Proceedings of the 89th Annual International Meeting of
the Psychometric Society, Prague, Czech Republic (2024).

Imputation Models for Special Subpopulations in Large-scale Survey Assessments

Usama S. Ali *†‡ and Frederic Robin†

†ETS Research Institute, ETS, Princeton, New Jersey, 08541, USA
‡Department of Educational Psychology, South Valley University, Qena, 83253, Egypt
*Corresponding author. Email: uali@ets.org

Abstract

Nonresponse in large-scale survey assessments can arise from factors such as language barriers, reading difficulties, or disabilities. Excluding these subpopulations may introduce bias into survey results. This study develops an imputation method for literacy-related nonresponse cases in the international adult survey (PIAAC). These cases completed a special background questionnaire—the doorstep interview—but did not proceed to the main cognitive assessment. Using such limited data from respondents across selected countries with varying proportions of such cases, we compared and evaluated multiple imputation models to improve proficiency estimation. The proposed approach provides a practical solution for enhancing inclusivity in educational measurement.

Keywords: missing data, latent regression models, literacy-related non-response

In this research, we explored the enhancement of reporting on special subpopulations in large-scale survey assessments with case study from an international adult survey.

1. Introduction

Large-scale survey assessments, such as the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), the Organisation for Economic Co-operation and Development's (OECD) Programme for the International Assessment of Adult Competencies (PIAAC) and Programme for International Student Assessment (PISA), as well as national assessments like the U.S. National Assessment of Educational Progress (NAEP), are critical tools for evaluating skills, knowledge, and competencies across diverse populations (Martin et al., 2020; National Center for Education Statistics (NCES), 2022; Organisation for Economic Co-operation and Development (OECD), 2019a). As participation in these assessments expands globally, the growing linguistic and cultural heterogeneity of test-takers presents unprecedented measurement challenges. A particularly pressing concern emerges when linguistic minorities, immigrant populations, and examinees with limited assessment language proficiency face test items that are linguistically or culturally inaccessible—resulting either in non-response patterns that produce missing data or in attempted responses that yield invalid measurement (Rubin, 1996; von Hippel, 2020). The challenge of accurately assessing the proficiency of such special subpopulations potentially compromises both the validity of cross-population comparisons and the equity of assessment outcomes. This paper addresses this critical issue in national and international assessment

contexts. In the following sections, we introduce the methodology for estimating and reporting proficiencies in large-scale survey assessments, followed by a case study of a special subpopulation in an international survey of adult skills where different models referred to in this manuscript as "imputation models" were proposed and compared to estimate the proficiencies of individuals with language barrier. The findings of this study were discussed and conclusions section is followed.

1.1 Plausible Values Methodology

Most modern large-scale survey assessments employ plausible values (PVs) methodology to estimate respondent proficiency while accounting for measurement error and missing data (Mislevy, 1991; von Davier et al., 2009). This methodology can be summarized as a three-step process that combines:

1. Item response theory (IRT) calibration of cognitive responses to estimate item parameters. Item parameters are estimated for each cognitive domain separately through unidimensional IRT models. Among these IRT models is the two-parameter logistic (2PL) model, where the probability correct response $X_j = 1$ is given by

$$p(X_j = 1|\theta) = \frac{\exp(\alpha_j(\theta - \beta_j))}{1 + \exp(\alpha_j(\theta - \beta_j))}. \tag{1}$$

2. A latent regression model that incorporates both responses from cognitive instruments and contextual variables from background questionnaires. This population-specific multivariate latent regression gives an expression for respondent's proficiency distributions on the multidimensional scales conditional on covariates (i.e., contextual information, y) in addition to the cognitive item responses (x). Based on Bayes' theorem, the posterior distribution of skills given the observed item responses and covariates is constructed as follows

$$P(\boldsymbol{\theta}_\nu|\boldsymbol{x}_\nu, \boldsymbol{y}_\nu, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto P(\boldsymbol{x}_\nu|\boldsymbol{\theta}_\nu)P(\boldsymbol{\theta}_\nu|\boldsymbol{y}_\nu, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}). \tag{2}$$

