

Novel Optical Character Recognition (OCR) as a Method of Capturing Food Expenditures from Point-of-Sale Receipts

Prepared and Presented by:

Adam Kaderabek

M.S. Survey & Data Science
University of Michigan

In Collaboration with:

Brady T. West, Institute for Social Research

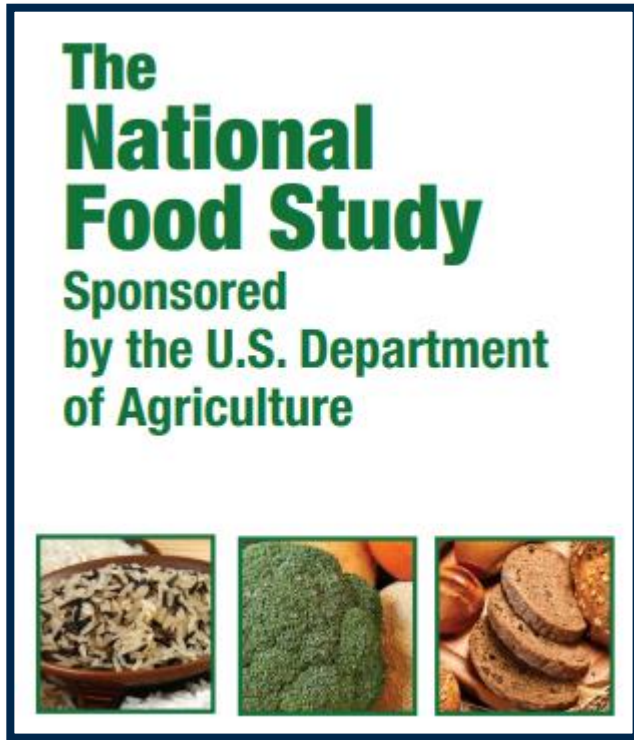
Jeffrey M. Gonzalez, USDA – Economic Research Service

Elina T. Page, USDA – Economic Research Service

This presentation was supported in part by the U.S. Department of Agriculture, Economic Research Service.

The findings and conclusions in this presentation are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy.

Background: FoodAPS-1

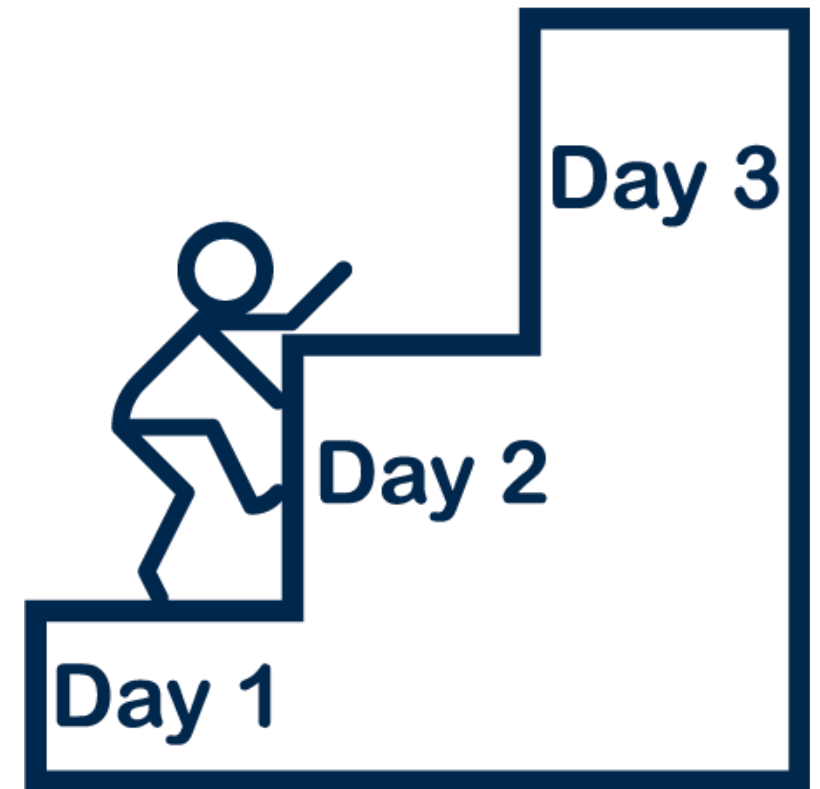


Informational brochure (The National Food Study)
<https://www.ers.usda.gov/media/8609/informationalbrochure.pdf>

- ❖ What was the Food Acquisition and Purchase Survey (FoodAPS-1) aka the National Food Study?
 - Diary Survey of Food Acquisition
 - Food Acquisition at the Item Level
 - Included Food without Expense
- ❖ Existing Research Places Emphasis on...
 - Food Security (e.g., USDA-Food and Nutrition Service's *Guide to Measuring Household Food Security*, 2000)
 - Food Expenditures (e.g., Consumer Expenditure Surveys: Diary Interview Survey – Diary Survey Form, 2012)

Challenges: FoodAPS-1

- ❖ Diaries attempt to capture frequent and often easily forgotten experiences over a prolonged period.
- ❖ Recording these details is burdensome (e.g., Dillman and House, 2013, Hu et al., 2017).
- ❖ Burden reduces compliance and leads to underreporting (e.g., Fricker et al., 2015, Maitland and Li, 2016).
 - Underreporting from Item-nonresponse (Hu et al., 2017)
 - Underreporting from Measurement Error (Stone et al. 1991, Shields and To 2005)



Addressing Underreporting

- ❖ External Benchmarks (Clay et al., 2016)
 - Consumer Expenditure Survey (CES) Diary Interview
 - National Health and Nutrition Examination Study (NHANES)
- ❖ Internal Benchmarks
 - None available
 - Reporting decreased at the day level however (Maitland and Li 2016, Petraglia et al. 2016, Yan and Maitland 2016, Hu et al. 2017, Hu et al. 2020)
- ❖ Record Check Studies
 - Follow-up interviews to verify receipts were not effective (Geisen et al., 2011)
 - 36% of 3,039 expenditures had a record with only 53% matching

ADCM & FoodAPS-2

- ❖ Even though underreporting was observed, FoodAPS-1 was successful as the first study of its kind.
- ❖ In preparation for FoodAPS-2 an Alternative Data Collection Method Study (ADCM) was conducted (2016).
- ❖ ADCM field tested a “FoodLogger” app to reduce burden and assist with data reporting.
- ❖ Participants were also asked if they were able to provide a receipt for the event.



ADCM Receipts

- ❖ ADCM respondents reported 4,906 food events and 1,598 events were reported as “having a receipt to upload”
- ❖ Locations where *a receipt is expected* accounted for 3,109 ADCM events with 1,247 (40%) corresponding to an itemized and legible receipt.
- ❖ 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.



ADCM Data Validation

❖ FAFH events:

- 18 missing, 5 images not receipts, 8 non-itemized receipts, and 1 partial receipt (missing total).

❖ FAH events:

- 9 missing or corrupted, 2 illegible, 1 missing total, and 1 non-itemized.

Table 1: Error Summary for ADCM Sampled Events (Manually Validated)

Error Type	Frequency	Absolute Difference		
		Minimum	Mean	Maximum
Food-Away-from-Home (FAFH, n=69)				
Incorrect Total (n=68)	12 events (17%)	\$0.08	\$8.73	\$51.05
Incorrect Tax (FAFH only) (n=68)	8 events (12%)	\$0.30	\$1.25	\$4.47
Incorrect Item Count (n=68) one receipt indicated “6 Eat & Play Combos” but report indicated 30 items	38 events (55%)	1.00	3.11	8.00
Food-at-Home (FAH, n=89)				
Incorrect Total (n=88)	15 events (17%)	\$0.04	\$7.06	\$24.11
Incorrect Item Count (n=88)	47 events (53%)	1.00	5.30	53.00

Sample Receipts Eligible for OCR

- ❖ FAFH events:
 - 8 receipts unsuitable for OCR
 - Added in 7 non-itemized receipts
- ❖ FAH events:
 - 6 receipts unsuitable for OCR
 - 4 “additional” images from top/bottom photos of receipts
- ❖ Image quality for these images was poor due to blur and/or poor lighting which leads to excessive image noise.

Table 2: Error Summary for OCR Eligible Events

Error Type	Frequency	Absolute Difference		
		Minimum	Mean	Maximum
Food-Away-from-Home (FAFH, n=68)				
Incorrect Total (n=67)	10 events (15%)	\$0.08	\$7.65	\$51.05
Incorrect Tax (FAFH only) (n=60)	4 events (7%)	\$0.31	\$1.59	\$4.47
Incorrect Item Count (n=59)	31 events (53%)	1.00	2.23	5.00
Food-at-Home (FAH, n=82)				
Incorrect Total (n=81)	13 events (16%)	\$0.06	\$6.94	\$24.11
Incorrect Item Count (n=82)	42 events (55%)	1.00	5.31	53.00

