# Measuring Consumption in 12 Minutes: Lessons from a Field Study in Rural Kenya

Natalie Naïri Quinn
with Geetika Nagpal, Kate Orkin

Lady Margaret Hall/OPHI
University of Oxford

7 November 2019

# Outline

# Background

A large-scale cluster-randomised RCT:

- AEA registration 996: *Direct and Interaction Effects of Cash Transfers and Psychological Interventions Promoting Future Orientation on Economic Outcomes*
- AEA registration 1484: *Impact of Cash Transfers on Income Distribution*
- Approx 10,000 households across 400 villages in rural parts of Siaya and Homa Bay counties in Western Kenya.

# Background

A large-scale cluster-randomised RCT:

- AEA registration 996: *Direct and Interaction Effects of Cash Transfers and Psychological Interventions Promoting Future Orientation on Economic Outcomes*
- AEA registration 1484: *Impact of Cash Transfers on Income Distribution*
- Approx 10,000 households across 400 villages in rural parts of Siaya and Homa Bay counties in Western Kenya.

A practical challenge:

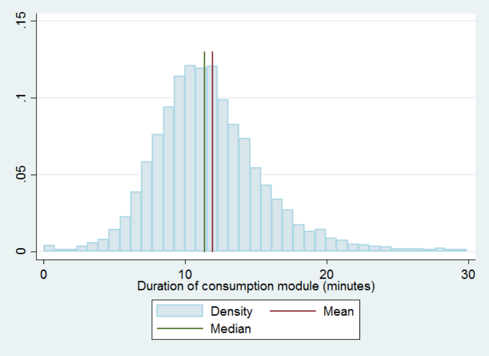
- We needed to measure consumption expenditure (aggregate)
- Survey time extremely tight (2hr total/household; many economic and psychological outcomes to measure)
- Clear information redundancy in consumption data: can we leverage that?

# Anticipating Results

- We got consumption module timing down to 12 minutes!
- Without entirely sacrificing data quality

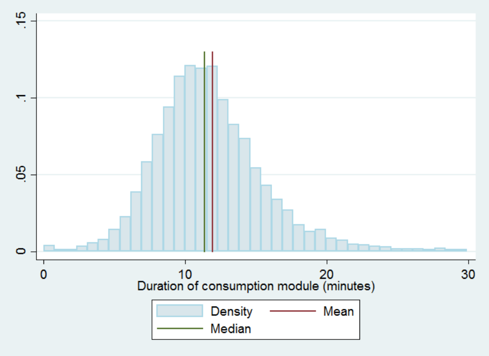
# Anticipating Results

- We got consumption module timing down to 12 minutes!
- Without entirely sacrificing data quality



# Anticipating Results

- We got consumption module timing down to 12 minutes!
- Without entirely sacrificing data quality




This paper:

- Assess validity to the extent possible
- Confident in our ability to estimate average treatment effects
- Still exploring implications for distributional inference

# Baseline

- Piloting survey (spring 2016): realisation of timing challenge
- Key literature: Beegle, De Weerdt, Friedman & Gibson (2012, JDE)
- Their subset approach appeared most promising
- Identification of max-consumption-share items using KIHBS 2005/06 (food, non-food non-durables)

# Baseline

- Piloting survey (spring 2016): realisation of timing challenge
- Key literature: Beegle, De Weerdt, Friedman & Gibson (2012, JDE)
- Their subset approach appeared most promising
- Identification of max-consumption-share items using KIHBS 2005/06 (food, non-food non-durables)

Other (untested) actions to speed up module:

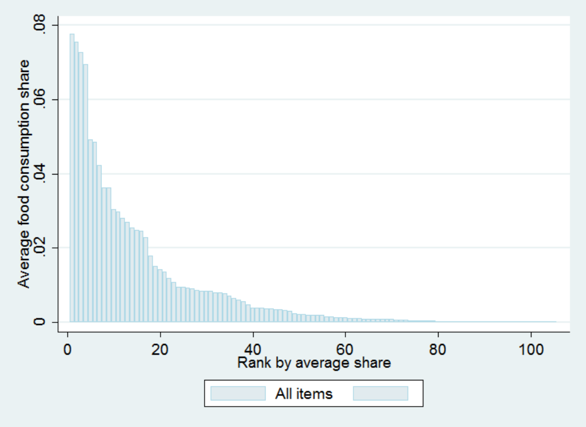
- Explored alternative ways to code SurveyCTO to most closely emulate paper-based implementation
- Explicit prompting of own-production, purchases, gifts and transfers but recorded collectively for each item
- Subset approach infeasible for durables, so implemented collapsed list

