Leveraging Paradata to Assess Respondent Behavior and Data Quality in the CE Online Diary Survey

Graham Jones & Parvati Krishnamurty, BLS

FedCASIC 2023
April 12, 2023

Agenda

- **What is Paradata?**

- **Background of CE Online Diary**

- **Use in measuring activity**

- **Analysis Results:**

  - Diary Use and Logins

  - Devices and Operating systems

  - Time spent in the Diary

  - Relation to CE Data Quality

- **Next Steps for CE Paradata use**

What are Paradata?

- Data generated as a by-product of the survey's a collection process

- Analysis of paradata can lead to…

  - ▶ Improved data collection efficiency

  - ▶ Reduced survey fielding costs

  - ▶ Better understanding of measurement error

CE Online Diary Background

Development of CE Online Diary Instrument

Large Scale Feasibility Test of Online Diaries (LSF) 10/2019 – 4/2020

CE Online Diary Supplemental Use 6/2020

CE Online Diary in Production 7/2022

CE Online Diary Paradata

- **Events and user actions recorded along with event details in tabular form.**
  - ▶ Each event (action) is recorded as its own row in the data
  - ▶ Each event (action) is time-stamped
  - ▶ All rows have a value describing the event "type"

- **Event Types:**
  - Login
  - Hyperlink (click)
  - Failed Login
  - Next Action
  - Logout

  - Previous Action
  - Entry
  - Field Change
  - Exit
  - Error Trigger

CE Online Diary Paradata

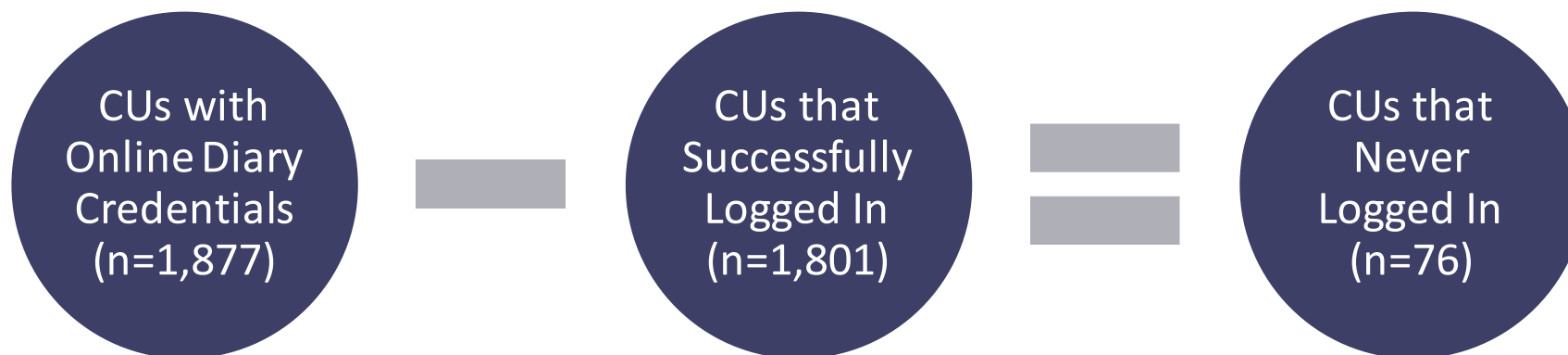|	respondent~l	event_ordi~l	event_time	type	event_id	event_page	environmen~l	environment_accept_language	environme~ht	environme~nt	environmen~h
1	1	1	2021-01-16 20:42:08	login			1	en-US,en;q=0.9	937	Mozilla/5.0…	1920
2	1	2	2021-01-16 20:42:11	entry	main	post_login	NA		NA		NA
3	1	3	2021-01-16 20:42:36	next_action	main	post_login	NA		NA		NA
4	1	4	2021-01-16 20:42:36	hyperlink	main	post_login	NA		NA		NA
5	1	5	2021-01-16 20:42:37	entry	main	diary	NA		NA		NA
6	1	6	2021-01-16 20:42:37	exit	main	post_login	NA		NA		NA
7	1	7	2021-01-16 20:42:42	hyperlink	main	diary	NA		NA		NA
8	1	8	2021-01-16 20:46:41	exit	main	diary	NA		NA		NA
9	1	9	2021-01-16 20:50:02	logout			NA		NA		NA
10	1	10	2021-01-18 05:53:49	login			2	en-US,en;q=0.9	937	Mozilla/5.0…	1920
11	1	11	2021-01-18 05:56:23	logout			NA		NA		NA
12	1	12	2021-01-20 02:15:25	login			3	en-US,en;q=0.9	937	Mozilla/5.0…	1920
13	1	13	2021-01-20 02:17:43	logout			NA		NA		NA
14	1	14	2021-01-22 06:04:57	login			4	en-US,en;q=0.9	937	Mozilla/5.0…	1920
15	1	15	2021-01-22 06:04:58	entry	main	diary	NA		NA		NA
16	1	16	2021-01-22 06:05:03	hyperlink	main	diary	NA		NA		NA
17	1	17	2021-01-22 06:05:05	field_change	main	diary	NA		NA		NA
18	1	18	2021-01-22 06:05:12	field_change	main	diary	NA		NA		NA
19	1	19	2021-01-22 06:05:19	field_change	main	diary	NA		NA		NA
20	1	20	2021-01-22 06:05:29	field_change	main	diary	NA		NA		NA
21	1	21	2021-01-22 06:05:35	field_change	main	diary	NA		NA		NA
22	1	22	2021-01-22 06:05:35	field_change	main	diary	NA		NA		NA

