

Simulating Topic–Audience Interactions in Sociolinguistic Style-Shifting: Speaker Heterogeneity, Missing Data and Inferential Performance

Thamer Marzouq

Assistant Professor, Department of Foreign Languages,
Taif University, Taif, Kingdom of Saudi Arabia

Approved: 08 October 2026

Posted: 10 October 2026

Copyright 2026 Author(s)

Under Creative Commons CC-BY 4.0

OPEN ACCESS

Cite As:

Marzouq, T. (2026). *Simulating Topic–Audience Interactions in Sociolinguistic Style-Shifting: Speaker Heterogeneity, Missing Data and Inferential Performance*. ESI Preprints. <https://doi.org/10.19044/esipreprint.10.2026.p1>

Abstract

Topic and audience can be examined jointly in repeated observations of linguistic variation, but evaluating an analysis requires a clearly defined interaction and an appropriate account of dependence and missingness. This computer experiment compares three estimators of a population-averaged log-odds interaction: a naive binomial analysis, an analysis with speaker-level sandwich uncertainty, and stratified inverse probability weighting. The factorial design crosses four speaker sample sizes, two token densities, three latent heterogeneity regimes, three interaction magnitudes and four observation regimes. It comprises 288 conditions with 2000 replications each. All observations are synthetic and computer-generated; no participants or empirical data were involved. Every numerical parameter is an illustrative assumption, not an estimate extracted from previous research. Conditional generating coefficients are calibrated to the same marginal target before bias, coverage, false-positive rates and positive detection are evaluated. With 60 hypothetical speakers, 100 tokens per cell and complete observations, the naive false-positive rate rises from 2.85% with a random intercept to 21.35% with additional random slopes; the corresponding speaker-level rate in the latter condition is 6.10%. With random slopes, 240 speakers and a target interaction odds ratio of 1.26, weighted positive detection is 99.95% with complete observations and 63.30% under covariate-dependent missingness. Weighting does not remove bias caused by latent selection. The contribution

is a transparent benchmark linking the interaction target, repeated-observation structure and missingness assumptions to inferential performance; it provides no empirical evidence about actual speakers or stylistic preferences.

Keywords: Style shifting; audience design; simulation; marginal interaction; clustered observations; missing data

1. Introduction

Studies of style shifting often ask whether a speaker's use of a linguistic feature changes with the topic of conversation, the audience, or their combination. Audience design gives interlocutors a central role in stylistic variation (Bell, 1984). Research on addressee and topic has also treated their contributions jointly (Rickford & McNair-Knox, 1994), while topic-specific phonetic variation illustrates why a conversation's subject cannot always be treated as background context (Love & Walker, 2013). These ideas motivate a repeated-measures design in which each speaker contributes observations under several topic and audience conditions.

Such a design poses a statistical question before it answers a sociolinguistic one. An interaction can describe a contrast for an individual with specified characteristics or a contrast between means averaged across a speaker population. With a nonlinear link, these are generally different quantities. Binary-response models for clustered data can therefore address different scientific questions even when their coefficients receive similar verbal labels (Neuhaus, 1992). A simulation that generates a conditional coefficient and evaluates a marginal estimate against that coefficient can mistake a difference in targets for estimator bias. Likewise, setting a conditional interaction coefficient to zero does not automatically create a marginal interaction null.

Dependence and missing observations introduce further complications. Many tokens from one speaker do not provide the same information as observations from many independently sampled speakers. At the same time, within-speaker comparisons can remove shared variation. Losing observations in some conditions changes that covariance structure, even when the missingness is unrelated to the outcomes. Accounting for dependence addresses uncertainty; it does not by itself correct a change in which speakers contribute to each condition mean.

The present study develops a reproducible benchmark for these problems using a crossed topic and audience design. Its methodological contribution is the alignment of a known marginal target with three transparent analyses across controlled variations in speaker count, token density, latent heterogeneity and missing observations. The estimators are

established constructions rather than a new statistical method. Their value here is that the source of a discrepancy can be traced to the target, the standard error calculation or the observation mechanism. The study follows the principle that a simulation should specify its aims, data-generating mechanisms, estimands, methods and performance measures explicitly (Morris et al., 2019).

All study observations are synthetic and computer-generated. There were no participants, recordings, field observations, elicitation sessions or empirical corpora. The numerical settings are illustrative assumptions selected to expose differences in performance. None was estimated from the cited literature. Consequently, the simulated pattern cannot establish that a real topic, audience or population produces more careful, formal or standard-associated speech. The labels give the statistical design a sociolinguistic interpretation without supplying empirical validation.

2. Conceptual scope and research questions

Style is not adequately represented by a single continuum from casual to careful speech. Coupland (2007) treats style in relation to identity, and Eckert (2008) explains how a variable's potential social meanings depend on its indexical field. Affect can also enter the organisation of style (Pratt, 2023). These perspectives limit the interpretation of the outcome used here. A binary indicator of a hypothetical standard-associated variant is a convenient response variable, not a comprehensive measure of style, communicative competence or linguistic quality.