   This model estimates the regression coefficients ($\boldsymbol{\Gamma}$) and the residual variance-covariance matrix ($\boldsymbol{\Sigma}$) using the estimated item parameters from step 1 as true values (Thomas, 1993).
3. Multiple imputation where a specific number of PVs (e.g., 5 to 20) were generated for each respondent on each cognitive domain from the estimated posterior distributions of proficiency using estimated $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ from Step 2 (Mislevy & Sheehan, 1987; von Davier et al., 2009).

1.2 The Challenge of Special Subpopulations in Reporting

While effective for general populations, this methodology faces limitations when applied to special subpopulations with systematic non-response patterns. The latent regression model assumes missingness can be explained by observed covariates (i.e., missing at random assumption; Rubin, 1987), which may not hold for groups with language barriers where no cognitive data are provided and almost all contextual variables (or predictors) are often omitted from standard background questionnaires.

The consequences of this limitation become evident when examining specific vulnerable groups:

1. English Language Learners in NAEP: Despite accommodations, ELL students' scores often reflect language barriers rather than content knowledge (Abedi, 2004). Standard PV generation may underestimate their true abilities without proper linguistic covariates.

2. Migrant Adults in PIAAC: First-generation immigrants frequently show non-response in literacy tasks, yet their occupational and educational backgrounds contain valuable information about latent proficiency (Organisation for Economic Co-operation and Development (OECD), 2019b).
3. Indigenous Students in TIMSS: When assessments aren't available in native languages, students may leave items blank, creating non-random missing patterns that standard PV approaches fail to address adequately (Wu, 2009).

2. The PIAAC Doorstep Interview Case Study

PIAAC is OECD's international survey of adult skills. In 2023, PIAAC in its second cycle measures adults' proficiency in literacy, numeracy, and adaptive problem solving (APS). For PIAAC, the sample represents the non-institutionalized population, age 16 to 65. Nonresponse in this survey occurs due to language barrier, reading/writing difficulty or disability. Literacy-related nonresponse (LRNR) is a subset of these cases – those persons with a language barrier. In the first cycle of PIAAC, LRNR cases were part of the target popoulation and selected sample was given sampling weights, but no plausible values reported. These cases had very little background data – often only estimated age and gender - and no cognitive data. Analyses determined that there was no sufficient information to estimate their PVs. Accordingly, these cases were not included in the country-level proficiency estimates. This provide a reporting issue: A sector of the population was left out. Moreover, as that sector of the population could not function in any of the major languages needed to exercise their skills as part of the country's workforce, expectations were that in that context their skills should be low. Thus not including them in the country's population estimate leads to some over-estimation.

Recognizing these limitations, PIAAC in its second cycle pioneered an enhanced data collection protocol for literacy-related non-respondents. To improve population estimates in the second cycle of PIAAC, an "abbreviated background questionnaire" known as the doorstep interview was created. The strategy behind the doorstep interview was to provide a specific instrument delivered by an interviewer to collect targeted background information with high predictive power from respondents unable to speak the assessment language. The doorstep background variables were:

1. Gender
2. Age
3. Educational level
4. Employment status
5. Country of birth
6. Number of years in the country if immigrant (i.e., not native-born)

These six variables embedded in the doorstep interview were also present in the full background questionnaire. Such doorstep interview was available in 28 PIAAC background questionnaire languages and 15 additional minority languages that participating countries selected the language(s) that would fit their minority groups. We also knew the full background questionnaire language is not their native language. This doorstep interview was administered if a translator was not available or if someone in the home cannot act as an interpreter; as it was always preferable to collect a full background questionnaire. This doorstep interview as an abbreviated background questionnaire was designed to potentially provide enough information to estimate PVs for these respondents.