Paradata Analysis Results

1. **Diary Use and Logins**

2. Devices and Operating systems

3. Time spent in the Diary

4. Relation to CE Data Quality

Online Diary Use

- **1,877 unique Consumer Units (CUs) with online diary credentials**
  - ▶ **1,801 successfully logged in**
  - ▶ 76 never successfully logged in

CUs with Online Diary Credentials (n=1,877) — CUs that Successfully Logged In (n=1,801) = CUs that Never Logged In (n=76)
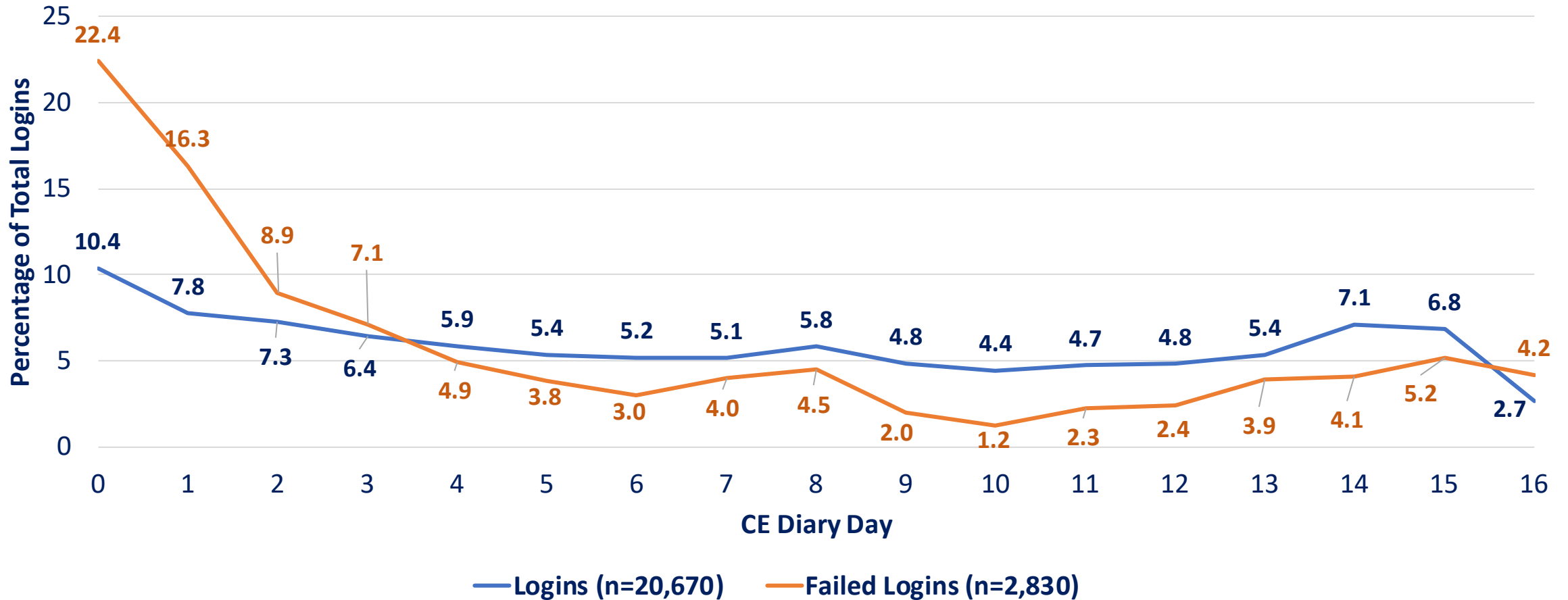
BLS