The four topic labels are Intimate, Local identity, Education and Language; the two audience labels are Familiar and Unfamiliar. They are design labels, with no claim that their meanings or effects are invariant across communities. The endpoint contrast between Language and Intimate is chosen before simulation because it provides a simple interaction estimand. It is not an omnibus test of all topics. The intermediate topics preserve the broader crossed design and contribute to the definition of overall missingness, but they do not contribute to the focal estimate.

This separation of a statistical target from social interpretation also matters for measurement. Recent work on multidialectal style shifting supports attention to individual speakers and mixed methods (Fernández-Mallat, 2025), while work on lexical variation argues for incorporating lexis into variationist analysis (Sandow et al., 2026). An empirical extension would therefore need locally defensible linguistic variables, observations of interaction and evidence about speakers' orientations. The present simulation evaluates none of those measurement choices.

Three research questions organise the analysis. RQ1 asks how calibrating the conditional generating model to a common marginal

interaction changes the definition of the null and the interpretation of bias. RQ2 asks how speaker count, token density and latent heterogeneity affect estimation error, interval coverage and false-positive rates. RQ3 asks how missingness that is random, explained by an observed covariate, or related to latent speaker characteristics changes these measures and positive detection, and what stratified weighting can recover under each mechanism.

3. Methods

3.1. Design and performance targets

Table 1 specifies the full factorial design. Four speaker counts are crossed with two token counts per speaker-condition cell, three heterogeneity regimes, three marginal interaction odds ratios and four observation regimes. Each of the resulting 288 conditions is replicated 2000 times, yielding 576000 synthetic data sets and 1728000 method evaluations. The varied quantities are design assumptions rather than empirically calibrated scenarios. A broad factorial comparison is used because performance can depend on combinations of assumptions, not only on one factor changed around a reference condition (Burton et al., 2006).

Table 1: Simulation factors and their assumed values

Factor	Values
Speakers per data set	30, 60, 120, 240
Eligible tokens per speaker and cell	20, 100
Latent standard deviations (intercept, audience, interaction)	(0, 0, 0); (0.55, 0, 0); (1, 0.4, 0.6)
Target marginal interaction odds ratio	1.00, 1.15, 1.26
Observation regime	Complete; MCAR; MAR given X; latent selection
Expected missing cells in incomplete regimes	30% across the eight cells
Replications per factorial condition	2000
Factorial conditions and data sets	288 conditions; 576000 data sets

All study observations are computer-generated. All numerical settings are illustrative assumptions. Latent effects are mutually independent, and measured covariate X is present in every regime.

An eligible token is a hypothetical opportunity for the binary variant to occur. Token density means a fixed 20 or 100 opportunities in every complete speaker-condition cell, not a rate per minute. Each hypothetical speaker is independently generated and has all eight intended cells. Observation loss removes an entire cell, including its outcomes. The design does not model loss of individual tokens within retained cells, complete speaker dropout or a varying eligible denominator.

3.2. Generating model and marginal interaction

For speaker i , topic t and audience a , the binary count has a binomial distribution with m eligible tokens and probability p . The unfamiliar audience indicator a is zero or one. The language-topic indicator L is one for Language and zero otherwise. A fully observed covariate X takes values -1 and 1 with equal probability; its coefficient is 0.6 . X is a generic measured predictor and is not assigned an age, gender, ethnicity or other empirical interpretation. Conditional on X and the speaker's latent effects, counts in different cells are independent.

$$\text{logit}(p_{ita}) = -0.82 + \beta_t + \log(1.41)a + \beta_{LA}L_t a + 0.6X_i \\ + b_{0i} + b_{Ai}a + b_{Li}L_t a$$

The topic coefficients are the logarithms of 1, 1.5, 3 and 4.1 in the listed topic order. The shared intercept is -0.82 and the unfamiliar-audience coefficient is $\log(1.41)$. These constants, the coefficient of X and all variance parameters are illustrative assumptions. Independent, mean-zero Gaussian latent effects represent a random intercept, a random audience slope and a random Language-by-Unfamiliar slope. Their standard deviations are respectively $(0, 0, 0)$, $(0.55, 0, 0)$ or $(1, 0.4, 0.6)$. The first regime has no latent effects but still contains measured heterogeneity through X . In the last regime, speakers can differ both in their overall use of the variant and in their response to a condition.

Let the marginal cell mean be the expectation of p over X and all latent effects in the hypothetical speaker superpopulation. The primary target is the Language-versus-Intimate difference in the unfamiliar-versus-familiar log-odds contrast:

$$\gamma = \text{logit}(\mu_{L1}) - \text{logit}(\mu_{L0}) - \text{logit}(\mu_{I1}) + \text{logit}(\mu_{I0})$$