# Baseline: focus on food items

- Reference data: KIHBS 2005/06 Rachuonyo District (N=170, 107 food items)
- Average food consumption shares for each food item:

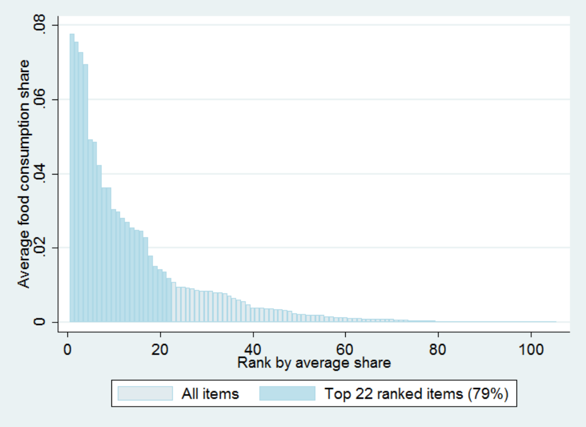
# Baseline: focus on food items

- Reference data: KIHBS 2005/06 Rachuonyo District (N=170, 107 food items)
- Average food consumption shares for each food item:



# Baseline: focus on food items

- Reference data: KIHBS 2005/06 Rachuonyo District (N=170, 107 food items)
- Average food consumption shares for each food item:




- Top 22 ranked items captured 79.35% of food consumption

# Baseline: focus on food items

No formal information analysis (yet), but some interesting observations:

- Included items
  - High-volume, high-frequency, low value (staples, eg maize, omena)
  - Lower-frequency, high value (luxuries, eg red meat)
- Excluded items
  - Low-volume, high-frequency, low value (eg salt, onions)
  - Very-low-frequency, high value (stronger luxuries, eg commercially processed foods)

# Baseline: focus on food items

No formal information analysis (yet), but some interesting observations:

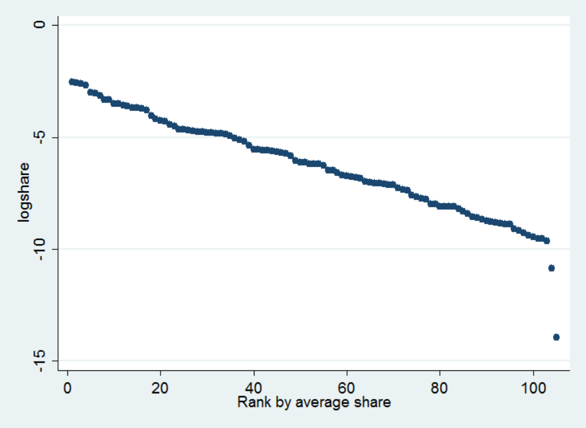
- Included items
  - High-volume, high-frequency, low value (staples, eg maize, omena)
  - Lower-frequency, high value (luxuries, eg red meat)
- Excluded items
  - Low-volume, high-frequency, low value (eg salt, onions)
  - Very-low-frequency, high value (stronger luxuries, eg commercially processed foods)

Some tweaks post piloting:

- Collapsing red meat; chicken and poultry; fresh milk
- Addition of mandazi and omena
- Anticipated coverage around 83%
- (Repeat analysis in Siaya added three further items)

# Aside

This is interesting:




- $\ln s = -2.736 - 0.066r$
- So measurement error varies with $e^{-\beta J}$ where $J$ is the number of items included

# Endline

- Time constraints even stronger: further cuts needed!
- Baseline data available to guide further cuts

# Endline

- Time constraints even stronger: further cuts needed!
- Baseline data available to guide further cuts

How to estimate scalefactors?

- KIHBS 2006/05 12 years old and samples not necessarily representative of the same population
- KIHBS 2015/16 not yet available
- Inspired by Pape & Mistiaen (2018), large sample size (>10,000 HH) allowed us to randomise excluded items, achieving a sample size of  300 for each
- Thus estimated scalefactor components

# Scale Factor Estimates

Food consumption

- Endline 18 core items
- Plus 3 randomly allocated extras
- Core captures 76.2% of food consumption
- Scalefactor 1.31

# Scale Factor Estimates

Food consumption

- Endline 18 core items
- Plus 3 randomly allocated extras
- Core captures 76.2% of food consumption
- Scalefactor 1.31

Non-food non-durable consumption

- Endline 9 core items
- Plus 1 randomly allocated extra
- Core captures 75.7% of non-food non-durable consumption
- Scalefactor 1.32

# Analysis: Estimation of Treatment Effects

- We did this in the context of a large-scale RCT...