Login Activity

CUs with successful logins (n=1,801)

- Accounted for 22,100 **logins** in 2021
- Had a **mean** login total of 12.3
- Had a **median** login total of 10

CUs with at least one login failure (n=583)

- Accounted for 3,160 **login failures** in 2021
- Had a **mean** login failure total of 5.4
- Had a **median** login failure total of 4

Percentage of Logins by Diary Day in 2021

Line chart showing Percentage of Total Logins (y-axis, 0 to 25) by CE Diary Day (x-axis, 0 to 16).

Logins (n=20,670) (blue): 10.4, 7.8, 7.3, 6.4, 5.9, 5.4, 5.2, 5.1, 5.8, 4.8, 4.4, 4.7, 4.8, 5.4, 7.1, 6.8, 2.7

Failed Logins (n=2,830) (orange): 22.4, 16.3, 8.9, 7.1, 4.9, 3.8, 3.0, 4.0, 4.5, 2.0, 1.2, 2.3, 2.4, 3.9, 4.1, 5.2, 4.2

Paradata Analysis Results

1. Diary Use and Logins

2. **Devices and Operating systems**
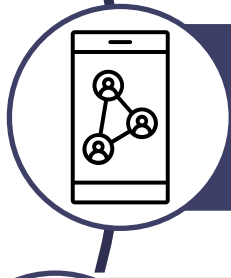
3. Time spent in the Diary

4. Relation to CE Data Quality

Paradata Analysis: Device Use in 2021

Desktop Only User

- All actions recorded in the paradata associated with a desktop
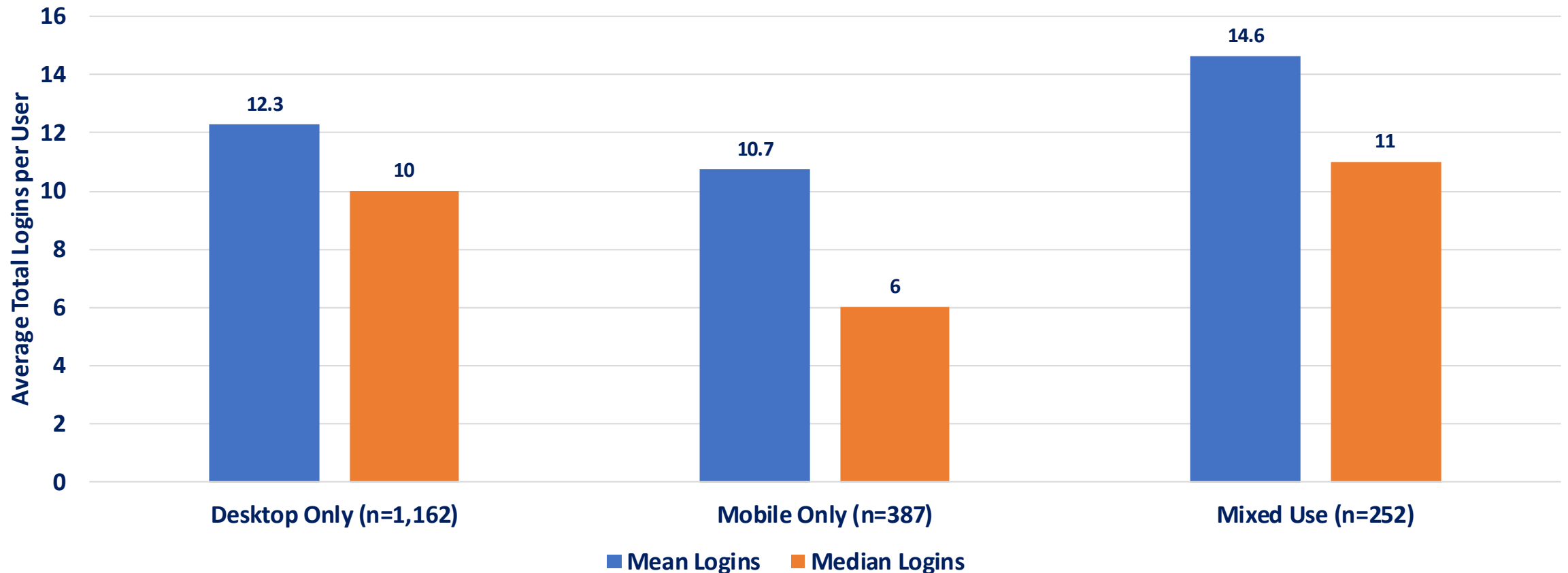- Accounted for 1,162 or 64.5% of CUs

