Preventing Data Falsification in Survey Research: Lessons from the Arab Barometer

Michael Robbins

Project Director
Arab Barometer
mdr7@princeton.edu

New Frontiers in Preventing, Detecting, and Remediating Fabrication in Survey Research

February 13, 2015

The Arab Barometer: About the Surveys

- 34,493 interviews
- 29 nationally representative surveys
- Three waves across 14 countries
  - Wave 1 (2006-7) in 7 countries
  - Wave 2 (2010-11) in 10 countries
  - Wave 3 (2012-14) in 12 countries

Surveys by Wave

Country	Wave 1	Wave 2	Wave 3
Algeria	✓	✓	✓
Bahrain	✓		
Egypt		✓	✓
Iraq		✓	✓
Jordan	✓	✓	✓
Kuwait			✓
Lebanon	✓	✓	✓
Libya			✓
Morocco	✓		✓
Palestine	✓	✓	✓
Saudi Arabia		✓	
Sudan		✓	✓
Tunisia		✓	✓
Yemen	✓	✓	✓

Methodolgy

- Face-to-face interviews in Arabic
- Area probability sampling
- Paper-and-Pencil Interviewing (PAPI)

Arab Barometer Goals

- Collect scientifically reliable survey data for academic and policy communities
- Increase knowledge about Arab publics and track political developments
- Build capacity for survey research in the region
- Collaborate with local scholars to design surveys covering important political, social, and regional topics
- Share best survey research practices

Problems of Duplication

The basic problem: Significant attempted data falsification via duplication in two countries in the 1st wave

Logic of Duplication

- Substantive items are more likely to be duplicated
  - Demographic variables can be falsified from census data
  - Geographic variables can be falsified from the sampling plan
  - Paradata can be falsified without effect on correlations

Measures to Avoid Duplication in Future Waves

- Changing our partners
- Structuring payments to disincentivize falsification by the firm
- Increased supervision from our regional headquarters
- Greater oversight of interviewer and supervisor training

Challenges of Preventing Data Falsification

- Face-to-face interviews pose challenges in oversight
  - Limitation of using computer-assisted personal interviewing (CAPI) in Arabic
  - Privacy concerns, especially in authoritarian contexts
  - Limitations of local teams
- Challenges in oversight of country teams
- Cooperative project with a goal of capacity-building

Further Tests for Data Falsification

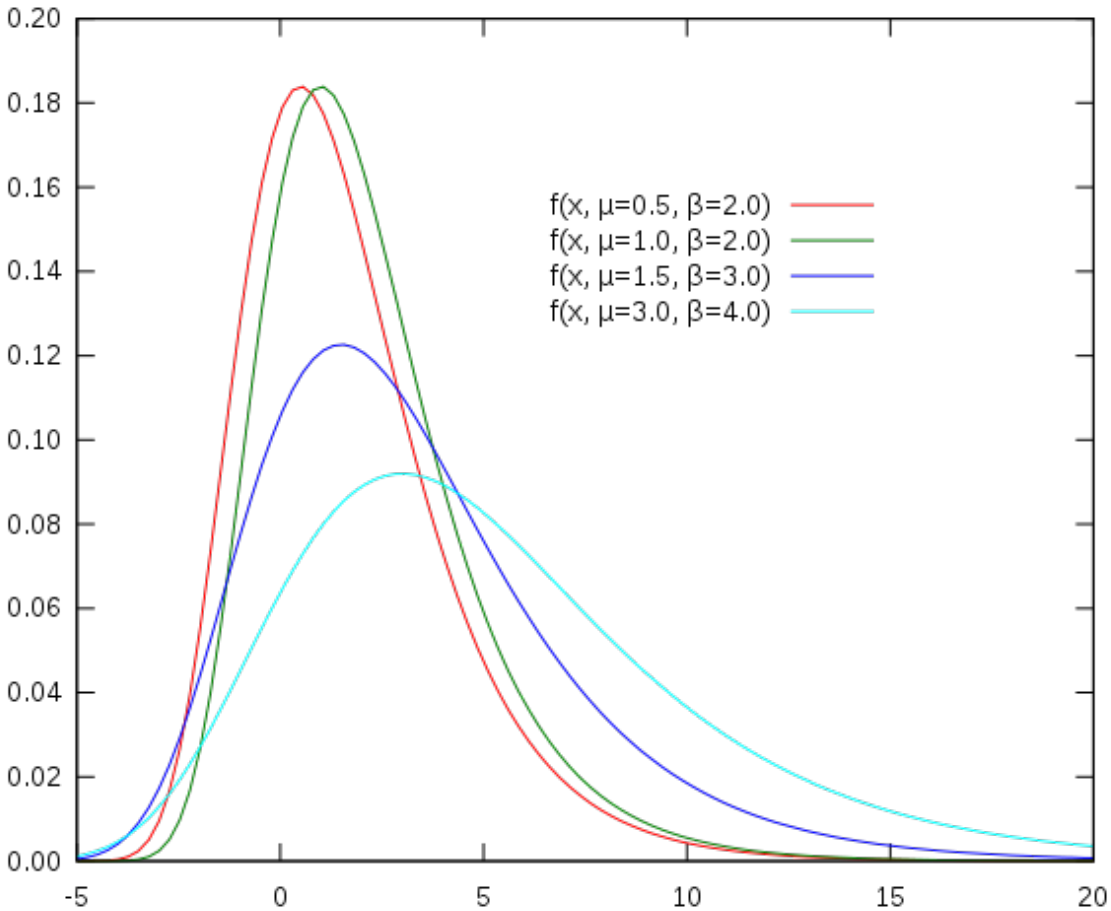
- Testing for exact duplicates is standard
- Testing for substantive duplicates is less common
- Even less common to test for near or partial duplicates