How it Works: Steps 1-3

Step 1: Select Text

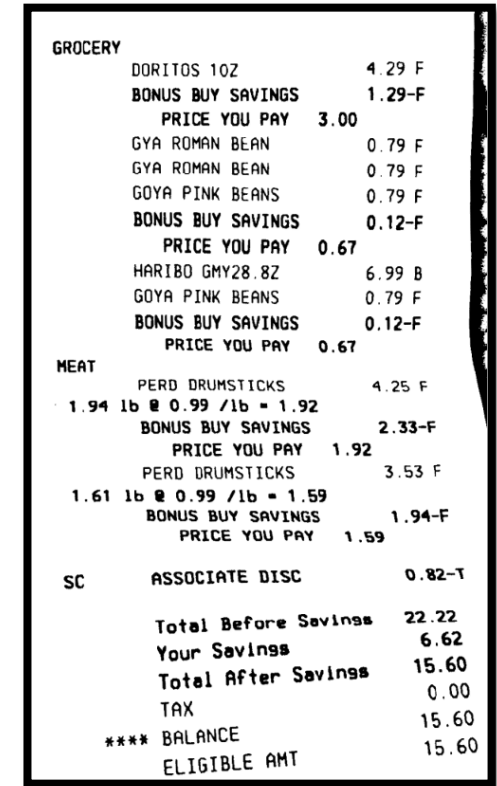


GROCERY	
DORITOS 10Z	4.29 F
BONUS BUY SAVINGS	1.29-F
PRICE YOU PAY	3.00
GYA ROMAN BEAN	0.79 F
GYA ROMAN BEAN	0.79 F
GOYA PINK BEANS	0.79 F
BONUS BUY SAVINGS	0.12-F
PRICE YOU PAY	0.67
HARIBO GMY28.8Z	6.99 B
GOYA PINK BEANS	0.79 F
BONUS BUY SAVINGS	0.12-F
PRICE YOU PAY	0.67
MEAT	
PERD DRUMSTICKS	4.25 F
1.94 lb @ 0.99 /lb = 1.92	
BONUS BUY SAVINGS	2.33-F
PRICE YOU PAY	1.92
PERD DRUMSTICKS	3.53 F
1.61 lb @ 0.99 /lb = 1.59	
BONUS BUY SAVINGS	1.94-F
PRICE YOU PAY	1.59
SC	
ASSOCIATE DISC	0.82-T
Total Before Savings	22.22
Your Savings	6.62
Total After Savings	15.60
TAX	0.00
**** BALANCE	15.60
ELIGIBLE AMT	15.60

Step 2: Give text to ImageMagick Wizard



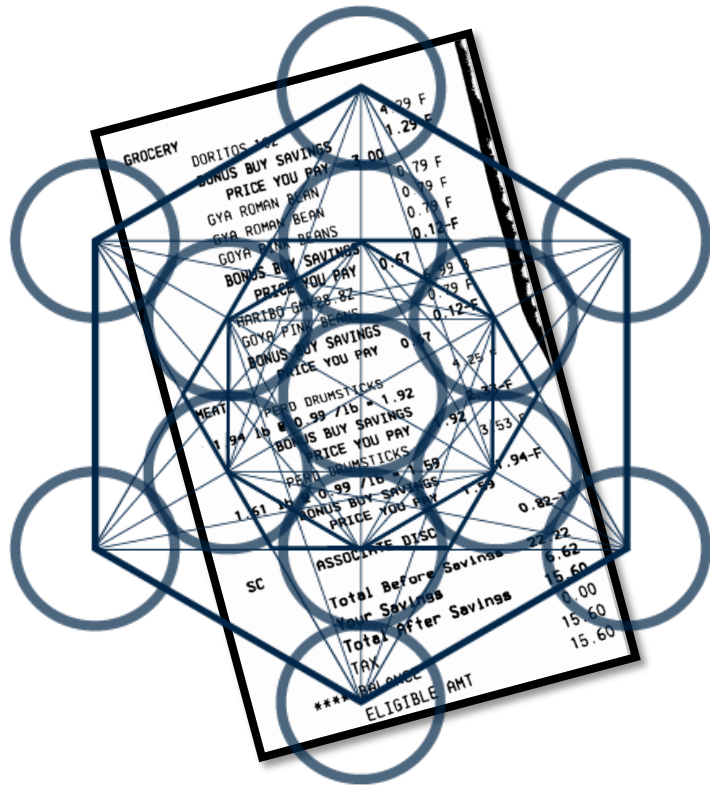
Step 3: Get Magick



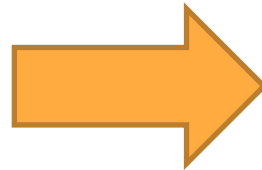
GROCERY	
DORITOS 10Z	4.29 F
BONUS BUY SAVINGS	1.29-F
PRICE YOU PAY	3.00
GYA ROMAN BEAN	0.79 F
GYA ROMAN BEAN	0.79 F
GOYA PINK BEANS	0.79 F
BONUS BUY SAVINGS	0.12-F
PRICE YOU PAY	0.67
HARIBO GMY28.8Z	6.99 B
GOYA PINK BEANS	0.79 F
BONUS BUY SAVINGS	0.12-F
PRICE YOU PAY	0.67
MEAT	
PERD DRUMSTICKS	4.25 F
1.94 lb @ 0.99 /lb = 1.92	
BONUS BUY SAVINGS	2.33-F
PRICE YOU PAY	1.92
PERD DRUMSTICKS	3.53 F
1.61 lb @ 0.99 /lb = 1.59	
BONUS BUY SAVINGS	1.94-F
PRICE YOU PAY	1.59
SC	
ASSOCIATE DISC	0.82-T
Total Before Savings	22.22
Your Savings	6.62
Total After Savings	15.60
TAX	0.00
**** BALANCE	15.60
ELIGIBLE AMT	15.60

How it Works: Steps 4-6

Step 4: Put “Magick” into Tesseract Engine



Step 5: Get Optically Recognized Text



OLID	OCRLines
1	GROCERY
2	DORITOS 102 4.29 F:
3	BONUS BUY SAVINGS 1.29-F
4	PRICE YOU PAY 3.00
5	GYA ROMAN BEAN 0.79 F
6	GYA ROMAN BEAN 0.79 F
7	GOYA PINK BEANS 0.79 F
8	BONUS BUY SAVINGS 0.12-F
9	PRICE YOU PAY 0.67
10	HARIBO GMY28.8Z 6.99 B
11	GOYA PINK BEANS 0.79 F q
12	BONUS BUY SAVINGS 0.12-F i
13	PRICE YOU PAY 0.67 :
14	MEAT
15	PERD DRUMSTICKS 4.25 F
16	- 1.94 lb @ 0.99 /lb = 1.92
17	BONUS BUY SAVINGS 2.33-F
18	PRICE YOU PAY 1.92
19	PERD DRUMSTICKS 3.53 F
20	1.61 lb @ 0.99 /lb = 1.59
21	BONUS BUY SAVINGS 1.94-F
22	PRICE YOU PAY 1.59
23	sc ASSOCIATE DISC Q.82-T
24	Total Before Savinss 22.22
25	Your Savinss . ye
26	Total After Savinss 0 00
27	TAK 15.60
28	xx%% BALANCE 15.60
29	ELIGIBLE AMT
30	

Step 6: Rinse and Shine

OLID	OCRLines2	amount
1	DORITOS 102	4.29
2	BONUS BUY SAVINGS	1.29
3	GYA ROMAN BEAN	0.79
4	GYA ROMAN BEAN	0.79
5	GOYA PINK BEANS	0.79
6	BONUS BUY SAVINGS	0.12
7	HARIBO GMY28.8Z	6.99
8	GOYA PINK BEANS	0.79
9	BONUS BUY SAVINGS	0.12
10	PERD DRUMSTICKS	4.25
11	1.94 LB @ 0.99 /LB	1.92
12	BONUS BUY SAVINGS	2.33
13	PERD DRUMSTICKS	3.53
14	1.61 LB @ 0.99 /LB	1.59
15	BONUS BUY SAVINGS	1.94
16	SC ASSOCIATE DISC	0.82
17	BALANCE	15.6

Where We Make the Sausage

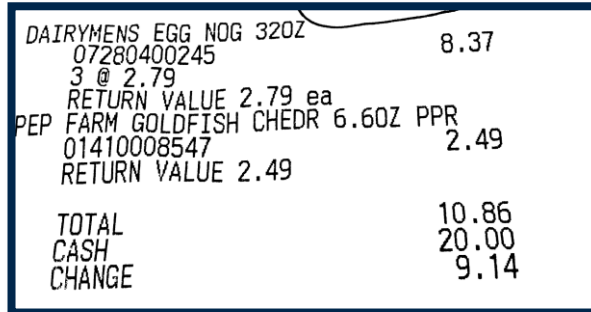
❖ Garbage Text

- Character/Word level
- Entire Line, e.g.,
“SUBTOTAL”

❖ Units

❖ Discounts

❖ Two-line Items



DAIRYMENS EGG NOG 32OZ	8.37
07280400245	
3 @ 2.79	
RETURN VALUE 2.79 ea	
PEP FARM GOLDFISH CHEDR 6.60Z PPR	2.49
01410008547	
RETURN VALUE 2.49	
TOTAL	10.86
CASH	20.00
CHANGE	9.14

```
57
58 for (i in 1:length(FAH4OCR)) { # for each receipt image in folder
59   tic("start")
60   # import image file i
61   image.scan<-image_read(FAH4OCR[i])
62   # scan image file i
63   image.text<-image_ocr(image.scan)
64   # un-list image text by line
65   image.lines<-data.frame(EventID=str_extract(FAH4OCR[i], "[0-9]+(?=_)"),
66                           OLID=1:length(unlist(str_split(image.text, "\n"))),
67                           OCRLines=unlist(str_split(image.text, "\n")))
68   # Convert all characters to UPPER case
69   image.lines$OCRLines2<-str_to_upper(image.lines$OCRLines)
70   # identify garbage lines and replace as empty line[]
71   image.lines$OCRLines2<-ifelse(str_detect(image.lines$OCRLines2,
72                                           regex(paste(garbageLine, collapse
73 = "|"))))==TRUE, "", image.lines$OCRLines2)
74
75   # remove single letters surrounded by whitespace or preceded by "-"
76   image.lines$OCRLines2<-str_remove_all(image.lines$OCRLines2,
77
78   regex("(?![\s])[:alpha:]{1}(?![\s])|(?<=\d-)[:alpha:]{1}(?![\s])|\\s0(?!
79   \s)|(?<=[A-Z0-9]{1})[:alpha:]{1}$"))
80   # replace any commas (,) as period/decimal point (.)
81   image.lines$OCRLines2<-str_replace_all(image.lines$OCRLines2, pattern =
82   ",", replacement = ".")
```