Mobile Only User

- All actions recorded in the paradata associated with a mobile device
- Accounted for 387 or 21.5% of CUs

Mixed Device User

- Actions recorded in the paradata split between desktop and mobile devices.
- Accounted for 252 or 14.0% of CUs

BLS

Login Activity Comparison by Device Use

Bar chart titled "Login Activity Comparison by Device Use" with y-axis labeled "Average Total Logins per User" ranging from 0 to 16.

- Desktop Only (n=1,162): Mean Logins 12.3, Median Logins 10
- Mobile Only (n=387): Mean Logins 10.7, Median Logins 6
- Mixed Use (n=252): Mean Logins 14.6, Median Logins 11

Legend: Mean Logins (blue), Median Logins (orange)

Paradata Analysis Results

1. Diary Use and Logins
2. Devices and Operating systems
3. **Time spent in the Diary**
4. Relation to CE Data Quality

Paradata Analysis: Time Spent in the Diary

■ Time Spent in the Diary Per User $= \dfrac{\sum_{t=1}^{n}(Event_t - Event_{t+1})}{60}$

■ Action/Event Types:

- Login
- Hyperlink (click)
- Failed Login
- Next Action
- Logout

- Previous Action
- Entry
- Field Change
- Exit
- Error Trigger
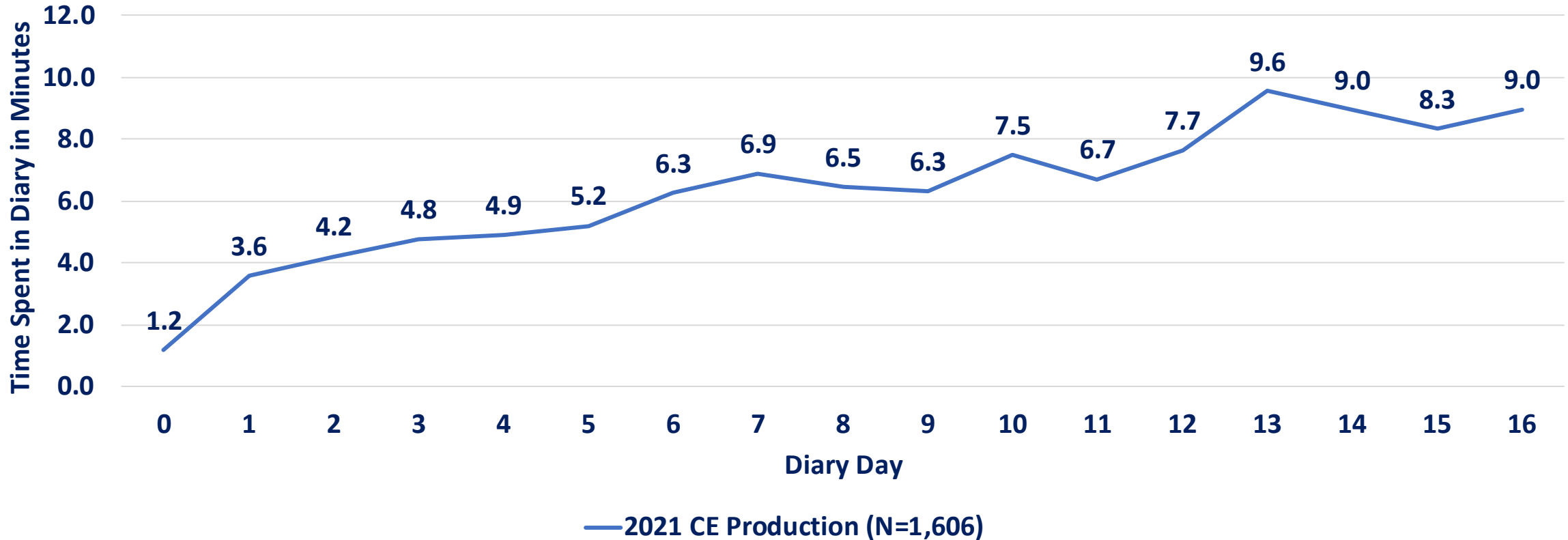
Paradata Analysis: Time Spent in the Diary

