Comparing sample estimates and population counts of payrolled employees to assess the representativeness of the Labour Force Survey

Daniel Ayoubkhani, Ted Dolby, Vahé Nafilyan
National Statistician's Analysis Unit

20 May 2026

Office for National Statistics

Purpose of this talk

- Outline methods under consideration for using admin data to improve estimates of UK labour market statistics

  → Specifically, identifying and accounting for employment-related non-response

- Research currently underway in ONS, aiming for publication later in 2026

- Work in progress – no results today

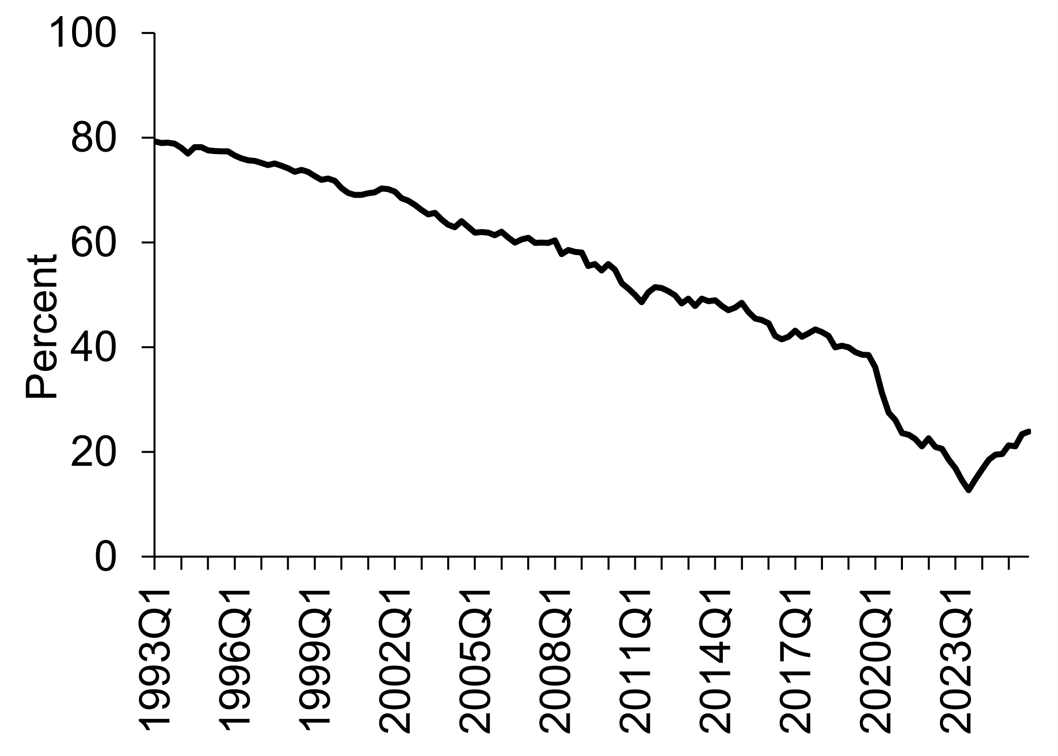
- We value your input and feedback!

The UK Labour Force Survey (LFS)