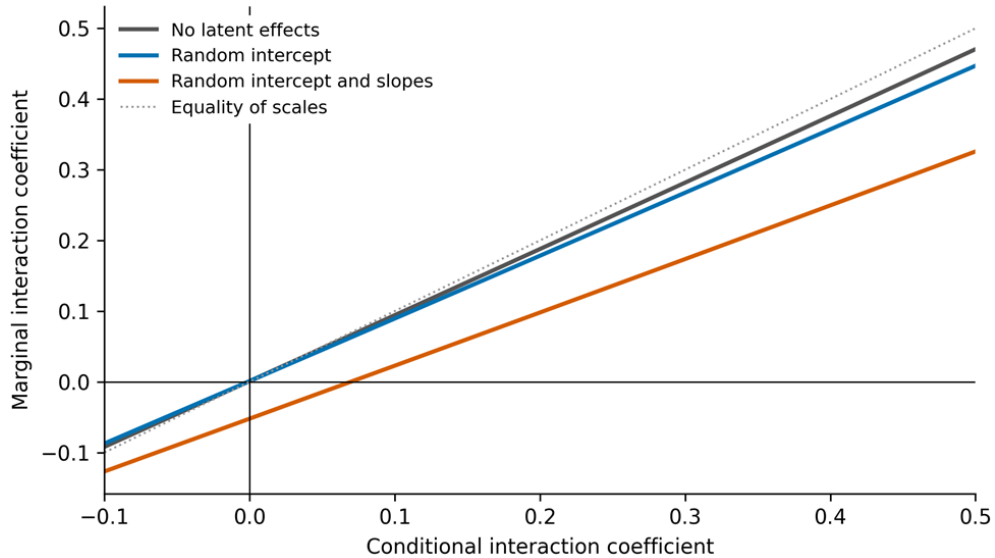
Here I and L denote Intimate and Language, and 0 and 1 denote Familiar and Unfamiliar. Exponentiating gamma gives a ratio of two marginal audience odds ratios. It is not a percentage-point difference in variant use. Bias and root mean squared error are evaluated on the gamma scale. The estimand averages over a newly sampled population of hypothetical speakers, rather than conditioning on the realised covariate distribution of one finite data set.

For each heterogeneity regime, 80-node Gauss-Hermite integration computes the marginal means. A scalar root search then chooses the conditional interaction coefficient that makes gamma equal to $\log(1.00)$, $\log(1.15)$ or $\log(1.26)$. Table 2 reports these calibrated values, and Figure 1 shows the distinction between the scales. Independent adaptive quadrature checks both the integration and the resulting targets. This calibration is essential for interpreting a false-positive rate: every null condition has gamma equal to zero on the scale being tested.

Table 2: Conditional coefficients calibrated to marginal interaction targets

Latent regime	Marginal OR	Marginal gamma	Conditional beta	Conditional OR
None	1.00	0.0000	-0.0011	0.9989
None	1.15	0.1398	0.1487	1.1603
None	1.26	0.2311	0.2461	1.2791
Intercept	1.00	0.0000	-0.0011	0.9989
Intercept	1.15	0.1398	0.1566	1.1696
Intercept	1.26	0.2311	0.2592	1.2959
Intercept and slopes	1.00	0.0000	0.0697	1.0721
Intercept and slopes	1.15	0.1398	0.2553	1.2908
Intercept and slopes	1.26	0.2311	0.3758	1.4562

OR = odds ratio. Marginal means average over X and latent effects. A conditional interaction null and a marginal interaction null need not coincide. Values are numerical properties of the stipulated model.

Figure 1: Conditional and marginal interaction scales in the stipulated model

Curves integrate over measured covariate X and the specified latent effects. The grey dotted line marks equality of scales. The horizontal and vertical zero lines distinguish the marginal and conditional nulls. These are exact numerical model calculations rather than estimates from empirical data.

3.3. Observation mechanisms

The complete regime retains every intended cell. Under missing completely at random (MCAR), each cell is independently retained with probability 0.70. In the covariate-dependent regime, retention depends on the fully observed X and the two condition indicators:

$$\text{logit } \Pr(R_{ita} = 1 | X_i, t, a) = \alpha - 1.2X_i - 0.5L_t - 0.25a$$

R denotes retention. A regime-specific intercept alpha makes the expected retained proportion equal to 0.70, averaged over all eight cells and

the hypothetical speaker population. This missing-at-random (MAR) mechanism permits retention to be associated with the binary outcome through X . It assumes that, after conditioning on X and the cell, retention provides no further information about that cell's outcome. X and the intended cell labels remain available for every hypothetical speaker, including speakers with few retained outcomes.

The latent-selection stress test adds -1.2 times the sum of the latent intercept and the latent interaction contribution to the retention logit. Thus, when latent variation is present, retention depends on quantities related to the outcome that are unavailable to the analysis. This violates the conditional independence assumption used for weighting on X alone. The intercept is recalibrated so that overall expected missingness remains 30%. With zero latent standard deviations this regime reduces to the MAR mechanism. Conditional on the generating variables, retention draws are independent across cells. The regimes are controlled examples; 30% is not an estimated rate of missing observations in sociolinguistic research.

3.4. Three analysis methods

All three methods use the four focal cells and estimate the same marginal log-odds contrast. The naive binomial method calculates an observed proportion for each cell, transforms the proportions to logits and applies the signed interaction contrast. Its delta-method variance treats all retained tokens as independent and ignores cross-cell covariance. A normal critical value gives its nominal 95% interval. This is a deliberately simple baseline for evaluating the consequences of a common independence assumption.