Maximum Percent Match Between Two Observations

- There is no statistical difference in the likelihood of an exact match and a 99% match between two observations
- A high percentage of near matches means data falsification is about as likely as with exact duplicates
- Plotting the maximum percent match for each observation across a survey should yield a known distribution

Maximum Percent Match Between Two Observations

- Maximum percent match should approximate a Gumbel distribution



Testing for Maximum Percent Match

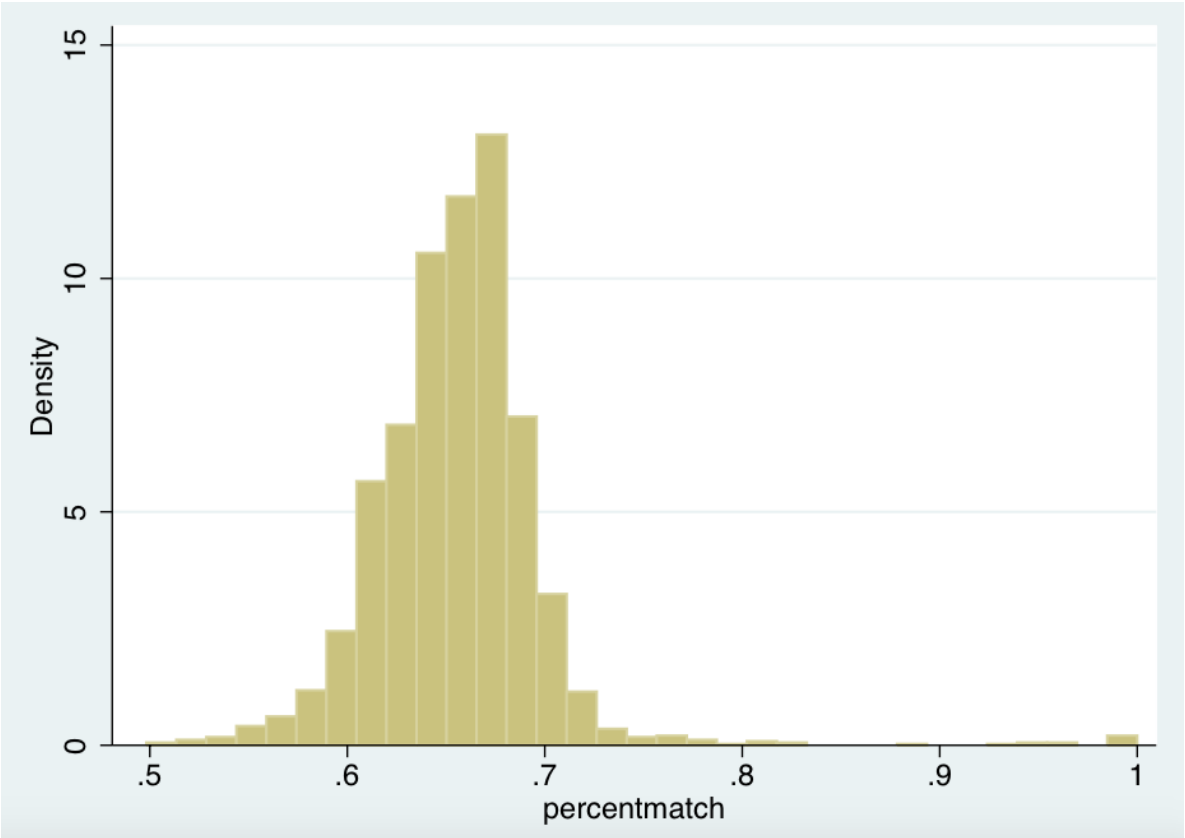
- **PercentMatch**: A publicly available Stata program to test for the percent match between two observations (developed with Noble Kuriakose)
  - Calculates the maximum percent match between two observations
  - Identifies the observation with which each observation has the highest percent match

Testing for Maximum Percent Match

- Potential Limitations
  - Items that were not administered to all respondents should be removed from analysis as they can bias the mean
  - Items with little to no variation may bias the mean
  - Homogeneous sub-populations can affect the distribution
  - More effective with a larger n and lengthy survey instrument