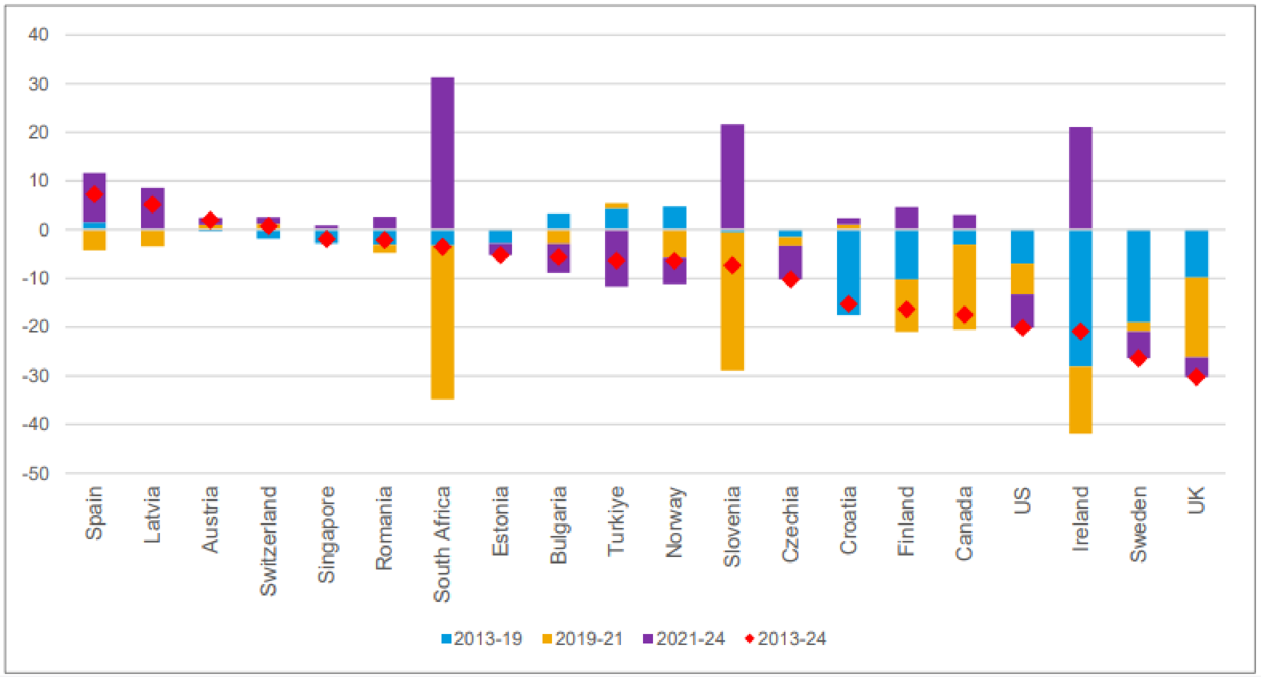
- Aim: provide timely indicators on labour market conditions (employment, earnings, working hours, etc) to inform economic and monetary policy
- UK coverage, run by ONS in Great Britain and NISRA in Northern Ireland
- Started in 1973 as biennial, annual from 1984 to 1991, quarterly from 1992
  → Scheduled to be replaced by the Transformed LFS (TLFS) in 2027
- UK's largest routine household survey: ~65k people from ~30k households in 2025 Q4
- Target population: private households + student halls and NHS staff accommodation
- Systematic random sample of addresses each quarter, all residents within sampled households are eligible
- Survey estimates take account of household clustering, implicit stratification by geography, and sampling weights

LFS response rates are in decline

LFS household response rates, all waves, GB[1]



Changes in LFS response rate for selected countries



Source: Fotopoulou et al (2026). Eroding Participation in Labor Force Surveys: Evidence, Drivers and Solutions. IMF WP/25/45.

While falling response rates do not necessarily mean increased non-response bias, they do suggest more *potential* for bias

[1] From 1993 to 2005: Q1 = Mar-May, Q2 = Jun-Aug, Q3 = Sep-Nov, Q4 = Dec-Feb. From 2006 onwards: Q1 = Jan-Mar, Q2 = Apr-Jun, Q3 = Jul-Sep, Q4 = Oct-Dec.

LFS weighting process

Design weights → **Non-response adjustment** → **Calibration**

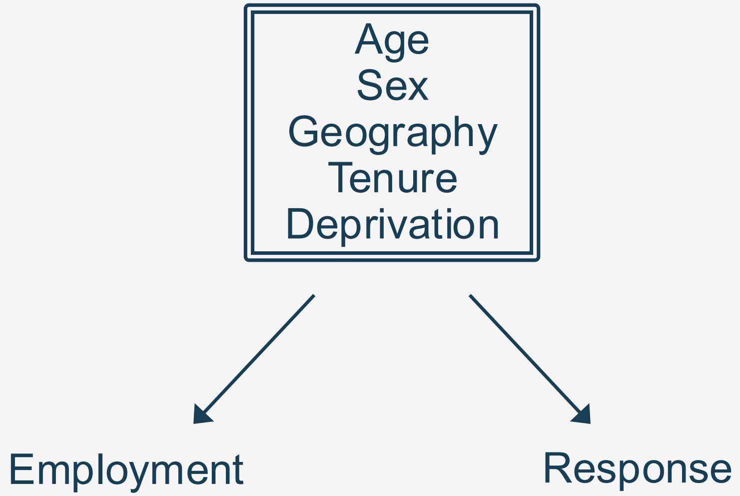
Design weights: Inverse-probability of household being selected into the sample