- 1,606 CUs had sufficient diary activity for the total time calculation in 2021.
  - Mean of 33.5 minutes
  - Median of 24.2 minutes

Time Spent in the Diary by Device Type

Desktop Only User
- Mean of **36.7 minutes**
- Median of **27.5 minutes**

Mobile Only User
- Mean of **19.7 minutes**
- Median of **10.3 minutes**

Mixed Device User
- Mean of **35.7 minutes**
- Median of **26.5 minutes**

Median Time Spent in Diary per Diary Day

Chart: Time Spent in Diary in Minutes (y-axis, 0.0 to 12.0) vs. Diary Day (x-axis, 0 to 16)

Data points for 2021 CE Production (N=1,606):
- Day 0: 1.2
- Day 1: 3.6
- Day 2: 4.2
- Day 3: 4.8
- Day 4: 4.9
- Day 5: 5.2
- Day 6: 6.3
- Day 7: 6.9
- Day 8: 6.5
- Day 9: 6.3
- Day 10: 7.5
- Day 11: 6.7
- Day 12: 7.7
- Day 13: 9.6
- Day 14: 9.0
- Day 15: 8.3
- Day 16: 9.0
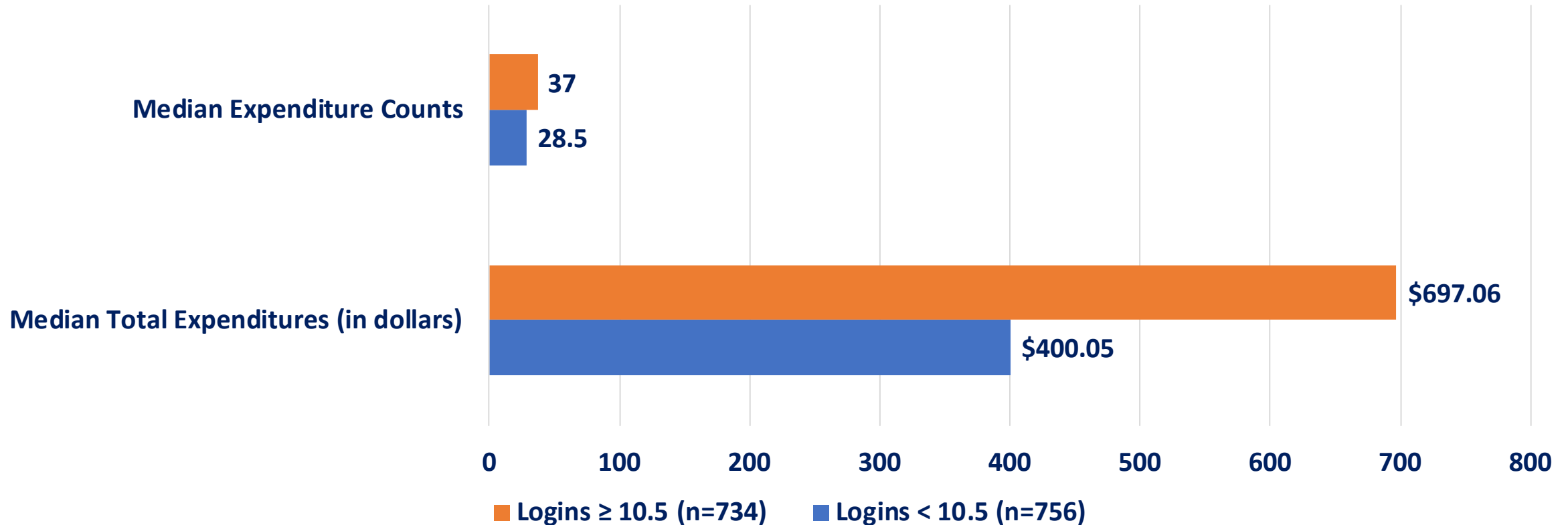
— 2021 CE Production (N=1,606)
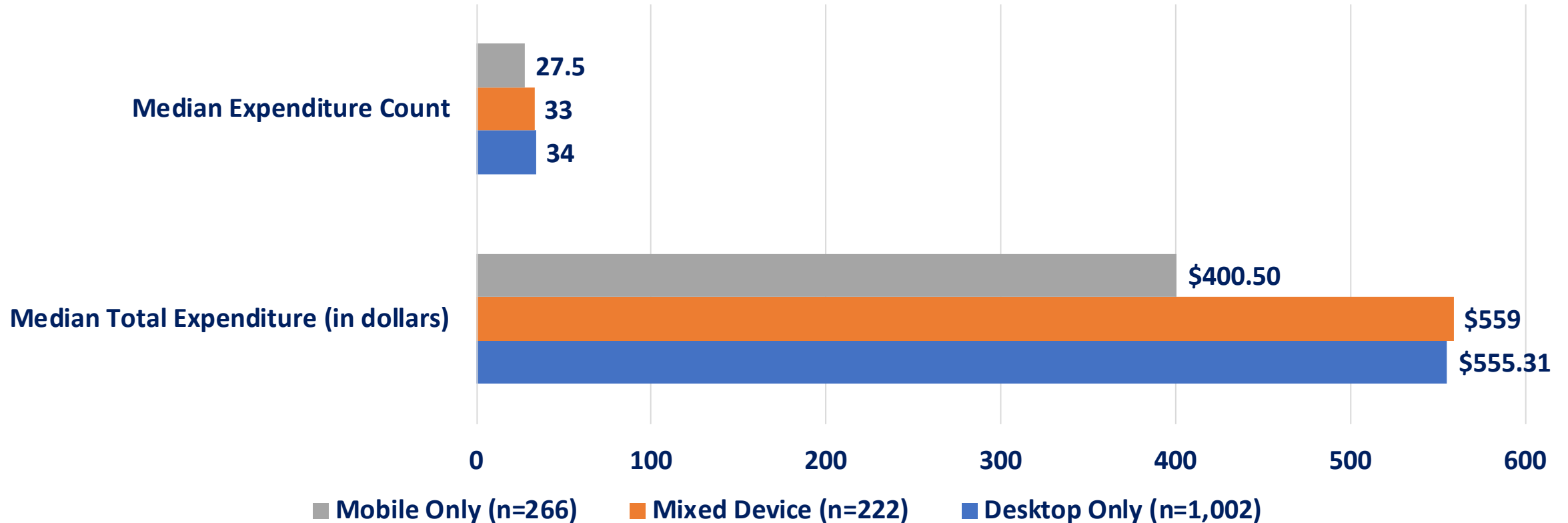
Paradata Analysis Results