Research demonstrated these variables significantly improved proficiency estimation for language-barrier populations (Paccagnella, 2021). By enriching the latent regression model

with these carefully selected covariates, PIAAC achieved more accurate population estimates while maintaining the integrity of the PV framework.

2.1 Focus and Contribution of This Study

In this study, we target different imputation models and compare their performance under the current features of the PIAAC assessment design to generate PVs for LRNR cases. The goal is to enhance the quality and inclusiveness of reporting by ensuring that no subpopulations are excluded. The study focuses on evaluating which imputation model best accounts for the unique characteristics of LRNR respondents while preserving the integrity and comparability of the assessment results.

This study makes three key contributions to the measurement literature:

1. Methodological Extension: We developed and compared alternative IRT and latent regression approaches to model and generate plausible values for a small subpopulation that had only very limited data (i.e., few key background variables and no cognitive information) but a-priori expectations.
2. Empirical Validation: focusing on PIAAC's language-barrier subgroup (i.e. those who administered the doorstep interview), we demonstrated that it was possible to generate plausible values to:
   - Reduces bias in population parameters
   - Improves the accuracy of proficiency estimates for non-respondents
   - Maintains reliability compared to external benchmarks
3. Practical Framework: We provide guidelines for assessment programs to:
   - Identify high-impact contextual variables during instrument development
   - Implement adaptive data collection protocols for special subpopulations
   - Integrate subgroup-specific modeling into standard PV methodology

3. Method

3.1 Data

This study used initial PIAAC main study data from four countries, varying in their percentages of doorstep cases. Since doorstep interview cases accounted for less than 2% of respondents across all countries, we selected two countries (C1 and C2) with higher doorstep interview rates (above 2% and 5%, respectively) and two (C3 and C4) with lower rates (below 1%). Table 1 provides further details, including:

- The number of doorstep interview cases and their percentages.
- Path 1 respondents who failed the Locator in both literacy and numeracy.
- Respondents who failed the locator in at least one domain (literacy or numeracy).

It is important to note that the theory is that, given that they have language barrier, they are expected to perform only at the level equivalent of someone with very low skill. As a consequence their path through the assessment (as described by Figure 1) would be equivalent to failing the locator (Path 1).

3.2 Targeted sample: Doorstep-like cases

Sampled persons with language barrier nonresponse were presumed to have distinct proficiency distributions in the cognitive domains (in the assessment language) from regular individuals in the target population. The doorstep interview cases are operationally comparable to Path 1 cases (i.e., those failing both literacy and numeracy sections of the Locator),

Table 1. Selected participating countries for study

Country ID	Unweighted sample N (%) of ... cases		
	DI[a]	P1[b]	F1[c]
C1	234 (3.5)	80 (1.2)	228 (3.4)
C2	897 (14.3)	25 (0.5)	300 (4.8)
C3	35 (0.6)	100 (1.6)	619 (9.8)
C4	36 (0.6)	98 (1.5)	633 (10.0)

Table note
a DI = Doorstep Interview
b P1 = Path 1
c F1 = Failed at least in one domain

with most expected to perform at the lowest proficiency levels (i.e., Proficiency Level 1 or Below). We used Path 1 cases as a benchmark for doorstep interview case performance across models. By design, Path 1 cases are routed to basic items of reading and numeracy components, resulting in cognitive data that includes literacy and numeracy performance data without any APS data (see Figure 1). This Path 1 data limitation would prevent APS proficiency estimation if Path 1 cases were used exclusively in the imputation models. To include APS data, we extended the sample to create a doorstep-like sample (denoted F1 in Table 1), comprising:

- Path 1 cases (failing both domains)
- Cases failing exactly one domain (literacy or numeracy Locator)

3.3 Imputation models in comparison

We examined three alternative latent regression IRT models (i.e., imputation models), each using different dataset conditions, for estimating the model and generating PVs for the doorstep cases. The studied models were the following:

