Conclusions

This analysis suggests that the Homescan data contain recording errors in several dimensions, but that the overall accuracy of self-reported data by Homescan panelists seems to be in line with many other surveys of this type. Our research fits into a broader literature of validation studies that have been conducted in the economics literature since responses to surveys and self-reported data are the foundation of many data sets used by economic researchers and policymakers. For example, the Panel Study of Income Dynamics (PSID), the Current Population Survey (CPS), and the Consumer Expenditure Survey (CEX) all include self-reported data and are used heavily by economists. One concern in using self-reported data is that information is recorded with error, and that the error is systematically related to the characteristics of the respondents or to the variables being recorded. To study the magnitude of measurement error and to document the distribution of the error, an empirical literature has emerged that compares the self-reported sample to a validation study.[17] While most of the literature has focused on data sets that record labor market decisions and outcomes, the Homescan data focuses on food purchase decisions. We compared the recording errors we find here to errors in these 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.

[17]Bound at el. (2001) provide a detailed review of this literature.

Having the unique opportunity to cross-validate a sub-sample of the Homescan data with retailer data allowed us to identify data misrecordings that would usually go undetected in most data sets and surveys. The most concerning issue we find relates to the way that prices are recorded by Nielsen for stores from which Nielsen uses its store-level data as an estimate of what households actually paid. This poses additional challenges when those stores have multiple possible prices in a given time period due to loyalty card or other shopper-specific price promotions.

We now offer some thoughts on what could be done to improve the data. Since prices are the variable most poorly recorded due, at least in part, to the way Nielsen imputes store-level prices, it seems that, at the very least, the data should include an indicator that an imputed price is used for a given shopping trip. Ideally, it would be best to know 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. There is still good reason to use store-level prices, when available as that additional information can be used, for example, to better identify purchases on deal. A deal can then be defined as any situation in which the price reported by the consumer is less than the store non-deal price reported by the store.

Nielsen could also probably improve the quality of the data by requiring the panelists to send in their receipts. The reported data could then be compared to those receipts (at least one other consumer panel-level data that uses this procedure). Random sampling of the receipts will both make the panelists more careful, and would also allow for quality control. As we find that certain households are more prone to mistakes along all the dimensions we analyzed, such random sampling may be used to design better sampling

weights, or even to drop out of the sample “negligent” panelists. The final analysis of the data can be improved, and bias potentially removed, by constructing a reliability index for the observations and weighting them 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.