The speaker sandwich method uses the same point estimate but calculates uncertainty from speaker-level influence contributions. This preserves covariance between outcomes contributed by one speaker. Its variance has the finite-sample factor $N/(N-1)$, and its interval uses a t critical value with $N-1$ degrees of freedom. A sandwich calculation is motivated by the distinction between a mean model and within-cluster dependence in longitudinal analysis (Liang & Zeger, 1986). Here it is implemented directly for saturated cell means, so it does not require fitting a random-effects model. The correction is approximate, particularly at small N .

The stratified inverse probability weighting method estimates retention probabilities separately in each focal cell and each X stratum, as the number retained divided by the number of intended speakers in that stratum. It then standardises observed stratum means to the empirical X distribution of all intended speakers. In this saturated setting, the result is also a poststratified mean. Its influence calculation includes the estimation of the retention probabilities and the sampling of X . The standard error and critical

value use the same speaker-level correction as the sandwich method. Under MAR given X , positivity and independent sampling of speakers, it targets the full-population cell means. Weighting cannot generally recover these means when retention still depends on unobserved outcome predictors (Seaman & White, 2013).

No method can use the latent effects, generating probabilities or masked outcomes. The implementation masks outcomes before arithmetic so that an unavailable value cannot affect an estimate even if stored as a number or a missing-value marker. An estimate is recorded as a failure when a required mean is undefined, a logit is nonfinite, or its standard error is nonfinite or nonpositive. Weighting also fails if an X stratum is absent or a focal cell has no observed outcomes in a required stratum. There is no continuity correction, weight truncation, replacement draw or selective retry. These are finite-data identification failures, not iterative model-convergence failures.

3.5. Performance measures and Monte Carlo uncertainty

For each method-condition combination, bias is the mean estimated gamma minus its known target. Root mean squared error combines systematic error and sampling variation. Empirical standard deviation is compared with the mean reported standard error, and mean interval width records precision. Coverage is the proportion of valid intervals containing the known gamma. At the null, the false-positive rate is the proportion excluding zero in either direction. For positive targets, positive detection is the proportion whose lower confidence limit exceeds zero. Detection therefore requires evidence in the correct direction; a significant negative estimate does not count as successful detection.

Bias, root mean squared error, coverage and conditional rejection measures use valid estimates, with their denominators reported. The main detection displays instead divide successful detections by all 2000 replications, counting a failed estimate as no detection. This distinction prevents a method from appearing more useful merely because difficult data sets were omitted. The package reports both detection denominators, failure counts and failure rates. Each test is a two-sided 5% test of the single prespecified contrast. There is no multiple-testing adjustment because the separate simulation conditions assess repeated operation of that test and do not form a family of empirical discoveries.

Monte Carlo standard errors (MCSEs) quantify simulation uncertainty. For a proportion p estimated from B valid replications, the MCSE is the square root of $p(1-p)/B$; the all-replication detection measure uses $B = 2000$. At a true proportion near 0.95 or 0.05, this is approximately 0.49 percentage points with 2000 valid replications. At 0.50 it is

approximately 1.12 points. Bias MCSE is the empirical standard deviation divided by the square root of B; a delta approximation is used for root mean squared error. Differences of a few tenths of a percentage point are consequently not interpreted as meaningful rankings. The complete metrics and their MCSEs are supplied for every condition.

3.6. Reproducibility and illustrative data

The base seed is 20260913. A distinct seed derived from each scenario identifier and replication index initiates an independent NumPy PCG64 stream. Changing computational batch size does not change a replication. All methods are applied to the same data within each replication, whereas different factorial conditions use separate streams. The package contains the full grid, replicate-level estimates and standard errors, analysis summaries, plot data, figure files, executable scripts, software versions and checksums. Exact commands and validation results appear in the accompanying documentation.

A separately generated example has 240 hypothetical speakers, 100 eligible tokens per cell, random intercepts and slopes, MAR retention and a target interaction odds ratio of 1.26. It contains 1920 intended records, of which 1,350 are observed and 570 are masked. Both the full synthetic truth and the analysis-ready masked data are supplied under clearly distinct filenames. The example is a data-format and workflow demonstration, not an additional empirical study or a Monte Carlo replication selected for its result.

The separate example additionally includes five auxiliary numerical proxies: lexical formality, lexical diversity, speech rate, filled-pause frequency and F0 range. They are generated directly by stated formulae and noise distributions, rather than extracted from texts or recordings. In particular, the lexical-diversity proxy is not a calculated MTLTD score. Their imposed co-patterning supplies no convergent validity for a construct of careful speech. They are documented and summarised in the package but play no role in the focal performance evaluation.

4. Results

4.1. A matched marginal null

Calibration gives materially different conditional interaction coefficients across heterogeneity regimes (Table 2). Under random intercepts and slopes, a marginal null requires a conditional coefficient of 0.0697, whereas a target marginal odds ratio of 1.26 requires 0.3758. Zero on the conditional scale is therefore not the null tested by the marginal estimators in this regime. The small departures from zero in the other regimes also reflect averaging through the nonlinear link, including averaging over X. These

quantities describe the stipulated generating model and do not estimate a substantive topic effect.