- Model 1 (Base Model): Uses the full sample and full background questionnaire variables (current reporting methodology). Note that for doorstep interview cases, data for background variables other than the six doorstep-specific variables are missing by design. In the latent regression model, all non-doorstep interview background variables were coded with a "missing" category (e.g., gender includes three response categories: male, female, and missing), while the six available doorstep interview variables retained their actual values. This means the model conditioned estimates on both the known doorstep interview variables and the missing responses of other background variables.
- Model 2: Uses only doorstep-like cases (e.g., the target cases as defined and justified in the previous section) with abbreviated background questionnaire (or doorstep interview) variables
- Model 3: Uses the full sample but only doorstep interview variables excluding all other contextual variables available in the full background questionnaire which means that for non-doorstep cases all these additional variables are turned off

Accordingly, the dataset conditions were defined by:

- Respondent sample: All cases, or only the targeted cases (e.g., P1 or F1 cases)
- Conditioning variables: Full background questionnaire variables, or only the doorstep interview variables

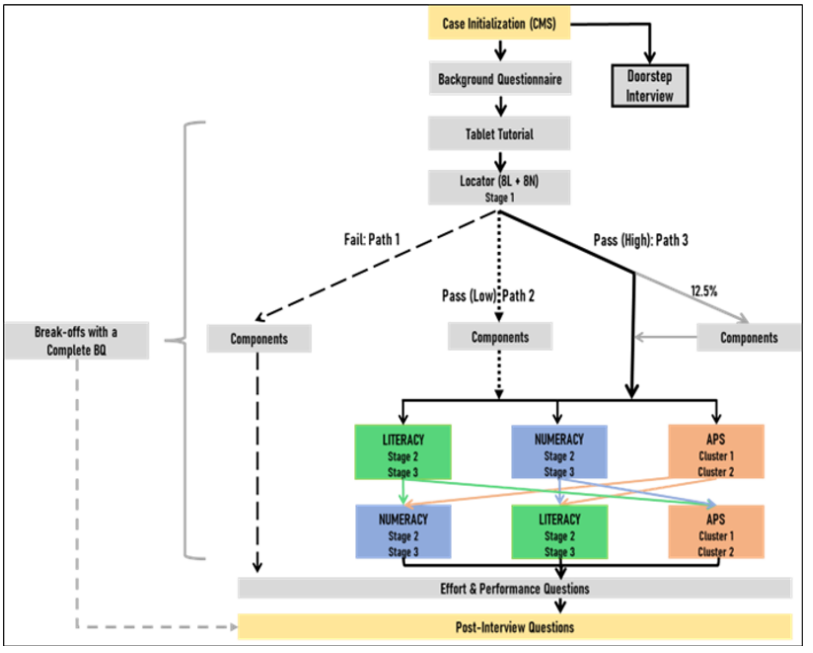

Figure 1. PIAAC general assessment design (Note. The horizontal dashed line indicates the cut score for Proficiency Level 1, set at 176.)

- Cognitive data for the doorstep interview sample: No cognitive data, or imputed cognitive data for doorstep interview cases

In the cognitive data imputation process, item scores were imputed (using single imputation) for all 16 literacy and numeracy locator items to mimic the responses provided by Path 1 respondents within each country. For each doorstep respondent, this was done by: (a) drawing a proficiency value from the Path 1 posterior theta distribution (averaged across all Path 1 respondents' posteriors); and (b) drawing correct or incorrect responses based on the drawn theta and the international IRT model for each item. Therefore, both the background information from the doorstep interview and the imputed cognitive items were used in estimating the proficiencies for those doorstep interview cases.