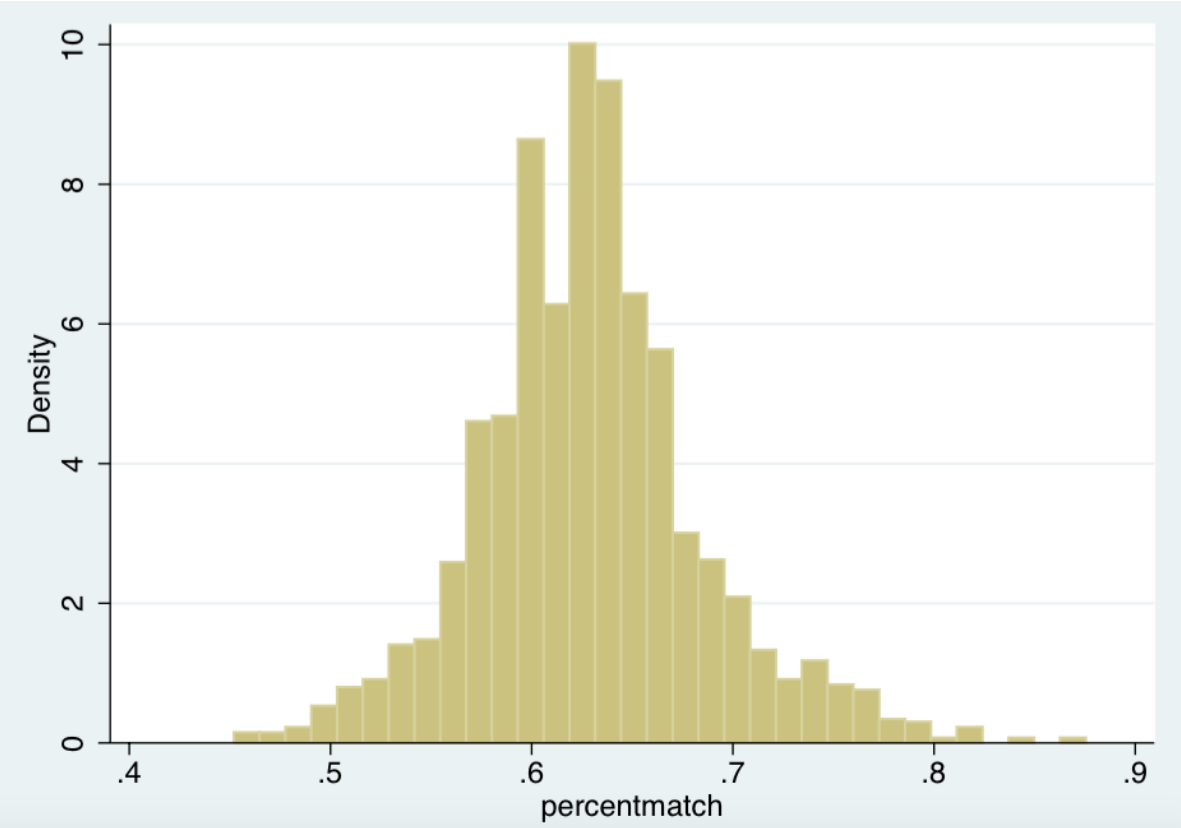
Percent Match in Practice

WVS6 for US (2011)



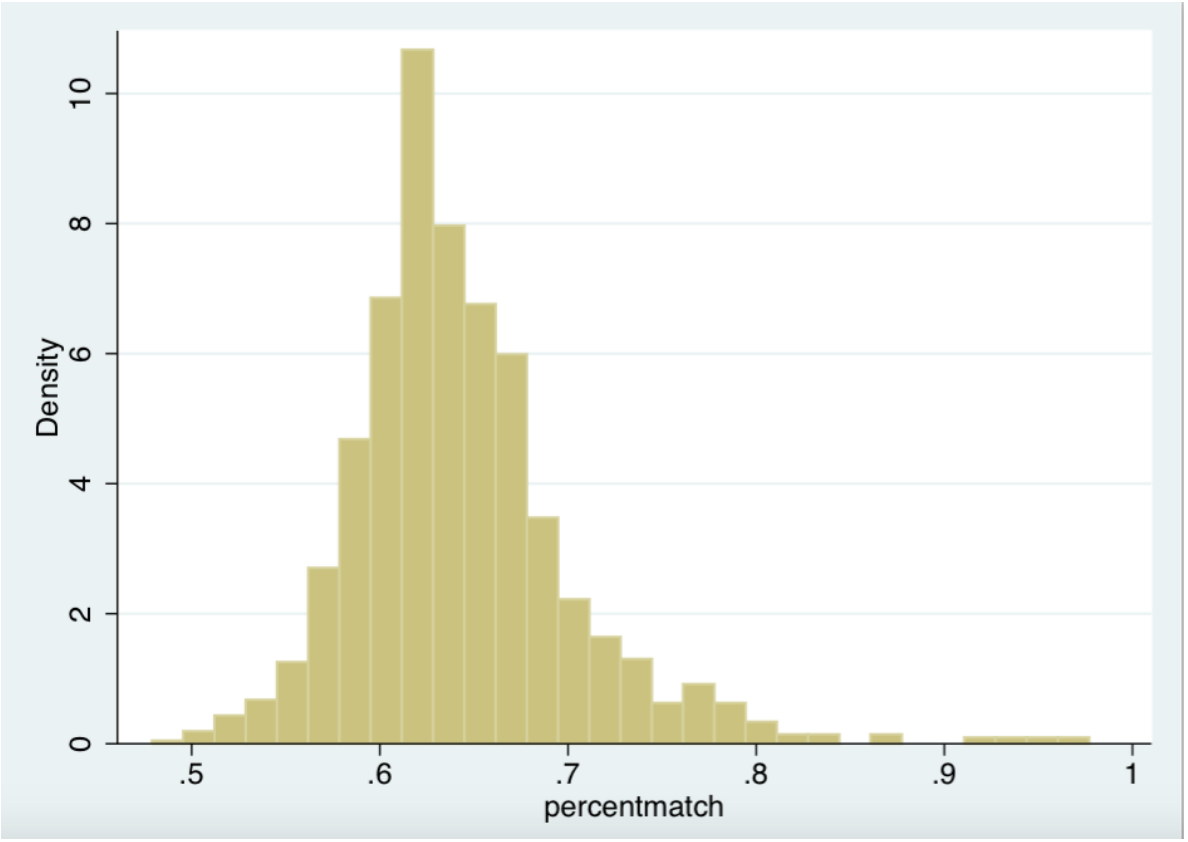
Percent Match in Practice

WVS6 for Germany (2013)