The independent numerical integration agrees with the quadrature used in the generating code to within 10^{-10} in absolute probability, and every calibrated target is recovered to within 10^{-10} on the log-odds scale. A separate full rerun reproduces every stored replicate estimate, standard error, missingness array and derived numerical data file. These checks establish computational consistency for this implementation; they do not establish realism of the assumed model.

4.2. Heterogeneity and uncertainty under complete observation

Table 3 holds the speaker count at 60 and token density at 100 under the marginal null. With no latent effects, the naive false-positive rate is 4.15%. With a random intercept it is 2.85%, indicating conservative inference in this condition. With random slopes added it is 21.35%, with only 78.65% coverage. The speaker sandwich method gives 6.10% false positives and 93.90% coverage in the latter condition. Because both methods use exactly the same estimate, the contrast concerns uncertainty rather than a change in point estimation.

Table 3: Null performance with complete observations and 60 speakers

Latent regime	Method	Bias	Empirical SD	Mean SE	Coverage %	FPR % (MCSE)
None	Naive	0.0003	0.0530	0.0546	95.85	4.15 (0.45)
None	Sandwich	0.0003	0.0530	0.0524	95.15	4.85 (0.48)
None	IPW	0.0003	0.0530	0.0524	95.15	4.85 (0.48)
Intercept	Naive	0.0002	0.0495	0.0544	97.15	2.85 (0.37)
Intercept	Sandwich	0.0002	0.0495	0.0508	95.85	4.15 (0.45)
Intercept	IPW	0.0002	0.0495	0.0508	95.85	4.15 (0.45)
Intercept and slopes	Naive	0.0006	0.0843	0.0539	78.65	21.35 (0.92)
Intercept and slopes	Sandwich	0.0006	0.0843	0.0809	93.90	6.10 (0.54)
Intercept and slopes	IPW	0.0006	0.0843	0.0809	93.90	6.10 (0.54)

Each condition has 100 tokens per cell, target marginal OR = 1 and 2000 valid replications. Bias, empirical SD and mean SE are on the log-odds scale. FPR = two-sided false-positive rate; MCSE is in percentage points. All nominal intervals are 95%. With a matched null, coverage and FPR sum to 100%.

The direction of the independence error is not uniform. Shared intercept variation can partly cancel in a within-speaker interaction, while additional interaction variation creates dispersion that token-level binomial uncertainty misses. Across all complete-data conditions, speaker sandwich coverage ranges from 93.70% to 96.20%. This is broadly near the nominal 95%, with some undercoverage remaining. The weighted estimator coincides with the unweighted speaker estimator in complete data because poststratification then reconstructs the original sample cell means and influence contributions.

4.3. Missing observations and bias

Table 4 compares missingness regimes at 120 speakers, 100 tokens per cell and random intercepts and slopes under the marginal null. Naive false-positive rates are 20.45% with complete data, 50.40% under MCAR and 51.65% under MAR. With MCAR, the speaker sandwich and weighted methods have coverage of 95.15% and 95.40%, respectively. These results show that a missingness mechanism can be ignorable for the mean while still changing the appropriate variance.

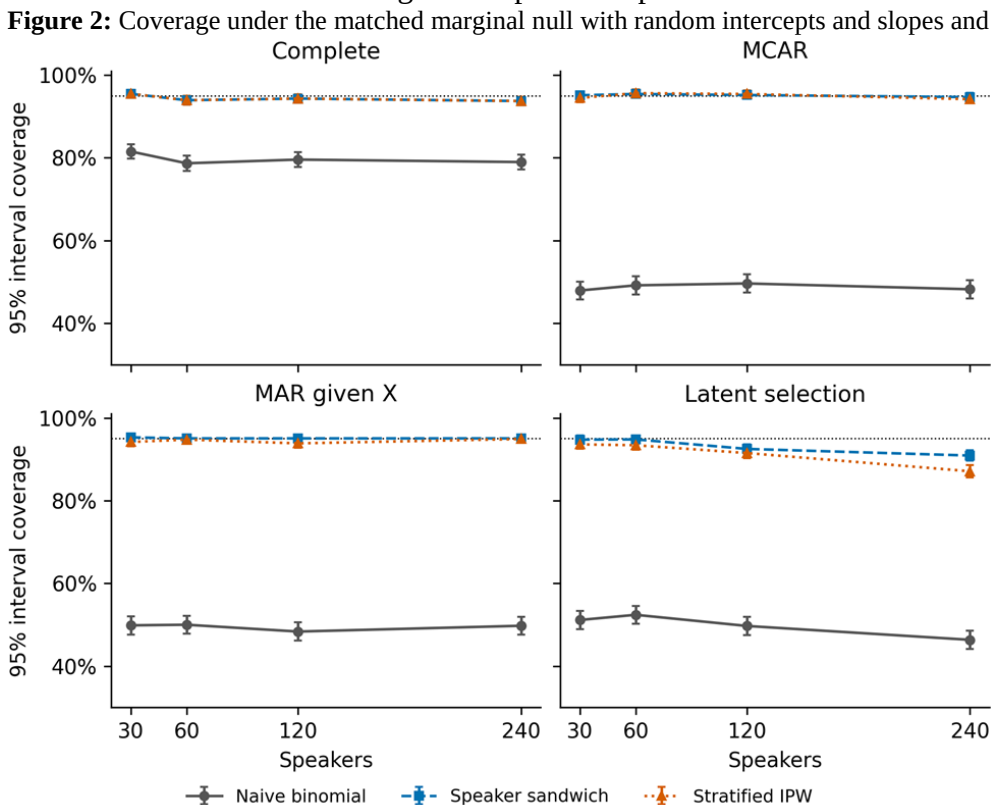
Table 4: Null performance across observation mechanisms