1. Diary Use and Logins

2. Devices and Operating systems

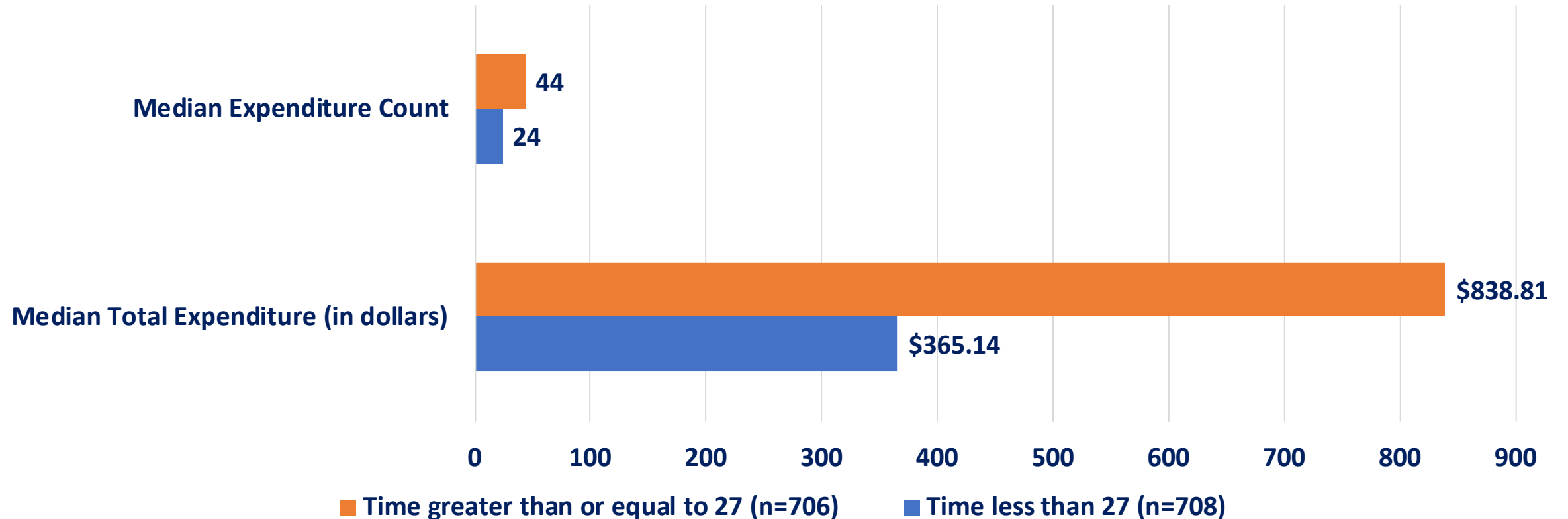3. Time spent in the Diary

4. **Relation to CE Data Quality**

Login Activity Relation to CE Data Quality

Median Expenditure Counts
- Logins ≥ 10.5 (n=734): 37
- Logins < 10.5 (n=756): 28.5

Median Total Expenditures (in dollars)
- Logins ≥ 10.5 (n=734): $697.06
- Logins < 10.5 (n=756): $400.05

Legend:
- Logins ≥ 10.5 (n=734)
- Logins < 10.5 (n=756)

Device Use Relation to CE Data Quality

Median Expenditure Count
- 27.5 (Mobile Only)
- 33 (Mixed Device)
- 34 (Desktop Only)

Median Total Expenditure (in dollars)
- $400.50 (Mobile Only)
- $559 (Mixed Device)
- $555.31 (Desktop Only)

0   100   200   300   400   500   600

■ Mobile Only (n=266)   ■ Mixed Device (n=222)   ■ Desktop Only (n=1,002)

BLS

Time Spent in Diary Relation to CE Data Quality

End of Paradata Analysis Results

Paradata Analysis Conclusions

- Granular detail of paradata allows researchers to look at diary keeping behavior from new perspective.

- Greater time spent in the diary and login activity were associated with higher quality data.

- Respondents using non-mobile devices outperformed those only using a mobile device.

Paradata Analysis Next Steps

- Incorporate average total logins and median total time spent in the diary into the CE Data Quality Profile as standard metrics.

- Continue to tweak the calculation of 'time spent in diary' for the most precise estimate.

- Revaluating values and naming convention for the variables in the paradata.

- Encouraging respondents to use non-mobile devices

Contact Information

Gray Jones

Economist

Consumer Expenditure Surveys

jones.graham@bls.gov