Non-response adjustment: Accounts for factors related to probability that household responds: region, area deprivation, output area classification

Calibration: Preserves “know” population distributions by age, sex, geography, tenure; also protects against non-response bias

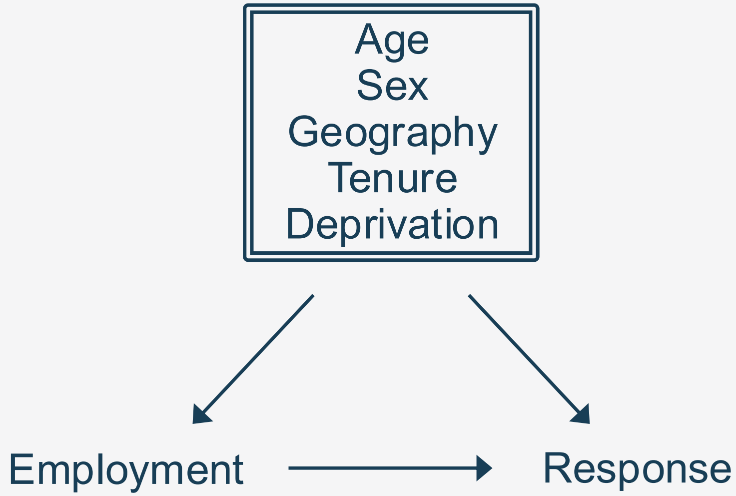
Non-response as a missing data problem

Missing at random (MAR)



H_0: P(response) is independent of employment status, conditional on weighting variables

Missing not at random (MNAR)



H_A: P(response) is dependent on employment status, even after applying weights

If H_A is true, employment rates estimated from the LFS will be biased

Calibration can shed light on size and direction of potential bias in pre-calibrated estimates

Generalised regression (GREG) calibration estimator:

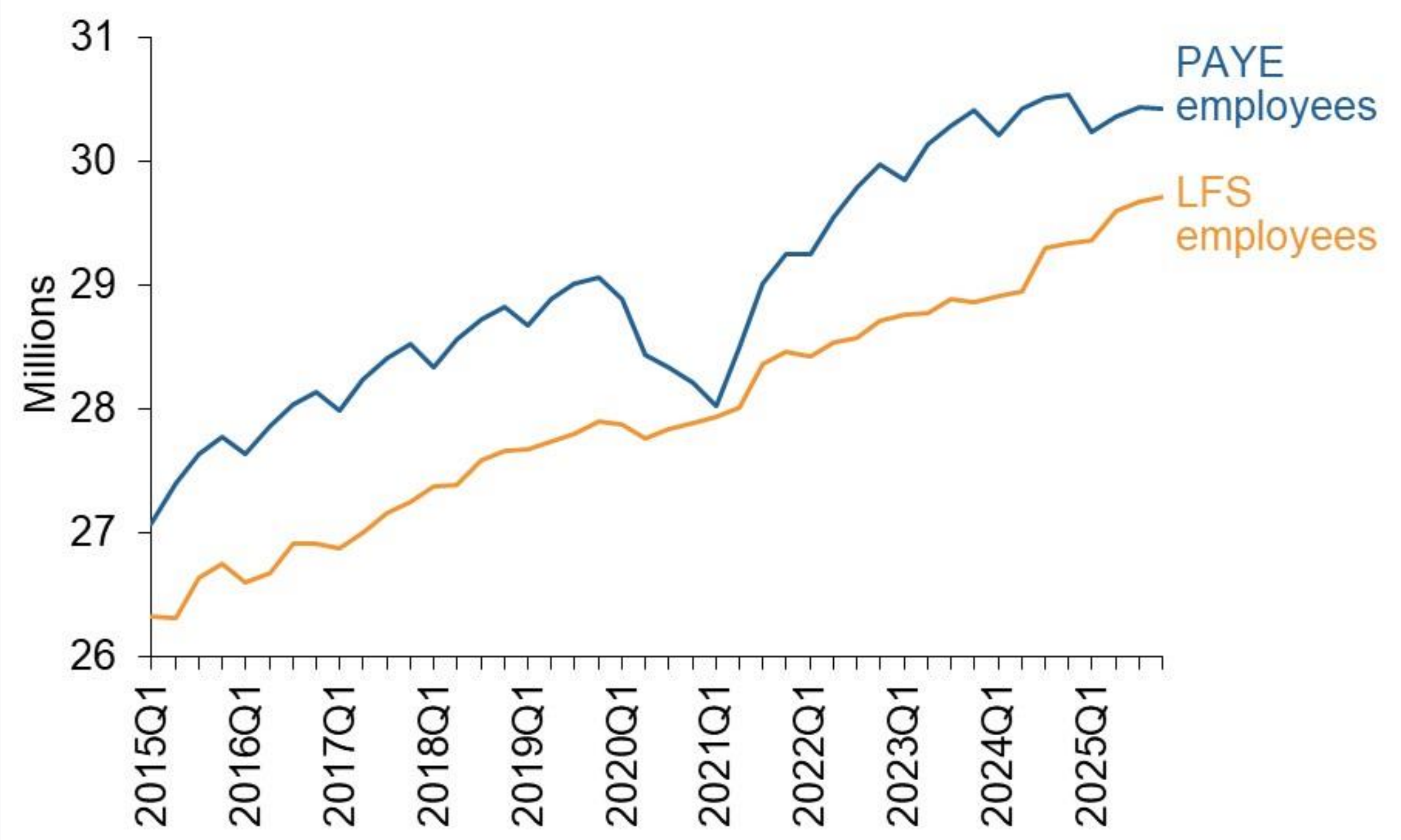
$$\hat{Y}_{CAL} = \hat{Y}_{ORIG} + \hat{\beta} \times \underbrace{(Z - \hat{Z}_{ORIG})}_{\text{Extent to which sample is representative of population in terms of calibration variable}}$$