Observation regime	Method	Bias	RMSE	Coverage %	FPR % (MCSE)
Complete	Naive	0.0011	0.0589	79.55	20.45 (0.90)
Complete	Sandwich	0.0011	0.0589	94.30	5.70 (0.52)
Complete	IPW	0.0011	0.0589	94.30	5.70 (0.52)
MCAR	Naive	0.0028	0.1352	49.60	50.40 (1.12)
MCAR	Sandwich	0.0028	0.1352	95.15	4.85 (0.48)
MCAR	IPW	0.0021	0.1213	95.40	4.60 (0.47)
MAR	Naive	0.0153	0.1369	48.35	51.65 (1.12)
MAR	Sandwich	0.0153	0.1369	95.05	4.95 (0.49)
MAR	IPW	0.0090	0.1470	93.85	6.15 (0.54)
Latent selection	Naive	-0.0504	0.1383	49.70	50.30 (1.12)
Latent selection	Sandwich	-0.0504	0.1383	92.50	7.50 (0.59)
Latent selection	IPW	-0.0654	0.1453	91.50	8.50 (0.62)

Random intercepts and slopes; 120 speakers; 100 tokens per cell; target marginal OR = 1; 2000 valid replications per method and condition. In incomplete regimes, expected missingness is 30% across all cells. Bias and RMSE are on the log-odds scale. MCSE is in percentage points. Full bias and RMSE MCSEs are in performance.csv.

Under MAR in this selected condition, the unweighted bias is 0.0153 and weighted bias is 0.0090 on the log-odds interaction scale. The corresponding root mean squared errors are 0.1369 and 0.1470. Weighting is therefore not uniformly more precise or more accurate in finite samples, even when its identifying assumption holds. Small residual bias in a contrast can coexist with substantial changes in individual cell means, and weighting introduces additional variability from sparsely observed strata. The full results are required to distinguish this finite-sample trade-off from a general claim about weighting.

With latent selection, the weighted bias is -0.0654, its coverage is 91.50%, and its false-positive rate is 8.50% in the same reference condition. The unweighted speaker method also loses coverage. Figure 2 shows that the shortfall persists and becomes larger as N increases in this stress test. More speakers narrow intervals around a selected population contrast; they do not remove the selection mechanism. By contrast, the MCAR results remain much closer to nominal coverage after speaker dependence is accounted for.



100 tokens per cell. Error bars are plus or minus 1.96 Monte Carlo standard errors; the horizontal dotted line marks 95%. Denominators are valid estimates. The speaker sandwich and IPW curves coincide with complete observations. MCAR and MAR regimes have 30% expected missingness; latent selection depends additionally on unavailable speaker effects.

4.4. Sample size token density and detection

Table 5 reports weighted positive detection for the two nonzero targets with 100 tokens per cell. For a target odds ratio of 1.26 and complete data, detection rises from 47.65% at 30 speakers to 79.35% at 60, 97.35% at 120 and 99.95% at 240. Under MAR, the corresponding values are 15.40%, 23.15%, 36.75% and 63.30%. For the smaller target of 1.15, detection at 240 speakers is 93.05% with complete observations and 29.65% under MAR. These are conditional properties of the simulated designs, not sample-size requirements for a real community.

Table 5: Positive detection by speaker count under complete observation and MAR

Target OR	Speakers	Complete % (MCSE)	MAR % (MCSE)	MAR failures
1.15	30	20.45 (0.90)	8.60 (0.63)	8
1.15	60	36.70 (1.08)	9.95 (0.67)	0
1.15	120	64.95 (1.07)	17.40 (0.85)	0
1.15	240	93.05 (0.57)	29.65 (1.02)	0
1.26	30	47.65 (1.12)	15.40 (0.81)	11
1.26	60	79.35 (0.91)	23.15 (0.94)	0
1.26	120	97.35 (0.36)	36.75 (1.08)	0
1.26	240	99.95 (0.05)	63.30 (1.08)	0

Stratified IPW; random intercepts and slopes; 100 tokens per cell. Positive detection means the lower 95% confidence limit exceeds zero on the log-odds scale. Percentages use all 2000 replications, including failures as no detection. MCSE is in percentage points. No complete-data estimate failed.

