Exploring Novel Optical Character Recognition (OCR) as a Method of Addressing Underreporting in Household Food Acquisition Surveys

Adam Kaderabek

Introduction

In 2012, the United States Department of Agriculture (USDA) Economic Research Service (ERS) and Food and Nutrition Service (FNS) co-sponsored the National Household Food Acquisition and Purchase Survey (FoodAPS-1) to fill a critical knowledge gap while informing decisions related to state and national initiatives on obesity, hunger, and nutritional assistance programs.[1] Prior to FoodAPS-1, national surveys on food and nutrition (e.g., National Health and Nutrition Examination Survey (NHANES)) prioritized estimating levels of food security/insecurity by focusing on behaviors indicative of food insecurity, for example, whether the respondent or household members were receiving assistance from a supplemental program such as the Supplemental Nutritional Assistance Program (SNAP) during a specified reference period.[2] In contrast to the emphasis on household food security, FoodAPS-1 was specifically interested in enumerating the number and type of food items procured by each household.

In order to collect item-level data related to food acquisition, FoodAPS-1 employed a seven-day food diary and asked participants to report all food that they procured each day. Diary surveys employ an extended reporting period which provides respondents opportunities to report frequent (and often less memorable) events as they occur and have been in use for over 100 years as a method of data collection (e.g., Bevans, 1913). A diary-based interview, conducted as part of the Consumer Expenditure Survey (CES), has been used to estimate household item-level food purchases.[3] Unlike FoodAPS-1, the CES – Diary Interview Survey focuses exclusively on food *expenditures* (Bureau of

[1] https://www.ers.usda.gov/foodaps

[2] (FSD170N), "[In the last 12 months], how many people in your household were authorized to receive Food Stamps?" (NHANES Food Security Section, published 2004)

[3] https://www.bls.gov/cex/csxsurveyforms.htm#diary

Labor Statistics, 2012). Additionally, the CES diary only collects data from a single respondent who reports on behalf of the entire household. FoodAPS-1 was uniquely different in that it asked *all* household members to report *any* food acquisition that occurred including garden produce, school meals, or food pantry contributions in addition to grocery stores and other food-for-purchase establishments.[4] By surveying all household members in this way FoodAPS-1 aimed to gather a complete picture of household food quantities and expenditures as well as consumption patterns for all household members.

Although diary surveys are an effective mode of data collection for such a task, they are also known to be burdensome and the level of burden in a diary survey can impact data quality and reporting behaviors (Maitland and Li 2012, Dillman and House 2013, Fricker et al. 2015, Hu et al. 2017). Respondents may cease to participate as data collection proceeds, resulting in item-nonresponse and attrition bias (Hu et al. 2017), or they may reduce their engagement with the survey leading to measurement error in reporting (e.g., omission/commission) (Stone et al. 1991, Shields and To 2005). Therefore, the adverse influence of survey burden on data quality in diary surveys can facilitate two forms of error associated with underreporting. Throughout this paper we will refer to these forms of underreporting error as a) underreporting due to nonresponse/item-nonresponse underreporting and b) underreporting due to measurement error/measurement error underreporting.

As a diary survey, FoodAPS-1 was not immune to underreporting (Maitland and Li 2016, Petraglia et al. 2016, Yan and Maitland 2016, Hu et al. 2017, Hu et al. 2020); however, identifying underreporting is less ambiguous than qualifying the cause and quantifying the impact. When available, external

[4] All adults and children ages 12+ who lived in the household were asked to complete a food diary. Adults were asked to report on behalf of children under 12 years old.

benchmarks can be used to identify the presence of underreporting but differences in survey designs severely limit the comparability of estimates between surveys (Fricker et al. 2015, Johnson and Li 2009, Li et al. 2010). FoodAPS-1 is unique in its design so there are currently no perfectly comparable benchmark surveys, however Clay et al. (2016) examined FoodAPS-1 data quality by comparing reported food expenditures to estimates from NHANES, CES, and the Information Resources, Inc (IRI) Consumer Network Panel. They found a general trend in which FoodAPS-1 estimates of household food expenditures were consistently lower than those found in NHANES but higher than CES and IRI estimates. Additionally, they found FoodAPS-1 estimates of food insecurity to be higher than those from the Current Population Survey (CPS) and the National Health Interview Survey (NHIS) (Clay et al., 2016).

In situations where no external comparison is available it can be possible to investigate underreporting through internal comparisons, i.e., using data from prior rounds of a survey to gain insight into reporting patterns over time, however it may remain unclear whether decreasing estimates for any given variable relate to underreporting errors or an actual change in behavioral patterns of respondents (Fricker et al. 2015). Here again, FoodAPS-1 was unique in being the first study of its kind. However, through day-to-day comparisons it has been possible to observe the effect of prolonged burden on reporting behavior and several studies have observed underreporting as a result of item-nonresponse and attrition in FoodAPS-1 (Maitland and Li 2016, Petraglia et al. 2016, Yan and Maitland 2016, Hu et al. 2017, Hu et al. 2020).

Perhaps the most promising, albeit the most intensive, method for validating underreporting in surveys of expenditure/consumption is through the use of record checks such as financial statements, bills, administrative records, and/or point-of-sale receipts. Receipts, as a record of food expenditures, are a particularly robust source of information wherein all expenses are typically identified in a line-item format with a clear description, quantity, and corresponding cost. Receipts conveniently identify the retailer, time/date of purchase, and the taxes and total cost for the purchase. Furthermore, existing research has shown respondents are willing to provide receipts as a record of food expenditure and receipts are useful for validating data quality (Rankin et al. 1998, Ransley et al. 2001, French et al. 2008).

Receipts were not collected as part of the FoodAPS-1 data collection; however, with the success of FoodAPS-1 the USDA has made plans to conduct FoodAPS-2. In preparation, an Alternative Data Collection Method (ADCM) pilot study was conducted to investigate ways to mitigate survey burden and improve data quality. As part of the ADCM, respondents were also asked if they would submit their food purchase receipts along with their completed food diaries.

This paper presents the culmination of research into the ADCM receipts. Specifically, we seek to evaluate the potential of optical character recognition (OCR) to capture receipt data and translate it into a reliable analytical format without the need for extensive manual coding and data capture. This research follows examinations of the most frequented establishments, general receipt properties, standardized elements, and a manual validation of 200 food events sampled from the ADCM reports corresponding to a receipt. It is clear that without an automated method to capture and parse receipt data it is impossible to leverage the receipts to reduce survey burden or estimate underreporting as a function of measurement error or nonresponse. To date there has been only limited commercial technology available to scan and collect data from purchase receipts but advancements in machine learning and computer vision techniques have made powerful tools available to researchers interested in working with text-based data. This paper investigates the effectiveness of using open-source software and novel methods of OCR to capture receipt data in hopes of estimating underreporting resulting from either measurement or nonresponse errors in diary surveys of food expenditures.

