uses this procedure. Random sampling of the receipts will both make the panelists more careful and allow for quality control. As we find that certain households are more mistake-prone along all the dimensions we analyze, such random sampling may be used to design better sampling weights, or even to drop some of the less accurate panelists. The final analysis of the data can be improved, and bias potentially removed, by constructing a reliability index and weighting observations accordingly. Given the current data available in Homescan, such an index might be hard to construct. But future data collection can be done with this goal in mind.

To users of the Homescan data (in its current form), our work provides a way to correct for measurement errors. In particular, we discussed how one could adjust parameter estimates to account for recording errors, and demonstrated how this works in one simple application. A sufficient statistic to almost any such adjustment is knowledge (through the validation study) of the distribution of the error conditional on variables observed in the primary data. To facilitate such corrections, we posted this distribution on our web pages. We hope that researchers using Homescan will use this distribution and one of the methods that correct for measurement errors to run plausible robustness checks of their results. As we emphasize throughout, the maintained assumption behind the correction procedure is that the conditional distribution of the measurement error is the same in the validation sample and the primary data. While we think that this is often a plausible assumption, researchers who use our posted distribution to adjust their estimates should evaluate the plausibility of this assumption in their particular context.

Appendix: Detailed description of the data construction

As mentioned in the text and sketched in Fig. 1, our data construction process involved two distinct steps. Below we describe each step in turn.

First step In principle, we could have asked the retailer to supply information on any of its stores visited at any point by a Homescan panelist. However, since generating the data involved some effort for the retailer we had to limit our data request, in the first step, to roughly fifteen hundred store-day transaction-level records.

We therefore proceeded as follows. First, we restricted the data set to two metropolitan areas, in which the retailer has high market share. This left us with 265 different retailer stores (147 in one area, and 118 in the other). Since we identify the store by the zip code of its location, we restricted attention to retailer stores that are the only retailer stores in the same zip code. This eliminated 76 stores (29%), and left us with 189 stores (101 in one area, 88 in the other). We then searched the Homescan data for shopping trips at these stores, with the additional conditions that: (i) the trip includes purchase of at least 5 distinct UPCs (to make a match easier); (ii) the trip occurred after February 15, 2004 (to guarantee that the retailer, who deletes transaction-level data older than two years, still had these data at the time we put in the data request); and (iii) the household shops at the retailer stores (according to Homescan) more than 20% and less than 80% of its trips. Our initial goal in generating the data was to study store choice; hence, we wanted consumers who visited the retailer's frequently, but not always. These trips were made by 342 distinct households in the Homescan data. For 240 of these households, we randomly selected a single trip for each of them. For the remaining 102 households, which included households with at least 10, and not more than 20, reported trips in Homescan data, we selected all their trips. We then requested from the retailer the full transaction records for the store-days that matched these 1,779 trips. Since 74 of these trips were to the same store on the same date, we expected to get 1,705 store-day transaction-level records.

We eventually got 1,603 of these 1,705 requested store-days (1,247 in the first area, 356 in the other), which account for 4,080,770 shopping trips. The missing stores were mostly due to random coding errors when generating the data. The retailer had little idea how we were going to match the data and had no way to systematically impact our results by dropping data. They include 122 distinct stores (74 in the first area, and 48 in the other). These 1,603 store-days are associated with 1,675 trips from the sample of 1,779 shopping trips described above. However, since the retailer enjoys high market share in both areas, it is not surprising that the 1,603 store-day transaction-level data records we obtained are associated with additional 904 trips in Homescan. These additional trips happened when two households in the Homescan panel visited the same store on the same day, which is somewhat likely since the market share of the retailer is high in the markets we studied. Given the way we constructed the sample, however, many of these additional trips include a small number of items, or households that rarely shop at the retailer's stores.

Second step After obtaining the data from the first step, we developed a simple algorithm to find likely matches between trips in the Homescan data with trips in the retailer's data. These likely matches were only used to speed up the data construction process (as described in the text, the data analysis in the paper uses a more systematic matching procedure.) The algorithm used the first five UPCs in the Homescan trip, and declared a match if at least three of these five were found in a given trip in the retailer's data. We used this algorithm with the data we obtained in the first step and found 1,372 likely matches that, according to Homescan, are associated with 293 distinct households. Of these households, 166 were associated with more than one likely match, and 105 with four or more.