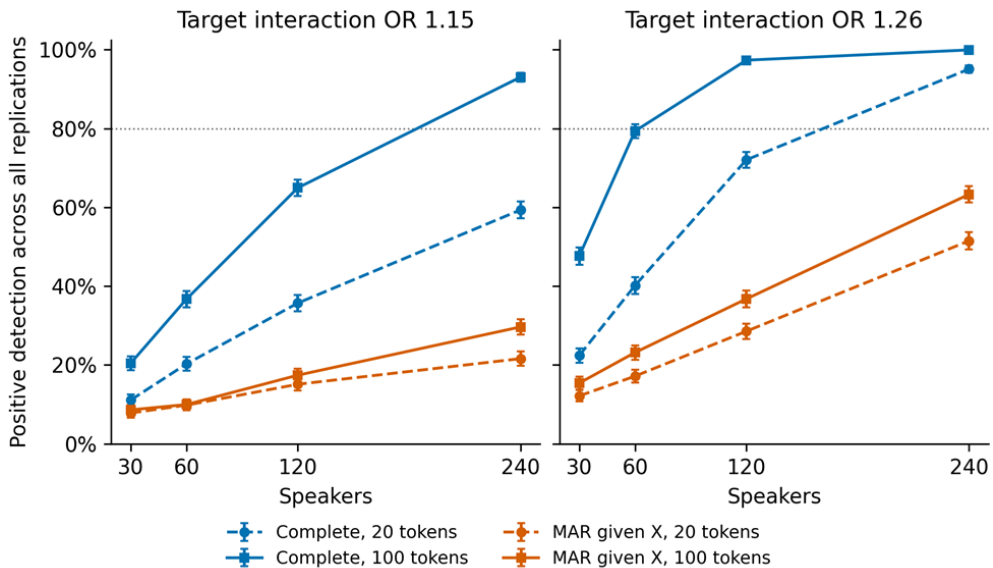


Figure 3: Weighted positive detection with random intercepts and slopes across speaker counts and token densities. Each percentage uses all 2000 replications, treating failures as no detection. Error bars are plus or minus 1.96 Monte Carlo standard errors. The dotted 80% line is an illustrative planning reference, not a prescribed threshold. Results are conditional on the synthetic generating and observation mechanisms.

Figure 3 shows how higher token density can improve detection, with its contribution depending on the number of speakers and retention mechanism. For the 1.26 target with complete observations, 120 speakers and 100 tokens yield 97.35% detection, compared with 95.10% for 240 speakers and 20 tokens. Thus, it would also be misleading to claim that increasing tokens can never be competitive. The study does not attach recruitment or annotation costs to these designs, and it does not identify a cost-optimal allocation. Under MAR, even the largest combination examined reaches only 63.30% for this target, below the illustrative 80% reference line.

4.5. Failures and the separate example

Neither the naive nor speaker sandwich method fails in the 576000 simulated data sets. The weighted method fails 229 times across 35 conditions, with at most 14 failures in one condition. The largest condition-specific failure proportion is therefore 0.70%. Failures are reported separately and included as unsuccessful outcomes in the principal detection displays. Their rarity in this grid should not be extrapolated to designs with smaller strata, lower retention or more covariates.

In the separate masked example, the weighted interaction odds ratio is 1.365 with a 95% interval of 1.140 to 1.634. This interval contains the generating target of 1.26. A single interval containing its target does not establish coverage: coverage is assessed by the repeated experiment above. The observed example contains 135,000 eligible opportunities and 66,244 occurrences of the hypothetical variant. Those values are generated counts, not speech tokens collected from people.

5. Discussion

5.1. The contribution of defining the interaction first

The first research question is resolved by an explicit calibration step. A conditional generating coefficient and a population-averaged interaction describe different contrasts in this binary design. Aligning the simulation truth with the estimand prevents a systematic mismatch from being reported as bias or false-positive behaviour. This principle extends the practical interpretation of clustered binary-response models to a topic and audience design. It does not make one target inherently preferable: a study concerned with a particular speaker's response could reasonably choose a conditional target, whereas a comparison of average variant use may require a marginal one.

The benchmark also changes what counts as a successful simulation. A programmed increase in variant use under a topic label is expected to appear in synthetic means. Recovering that ordering does not validate audience design, attention to speech or a theory of careful speech. Here the

evidence concerns properties of analysis procedures under known assumptions: whether an estimate is centred on its target, whether uncertainty is calibrated, how often a null is rejected and whether a specified nonzero contrast is detected. The stipulated direction of topic coefficients has no independent evidential status.

5.2. Speaker dependence and the value of retained contrasts

The second research question shows why a large token denominator is insufficient to justify a binomial standard error. Speaker-specific responses leave variation after tokens are aggregated. Yet the conservative random-intercept result also cautions against treating every dependence problem as a simple inflation factor. What matters is the covariance of the particular contrast. A component common to several cells can cancel, whereas variation specific to the interaction remains.