Receipt Data

Alternative Data Collection Method (ADCM) Study

The ADCM Pilot sample was conducted in 2016. The sample was selected using an address-based sampling frame created from 12 Primary Sampling Units (PSUs) in nine states. The survey aimed to collect nationally representative data from 500 households, including 150 households participating in the Supplemental Nutrition Assistance Program (SNAP, formerly the Food Stamp Program). Data Collection was conducted in both English and Spanish. In total, ADCM respondents reported 4,906 food events and 1,598 events were reported as "having a receipt to upload" as indicated by the respondent. A food event is defined as any occurrence of the respondent procuring some amount of food from an observable source or location (similar to FoodAPS-1, the ADCM collected data about all sources of food acquisition). The ADCM provided thirteen options for reporting the event location. We focus our attention on five location categories specifically which include those establishments that regularly provide receipts for purchased items, i.e., grocery stores, restaurants/bars, convenience stores, club stores, and superstores/big box stores.[5] These "receipt expected" locations accounted for 3,109 ADCM events with 1,247 (40%) corresponding to an itemized and legible receipt. They also serve as the population of events for which ADCM receipt validation was possible.

[5] The locations remaining locations are not expected to provide transactional receipts and are therefore excluded from this analysis. They include: 1) School, daycare, before/after school care, 2) Work, 3) Vending machine, 4) Family or friend's place, 5) Farmer's market, 6) Food pantry, 7) Soup kitchen, 8) Other.

ADCM Validation Sample

An initial random sample of 100 Food-Away-from-Home (FAFH) and 100 Food-at-Home (FAH) events was selected from the 1,247 events. The accuracy of total cost, number of items reported, and item prices were key variables of interest. It is important to note that the receipts corresponding to the sampled events varied greatly in quality and multiple factors were found to influence the effectiveness of the receipts for data validation (e.g., missing/corrupted files, non-receipt images, non-itemized receipts, and images of insufficient quality to be legible). Any limitation to manual interpretation of receipt data will similarly prohibit automated interpretation and issues related to receipt quality should be kept in mind with either approach.

Within the 100 sampled FAFH events, 18 images were unlocatable, and 5 images were not of receipts. There were an additional 8 receipts which were not itemized and only indicated the total amount paid with no indication of subtotals or tax. We were able to manually validate the number of items and the total cost for 69 of the 100 FAFH events. The average number of items across these 69 receipts was 3.78 compared to an average of 3.55 based on the reported data which indicates bias arising from item-nonresponse underreporting. The average FAFH receipt total was $17.20 compared to an average reported event total of $16.05, indicating measurement error underreporting occurred to some extent as well. Only 28 of the FAFH reports (40.58%) included both accurate total expense and accurate item totals and 9 reports (13.04%) were errant in both the reported total expense and item count.

Of the 100 sampled FAH events, 9 receipt images were missing, 2 receipts were illegible, and there was 1 non-itemized receipt which appeared to be a misclassified FAFH event. This allowed for the manual validation of 88 FAH receipts in total. Both forms of underreporting errors were found to be present in the 88 FAH events, however in contrast to FAFH events, the prevalence of item-nonresponse underreporting was higher than measurement error underreporting. The average number of items per receipt for FAH receipts was 10.89 items compared to 9.03 for reported items and an average receipt total of $36.81 compared to $35.85 for reported events. Accuracy/error rates among FAH reports were marginally better than FAFH reports with 37 reports (42.05%) including accurate expense totals and item counts and only 8 reports (9.10%) were inaccurate across both dimensions.

ADCM OCR Sample

Data for our analysis of this novel OCR method came from a subset of the sampled ADCM receipts discussed above. The sampled receipts were reviewed for their completeness and potential for OCR processing. Following the review, 8 of the 69 FAFH receipts and 6 of the 88 FAH receipts were found to be unsuitable for OCR. The evaluation of quality was made based on the clarity of the image and consistency of the lighting. The intention of the ADCM was primarily to investigate whether respondents would be willing to submit receipts; as such, there was no formal methodology to establish standards for image resolution, file formats, and image composition prior to data collection. In order to bolster our available sample 7 of the non-itemized FAFH were included to identify the value of capturing partial information when available, additionally 4 of the FAH receipts were submitted using two images (due to length of receipt) which must be processed individually with OCR. The resulting OCR test sample includes 154 images corresponding to 150 unique food expenditures.

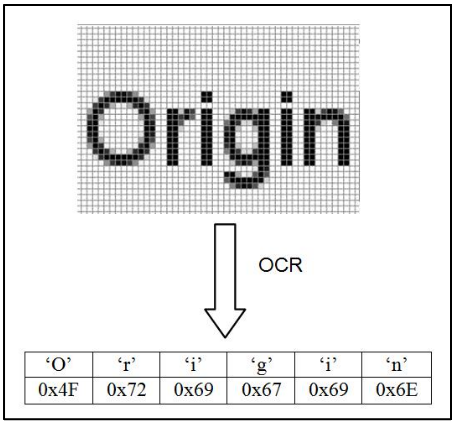
Table 1 presents a summary of the reporting errors associated with the event totals and reported item counts for FAFH and FAH events selected for the OCR examination. We also include an error summary for FAFH event tax (which was collected separately from the event total for FAFH events) as an additional indication of reporting errors. In whole, 66% of the sampled FAFH events and 67% of sampled FAH reports included some form of error. Among each error type, there were instances of both underreporting and overreporting; therefore, the table presents the absolute difference between reported (errant) data and receipt data. Although both over and underreporting contribute to biasing estimates, underreporting is generally more prevalent in expenditure research and has a greater impact on the accuracy of survey estimates; as such, we focus our attention on it here. However, it is important to recognize the effective capturing of receipt data could ultimately mitigate overreporting as well.

Table 1: Error Summary for ADCM Sampled Events

Error Type	Frequency	Absolute Difference		
Food-Away-from-Home (FAFH, N=68)		**Minimum**	**Mean**	**Maximum**
Reported Total Incorrect or Missing (n=67)[a]	10 events (15%)	$0.08	$7.65	$51.05
Reported Tax Incorrect or Missing (FAFH only) (n=60) [b]	4 events (7%)	$0.31	$1.59	$4.47
Incorrect Item Count or Missing (n=59) [c]	31 events (53%)	1.00	2.23	5.00
(a) one FAFH receipt did not include the portion containing the total and could not be verified against report. (b) in addition to (a), seven FAFH receipts where not itemized, i.e., no tax or subtotal. (c) one additional receipt for "6 Eat and Play Combos" was removed, report indicated 24 food items.				
Food-at-Home (FAH, N=82)		**Minimum**	**Mean**	**Maximum**
Reported Total Incorrect or Missing (n=81)[d]	13 events (16%)	$0.06	$6.94	$24.11
Reported Item Count Incorrect or Missing (n=82)	42 events (51%)	1.00	5.31	53.00
(d) one FAH receipt did not include the portion containing the total and could not be verified against report.				