Extent to which calibration variable is correlated with variable of interest

Implication: to test for bias in the LFS, we need to find a variable that is well correlated with employment and is available for everyone in the population

LFS consistently estimates fewer employees than HMRC PAYE records

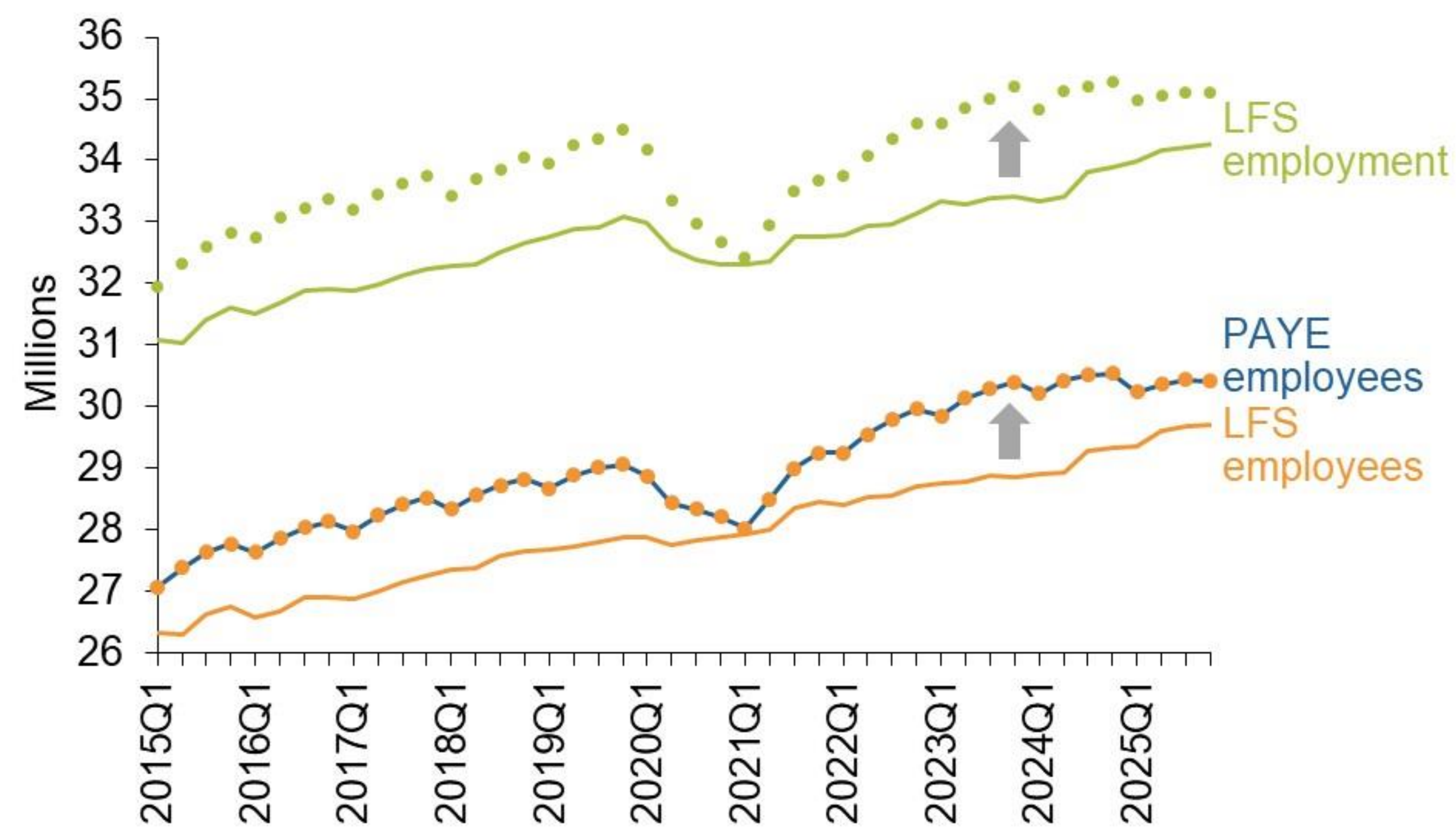
No. employees estimated by LFS and PAYE records, UK