Huge Success

MOH CEREAL	004240090478 F	4.98 0
GU BRD ROUND	007074228544 F	1.08 0
GU BRD ROUND	007074228544 F	1.08 0
EQUAL 28.22Z	030025890007 F	10.48 0
POPCORN	004166705112 F	4.98 N
CHOC CHECKOU	003400000241 F	0.78 N
CHOC CHECKOU	003400000241 F	0.78 N
6468 USB	061965909195	19.97 X
EGGS 18CT	007074212708 F	1.08 0
RANCH DRSNG	004132126146 F	1.98 0
BISCUITS	001800000918 F	1.98 0
BISCUITS	001800000918 F	1.98 0
HAZELNUT CRM	007074212203 F	2.48 0
KFT SINGLES	002100061526 F	3.98 0
SLICED HAM	007590000088 F	8.72 0
TK CL BACON	004470001990 F	5.48 0
O.M. BOLO	004470000857 F	2.00 0
TURKEY WINGS	022099100767 F	7.67 0
ECK NCSHKD S	002781530044 F	6.98 0
GU VINEGAR	007074235257 F	1.50 0
GRN CABBAGE	000000004069KI	1.33 N
2.30 lb @ 1 lb /0.58		
GRN CABBAGE	000000004069KI	1.35 N
2.32 lb @ 1 lb /0.58		
BROC CROWNS	000000003082KI	0.77 N
0.50 lb @ 1 lb /1.54		
CLEMENTINE	005410722101 I	3.97 N
CHEESE SHRED	002100005457 F	2.00 0
COOKED SHRMP	007539102438 F	5.98 0
CKN WING LG	025119881024 F	10.24 0
CHICKEN PAUS	025977250214 F	2.14 0
TURNIPS	002076400044 F	3.97 N
BAG KALE	002076465222 I	2.48 N
RLHON JCE	001480058223 F	1.98 N
PKG SALAD	068113132895 I	1.96 N
GRN CABBAGE	000000004069KI	1.39 N
2.40 lb @ 1 lb /0.58		
BELL PEPPER	000000004065KI	0.84 N
CUCUMBER	000000004062KI	0.68 N
SUBTOTAL		131.04
TAX 1 6.500 %		1.30
TOTAL		132.34

MOH CEREAL	004240090478 F	4.98 0
GU BRD ROUND	007074228544 F	1.08 0
GU BRD ROUND	007074228544 F	1.08 0
EQUAL 28.22Z	030025890007 F	10.48 0
POPCORN	004166705112 F	4.98 N
CHOC CHECKOU	003400000241 F	0.78 N
CHOC CHECKOU	003400000241 F	0.78 N
6468 USB	061965909195	19.97 X
EGGS 18CT	007074212708 F	1.08 0
RANCH DRSNG	004132126146 F	1.98 0
BISCUITS	001800000918 F	1.98 0
BISCUITS	001800000918 F	1.98 0
HAZELNUT CRM	007074212203 F	2.48 0
KFT SINGLES	002100061526 F	3.98 0
SLICED HAM	007590000088 F	8.72 0
TK CL BACON	004470001990 F	5.48 0
O.M. BOLO	004470000857 F	2.00 0
TURKEY WINGS	022099100767 F	7.67 0
ECK NCSHKD S	002781530044 F	6.98 0
GU VINEGAR	007074235257 F	1.50 0
GRN CABBAGE	000000004069KI	1.33 N
2.30 lb @ 1 lb /0.58		
GRN CABBAGE	000000004069KI	1.35 N
2.32 lb @ 1 lb /0.58		
BROC CROWNS	000000003082KI	0.77 N
0.50 lb @ 1 lb /1.54		
CLEMENTINE	005410722101 I	3.97 N
CHEESE SHRED	002100005457 F	2.00 0
COOKED SHRMP	007539102438 F	5.98 0
CKN WING LG	025119881024 F	10.24 0
CHICKEN PAUS	025977250214 F	2.14 0
TURNIPS	002076400044 F	3.97 N
BAG KALE	002076465222 I	2.48 N
RLHON JCE	001480058223 F	1.98 N
PKG SALAD	068113132895 I	1.96 N
GRN CABBAGE	000000004069KI	1.39 N
2.40 lb @ 1 lb /0.58		
BELL PEPPER	000000004065KI	0.84 N
CUCUMBER	000000004062KI	0.68 N
SUBTOTAL		131.04
TAX 1 6.500 %		1.30
TOTAL		132.34

RecLine	LineItem	UPC	size	quantity	unitprice	itemprice
1	MOH CEREAL	904240000078	#N/A	1	#N/A	4.98
2	GU BRD ROUND	007074279544	#N/A	1	#N/A	1.08
3	GU BRD ROUND	007074278544	#N/A	1	#N/A	1.08
4	EQUAL 28.22Z	030025890007	#N/A	1	#N/A	10.48
5	POPCORN	004166705112	#N/A	1	#N/A	4.98
6	CHOC CHECKOU OOA41	#N/A	40000 Z	1	#N/A	0.78
7	CHOC CHECKOU OO3D000N0241	#N/A	#N/A	1	#N/A	0.78
8	6468 USB	061965909195	#N/A	1	#N/A	19.97
9	EGGS 16CT	007074212708	#N/A	1	#N/A	1.08
10	RANCH DRSNG	004132176146	#N/A	1	#N/A	1.98
11	BISCUITS	001800000918	#N/A	1	#N/A	1.98
12	BISCUITS	001800000918	#N/A	1	#N/A	1.98
13	HAZELNUT CRM	007074212203	#N/A	1	#N/A	2.48
14	KFT SINGLES	002100061526	#N/A	1	#N/A	3.98
15	SLICED HAH	007490000088	#N/A	1	#N/A	8.72
16	TKBACON	004470001990	#N/A	1	#N/A	5.48
17	O.M. BOLO	004470000857	#N/A	1	#N/A	2.00
18	TURKEY WINGS	022099100767	#N/A	1	#N/A	7.67
19	ECK NCSHKD \$	002781540044	#N/A	1	#N/A	6.98
20	GU VINEGAR	007874215257	#N/A	1	#N/A	1.50
21	GRN CABBAGE	000000004069	#N/A	1 1		#N/A
22	GRN CABBAGE	000000004069	#N/A	1 1		#N/A
23	BROC CROWNS	000000003082	#N/A	1 1		#N/A
24	CLEHENTINE	005410722101	#N/A	1	#N/A	3.97
25	CHEESE SHRED	002100005457	#N/A	1	#N/A	2.00
26	COOKED SHRMP	007539102438	#N/A	1	#N/A	5.98
27	CKN WING10.24	025119881024	#N/A	1	#N/A	#N/A
28	CHICKEN PAWS	025977250214	#N/A	1	#N/A	2.14
29	TURNIPS	002076400044	#N/A	1	#N/A	3.97
30	BAG KALE	00287645222	#N/A	1	#N/A	2.48
31	RLHON JCE	001480058223	#N/A	1	#N/A	1.98
32	PKG SALAD1.96	068113132895	#N/A	1	#N/A	#N/A
33	GRN CABBAGE	000000004069	#N/A	1 1		#N/A
34	BELL PEPPER QKI	000000004065	#N/A	1	#N/A	0.84
35	CUCUMBER0.68	000000004052	#N/A	1	#N/A	#N/A
36	TOTAL	#N/A	#N/A	1	#N/A	132.34

Huge Failure

```

GV WATER 007874235191 F 0.88 N
TP SNDWCH 007229000065 F 8.50 0
SF ADV RTD 000834674008 F 5.93 N
BREAD 007294561241 F 2.82 0
GV BOWL 007874202865 F 1.87 0
JD SNDWCH 007790031090 F 5.98 0
FF GL C L 005000029261 0.57 X
FF GL CHK 005000057845 0.57 X
GVY LVS TKY 005000058004 0.57 X
FF GL CHK 005000057845 0.57 X
LNF YOGURT 003663203255 F 3.47 0
LNF GK CRNCH 003663201108 F 1.00 0
LNF YOGURT 003663203738 F 1.00 0
APOTHIC 008500001774 7.98 T
BANANAS 000000004011KI 1.34 N
2.79 lb @ 1 lb /0.48 2.98 N
PKG SALAD 068113102786 I 2.76 N
PKG SALAD 068113132895 I 2.00 0
DONUTS 007087000908 F 2.00 0
DONUTS 007087000909 F 2.00 0
SUBTOTAL 52.79
TAX 1 8.250 % 0.85
TOTAL 53.64
DEBIT TEND 53.64
DEBIT CASH BACK 20.00
TOTAL DEBIT PURCHASE 73.64
CHANGE DUE 20.00
    
```

```

GV WATER 007874235191 F 0.88 N
TP SNDWCH 007229000065 F 8.50 0
SF ADV RTD 000834674008 F 5.93 N
BREAD 007294561241 F 2.82 0
GV BOWL 007874202865 F 1.87 0
JD SNDWCH 007790031090 F 5.98 0
FF GL C L 005000029261 0.57 X
FF GL CHK 005000057845 0.57 X
GVY LVS TKY 005000058004 0.57 X
FF GL CHK 005000057845 0.57 X
LNF YOGURT 003663203255 F 3.47 0
LNF GK CRNCH 003663201108 F 1.00 0
LNF YOGURT 003663203738 F 1.00 0
APOTHIC 008500001774 7.98 T
BANANAS 000000004011KI 1.34 N
2.79 lb @ 1 lb /0.48 2.98 N
PKG SALAD 068113102786 I 2.76 N
PKG SALAD 068113132895 I 2.00 0
DONUTS 007087000908 F 2.00 0
DONUTS 007087000909 F 2.00 0
SUBTOTAL 52.79
TAX 1 8.250 % 0.85
TOTAL 53.64
DEBIT TEND 53.64
DEBIT CASH BACK 20.00
TOTAL DEBIT PURCHASE 73.64
CHANGE DUE 20.00
    
```