Methods

Figure 1: Example of OCR Interpretation



'O'	'r'	'i'	'g'	'i'	'n'
0x4F	0x72	0x69	0x67	0x69	0x6E

As the name suggests, optical character recognition is a method by which the visual representations of text (e.g., characters in this sentence on a printed copy of this page) are recognized by a computer algorithm and interpreted based on the image geometry, specifically the distribution and density of image pixels. Figure 1 illustrates this process by displaying how the pixel orientations of the letters in the word "origin" are identified and interpreted as specific hexadecimal codes that represent the lexical form of the letter in a machine-readable format.[6] Motivation for exploring this OCR methodology come from a 2018 paper published in the International Journal of Scientific & Engineering Research titled, "OCR Engine to Extract Food-Items, Prices, Quantity, Units from Receipt Images, Heuristics Rules Based Approach" (Ullah et al. 2018).

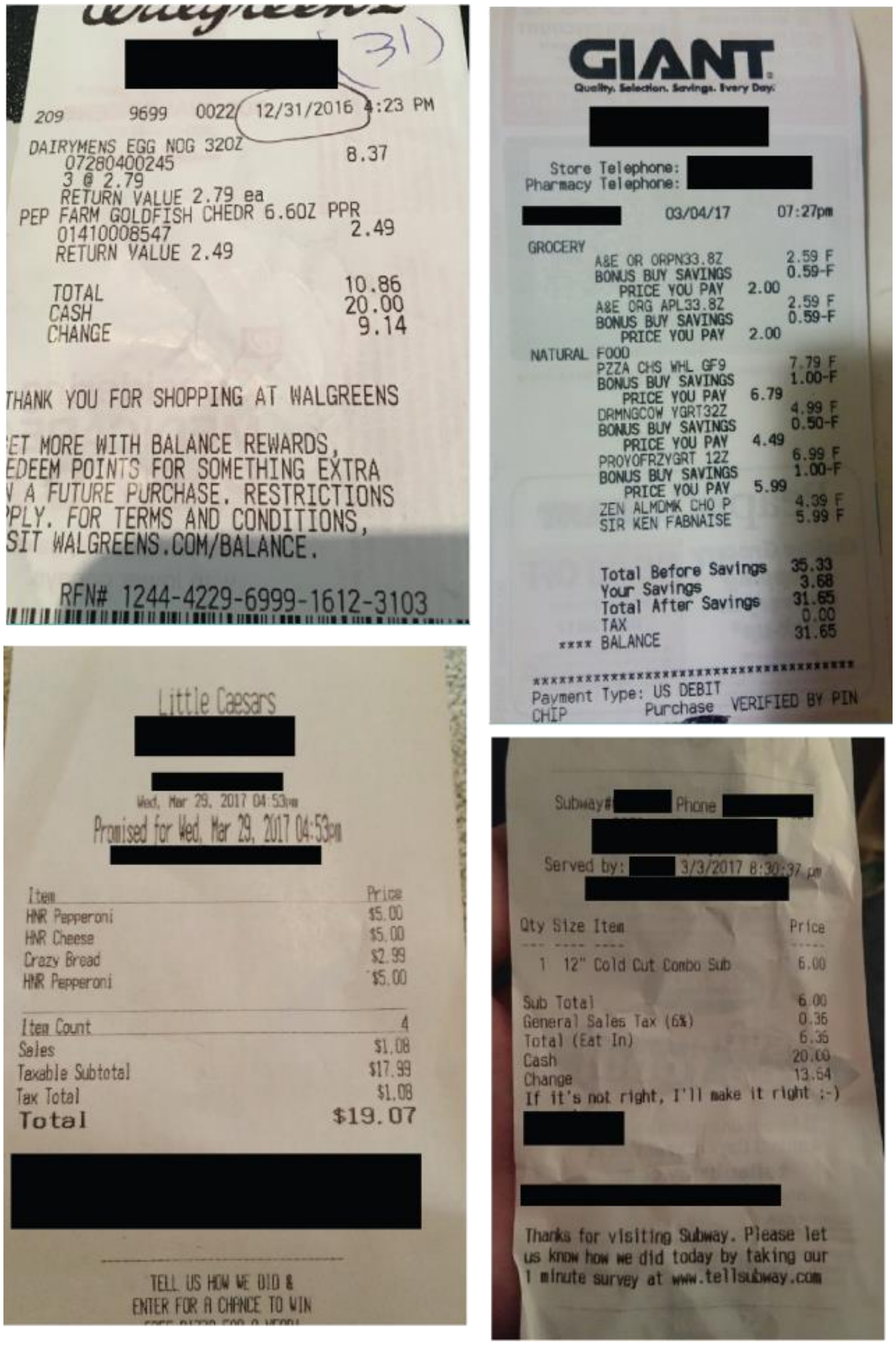
The manual validation of the sampled FAFH and FAH events discussed above will be used as a benchmark for the rates of underreporting in the sample. Using the open-source Tesseract OCR[7] engine, the ImageMagick[8] application and R[9], we then employ an image pre-processing protocol before the OCR processing of the receipts. Once text data has been captured, specified heuristics are employed to parse the data. The capture of text data is highly efficient under ideal conditions but the parsing of data into a useful analytical format is more complicated.

The selected events represent an array of companies including several of the most popular FAFH and FAH establishments as well as small grocers and restaurants. The receipts also represent differing levels of standardization in the listing of items, quantities, sizes, and prices. Figure 2 provides two FAFH and two FAH examples which are generally representative of the receipts selected for OCR processing. The requisite receipt elements to assess underreporting include the number of items listed on the receipt as defined by the total number of line-items (including quantity multipliers), the line-item price, and the total expense (less tax for FAFH and including tax for FAH expenditures). Generally speaking, the variation of receipt formats is greater among FAFH establishments and this

[6] Image created by AerodyneSr and made available by Wikimedia Commons under CC BY SA 4.0 https://creativecommons.org/licenses/by-sa/4.0/legalcode
[7] https://github.com/tesseract-ocr
[8] https://imagemagick.org
[9] https://www.R-project.org/

increased variation did ultimately require using different heuristics to parse the data relative to the FAH process.