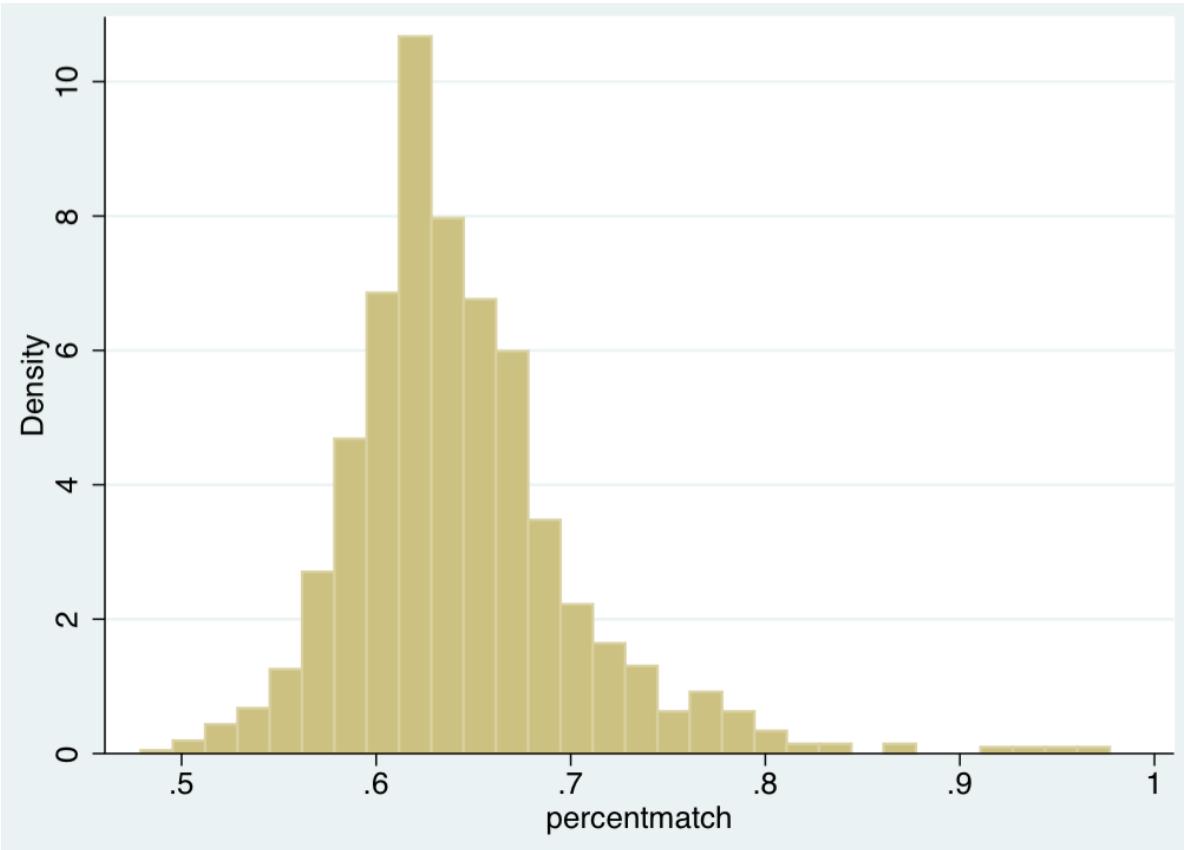
Percent Match in Practice

AB3 for Palestine (2012)