RecLine	LineItem	UPC	size	quantity	unitprice	itemprice
1	GV WAT	#N/A	#N/A	1	#N/A	0078
2	LA SERASSOEU	#N/A	#N/A	1	#N/A	#N/A
3	RESSOGB0GS	#N/A	#N/A	1	#N/A	#N/A
4	FF SNDWC ERIE	#N/A	#N/A	1	#N/A	#N/A
5	FF4	#N/A	#N/A	1	#N/A	#N/A
6	VYTE	#N/A	#N/A	1	#N/A	#N/A
7	ESTEIH	#N/A	#N/A	1	#N/A	#N/A
8	NF GUR	#N/A	#N/A	1	#N/A	000
9	3	#N/A	#N/A	1	#N/A	#N/A
10	EAE	#N/A	#N/A	1	#N/A	#N/A
11	53	#N/A	#N/A	1	#N/A	255
12	DOS668705758	#N/A	#N/A	1	#N/A	#N/A
13	ITEAD	#N/A	#N/A	1	#N/A	#N/A
14	INS MAB11KI	#N/A	#N/A	1	#N/A	1.00
15	SE	#N/A	#N/A	1	#N/A	#N/A
16	REE	#N/A	#N/A	1	#N/A	#N/A
17	3 UBTO	#N/A	#N/A	1	#N/A	3.09
18	OVE EET BR.	#N/A	#N/A	1	#N/A	#N/A
19	RTBA	#N/A	#N/A	1	#N/A	#N/A
20	PURC	#N/A	#N/A	1	#N/A	#N/A
21	WONGE	#N/A	#N/A	1	#N/A	#N/A
22	DU	#N/A	#N/A	1	#N/A	#N/A
23	13	#N/A	#N/A	1	#N/A	#N/A
24	0.00	#N/A	#N/A	1	#N/A	#N/A

Results – Comparison Summary

❖ FAFH events:

- Difficult to capture item quantities due to variation of formatting.
- OCR compared to item count based on number of receipt “food-lines”.
- OCR totals further from truth than reports but OCR tax very close to true amount.

❖ FAH events:

- OCR item counts closer to truth than reported data.
- OCR totals further from truth than reports and extreme variation arising from errors capturing the data.

Table 3	Method		
FAFH (N=68)	Report Data	Manual Validation	OCR Results
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 Total (Subtotal)	\$16.42	\$17.36	\$19.92
Mean Event Tax (FAFH only)	\$0.95	\$1.18	\$1.20*
FAH (N=82)			
Mean Item Count	7.21	9.13	10.43
Mean Event Total**	\$24.93	\$25.47	\$23.75

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

** Excluding 4 events corresponding to largest OCR mis-captures of over \$100 each.

OCR Performance Summary – FAFH Totals

- ❖ Output included indicators for:
 - Subtotal
 - Tax
 - Total
- ❖ If regex patterns identified control text and found an amount then indicator=1, otherwise 0.
- ❖ Accurate indicator 68% but 1/3 of corresponding totals were incorrect.
- ❖ ~40% of all OCR totals inaccurate or missing

Table 4: FAFH Totals		Absolute Difference		
OCR Indicates & Captures Total (n=65)*	Frequency	Minimum	Mean	Maximum
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%) 10 events (15%)	- \$6.40	\$8.00 \$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

* one FAFH receipt did not include portion with total and two of the non-itemized receipts did not include a total line in their format so indication was not possible.

** italicized values indicate “true” receipt values where corresponding OCR results are NA.

OCR Performance Summary – FAFH Tax

- ❖ Accurate tax indicator
~70% of time but ~1/5 of corresponding amounts were incorrect.
- ❖ ~40% of taxes inaccurate or missing
- ❖ Extreme maximum outlier of \$831.31, additional outlier of \$80.00.

Table 5: FAFH Tax		Absolute Difference		
OCR Indicates & Captures Tax (n=59)*	Frequency	Minimum	Mean	Maximum
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

* 1 partial receipt, 7 non-itemized receipts, and one without a tax field excluded

** italicized values indicate “true” receipt values where corresponding OCR results are NA.

OCR Performance Summary – FAFH Subtotals

- ❖ Accurate subtotal indicator ~78% of time but ~1/5 of corresponding amounts were incorrect.
- ❖ 34% of subtotals inaccurate or missing

Table 5: FAFH Subtotals		Absolute Difference		
OCR Indicates & Captures Subtotal (n=55)*	Frequency	Minimum	Mean	Maximum
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

* In addition to the 9 ineligible receipts from prior two slides, 4 additional receipts did not incorporate subtotals.

** italicized values indicate “true” receipt values where corresponding OCR results are NA.

OCR Performance Summary – FAH Totals

- ❖ Accurate total indicator
70% of time only 1/10 of corresponding amounts were incorrect. (~1/2 the errors of FAFH indicators)
- ❖ 26% of totals inaccurate or missing

Table 7: FAH Totals		Absolute Difference		
OCR Indicates & Captures Total (n=81)*	Frequency	Minimum	Mean	Maximum
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%)	\$0.02	\$25.58	\$250.07
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

* 1 FAH receipt was missing bottom portion containing total.

** italicized values indicate “true” receipt values where corresponding OCR results are NA.

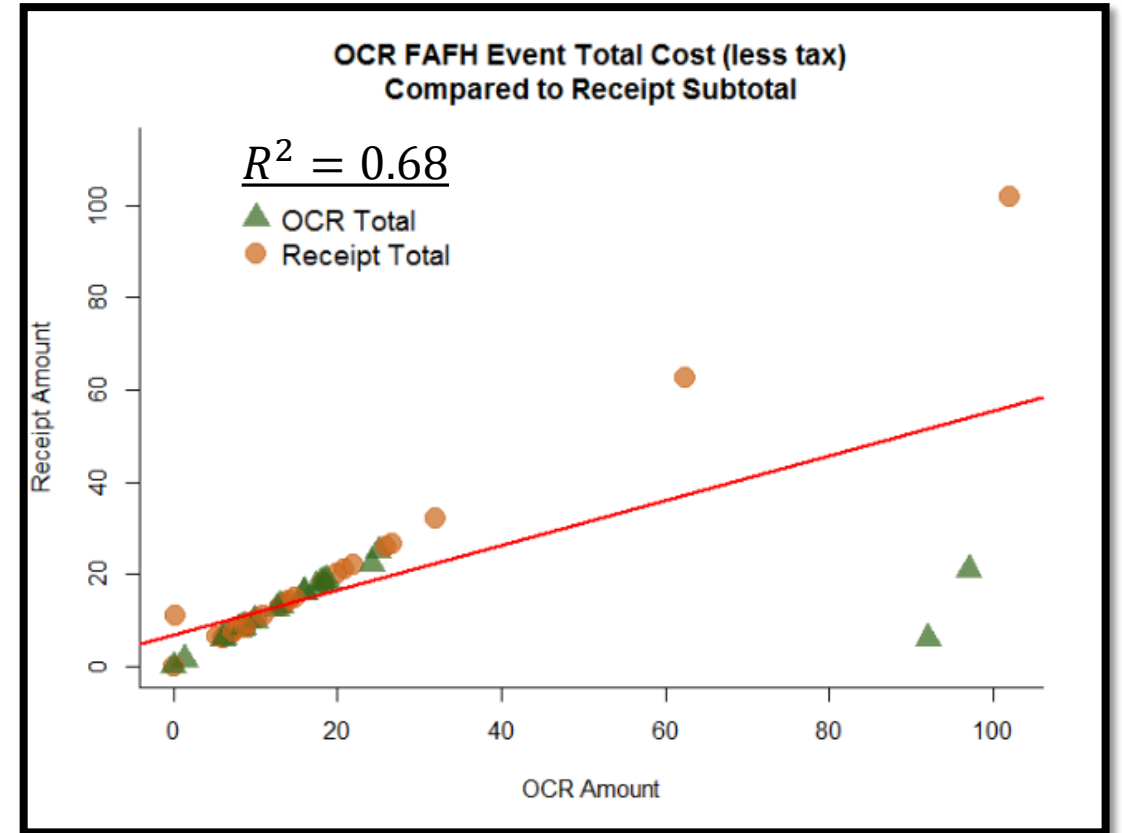
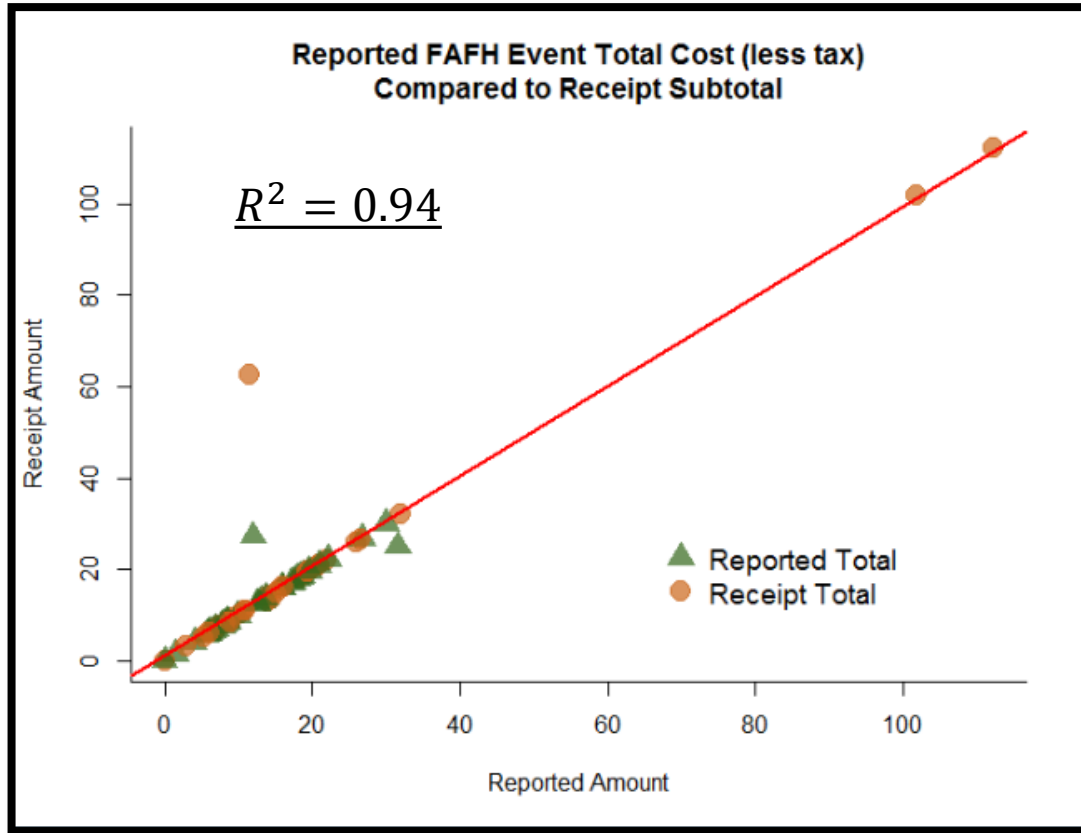
OCR Performance Summary – Item Counts