Figure 2: Example Receipts



```
[redacted]
                                  31)
209        9699    0022   12/31/2016 4:23 PM
DAIRYMENS EGG NOG 32OZ
    07280400245                       8.37
    3 @ 2.79
    RETURN VALUE 2.79 ea
PEP FARM GOLDFISH CHEDR 6.6OZ PPR
    01410008547                       2.49
    RETURN VALUE 2.49

    TOTAL                            10.86
    CASH                             20.00
    CHANGE                            9.14

THANK YOU FOR SHOPPING AT WALGREENS

ET MORE WITH BALANCE REWARDS,
EDEEM POINTS FOR SOMETHING EXTRA
N A FUTURE PURCHASE. RESTRICTIONS
PPLY. FOR TERMS AND CONDITIONS,
SIT WALGREENS.COM/BALANCE.

      RFN# 1244-4229-6999-1612-3103
```

```
            Little Caesars
[redacted]
[redacted]
       Wed, Mar 29, 2017 04:53pm
  Promised for Wed, Mar 29, 2017 04:53pm
[redacted]

Item                                Price
HNR Pepperoni                       $5.00
HNR Cheese                          $5.00
Crazy Bread                         $2.99
HNR Pepperoni                       $5.00

Item Count                              4
Sales                               $1.08
Taxable Subtotal                   $17.99
Tax Total                           $1.08
Total                              $19.07

[redacted]

        TELL US HOW WE DID &
       ENTER FOR A CHANCE TO WIN
```

```
GIANT
Quality. Selection. Savings. Every Day.
[redacted]

    Store Telephone: [redacted]
 Pharmacy Telephone:

[redacted]        03/04/17        07:27pm

GROCERY
          A&E OR ORPN33.8Z                2.59 F
          BONUS BUY SAVINGS               0.59-F
             PRICE YOU PAY      2.00
          A&E ORG APL33.8Z                2.59 F
          BONUS BUY SAVINGS               0.59-F
             PRICE YOU PAY      2.00
NATURAL   FOOD
          PZZA CHS WHL GF9                7.79 F
          BONUS BUY SAVINGS               1.00-F
             PRICE YOU PAY      6.79
          DRMNGCOW YGRT32Z                4.99 F
          BONUS BUY SAVINGS               0.50-F
             PRICE YOU PAY      4.49
          PROYOFRZYGRT 12Z                6.99 F
          BONUS BUY SAVINGS               1.00-F
             PRICE YOU PAY      5.99
          ZEN ALMDMK CHO P                4.39 F
          SIR KEN FABNAISE                5.99 F

          Total Before Savings           35.33
          Your Savings                    3.68
          Total After Savings            31.65
          TAX                             0.00
     **** BALANCE                        31.65

****************************************
Payment Type: US DEBIT
CHIP         Purchase   VERIFIED BY PIN
```

```
    Subway#[redacted] Phone [redacted]
    [redacted]
  Served by: [redacted] 3/3/2017 8:30:37 pm
    [redacted]

Qty Size Item                        Price
--- ---- ----                         -----
 1  12" Cold Cut Combo Sub            6.00

Sub Total                             6.00
General Sales Tax (6%)                0.36
Total (Eat In)                        6.36
Cash                                 20.00
Change                               13.64
If it's not right, I'll make it right ;-)
[redacted]

[redacted]

Thanks for visiting Subway. Please let
us know how we did today by taking our
1 minute survey at www.tellsubway.com
```

Tesseract OCR

Tesseract is an open-source text recognition (OCR) engine, available under the Apache 2.0 license. Tesseract can be used directly via command line, or (for programmers) by using an API to extract printed text from images. It supports a wide variety of languages, including R and Python. Tesseract was originally developed by Hewlett Packard in 1985, and in 2005 Hewlett Packard released the code as open-source. Beginning in 2006 Tesseract development has been sponsored by Google (Google 2021). Interested readers can find information on existing applications of Tesseract through the GitHub repository, https://github.com/tesseract-ocr.

Tesseract has been adapted to work with multiple languages and alphabets. Its accuracy is dependent on the quality of the image as a composite of several properties. Of critical importance are image clarity and resolution which influence the level of noise present in the image. Resolution is defined by the image capture device, but the clarity is dependent on the level of focus and ambient lighting. Geometric properties of the image are also very important. Tesseract works by identifying pixels in relation to each other and associates the identified shapes with known characters. If the image is skewed or the perspective exaggerates the depth-of-field (e.g., positioning the top of the receipt further away from the camera than the bottom) then Tesseract will have difficulty accurately capturing the text from the pixel data.

ImageMagick

ImageMagick is free software delivered as a ready-to-run binary distribution or as source code that you may use, copy, modify, and distribute in both open and proprietary applications (ImageMagick Development Team, 2021). It is distributed under a derived Apache 2.0 license (https://imagemagick.org/index.php). Similar to Tesseract, ImageMagick has been adapted to work with multiple programming languages including R and Python. ImageMagick utilizes multiple computational threads to increase performance and can read, process, or write mega-, giga-, or tera-pixel image sizes.

ImageMagick enhances the performance of OCR by efficiently pre-processing images. Pre-processing for OCR generally includes cropping out irrelevant portions of the image, binarizing the image into black and white pixels only, adjusting image resolutions, and correcting issues with image skew and perspective. ImageMagick has additional capabilities to remove image noise, straighten wavy lines of text, mirroring and scaling of images, and many other transformations which are not necessary for the OCR process. The preprocessing steps can be augmented through a wide array of functions and parameters available within ImageMagick; however, for the purposes of this analysis the default parameters were used throughout each step.

Heuristics and Data Parsing

Although receipts are not completely standardized there are a number of properties that make receipt data highly recognizable. For example, in the United States, receipts read left to right, most commonly with the item description appearing first, possibly followed by an indication of quantity or size, and the item specific price directly to the right. Prices are generally aligned and are summed to provide subtotals, tax, and total cost below the listing of purchased items. There are often other elements that are regularly occurring such as a Universal Product Code (UPS), or loyalty-member discounts following sale items. Dollar signs are commonly associated with prices but not always, and within our ADCM receipts, prices are universally listed as a numeric value including two decimal places. Combinations of letters and numbers are seldom used outside of item unit-of-measure, e.g., 12 oz, 3 lb, etc. Developing a set of rules around these patters allows the use of regular expressions (commonly referred to as “regex”) and Boolean logic to identify pertinent elements of the receipt as well as removing captured text which is not informative.

For this analysis, we define a “line item” as a scanned line of text that remained after cleaning/parsing and was not indicated to be a discount, quantity line, subtotal, tax, or total. Additionally, we exclude from our count any line following an indicated subtotal, tax, or total. Although lines following the receipt subtotal, tax, or total are in excess of the desired results, there are no instances of a receipt containing data of interest (in this examination) beneath the actual receipt total and successfully identifying a receipt total is tantamount to identifying the “end” of the receipt.

