Introduction

Nielsen Homescan data provide rich information about household purchasing patterns that allows researchers to study questions that cannot be addressed using other forms of data. For example, Homescan data cover purchases at retailers such as Wal-Mart and Whole Foods that traditionally do not cooperate with scanner data collection companies. In addition, due to their national coverage, Homescan data provide wide variation in household location and demographics, in contrast to other retail research panels, in which most households are from a small number of markets with relatively limited variation in demographics.

[1]See Aguiar and Hurst (2007), Broda and Weinstein (2007), and Hausman and Leibtag (2007) for academic research that relied heavily on Nielsen Homescan data

However, questions have been raised about the credibility of the Homescan data since the data are self-recorded and the recording process is time-consuming. One concern is potential sample selection. Given the time commitment, the households who agree to participate in the sample might not be representative of the population of interest. A second concern is that the households who agree to participate in the sample might record their purchases incorrectly.

In this study, we used a unique data set from a single retailer to examine the second concern. We constructed a data set to allow the matching of records from the Nielsen Homescan data with detailed transaction-level data from the retailer. Thus, we were able to observe the same transaction twice—as it was recorded by the retailer, just before the items left the store, and as it was recorded by the Homescan panelist, just after the items reached the house. By comparing the two data sources, we are able to identify three types of potential inaccuracies in the Homescan data:

- if the household did not report a trip to the retailer or misrecorded the trip information (store and date)
- within a trip, if the household did not record, or misrecorded, the product (universal product code, or UPC) information
- for a given product, document misreporting of the price, quantity, and deal information.

This study has multiple purposes. First, we documented the accuracy of Homescan data, by describing the magnitude of mistakes for each of the aforementioned potential recording errors. Second, we investigated whether and how errors are correlated with household or trip characteristics, which would be suggestive of which type of analysis would be more sensitive to such errors, and how. For example, we ask whether a correlation between a price "paid" and demographics could be driven by systematic measurement errors. Third, we plan to use the results of this analysis to suggest adjustments to the use of the data that would make analysis less prone to recording errors[2].

[2]More generally, the study offers an opportunity to cross validate self-reported data. This has been done for other data sets. Bound et al. (2001) surveyed similar work on validation studies, primarily in the context of the PSID (Panel Study of Income Dynamics).

Before looking at the data in more detail, it is important to clarify terminology. First, the retailer's data are treated as the "truth," allowing any differences between the data sets to be attributed to "errors" or "mistakes." 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. The second terminology issue relates to what is meant by “errors” or “misrecording” in the Homescan data. This issue could be driven by various mechanisms: recording errors by the Homescan panelists themselves, misunderstanding of instructions or errors that are generated due to the way Nielsen puts together its data. This latter case seems most important for the price variable, but overall, the use of “errors,” “mistakes,” or “misrecording” means any of the possible mechanisms.