- ❖ FAFH item counts not terrible but not able to capture quantities at this point.
- ❖ Range of FAFH errors smaller than that of reported data (maximum diff=5, Table 2)
- ❖ FAH item counts not as successful.
- ❖ Range of FAH errors was greatly reduced compared to reported data (maximum diff=53, Table 2)

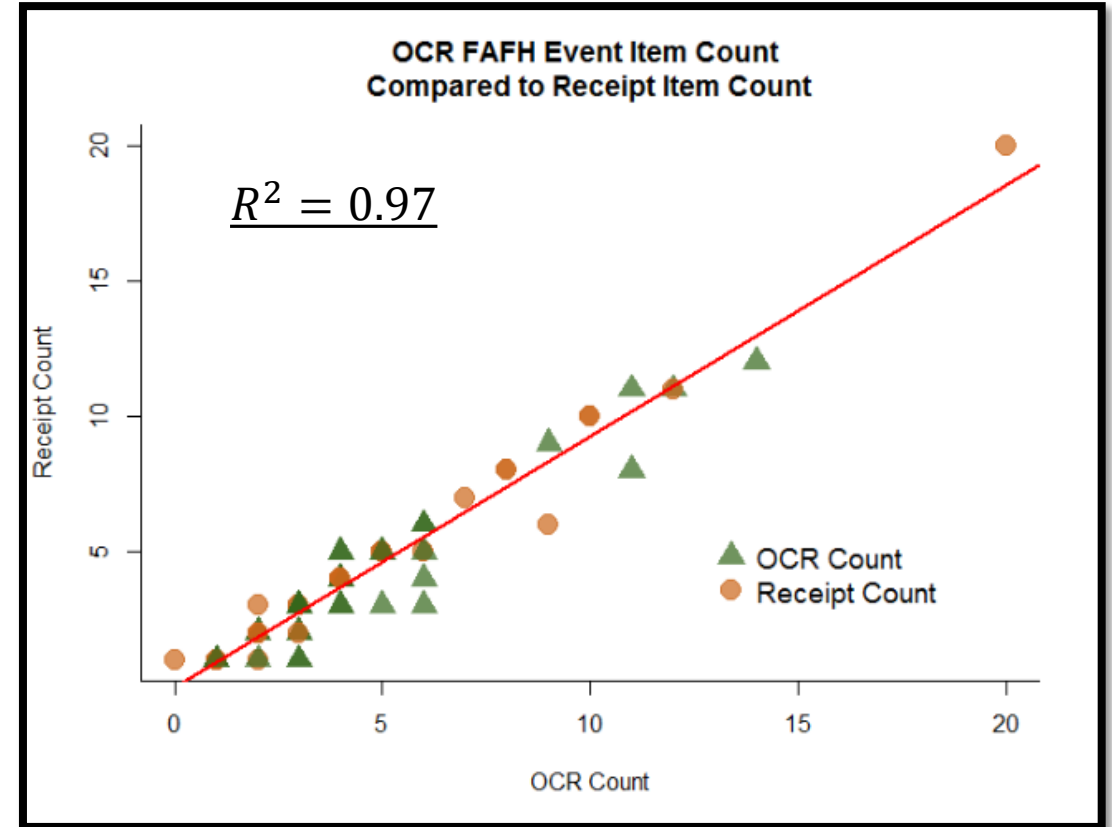
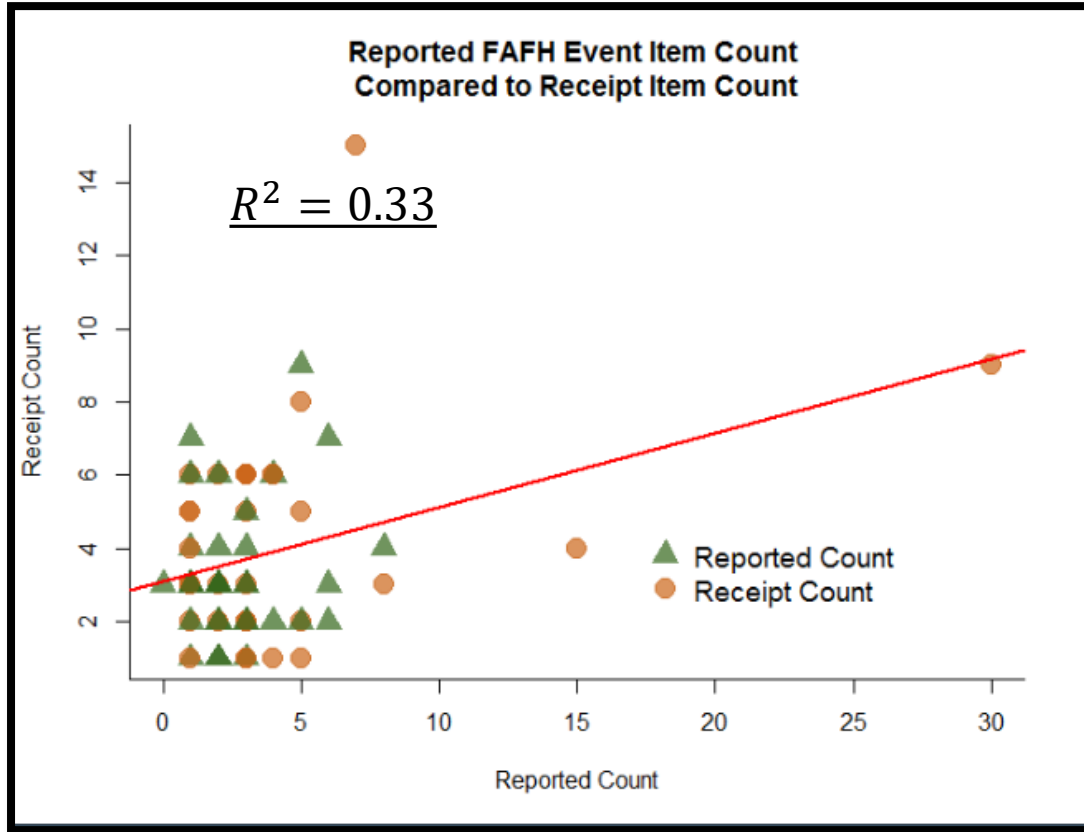
Table 8: Item Counts		Absolute Difference		
FAFH OCR Item Counts (n=65)*	Frequency	Minimum	Mean	Maximum
Item Count Correct	42 events (65%)	N/A	N/A	N/A
Item Count Incorrect	23 events (35%)	1.00	1.48	3.00
FAH OCR Item Counts (n=82)				
Item Count Correct	32 events (39%)	N/A	N/A	N/A
Item Count Incorrect	50 events (61%)	1.00	3.00	11.00

* 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.