The total indicator was programmed to identify regex patterns related to the receipt total in the string (e.g., “total”, “balance”, “amount due”). If a regex pattern was found with a corresponding value (representing the expense), then the total indicator was coded as 1, otherwise it was coded as 0. Tax was indicated using regular expressions targeting the word “tax” within a string including patterns that were highly likely to be representations of “tax”, for example “tak” or “ax” alone on a line. Subtotals, discounts, and lines denoting the quantity of an item where similarly denoted during parsing. Discounts were rare in the FAFH sample and were identified directly using a list of identified strings such as “EMP DISC”, “MANAGER MEAL”, or “DISCOUNT”. Discounts among FAH receipts were far more prevalent and varied in their descriptions but within the sample receipts they were universally indicated by a hyphen immediately before or after the price.

It is possible that any indicator could fail for either not recognizing the specified regular expressions, or by recognizing the expression but failing to identify an associated dollar value. Instances where the control expressions were recognized but dollar values were not present are coded separately to identify possible causes of technical failure. It is also important to note that differences in the original data collection protocols result in key differences in the use of indicators across event types. FAFH reports asked participants to report the total before tax (excluding tips and surcharges) and the amount of tax separately. FAH events requested the expense total including tax.

We include a general outline of the parsing algorithm here, although as noted above there were aspects of capturing FAFH and FAH data that required some customization by event type, e.g., the capturing of discounts, totals, and tax.

1. Import Image into Tesseract
2. Parse Tesseract output into separate lines
3. Convert all text to upper case
4. Replace commas (,) with period/point (.) (Commas are seldom used in receipts and periods are easily misinterpreted depending on image quality)
5. Remove special characters excluding ".", "@" and "-" (remaining special characters serve as control indicators)
6. Remove instances of single letters (Commonly found in FAH receipts but also frequently a result of poor text capture)
7. Remove any strings of 5 or more consecutive numbers
8. Extract discounts indicated by a hyphen preceding or following the price (for FAH only, discounts for FAFH were identified with custom dictionary)
9. Create discount line indicator to identify discount as "non-line item"
10. Create amount indicator for FAH lines containing "@" to indicate correspondence to a line item.
11. Remove whitespace from beginning and end of lines and replace multiple whitespaces with single space
12. Identify garbage lines of text using pre-defined regex dictionary and delete (e.g., "VALUED CUSTOMER", department headers such as "PRODUCE", or text including ".COM" indicating a website)
13. Extract prices
14. Create total indicator
15. Remove remaining garbage text within lines
16. Remove empty lines from data and export results

Comparative Analysis

In order to assess the performance of OCR as a method of data capture and validation, the results will be compared to the benchmark established by the manually coded review. First, we summarize the differences in the reported data, the OCR-extracted data, and receipt-provided total costs and item counts. Then we examine the correlation of the OCR results to the manually coded data and compare those results to the correlation between reported data and the receipts. If the OCR produces price and item information that is highly correlated to the manually coded information there will be support for OCR's ability to accurately capture the critical expenditure information which could potentially be used to reduce respondent burden. We also investigate the errors from OCR more generally in order to understand the influence of technical issues in processing. That is, if the internal OCR error rate is low there will be evidence supporting the efficacy of novel OCR methods to reliably capture non-standardized text data.

Results

Among the 154 images deemed suitable for OCR processing, all images were successfully scanned and parsed into a digital format, however not all results were informative of the true receipt data. Generally speaking, OCR processing took about 1 second per receipt to scan and parse. Some difficulty arose in comparing item counts within FAFH events because the condition of the receipts made it impossible to efficiently capture quantity indicators. Because of this the OCR results for number of items were compared to a count of all food related lines evident on the receipt instead of a combining the number of lines and corresponding quantity indicators.

Table 2 provides a summary comparison of the respondent reported data, the manual receipt validation, and the results of the OCR processing. We see, within the sampled events, the mean item count for the OCR results is slightly higher than the receipt data and without being able to compare the number of line items captured to the quantity of items reported we are unable to identify if the OCR results were more or less accurate than the respondent data. The average subtotal amount from the OCR results is also higher than the manually validated mean and the corresponding difference is over 2.5 times that of the respondent data. We do see that OCR was more effective in capturing the

FAFH tax although the comparison did require excluding two extreme outliers from the OCR results. Similar issues were found when looking at the FAH receipt totals where 4 totals were mis-captured and resulted in differences of over $100 from the true amount. These extreme deviations were found to be one of the more challenging OCR errors. When a person misreports an amount, it is usually an error of estimation and results are often close to what the truth should be; however, when OCR mis-captures an amount the result may be missing integers or including errant digits altogether. The most extreme example of this was the result of an event tax of $833.00 for a receipt indicating a total of $26.64 and a true tax of only $1.69. Interestingly enough for the same event, the OCR results successfully captured all items and corresponding prices accurately, the sum of which matched the receipt subtotal perfectly.

Table 2: Summary of FAFH Reports, Receipt Data, and OCR Results

FAFH Events (N=68)	Report Data	Manual Validation	OCR
Mean Item Count (OCR vs Receipt)	N/A	4.46	5.00
Mean Item Quantity (Report vs Receipt)	3.41	3.67	N/A
Mean Event Subtotal	$16.42	$17.36	$19.92
Mean Event Tax	$0.95	$1.18	$1.20*