We then asked the retailer to use the loyalty card used in these 1,372 shopping trips and to provide us with all the transactions available for the households associated with these cards (in the retailer's data during the year 2004). Only two of the requested trips were not associated with loyalty cards. For the rest, we obtained all the transactions associated with the same loyalty card, and additional transactions that are associated with loyalty cards used by the same household, as classified by the retailer. Since associating multiple cards with the same household may not be perfect, in the analysis we experimented with both the card-level and the household-level matching.

In this step we obtained a total of 40,036 shopping trips from the retailer. These 40,036 trips are associated with 384 distinct stores (139 in the first area, 109 in the second, and 136 in other areas), with 682 distinct loyalty cards (472 in the first area, 203 in the second, and 7 in other areas), and with 529 distinct households, according to the retailer's definition (380 in the first area, 140 in the other). Finally, the 40,036 trips are associated with 34,316 unique store-date-loyalty card combinations, 33,744 unique store-data-household combinations (using the retailer's definition of a household), and 27,746 unique store-date-household combinations, using the Homescan definition. Of these trips, 3,884 (9.7%) occurred in a store-day already appearing in the data we obtained earlier, and therefore are one of the 4,080,770 trips obtained in the first step. Recall that the algorithm we used to request these data was geared to find likely matches, and therefore may have also found wrong matches. This is one reason that the number of households we intended to match (291, the original 293 minus two that had no associated loyalty cards) is less than the number of households associated with these trips. A second reason may be multiple cards used by the same household that are not linked to each other by the retailer.

Summary To summarize, we have two different types of data from the retailer. The first data set includes full transaction record of 1,603 distinct store-days. In the data set trips are not associated with a loyalty card. The second data set includes 40,036 trips, which are associated with particular loyalty cards and households. 3,884 of these trips overlap and appear in both data sets. The first data set is designed to match multiple transactions of 102 households in the Homescan data, and isolated transactions of other households. The second data set is designed to match all transactions of almost 300 households.

References

Aguiar, M., & Hurst, E. (2007). Life-cycle prices and production. *American Economic Review, 97*(5), 1533–1559.

Ashenfelter, O., & Krueger, A. B. (1994). Estimates of the economic returns to schooling from a new sample of twins. *American Economic Review, 84*(5), 1157–1173.

Bound, J., Brown, C. C., & Mathiowetz, N. (2001). Measurement error in survey data. In E. E. Learner & J. J. Heckman (Eds.), *Handbook of econometrics* (pp. 3705–3843). New York: North Holland.

Bound, J., & Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics, 9*(1), 1–24.

Broda, C., & Weinstein, D. E. (2009). Product creation and destruction: Evidence and price implications. *American Economic Review*, in press.

Broda, C., & Weinstein, D. E. (2008). Understanding international price differences using barcode data. NBER Working Paper No. 14017.

Chen, X., Hong, H., & Tamer, E. (2005). Measurement error models with auxiliary data. *Review of Economic Studies, 72*(2), 343–366.

Dube, J-P. (2004). Multiple discreteness and product differentiation: Demand for carbonated soft drinks. *Marketing Science, 23*(1), 66–81.

Einav, L., Leibtag, E., & Nevo, A. (2008). *On the accuracy of Nielsen Homescan data.* USDA Economics Research Report Number 69.

Gupta, S., Chintagunta, P., Kaul, A., & Wittink, D. R. (1996). Do household scanner data provide representative inferences from brand choices: A comparison with store data. *Journal of Marketing Research, 33*(4), 383–398.

Hausman, J., & Leibtag, E. (2007). Consumer benefits from increased competition in shopping outlets: Measuring the effect of Wal-Mart. *Journal of Applied Econometrics, 22*(7), 1157–1177.

Katz, M. (2007). *Estimating supermarket choice using moment inequalities.* Ph.D. Dissertation, Harvard University.