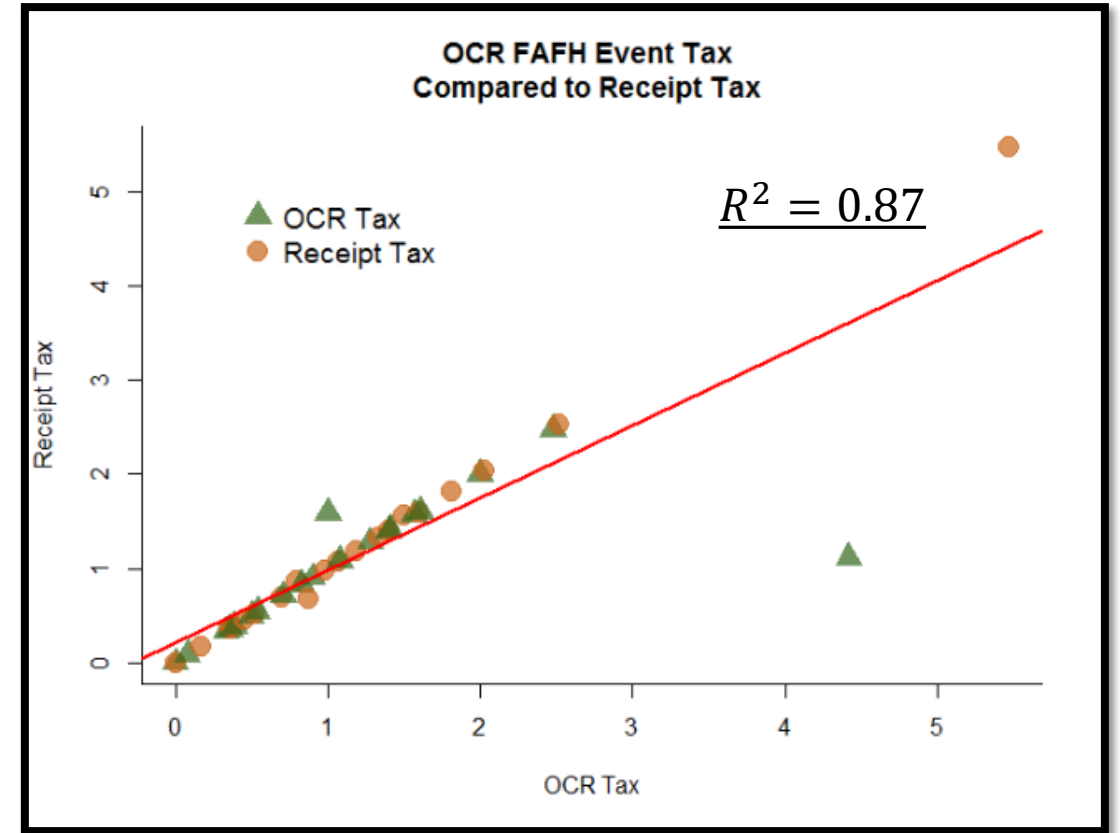
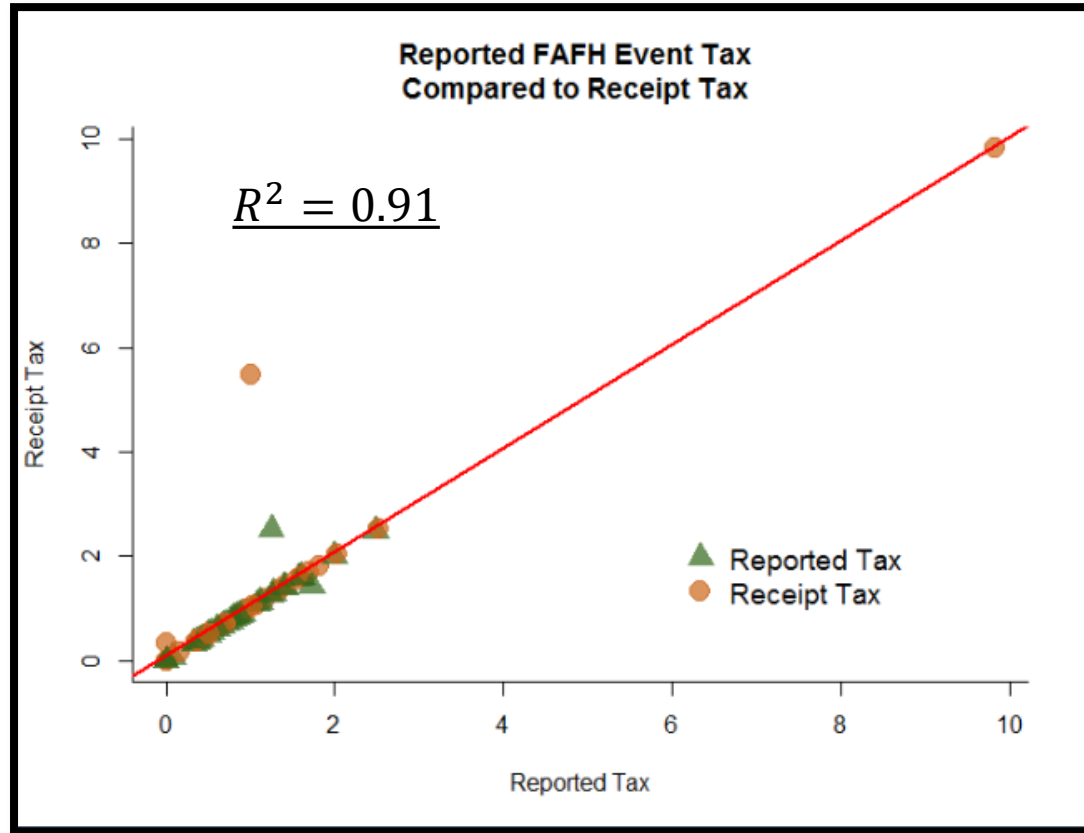
Visualizing FAFH Totals



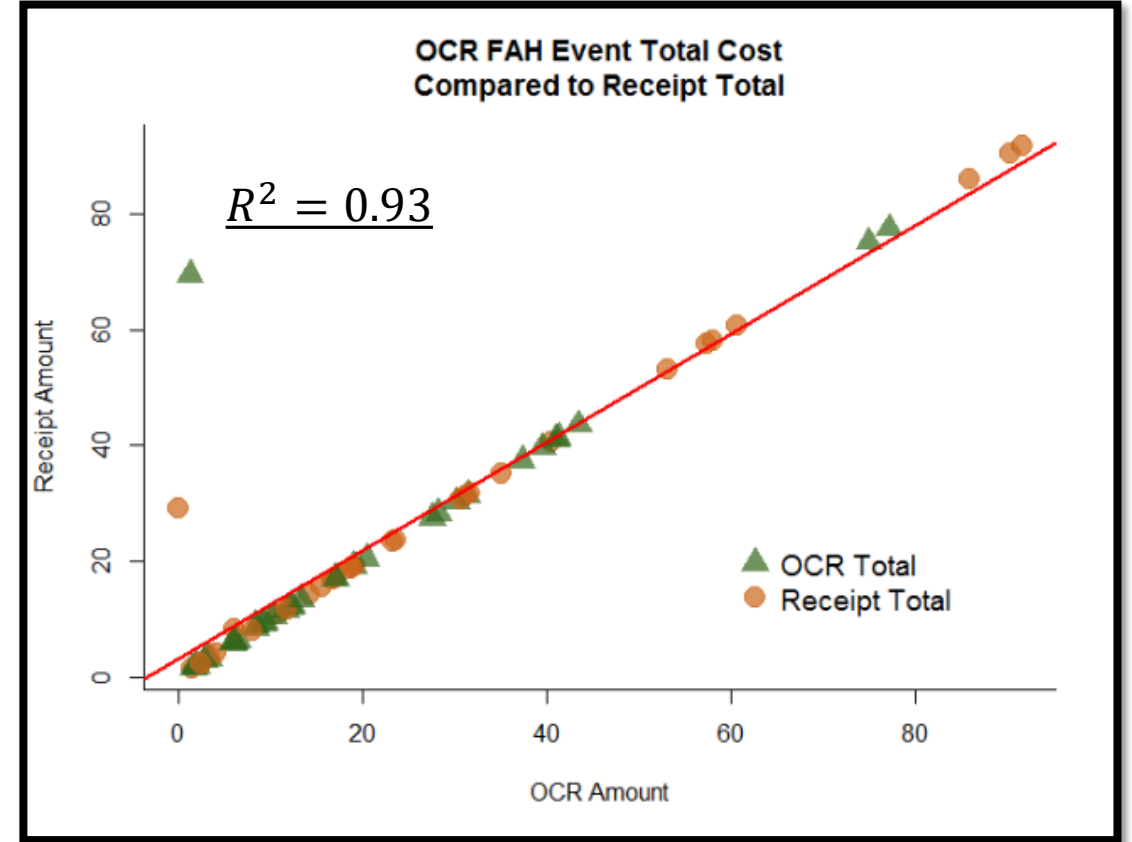
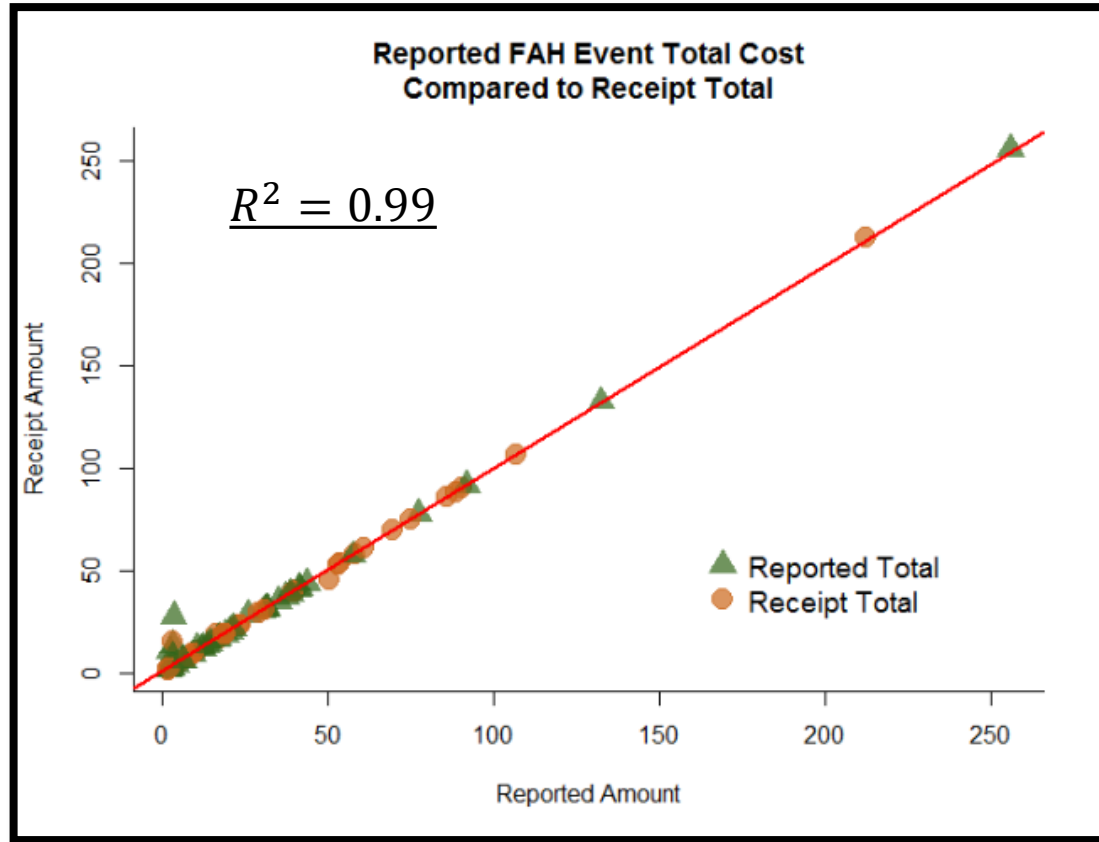
Visualizing FAFH Item Counts