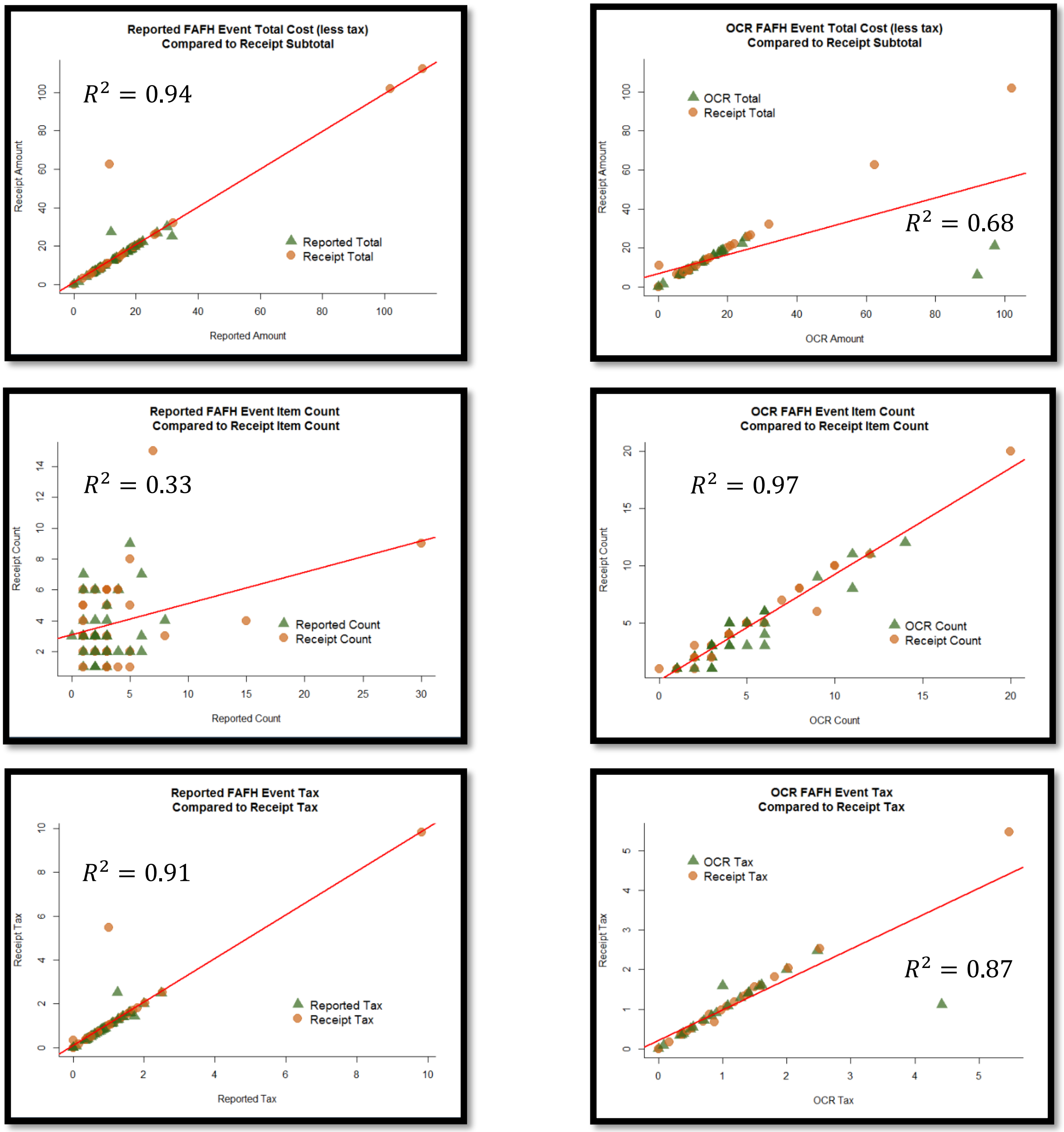
* Excluding two outliers with errant values of $80.00 & $833.00

FAH Events (N=82)	Report Data	Manual Validation	OCR
Mean Item Count	7.21	9.13	10.43
Mean Event Total**	$24.93	$25.47	$23.75

** 4 Largest receipt totals also resulted in the 4 largest OCR mis-captures of over $100 each are removed for comparison.

Figure 3 illustrates the correlations of the reported data to the true receipt values as well as the correlations of the OCR data to the receipt values. We see a weaker correlation for the OCR captured totals but the correlation for OCR item counts is much stronger than that of the reported item totals. When looking at the correlation of reported and OCR-captured tax we see the correlation for the OCR values was slightly lower but close to the correlation of the reported data to the truth.

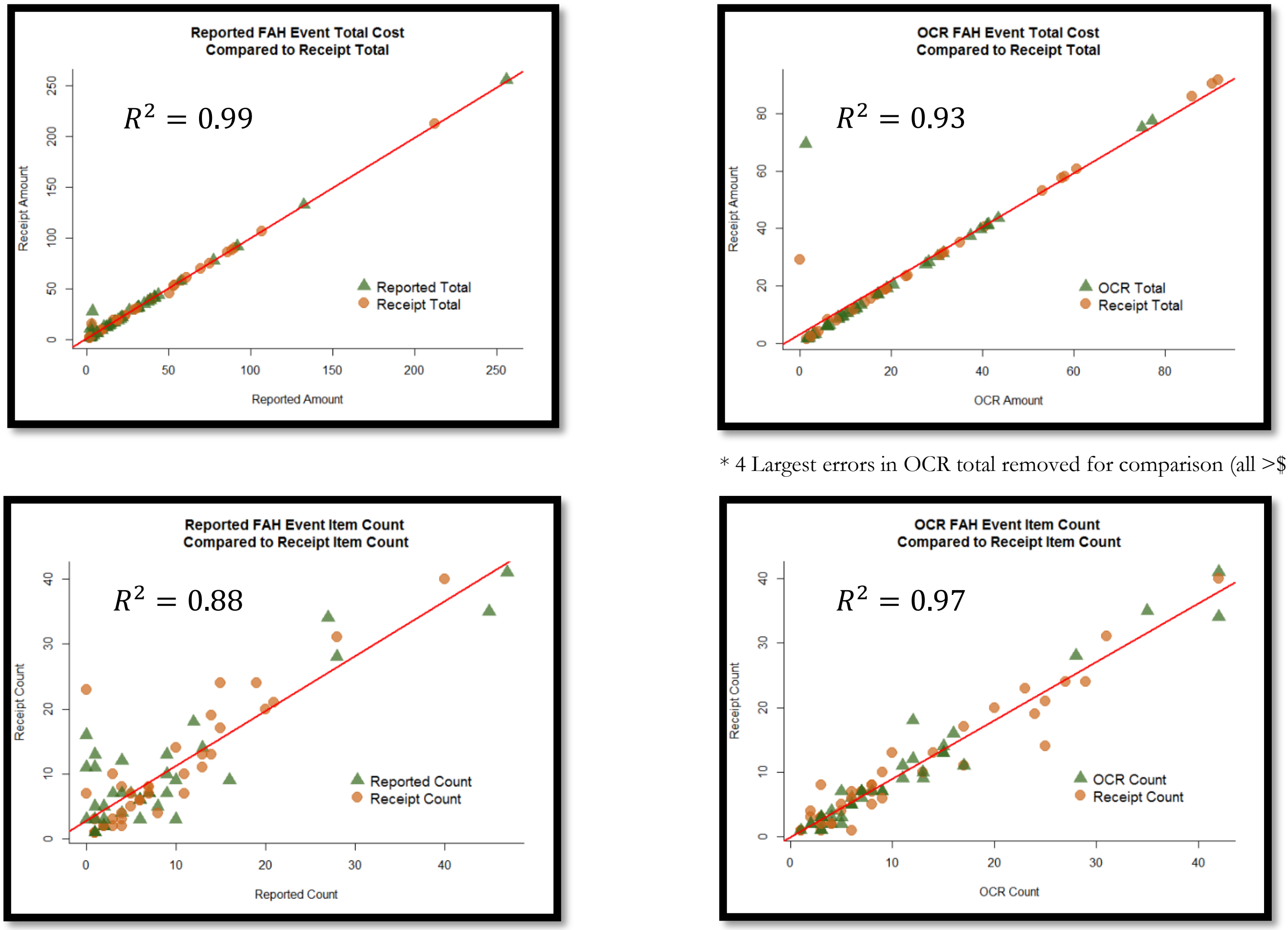
Figure 3: Comparison of FAFH Report to Receipt & OCR to Receipt Correlations



* Tax plot excludes outliers with errors of $80.00 & $833.00

Figure 4 illustrates the same comparisons for FAH events. Again, we see that the correlation of total cost to the truth is weaker for the OCR results, although it is much stronger than the correlation among FAFH OCR results. We also see that the correlation of the OCR item totals is stronger than that of the reported item counts.