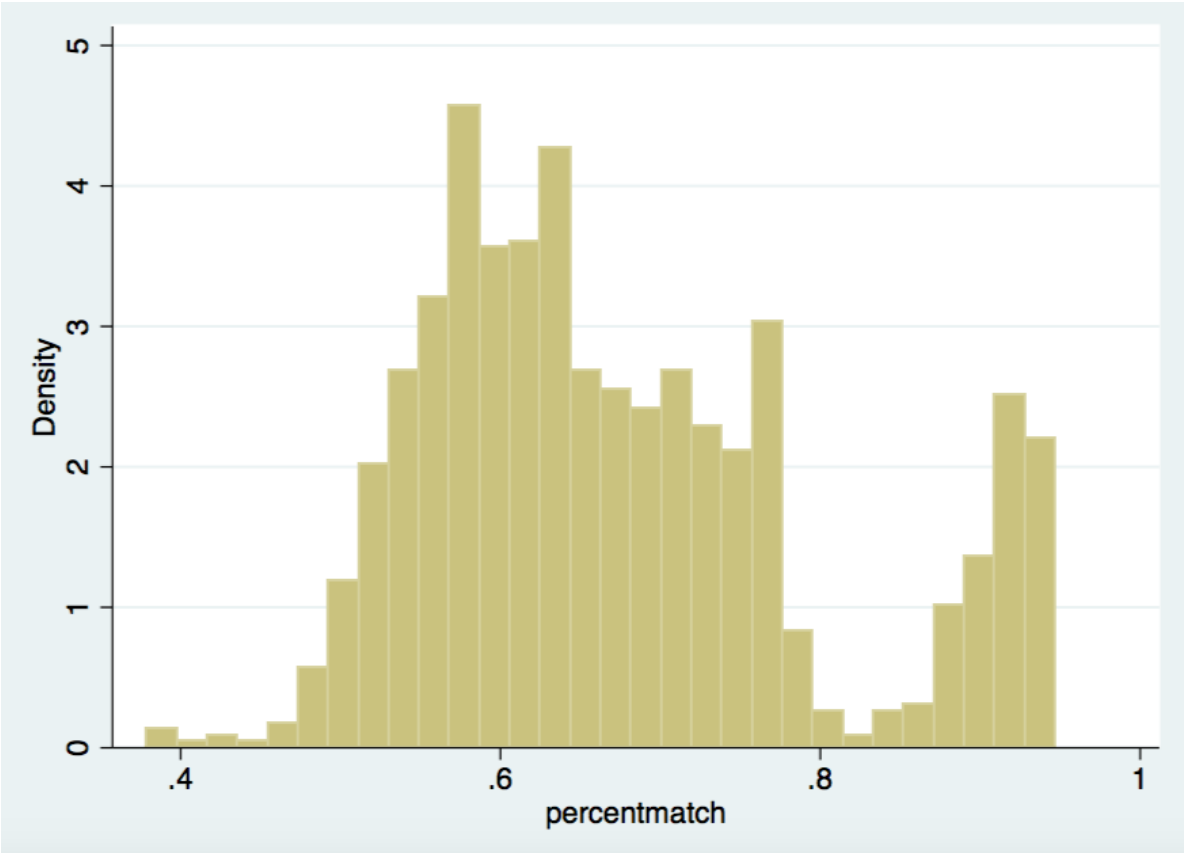
Percent Match in Practice

AB3 for Libya (2014)



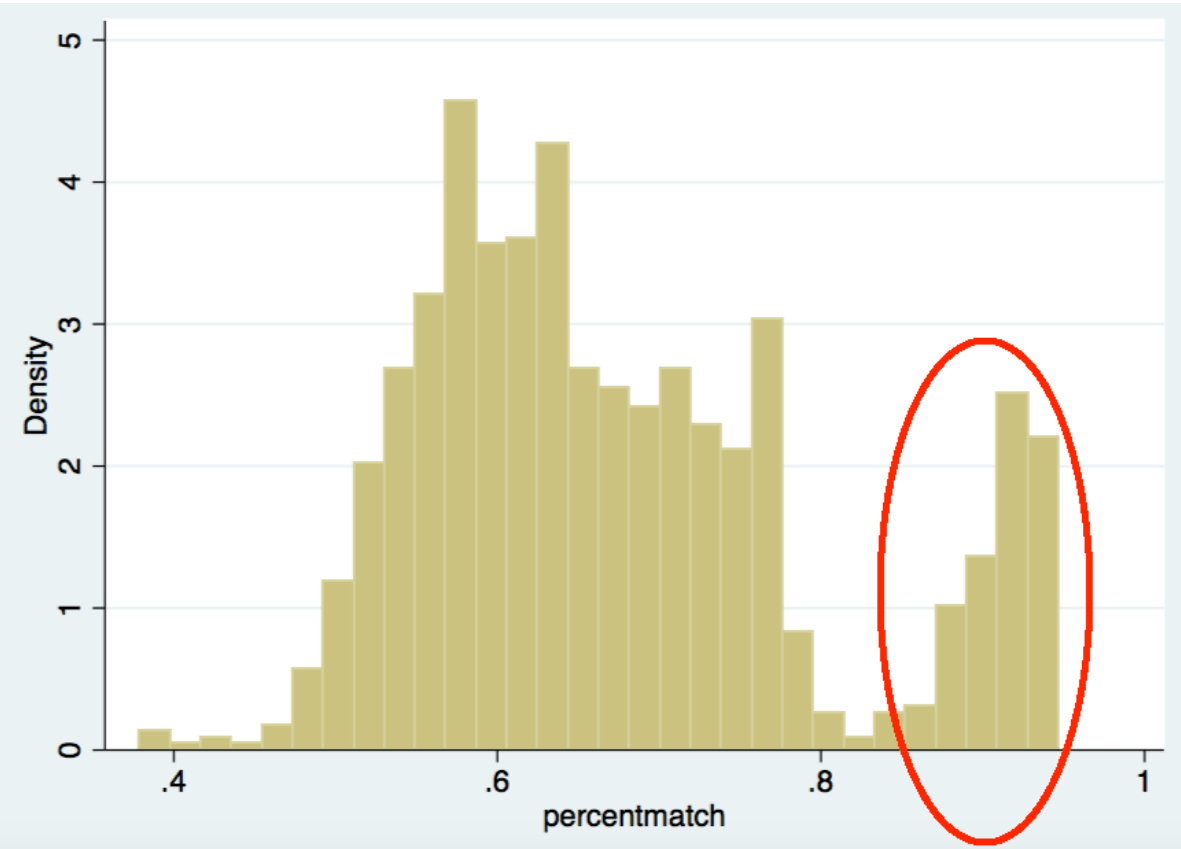
Percent Match in Practice

AB3 for Kuwait (2014)