# Analysis: Estimation of Treatment Effects

- We did this in the context of a large-scale RCT...
- ...Can we recover well-identified treatment effects?

# Analysis: Estimation of Treatment Effects

- We did this in the context of a large-scale RCT...
- ...Can we recover well-identified treatment effects?
- Households indexed $i$, items indexed $j$
- If 'true' consumption $c_i = \sum_j x_{ij} + \sum_j u_{ij} = x_i + \sum_j u_{ij}$

# Analysis: Estimation of Treatment Effects

- We did this in the context of a large-scale RCT...
- ...Can we recover well-identified treatment effects?
- Households indexed $i$, items indexed $j$
- If 'true' consumption $c_i = \sum_j x_{ij} + \sum_j u_{ij} = x_i + \sum_j u_{ij}$
- Let $u_{ij} = \alpha_{ij} x_i$, so $c_i = x_i(1 + \sum_j \alpha_{ij})$

# Analysis: Estimation of Treatment Effects

- We did this in the context of a large-scale RCT...
- ...Can we recover well-identified treatment effects?
- Households indexed $i$, items indexed $j$
- If 'true' consumption $c_i = \sum_j x_{ij} + \sum_j u_{ij} = x_i + \sum_j u_{ij}$
- Let $u_{ij} = \alpha_{ij} x_i$, so $c_i = x_i(1 + \sum_j \alpha_{ij})$
- $ATE = \mathbb{E}(c_i(1) - c_i(0)) =$
  $\mathbb{E}\left(x_i(1)(1 + \sum_j \alpha_{ij}(1)) - x_i(0)(1 + \sum_j \alpha_{ij}(0))\right)$

# Analysis: Estimation of Treatment Effects

- We did this in the context of a large-scale RCT...
- ...Can we recover well-identified treatment effects?
- Households indexed $i$, items indexed $j$
- If 'true' consumption $c_i = \sum_j x_{ij} + \sum_j u_{ij} = x_i + \sum_j u_{ij}$
- Let $u_{ij} = \alpha_{ij} x_i$, so $c_i = x_i(1 + \sum_j \alpha_{ij})$
- $ATE = \mathbb{E}(c_i(1) - c_i(0)) =$
  $\mathbb{E}\left(x_i(1)(1 + \sum_j \alpha_{ij}(1)) - x_i(0)(1 + \sum_j \alpha_{ij}(0))\right)$
- If $x_i$ is uncorrelated with $\sum_j \alpha_{ij}$ (testable) then
- $ATE = \mathbb{E}(x_i(1))\mathbb{E}(1 + \sum_j \alpha_{ij}(1)) - \mathbb{E}(x_i(0))\mathbb{E}(1 + \sum_j \alpha_{ij}(0))$
- Each part of which may be estimated.

# Analysis: Efficient Survey Design

- Sample size $= N$, number of items included $= J$.
- Fixed cost per household $F$
- Survey cost varies in proportion to $N(J + F)$
- Standard error of estimates varies in proportion to $e^{-J}/\sqrt{N}$
- Researcher's objective: choose $N$ and $J$ to maximise power subject to a budget constraint
- Solution of the constrained optimisation problem identifies a fixed (efficient) $J$, independent of sample size!

# Analysis: Efficient Survey Design

- Sample size = $N$, number of items included = $J$.
- Fixed cost per household $F$
- Survey cost varies in proportion to $N(J + F)$
- Standard error of estimates varies in proportion to $e^{-J}/\sqrt{N}$
- Researcher's objective: choose $N$ and $J$ to maximise power subject to a budget constraint
- Solution of the constrained optimisation problem identifies a fixed (efficient) $J$, independent of sample size!

Extension:

- The relation between rank and shares is weaker as sample becomes more varied
- Efficient $J$ increases in this situation

# Concluding Remarks

- Work in progress
- Attempting to extend to validation of distributional analysis
- Reference datasets small so exploring simulation methods
- Joint distribution of components of consumption vital
- But facing the curse of dimensionality...
- Comments and suggestions very welcome!

# Acknowlegements


- Funding from the John Fell Fund, the Bill and Melinda Gates Foundation and an anonymous donor.
- Project coordination from Mahreen Mahmud
- Research assistance from Victoria Schreitter, Axel Eizmendi Larrinaga
- Kenya Integrated Household Budget Survey 2005-2006, Version 1.0 of the KNBS dataset