Can we simply calibrate LFS to PAYE employees?

- Adjust survey weights such that weighted LFS employees = population PAYE employees
- Differences between original and recalibrated estimates might imply non-response bias in LFS
- Same concept as ONS business surveys (e.g. ABS turnover calibrated to IDBR turnover)

LFS & PAYE employees and LFS (ILO) employment, UK



However, there are several possible reasons for differences between LFS and PAYE employees:

1) Non-response bias in LFS
2) Definitional/reporting differences
3) Population coverage differences

Aim: account for (2) and (3) to isolate (1)

Accounting for definitional/reporting differences

- **Challenge:** various sources of difference between LFS and PAYE employee status

    → Definitions: ILO VS. administrative
    → Timing effects: week-in-month vs. through-the-month
    → Measurement errors: e.g. due calendarization of PAYE, respondent error in LFS

- **Solution:** use PAYE records to derive a common employee variable for *both* the LFS sample (what is being calibrated) *and* the PAYE population (what we're calibrating to)

    → Relies on linkage of LFS & PAYE via the ONS Demographic Index (DI)
    → PAYE employee status is well correlated with LFS employee status (Phi coefficient ~0.9), employment (~0.8) and economic activity (~0.8)

Deriving PAYE status on the LFS sample for a single month

LFS respondent sample

DI	ILO employee (self-reported)
FIEX8415	1
BAKP4385	1
?	0
ZPTQ7721	1
RJHS9629	0
KCYM3034	0

PAYE population

DI	Gross pay
CQMP6006	£2,820
IBXI1516	£7,266
BAKP4385	£4,089
YPVX0624	£990
TYER9602	£3,565
ODAJ8656	£4,241
ZPTQ7721	£5,875
PPCV0726	£4,180
RJHS9629	£2,741
AYZD3652	£5,386
USCY8314	£1,387
REOT7553	£4,260

Note: dummy data

Deriving PAYE status on the LFS sample for a single month

LFS respondent sample

DI	ILO employee (self-reported)	PAYE employee (derived)
FIEX8415	1	0
BAKP4385	1	1
?	0	?
ZPTQ7721	1	1
RJHS9629	0	1
KCYM3034	0	0

Join

PAYE population

DI	Gross pay
CQMP6006	£2,820
IBXI1516	£7,266
BAKP4385	£4,089
YPVX0624	£990
TYER9602	£3,565
ODAJ8656	£4,241
ZPTQ7721	£5,875
PPCV0726	£4,180
RJHS9629	£2,741
AYZD3652	£5,386
USCY8314	£1,387
REOT7553	£4,260

Note: dummy data

Deriving PAYE status on the LFS sample for a single month

LFS respondent sample

DI	ILO employee (self-reported)	PAYE employee (derived)
FIEX8415	1	0
BAKP4385	1	1
?	0	0
ZPTQ7721	1	1
RJHS9629	0	1
KCYM3034	0	0

Impute (~2%)