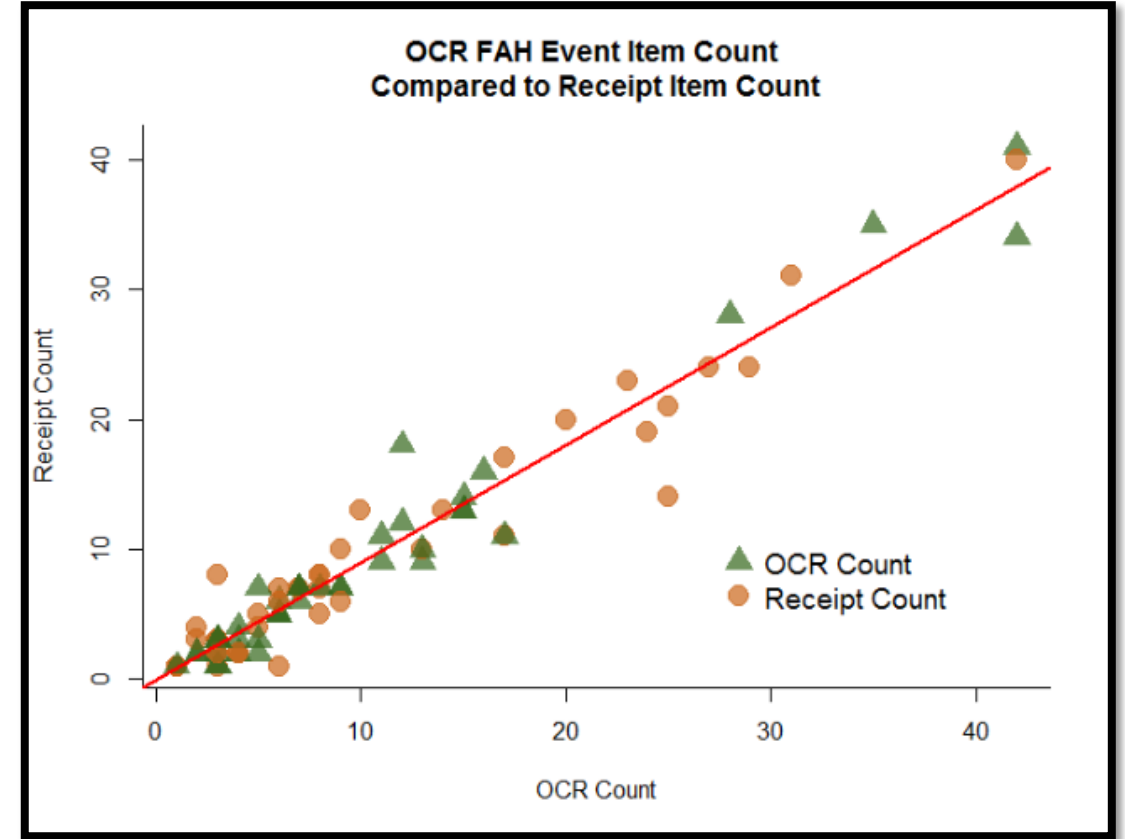
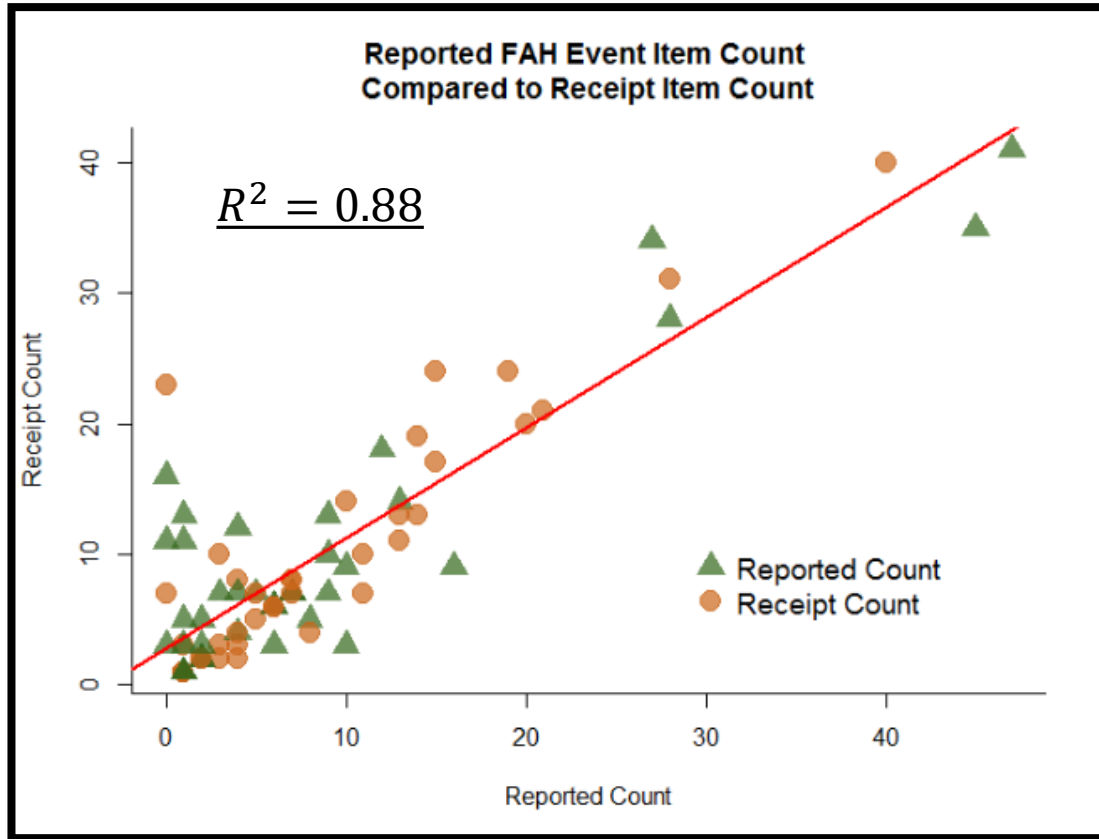
Visualizing FAFH Tax