Percent Match in Practice

AB3 for Kuwait (2014)



Common Link

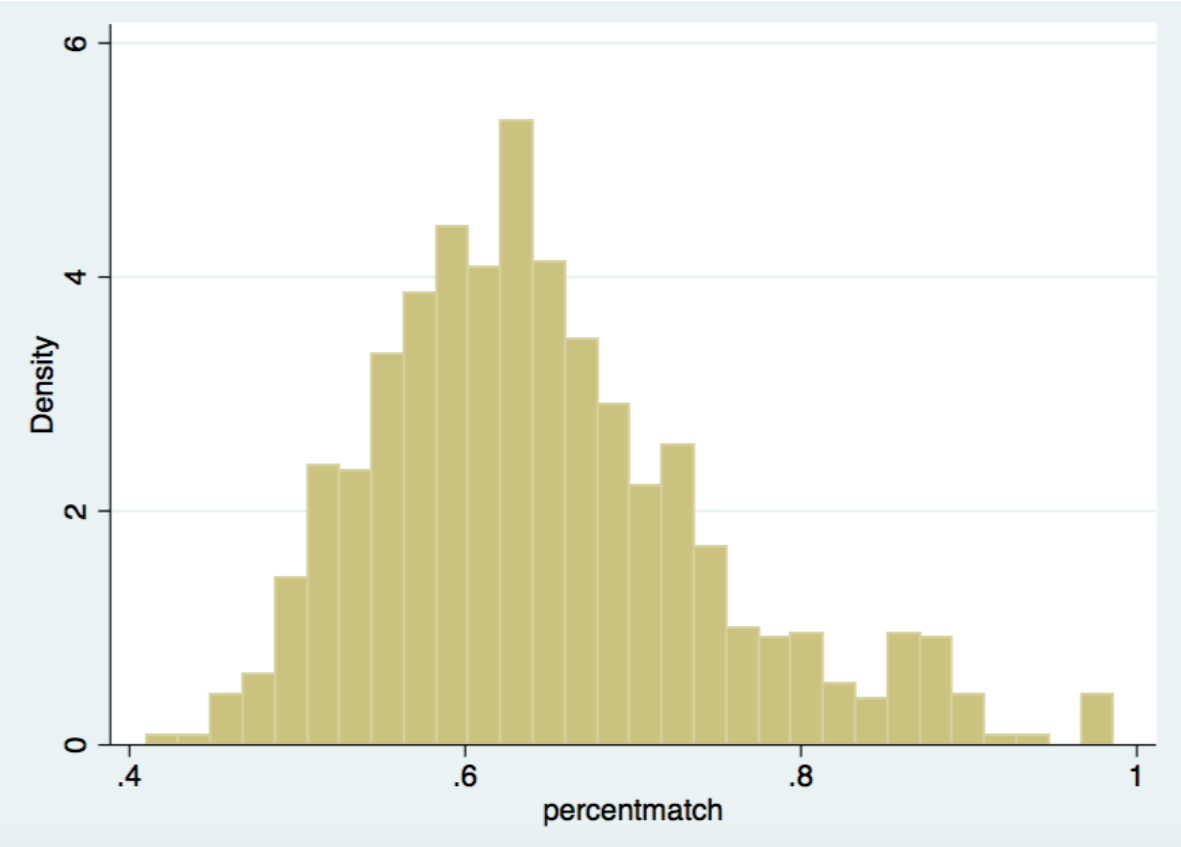
Interviewer	**Match $<$ 80%**	**Match $\geq$ 80%**
1310	*1*	*122*
1311	127	0
1312	72	0
1313	63	0
1314	113	0
1315	77	28
1316	115	0
1317	76	0
1318	37	0
1319	34	0
1320	61	4
1321	3	0
1322	84	0
1323	58	0
1324	62	11
1325	39	13
Total	**1,022**	**178**

Further Examination

- Interviewer 1310 had duplicated interviews but randomized every ≈10th variable
  - None of 1310's interviews had a maximum percent match greater than 95%
- Many expected correlations did not hold across his interviews unlike those of other interviewers
- Result was to eliminate all of 1310's interviews and those of others with a percent match of 80% or above