PAYE population

DI	Gross pay
CQMP6006	£2,820
IBXI1516	£7,266
BAKP4385	£4,089
YPVX0624	£990
TYER9602	£3,565
ODAJ8656	£4,241
ZPTQ7721	£5,875
PPCV0726	£4,180
RJHS9629	£2,741
AYZD3652	£5,386
USCY8314	£1,387
REOT7553	£4,260

Note: dummy data

Accounting for population coverage differences

- Over-coverage in PAYE population vs. LFS target population will manifest as under-representation of employees, even if the sample is perfectly representative
- The extent of over-coverage is currently unknown and therefore must be estimated

Over-coverage due to non-domestic residency	Over-coverage due to residency in CEs other than student halls and NHS staff accommodation
Problem: No overseas residence indicator in PAYE dataset; missing geography info could be due to living abroad *or other reasons* (e.g. postcode not supplied)	**Problem:** No info on residential circumstances recorded in PAYE dataset
Solution: Calculate missing geography % among PAYE records linked to LFS respondents (i.e. definitely inside the target population)	**Solution:** Link March 2021 PAYE records to 2021 Census of England & Wales responses to obtain residency info
Estimated scale of PAYE over-coverage due to non-domestic residency = 0.4% (GB only)	**Estimated scale of PAYE over-coverage due to living in non-LFS CEs = 0.05% (age 50-64) to 0.24% (age 16-24)**

Estimation

- Calibration to PAYE totals added as an additional constraint to existing partitions:

Partition 1:	local authority districts
+ Partition 2:	sex * GB/NI-specific age groups
+ Partition 3:	region/country * sex * age groups
+ Partition 4:	GB/NI * sex * age groups
+ Partition 5:	housing tenure groups
+ Partition 6:	PAYE employee status * sex * age groups

Compare

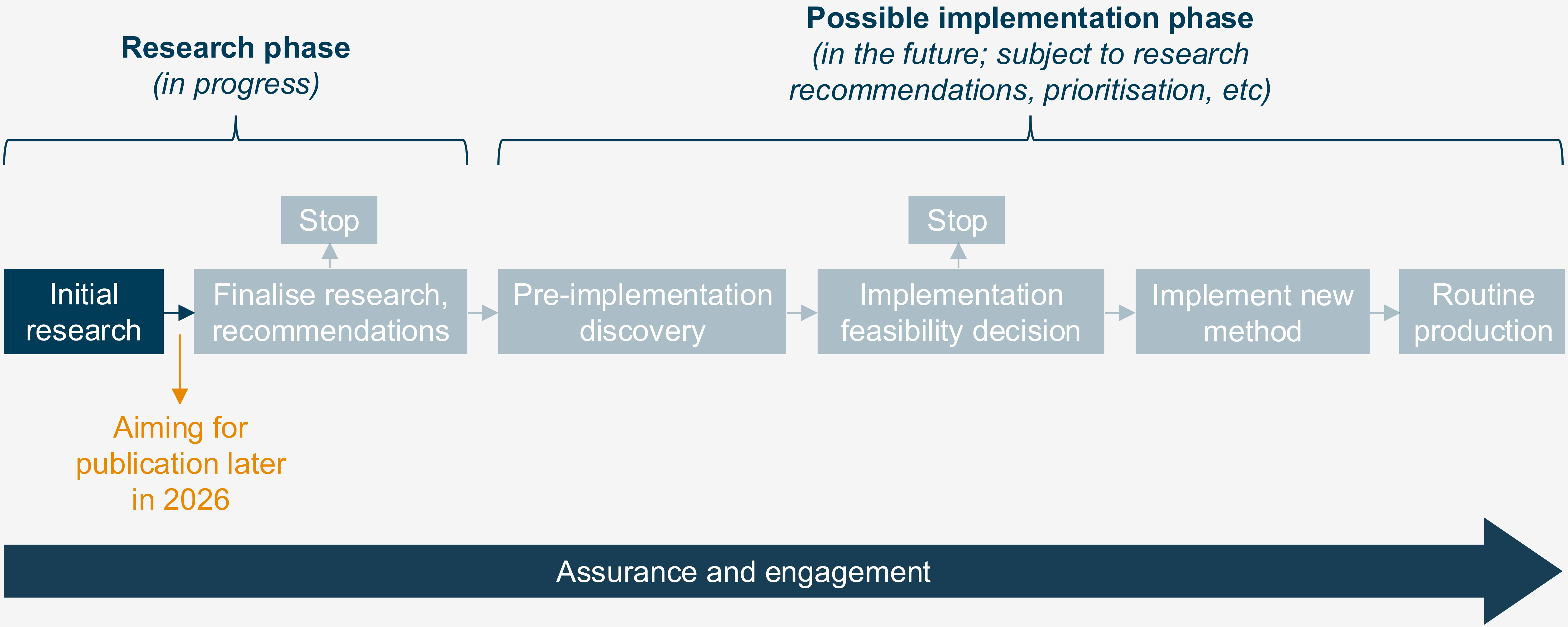
- New calibration groups reflected in estimation of standard errors (via linearization)