Table 2. Sample and conditioning variables used in the studied models

	Full BQ[a]	Abbreviated DI[b]
All Cases	Model 1	Model 3
Targeted Cases	N/A	Model 2

Table note
a BQ = Background Questionnaire
b DI = Doorstep Interview

The studied models are illustrated in Table 2. Each model can be implemented with or without cognitive data imputation for doorstep interview cases. For cognitive data imputation, we used Path 1 respondents (failing in both literacy and numeracy sections of the Locator). Country-specific Path 1 posterior proficiency distributions were applied to impute responses for all Locator items (eight literacy and eight numeracy) for each doorstep interview case. The PIAAC assessment design is shown in Figure 1.

4. Results

Figure 2 provides the country-level proficiency mean plus and minus the standard deviation (+/- 1 SD) for doorstep interview cases with (right panels) and without imputed cognitive data (left panels) in Models 1, 2 and 3 for the three cognitive domains: literacy, numeracy, and APS. The results of comparing the three imputation models with and without imputation of cognitive data are summarized as follows:

- Model 1: As the base model for regular (non-doorstep) respondents, it produced extremely low scores for doorstep interview cases regardless of imputation. This was expected because Model 1 is unsuitable for doorstep interview cases due to extensive missing covariates (six available variables versus 240+ in the full background questionnaire). The severe missingness prevents reliable PV estimation for doorstep interview cases.
- Model 2 with imputed cognitive data and Model 3 without imputed cognitive data yielded unsatisfactory results because:
  - Model 2 with imputed cognitive data involves "double dipping" (i.e., using two features that would limit the performance of the doorstep interview cases: using a sample of doorstep-like cases in addition to imputing cognitive data based on Path 1 cases) and
  - Model 3 without imputed cognitive data fails to distinguish doorstep cases from other respondents
- Model 2 without imputed cognitive data: Produced reasonable results but required inclusion of higher-performing non-Path 1 cases to obtain APS data, which biased doorstep case estimates (based solely on demographics).
- Model 3 with imputed cognitive data: Emerged as the recommended approach, providing:
  - More consistent cross-country/domain results
  - Proper utilization of Path 1 proficiency distributions for imputation

Both imputation models—Model 2 without imputed cognitive data and Model 3 with imputed cognitive data—yielded substantively reasonable results. Figure 3 compares their performance in estimating proficiency distributions for doorstep interview cases relative to Path 1 cases, revealing two key insights:

- Model 3 with imputed cognitive data produced proficiency distributions for doorstep interview cases that closely aligned with Path 1 cases, suggesting successful recovery of latent ability patterns through cognitive data imputation.
- Model 2 without imputed cognitive data showed divergent distributions, indicating that having no cognitive data leads to meaningfully different proficiency estimates.

These results demonstrate the value of incorporating cognitive data through imputation when analyzing incomplete assessment records.

Based on these findings, the PIAAC Technical Advisory Group recommended evaluating a modified approach that maintains Model 3's core structure while addressing concerns about imputation. The proposed alternative, referred to as Model 3 with DI-like variable, eliminates cognitive data imputation for doorstep interview cases but introduces a new binary conditioning variable (coded 1 for Doorstep Interview/Path 1 cases and 0 otherwise) alongside the original six demographic variables. This modified approach preserves the seven-variable framework while offering a distinct methodological solution.

Figure 4 compares country-level means (±SD) for both versions of Model 3, demonstrating that the original imputation-based approach yields superior results. Specifically, Model 3 with cognitive data imputation provides more consistent cross-country and cross-domain estimates by leveraging country-specific Path 1 proficiency distributions to inform the imputation

Figure 2. Proficiency Mean (+/- SD) for Doorstep Interview Cases with and without Imputed Cognitive Data in Models 1, 2 and 3 (Note. The horizontal dashed line indicates the cut score for Proficiency Level 1, set at 176.)