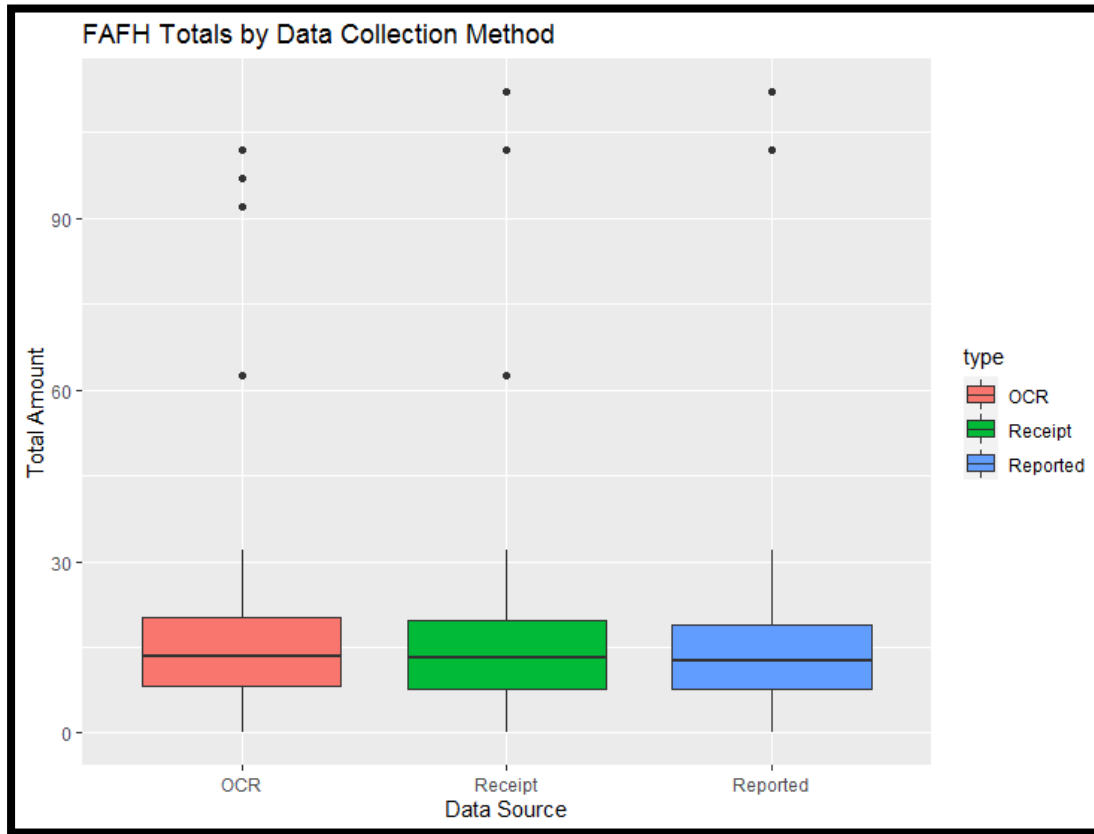
Visualizing FAH Totals



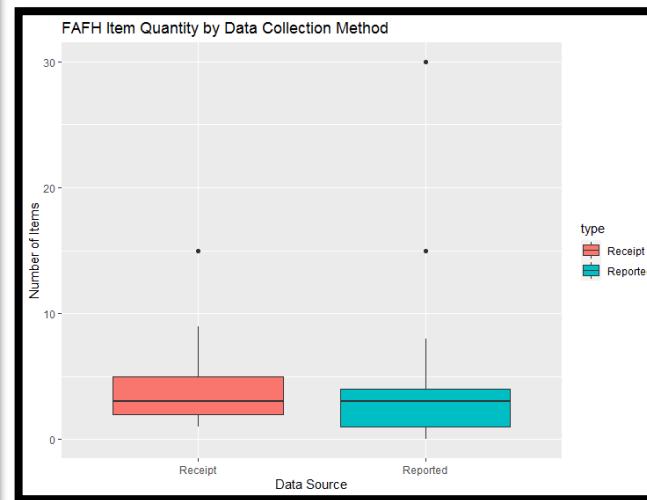
Visualizing FAH Item Counts



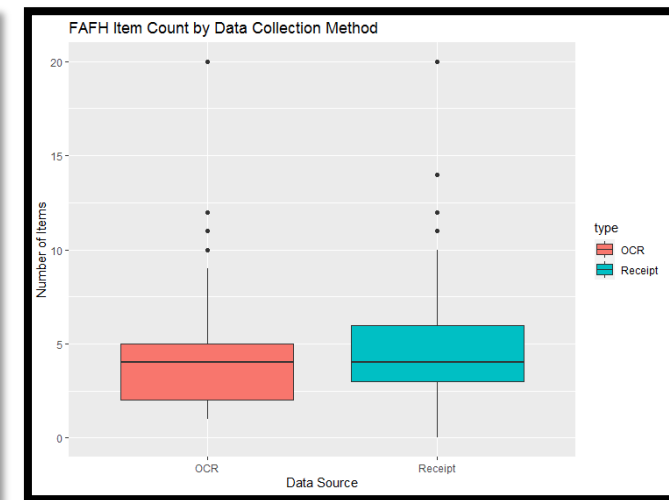
Visualizing Variations FAFH



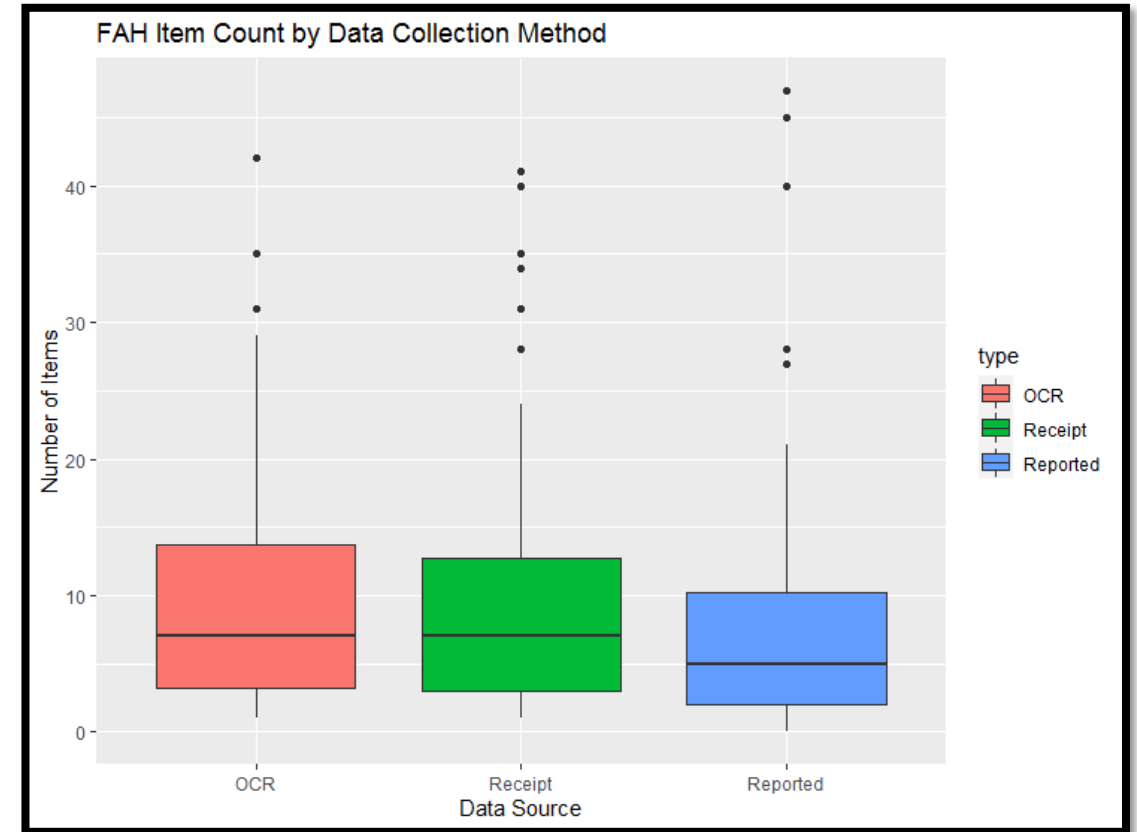
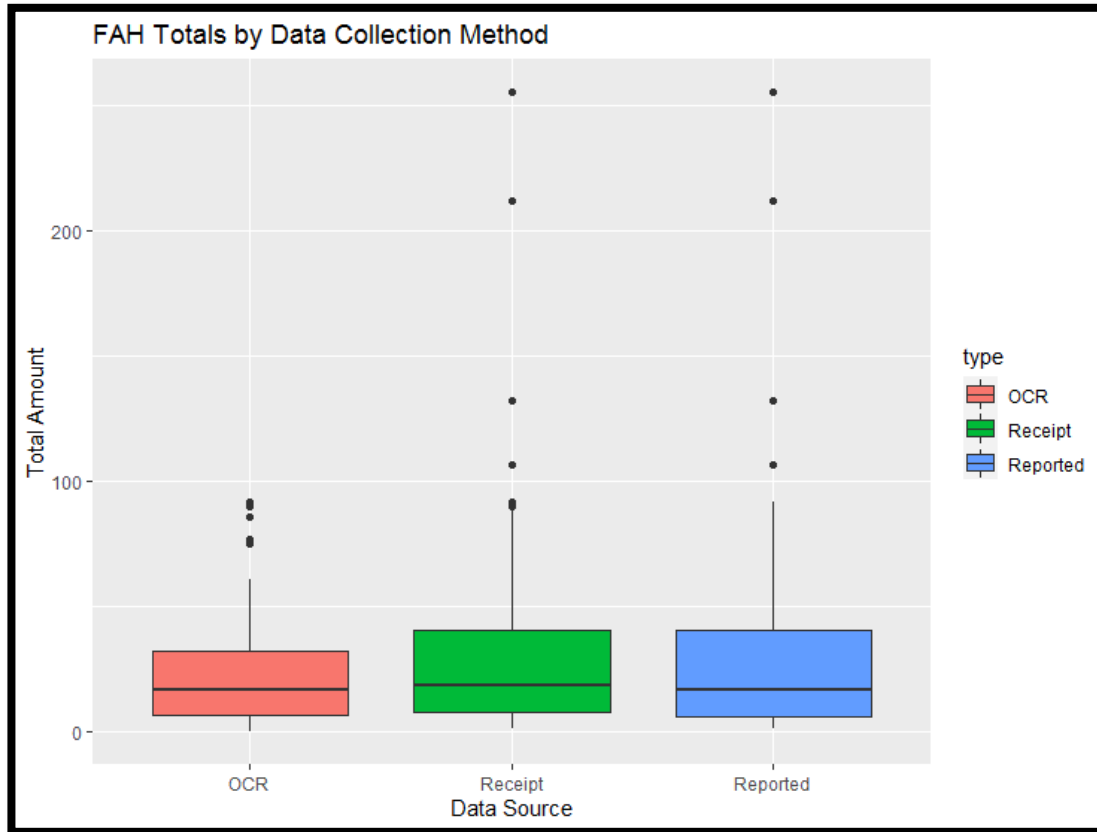
Total Item Quantity
(Reports vs Receipts)



Total Item Counts
(OCR vs Receipts)



Visualizing Variations FAH



Conclusions

- ❖ 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.
- ❖ We found the technology to be capable of capturing text data, and in many cases accurately so.
- ❖ instances where OCR may not be as accurate as respondents' reported data and this is highly dependent of the quality of the original image.
- ❖ 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.
- ❖ 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.

Thank You!
amkad@umich.edu

References

- 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.
- 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.
- Geisen, E., A. Richards, C. Strohm, and J. Wang (2011), "U.S. Consumer Expenditure Records Study Final Report," DCES Report.
- Google 2021. Tesseract OCR v 5.0.0. Lead Developer, Ray Smith. Repository Maintainer, Zdenko Podobny. <https://github.com/tesseract-ocr> (Accessed January, 2022).
- 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.

References Continued

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

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

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)

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 Department of Agriculture, Food and Nutrition Service. *Guide to Measuring Household Food Security, Revised 2000*. Gary Bickel, Mark Nord, Cristofer Price, William Hamilton, and John Cook. 2000. U.S. Department of Agriculture, Food and Nutrition Service, Alexandria VA.

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).