Figure 4: Comparison of FAH Report to Receipt & OCR to Receipt Correlations



* 4 Largest errors in OCR total removed for comparison (all >$100).

The variation in the OCR plots and correlations indicates some of the difficulty alluded to earlier, that is, when OCR misinterprets data, the distortion is unlike the distortion caused by more traditional reporting errors. Table 3 articulates the internal aspects of the OCR processing for FAFH events to help understand where and why OCR was successful or unsuccessful. We see that the process was able to make effective indications of subtotals, tax and totals between 51% - 65% of the time depending on the indicator. We see OCR was most successful in capturing line-item counts (65% of events) and it also correctly indicated and captured subtotals in 65% of events.

Table 3: Performance Summary for FAFH OCR Scanned Receipts

Food-Away-from-Home Receipts (FAFH, N=68)		**Absolute Difference**		
	Frequency	**Minimum**	**Mean**	**Maximum**
OCR Indicates & Captures Total (n=65)[a]				
Indicator=1 & Total Correct (i.e., “true” receipt total captured)	33 events (51%)	N/A	N/A	N/A
Indicator=1 & Total Incorrect (i.e., total found but amount incorrect)	11 events (17%)	$0.02	$25.58	$250.07
Indicator=0 & Total Correct (i.e., regex failed to recognize indicator pattern)	7 events (11%)	N/A	N/A	N/A
Indicator=0 & Total Incorrect/Missing** (i.e., OCR results in error)	1 event (2%)	-	$8.00	-
	10 events (15%)	*$6.40*	*$15.88*	*$29.79*
Indicator Regex Found – Amount Missing** (i.e., OCR found “TOTAL” but not amount)	3 events (5%)	*$9.29*	*$18.34*	*$28.33*
OCR Indicates & Captures Tax (n=59)[b]				
Indicator= 1 & Tax Correct (i.e., “true” receipt tax captured)	34 events (58%)	N/A	N/A	N/A
Indicator=1 & Tax Incorrect (i.e., tax found but amount incorrect)	7 events (12%)	$0.06	$130.68	$831.31
Indicator=0 & Tax Correct (i.e., regex failed to recognize indicator pattern)	1 event (2%)	N/A	N/A	N/A
Indicator=0 & Tax Missing** (i.e., OCR results in error)	10 events (16%)	*$0.00*	*$1.74*	*$9.82*
Indicator Regex Found – Amount Missing** (i.e., OCR found “TAX” but not amount)	7 events (12%)	*$0.00*	*$0.49*	*$1.42*
Subtotal Indicator (n=55)[c]				
Indicator= 1 & Subtotal Correct (i.e., “true” receipt subtotal captured)	36 events (65%)	N/A	N/A	N/A
Indicator=1 & Subtotal Incorrect (i.e., subtotal found but amount incorrect)	7 events (13%)	$0.06	$25.21	$86.02
Indicator=0 & Subtotal Correct (i.e., regex failed to recognize indicator pattern)	1 event (2%)	N/A	N/A	N/A
Indicator=0 & Subtotal Missing** (i.e., OCR results in error)	5 events (10%)	*$6.99*	*$14.56*	*$27.27*
Indicator Regex Found – Amount Missing** (i.e., OCR found “SUBTOTAL” but not amount)	6 events (11%)	*$6.19*	*$28.30*	*$112.25*
Item Counts (n=65)*				
Item Count Correct	42 events (65%)	N/A	N/A	N/A
Item Count Incorrect	23 events (35%)	1.00	1.48	3.00

(a) one FAFH receipt did not include portion with total so indication would not be possible and two of the non-itemized receipts did not include a total line in their format.
(b) in addition to the partial receipt, seven FAFH receipts where not itemized, i.e., no tax or subtotal, and one receipt did not list tax.
(c) In addition to 9 FAFH events in (b), 4 additional receipts did not incorporate subtotals in the format.
* Line-item counts for OCR/Receipt comparisons had to be measured relative to lines of text as opposed to qualified food items and hence results for all but 3 of the non-itemized receipts were able to be compared.
** italicized values indicate “true” receipt values where corresponding OCR results are NA

Table 4 displays a similar breakdown of OCR performance within FAH events. We see that the receipt total was successfully captured in almost 75% of events although the indication of a total being present was not successful for 9 of the events, indicating the regex patterns were not comprehensive enough to recognize all total formats. We also see the extremity of mis-captured data with differences between the OCR results and the true value exceeding $200. In a positive light, we see that OCR was successful in narrowing the range of values for incorrect item totals. The range of errant reports included a maximum difference of 54 items (from Table 1) whereas OCR differences maxed out at 11 items.

Table 4: Performance Summary for FAH OCR Scanned Receipts

Food-at-Home (FAH, N=82)		**Minimum**	**Mean**	**Maximum**
OCR Indicates & Captures Total (n=81)[d]				
Indicator=1 & Total Correct (i.e., "true" receipt total captured)	51 events (63%)	N/A	N/A	N/A
Indicator=1 & Total Incorrect (i.e., total found but amount incorrect)	6 events (7%)	$0.50	$80.94	$214.95
Indicator=0 & Total Correct (i.e., regex failed to recognize indicator pattern)	9 events (11%)	N/A	N/A	N/A
Indicator=0 & Total Incorrect/Missing** (i.e., OCR results in error)	4 events (5%)	$0.14	$57.34	$200.00
	11 events (14%)	*$1.99*	*$26.70*	*$88.39*
Item Counts (n=82)				
Item Count Correct	32 events (39%)	N/A	N/A	N/A
Item Count Incorrect	50 events (61%)	1.00	3.00	11.00

(d) one FAH receipt did not include portion with total.
** italicized values indicate "true" receipt values where corresponding OCR results are NA

Discussion

We sought to evaluate the potential of optical character recognition (OCR) to capture receipt data and translate it into a reliable analytical format without the need for extensive manual coding and data capture. It was our aim to identify whether these methods could sufficiently provide estimates for underreporting in surveys of food expenditures. Receipts are a very appealing source of record data for these types of estimates because, as a record of food expenditures, receipts are a particularly robust source of information wherein all expenses are typically identified in a line-item format with a clear description, quantity, and corresponding cost. Furthermore, respondents are willing to provide receipts as a record of food expenditures along with reported data.