process. As expected, both models showed consistent performance for literacy and numeracy. However, the modified version of Model 3 failed to limit the performance of the doorstep interview cases as intended because, by design, the doorstep-interview-like cases have cognitive data only for literacy and numeracy but not APS. Consequently, the doorstep-interview-like variable did not effectively constrain the performance of doorstep cases, as evidenced by the low variability in outcomes (i.e., the performance estimates with and without doorstep inter-

view cases were very close). Figure 5 reveals the overestimation (reaching Proficiency Level 3) of doorstep interview cases and Path 1 cases under the modified version of Model 3, compared to the expected performance of Path 1 cases under more robust model specifications like Model 3 with imputed cognitive responses (see Figure 3). These results confirm the value of carefully implemented cognitive data imputation for maintaining estimation accuracy in large-scale assessments.

5. Conclusion

This study examined methodological approaches for addressing literacy-related nonresponse (LRNR) in large-scale survey assessments, with three key findings:

First, standard imputation procedures relying on the missing-at-random assumption prove inadequate for LRNR cases, as language barriers create missing-not-at-random patterns that correlate with the assessed competencies. Our analysis demonstrates that conventional models like Model 1 (using full background questionnaire) produce unreliable estimates for these special subpopulations due to extensive missing covariates.

Second, among alternative approaches, Model 3 with cognitive data imputation emerged as the most effective solution, providing:

- Consistent proficiency estimates across countries and domains
- Proper utilization of Path 1 respondent distributions
- Reduced bias compared to non-imputation approaches

Third, the study highlights the critical trade-off between methodological limitations and representation - while no current approach is ideal, excluding LRNR cases introduces greater bias than model-based inclusion. This work advances assessment practice by:

- Validating an imputation framework for language-barrier cases
- Demonstrating how demographic data can support more inclusive scoring
- Establishing principles for handling non-ignorable nonresponse

Future research should explore hybrid designs combining doorstep interviews with refined imputation techniques. Nevertheless, this study provides actionable solutions for maintaining both validity and inclusivity in international assessments facing growing linguistic diversity.

Acknowledgement

The author acknowledges the support of this research by my psychometric and data analysis team colleagues at ETS Research Institute: Lokesh Kapur, Wei Zhao, and Mathew Kandathil.

Funding Statement This work was conducted as part of a contract between ETS and the OECD (Organisation for Economic Co-operation and Development) for the implementation of the second cycle of the Programme for the International Assessment of Adult Competencies. Any views expressed in the paper is solely of the authors and do not necessarily reflect those of the OECD or its member countries.

Competing Interests None.

References

Abedi, J. (2004). The No Child Left Behind act and English language learners: Assessment and accountability issues. Educational Researcher, 33(1), 4–14. https://doi.org/10.3102/0013189X033001004

Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020). Methods and procedures: TIMSS 2023 technical report. TIMSS & PIRLS International Study Center.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. Implementing the new design: The NAEP 1983-84 technical report, 361–380.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. Psychometrika, 56(2), 177–196. https://doi.org/10.1007/BF02294457

National Center for Education Statistics (NCES). (2022). NAEP technical documentation (tech. rep.). U.S. Department of Education. Washington, DC.

Organisation for Economic Co-operation and Development (OECD). (2019a). PISA 2018 technical report. OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2019b). Technical report of the Survey of Adult skills (PIAAC) (3rd, tech. rep.). Paris, France, OECD Publishing.

Paccagnella, M. (2021). Literacy and numeracy proficiency in IALS, ALL and PIAAC. OECD Education Working Papers, (257). https://doi.org/10.1787/3f075a07-en

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434), 473–489. https://doi.org/10.1080/01621459.1996.10476908

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. Journal of Computational and Graphical Statistics, 2(3), 309–322.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? IERI Monograph Series, 2, 9–36.

von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. Sociological Methods & Research, 49(3), 699–718. https://doi.org/10.1177/0049124117747303

Wu, M. (2009). Issues in large-scale assessments and reporting for indigenous students. Educational Research for Policy and Practice, 8(1), 61–73. https://doi.org/10.1007/s10671-009-9064-7