- Potential bias-variance trade-off: check design effects aren't substantially inflated vs. current calibration groups due increased variability of weights

Further considerations and challenges

- Calibration to PAYE-derived quantities other than employee status (e.g. pay quantiles)?
- PAYE data not sufficiently timely for processing monthly LFS estimates, and early data are subject to revision
- Linkage precision high but imperfect (~99.5%); need to simulate effects of linkage error
- Use PAYE to model sample attrition in longitudinal follow-up and improve estimates of employment flows?
- PAYE only covers employees so incorporating HMRC data on the self-employed may improve the calibration (but not a necessity) – impact assessment being planned using historical data
- LFS planned to be replaced with TLFS in 2027, but limited data collected so far and survey methodology still being refined

Next steps



Research phase
(in progress)
Possible implementation phase
(in the future; subject to research recommendations, prioritisation, etc)
Stop
Stop
Initial research
Finalise research, recommendations
Pre-implementation discovery
Implementation feasibility decision
Implement new method
Routine production
Aiming for publication later in 2026
Assurance and engagement

- **Questions?**
- **Suggestions?**
- **Feedback?**

daniel.ayoubkhani@ons.gov.uk

Appendices

What about under-coverage in PAYE?

- Some employees will not appear in the PAYE dataset, e.g. those who are:
  - In PAYE schemes for which all employees are below the NI threshold
  - Working for employers that are exempt from digitally filing PAYE returns
  - Not currently being paid or paying tax (e.g. long-term sickness, career break)
  - Subject to special data protection measures by HMRC

- However, this represents classification error rather than coverage error:

LFS sample:	Absence of linked PAYE record → coded as non-employee
Control population:	Non-employee population total derived by subtraction (mid-year population estimate minus PAYE employees)

- Thus, the same classification error present in both the sample and the control population; no problem for proposed methodology (though will reduce correlation between y_i and z_i)

Calibration methodology in detail

Aim: produce new weights w_i that minimally adjust the design weights d_i such that for a set of calibration variables $\mathbf{z}_i$, estimated totals from the sample $\hat{\mathbf{Z}}$ equal known totals in the population $\mathbf{Z}$:

$$w_i = d_i(1 + \mathbf{z}_i'\boldsymbol{\lambda})$$ such that: $$\sum_{i \in s}^{n} w_i \, \mathbf{z}_i = \hat{\mathbf{Z}} = \sum_{i=1}^{N} \mathbf{z}_i = \mathbf{Z}$$

The distance function to be minimised is the Chi square distance: $$D(w, d) = \sum_{i \in s}^{n} \frac{(w_i - d_i)^2}{2d_i}$$

The parameter vector $\boldsymbol{\lambda}$ is found using Lagrange multipliers: $$\boldsymbol{\lambda} = \left(\sum_{i \in s}^{n} d_i \mathbf{z}_i \mathbf{z}_i^{i}\right)^{-1} \left(\mathbf{Z} - \sum_{\boldsymbol{i \in s}}^{\boldsymbol{n}} d_i \mathbf{z}_i\right)$$

The calibrated total estimate of the variable of interest is then: $$\hat{Y}_{CAL} = w_i y_i$$

Equivalently, this can be expressed as the GREG estimator: $$\hat{Y}_{CAL} = d_i y_i + \left(\mathbf{Z} - \sum_{\boldsymbol{i \in s}}^{\boldsymbol{n}} d_i \mathbf{z}_i\right) \widehat{\boldsymbol{\beta}}$$

with the parameter vector $\widehat{\boldsymbol{\beta}}$ estimated from the design-weighted regression: $$y_i = \mathbf{z}_i \widehat{\boldsymbol{\beta}} + \varepsilon_i$$