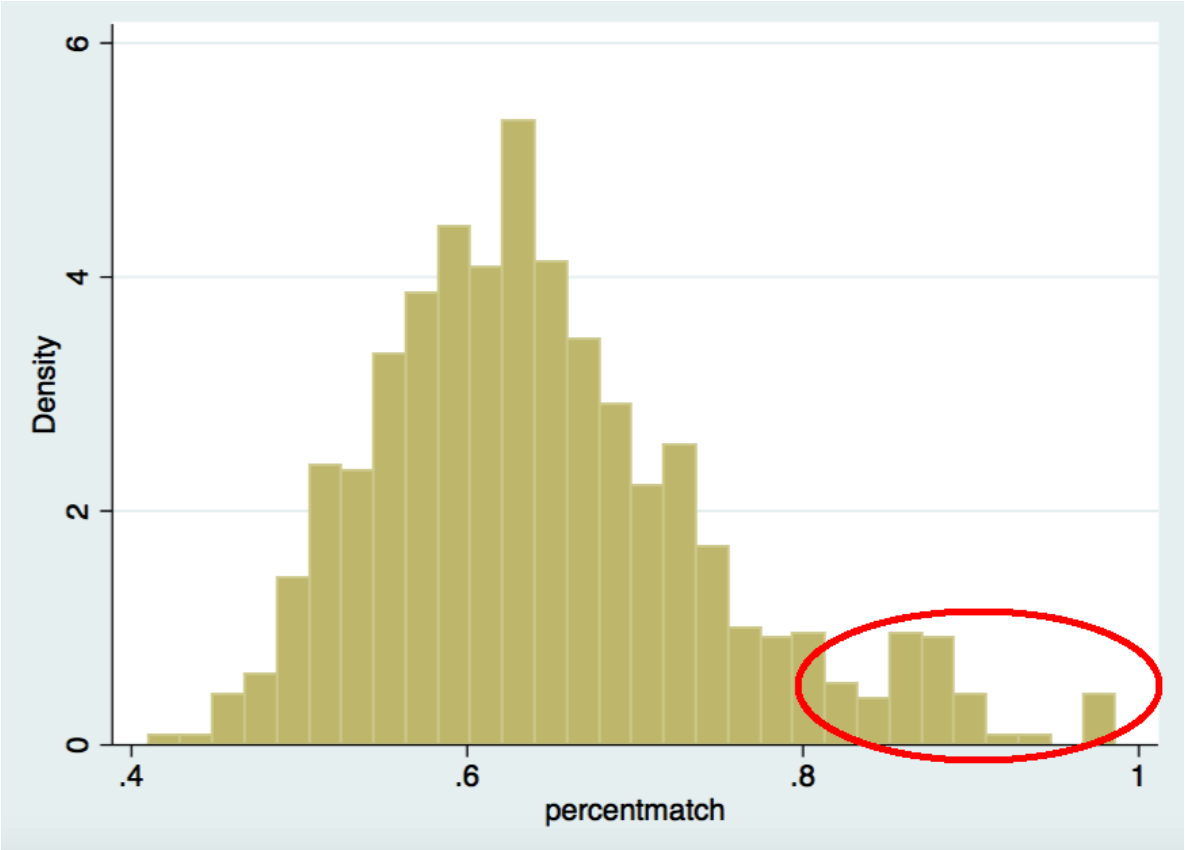
The Case of Morocco

AB3 for Morocco



The Case of Morocco

AB3 for Morocco



Percent Match by Interviewer

Interviewer	**% Match $<$ 80%**	**% Match $\geq$ 80%**
4	41	9
8	31	3
10	28	2
12	39	11
13	45	5
14	91	4
15	52	31
20	**21**	**19**
21	**23**	**17**
25	44	6

Note: Table includes only interviewers with a percent match $\geq$ 80%

Closer Examination of Interviews by 20 & 21

Support for Islamist Party by Interviewer

Interviewer	*Average % PJD Support*	*Max % PJD Support*
20 & 21	66.3	77.5
All others	10.1	24.1

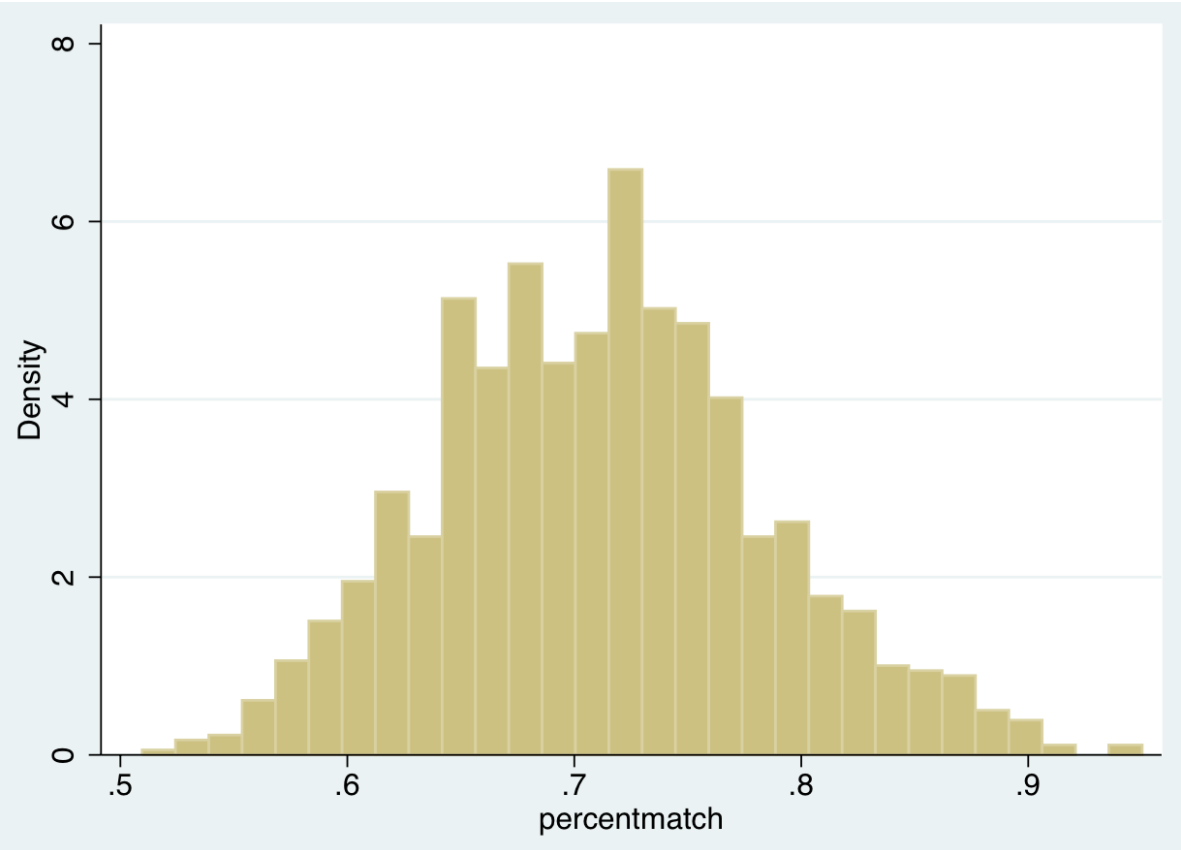
- Result was to eliminate all of their interviews conducted by these two interviewers