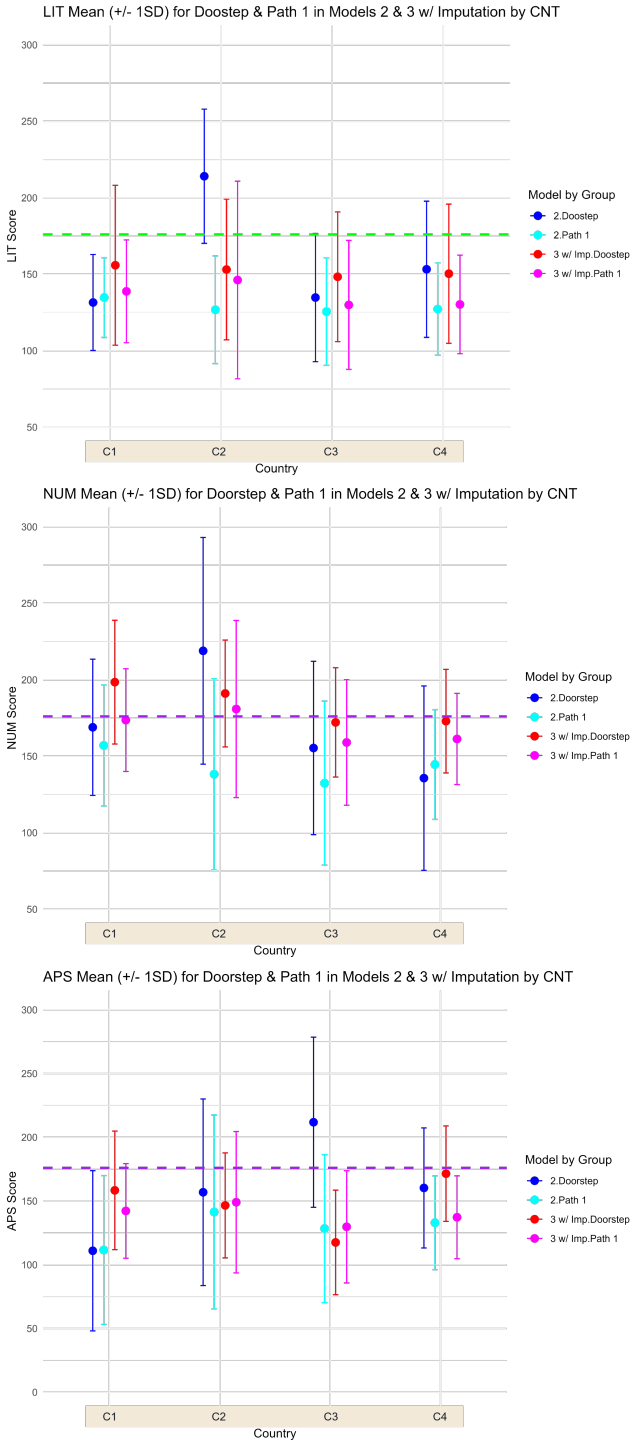

Figure 3. Proficiency Mean (+/- SD) for Doorstep Interview and Path 1 cases under Recommended Settings of Models 2 and 3 (Note. The horizontal dashed line indicates the cut score for Proficiency Level 1, set at 176.)