We found the technology to be capable of capturing text data, and in many cases accurately so. However, there are a number of instances where OCR may not be as accurate as respondents' reported data and this is highly dependent on the quality of the original image. It seems evident that the deficiencies of OCR in this examination are more indicative of the procedure than the potential of OCR. The experience of developing and testing this methodology indicates that image quality and composition are the most influential factors in the facilitation of novel methods of OCR for data collection.

Future research should emphasize standardized methods for capturing receipt images. Additionally, it would be useful to employ machine learning methods to assist in text prediction when the Tesseract OCR results are imperfect. Our design relied on logical interpretation of scanned data using heuristics based on English language rules and conventions and they were found to be successful in identifying, capturing, and eliminating a wide variety of data. Incorporating machine learning into the prediction of text elements that are not correctly discerned by the OCR process could greatly increase the flexibility for programming and processing. As was found in the Ullah et al. (2018) paper, it seems evident that the technology can reliably accomplish the task to some extent but researchers investigating novel OCR methods will have to make some considerations regarding the type of information they intend to capture. It is our belief that incorporating novel OCR methods into data collection using images of dependable quality would produce results that can be reliably used for data validation with some manual intervention and review.

References

Bevans, G. E. 1913. How workingmen spend their spare time. New York: Columbia University Press.

Economic Research Service (ERS), U.S. Department of Agriculture (USDA). n.d. National Household Food Acquisition and Purchase Survey (FoodAPS). https://www.ers.usda.gov/foodaps (Accessed January, 2022).

Clay, M., M. Ver Ploeg, A. Coleman-Jensen, H. Elitzak, C. Gregory, D. Levin, C. Newman, and M. P. Rabbitt (2016), "Comparing National Household Food Acquisition and Purchase Survey (FoodAPS) Data With Other National Food Surveys' Data," EIB-157, US Department of Agriculture, Economic Research Service, July 2016, [online], Available at https://www.ers.usda.gov/webdocs/publications/79893/eib-157.pdf (Accessed January, 2022).

Dillman, D. A., and C. C. House (Eds.) (2013), Measuring What We Spend: Toward a New Consumer Expenditure Survey. Washington, DC: National Academies Press.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren and Michael Westberry. Integrated Public Use Microdata Series, Current Population Survey: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D030.V9.0 (Accessed January, 2022)

French, Simone A., Scott T. Shimotsu, Melanie Wall, and Anne Faricy Gerlach. 2008. "Capturing the spectrum of household food and beverage purchasing behavior: a review." Journal of the American Dietetic Association 108, no. 12: 2051-2058.

Fricker, Scott, Brandon Kopp, Lucilla Tan, and Roger Tourangeau. 2015. "A Review of Measurement Error Assessment in a US Household Consumer Expenditure Survey." *Journal of Survey Statistics and Methodology* 3, no. 1: 67-88.

Garner, Thesia I., Robert McClelland, and William Passero. 2009. "Strengths and Weaknesses of the Consumer Expenditure Survey from a BLS Perspective." Presented at National Bureau of Economic Research Summer Institute Conference on Research on Income and Wealth.

Hu, Mengyao, Garrett W. Gremel, John A. Kirlin, and Brady T. West. 2017. "Nonresponse and Underreporting Errors Increase Over the Data Collection Week Based on Paradata from the National Household Food Acquisition and Purchase Survey." *The Journal of Nutrition* 147, no. 5: 964-975.

Hu, Mengyao, Edmundo Roberto Melipillán, Brady T. West, John A. Kirlin, and Ilse Paniagua. 2020. "Response patterns in a multi-day diary survey: implications for adaptive survey design." *In Survey Research Methods*, vol. 14, no. 3, pp. 289-300. 2020.

The ImageMagick Development Team, 2021. ImageMagick, Available at: https://imagemagick.org (Accessed January, 2022).

National Health and Nutrition Examination Survey (NHANES). Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). 2004. “1999-2000 Data Documentation, Codebook, and Frequencies – Food Security Section (FSQ)”. https://wwwn.cdc.gov/nchs/data/nhanes/1999-2000/questionnaires/fq-fs.pdf (Accessed January, 2022)

Johnson, Kathleen W., and Geng Li. 2009. "Household Liability Data in the Consumer Expenditure Survey." *Monthly Labor Review*. 132: 18.

Li, Geng, Robert F. Schoeni, Sheldon Danziger, and Kerwin Kofi Charles. 2010. “New Expenditure Data in the PSID: Comparisons with the CE.” *Monthly Labor Review*, 29–39.

Maitland, Aaron, and Lin Li. 2016. "Review of the Completeness and Accuracy of FoodAPS 2012 Data." *Washington (DC): Economic Research Service, USDA*.

Petraglia, Elizabeth, Wendy Van de Kerckhove, and Tom Krenzke. 2016. "Review of the Potential for Nonresponse Bias in FoodAPS 2012." *Prepared for the Economic Research Service, US Department of Agriculture*.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rankin, Janet Walberg, Richard A. Winett, Eileen S. Anderson, Patricia G. Bickley, John F. Moore, Michael Leahy, Carl E. Harris, and Robert E. Gerkin. 1998. "Food purchase patterns at the supermarket and their relationship to family characteristics." Journal of Nutrition Education 30, no. 2: 81-88.

Ransley, Joan K., Judith K. Donnelly, Tanya N. Khara, Helen Botham, Heidy Arnot, Darren C. Greenwood, and Janet E. Cade. 2001. "The use of supermarket till receipts to determine the fat and energy intake in a UK population." Public health nutrition 4, no. 6: 1279-1286.

Shields, Jennifer, and Nhien To. "Learning to Say No: Conditioned Underreporting in an Expenditure Survey." American Association for Public Opinion Research-American Statistical Association, Proceedings of the Section on Survey Research Methods (2005): 3963-8.

Google 2021. Tesseract OCR v 5.0.0. Lead Developer, Ray Smith. Repository Maintainer, Zdenko Podobny. https://github.com/tesseract-ocr (Accessed January, 2022).

Stone, A. A., Kessler, R. C., & Haythomthwatte, J. A. (1991). Measuring Daily Events and Experiences: Decisions for the Researcher. *Journal of Personality*, 59(3), 575–607.

Ullah, Rafi, Ali Sohani, Athaul Rai, Faraz Ali, and Richard Messier. 2018. "OCR Engine to Extract Food-Items, Prices, Quantity, Units from Receipt Images, Heuristics Rules Based Approach." International Journal of Scientific & Engineering Research 9, no. 2: 1334-1341.

United States Bureau of Labor Statistics (BLS). United States Department of Labor. n.d. Consumer Expenditure Surveys: Diary Interview Survey – Diary Survey Form. 2012. https://www.bls.gov/cex/csx801p.pdf (Accessed January, 2022).

Yan, Ting, and Aaron Maitland. 2016. "Review of the FoodAPS 2012 Instrument Design, Response Burden, Use of Incentives, and Response Rates." *Washington (DC): Economic Research Service, USDA.*