Gains in Data Quality

- Focus on falsification via duplication has yielded significant gains in data quality
  - Partners are incentivized to improve oversight of fieldworkers
  - Gains can be seen even in cases where other means of oversight are limited
    - Example: Comparison of two data sets from Algeria with the same Principal Investigator

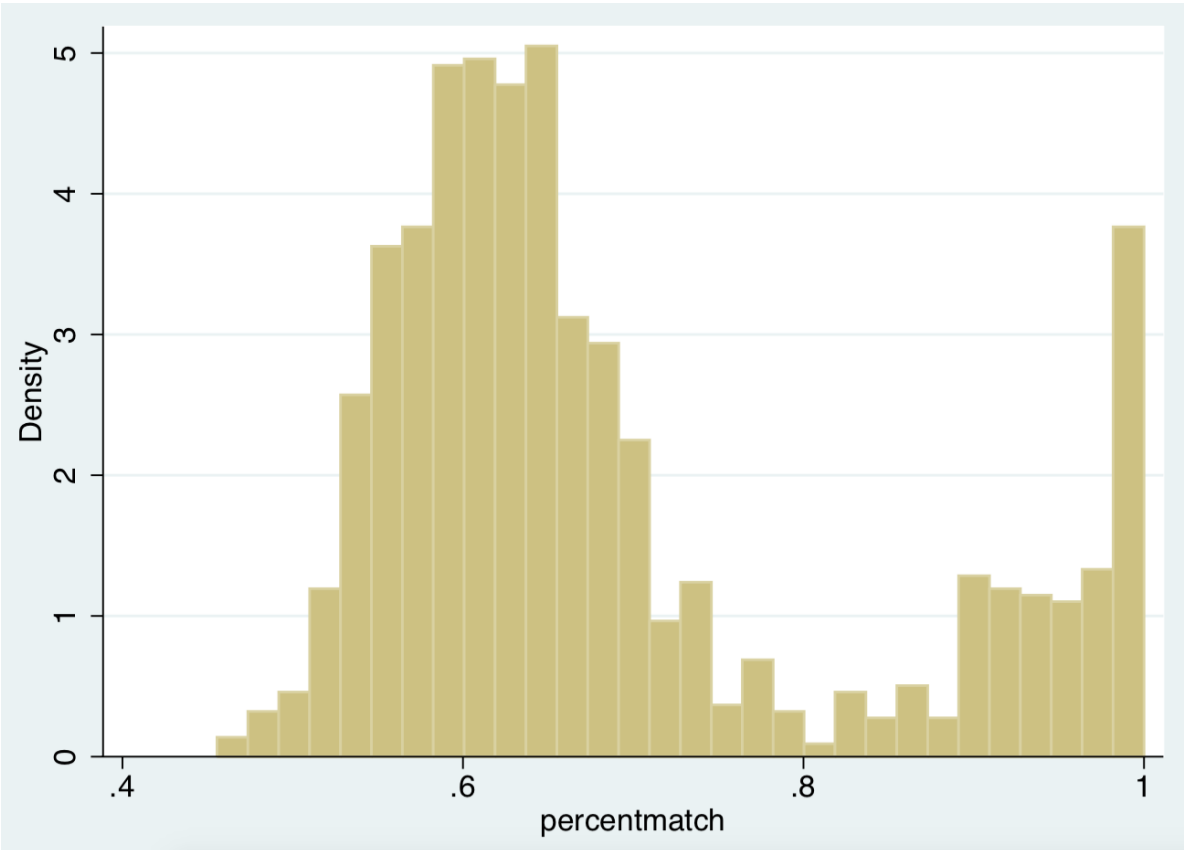
Percent Match in Algeria

Algeria in AB3 (2013)



Percent Match in Algeria

Algeria in WVS6 (2014)



Takeaways

- Program can be used to identify observations that have a greater risk of falsification
- Three proposed tests to limit likelihood of falsification
  - Maximum percent match approximates a Gumbel distribution with a mean of ≈ 0.7 or less
  - Expected correlations hold within the substantive variables and with demographic variables
  - Demographic variables approximately match sampling frame
- Combined these tests raise the costs of data falsification for firms to disincentivize cheating