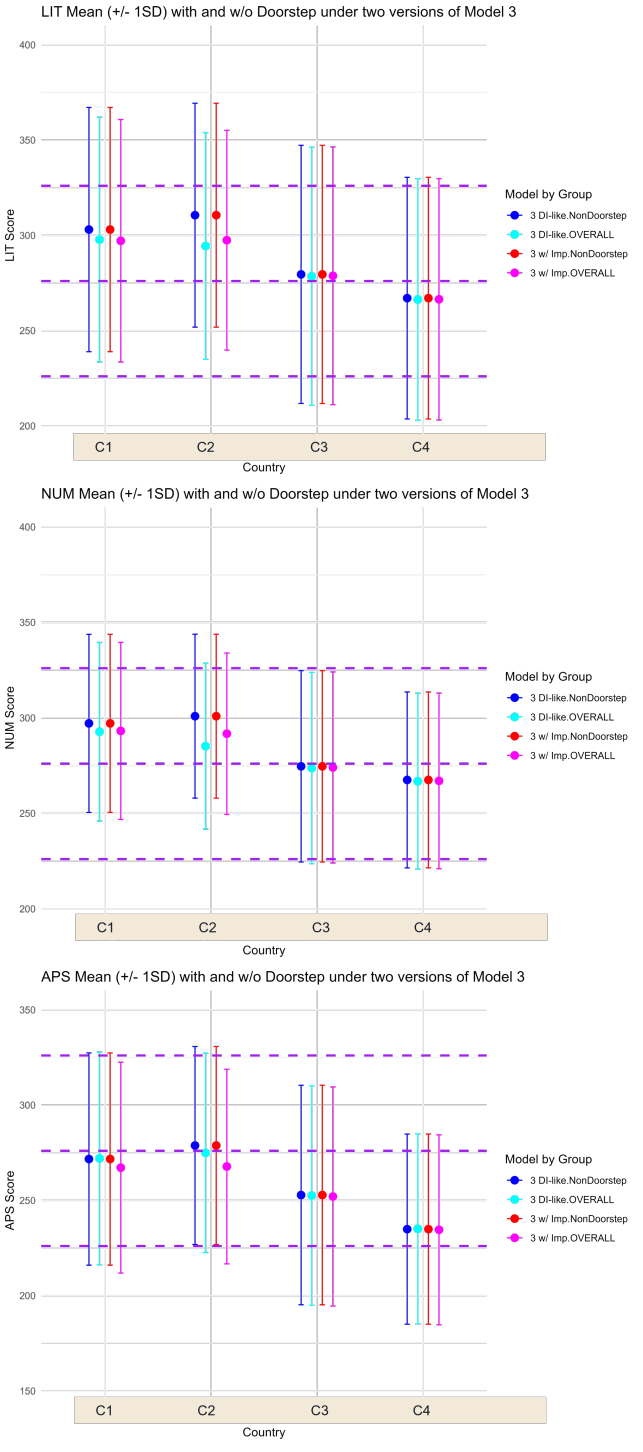

Figure 4. Country Mean (+/- SD) with and without Doorstep Interview Cases under Two Versions of Model 3 (Note. The horizontal dashed line indicates the cut score for Proficiency Level 1, set at 176.)

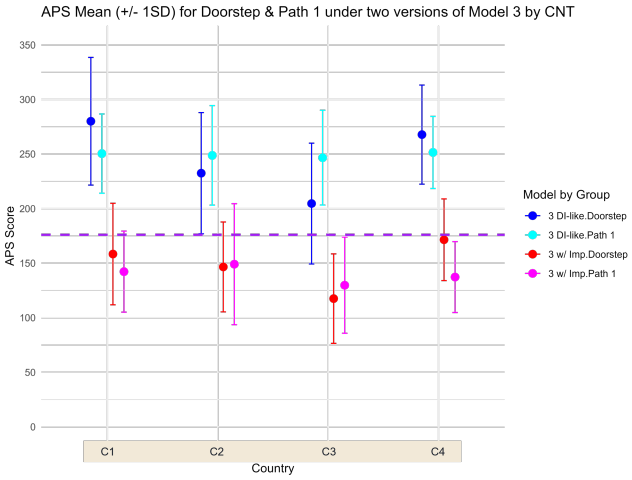

Figure 5. Proficiency Mean (+/- SD) for Doorstep Interview and Path 1 Cases under Two Versions of Model 3 (Note. The horizontal dashed line indicates the cut score for Proficiency Level 1, set at 176.)