The missingness comparisons sharpen this point. Under complete observation, the same speakers contribute to every term of the interaction. Under cell loss, the effective speaker composition differs across terms, so the balance that allowed common variation to cancel is weakened. The marked increase in naive false positives under MCAR cannot be attributed to outcome-dependent selection. It follows from applying an unsuitable independence variance to a changed repeated-observation structure. Retaining observations across all intended conditions can therefore matter for precision as well as for avoiding bias.

These findings support planning with both speaker count and token count visible. Additional tokens reduce binomial noise within a cell; additional speakers average over independently generated speaker responses. The relative benefit depends on their variances and on what is observed. Because the current parameter values and costs are hypothetical, the results supply a framework for sensitivity analysis rather than a universal minimum number of speakers.

5.3. Missingness assumptions govern what weighting can recover

The third research question separates uncertainty correction from selection correction. A speaker sandwich standard error can reflect repeated observations while leaving the point estimate centred on the observed cases. Weighting uses the fully observed covariate to restore the intended population composition under the specified MAR mechanism. However, estimating retention in small cell-strata can increase variance and occasionally make the estimator undefined. The reported finite-sample results therefore do not establish uniform dominance of weighting over the unweighted analysis.

The latent-selection condition makes a different limitation visible. Its observation model uses unobserved speaker variation related to the outcome. Weighting on X alone cannot reconstruct that missing information. As sample size grows, greater precision can make this bias more consequential for coverage. In an empirical application, an assertion that data are MAR would need substantive support from the collection process and observed predictors; robust standard errors would not supply that support. Sensitivity analyses would need mechanisms appropriate to the actual source of missing observations.

5.4. Sociolinguistic interpretation and limits

The findings are about inference for one hypothetical binary feature. They do not show that discussing language makes real speakers more standard, that an unfamiliar audience induces careful speech, or that standard-associated forms communicate better. An empirical study would have to justify the eligible context, identify the socially meaningful variant and examine whether the topic and audience categories were recognised in interaction. The history of stylistic analysis, including Labov (1972), supplies conceptual questions, not numerical calibration for this computer experiment.

Several restrictions delimit the benchmark. The generating distribution has independent Gaussian latent components, a binary covariate with fixed probabilities, fixed token counts and only one nonzero missingness rate. Tokens are independent conditional on the speaker effects; there are no item effects, interviewer effects, serial dependence, carryover or uncertain coding. Correlated latent effects, different baseline prevalences, skewed speaker distributions, variable denominators and whole-speaker dropout could change performance. These conditions were not simulated and no robustness claim is made for them.

The comparison also omits mixed-effects logistic models, multiple imputation, augmented weighting, penalised estimation and alternative finite-cluster corrections. A mixed-effects model could answer a conditional question or yield a marginal target after integration, but either use would require a separately specified evaluation. The present saturated estimators were chosen for transparency. The t correction reduces neither selection bias nor every small-sample coverage error. Finally, the endpoint interaction is scale dependent and does not represent the entire four-topic pattern. A probability-scale contrast or an omnibus test could produce a different planning conclusion.

Conclusion

This synthetic experiment supplies an auditable way to evaluate a topic by audience analysis against a matched marginal target. Under the stipulated conditions, speaker-level uncertainty substantially improves calibration over token independence, while missing observations can reduce detection and expose selection bias. Stratified weighting addresses the specified observed-covariate mechanism but cannot resolve the latent-selection stress test. The substantive value is a reusable design-evaluation procedure and a clear account of its assumptions. Every observation remains computer-generated, and every numerical setting remains illustrative rather than evidence about actual language use.

Generative artificial intelligence use: During preparation and revision of this manuscript, the author used OpenAI ChatGPT in a Codex environment, accessed in September 2026, to assist with language development, document structuring, methodological specification, Python code generation and execution, numerical checking, reference checking and preparation of the transparent synthetic-data workflow. The author accepts full responsibility for the accuracy, originality, integrity and journal-policy compliance of the manuscript and all associated materials. No artificial-intelligence output is presented as human-participant or empirical data. All numerical parameters are illustrative assumptions rather than estimates extracted from previous research.

Conflict of Interest: The author reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: The author did not obtain any funding for this research.

References:

1. Bell, A. (1984). Language style as audience design. *Language in Society*, 13(2), 145-204.
<https://doi.org/10.1017/S004740450001037X>
2. Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279-4292. <https://doi.org/10.1002/sim.2673>
3. Coupland, N. (2007). *Style: Language variation and identity*. Cambridge University Press.
<https://doi.org/10.1017/CBO9780511755064>

4. Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, 12(4), 453-476. <https://doi.org/10.1111/j.1467-9841.2008.00374.x>
5. Fernández-Mallat, V. (2025). Style-shifting in multidialectal spaces: The case for speaker-based, mixed-methods approaches to dialect contact. *Languages*, 10(12), Article 299. <https://doi.org/10.3390/languages10120299>
6. Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.
7. Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22. <https://doi.org/10.1093/biomet/73.1.13>
8. Love, J., & Walker, A. (2013). Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech*, 56(4), 443-460. <https://doi.org/10.1177/0023830912453132>
9. Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074-2102. <https://doi.org/10.1002/sim.8086>
10. Neuhaus, J. M. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research*, 1(3), 249-273. <https://doi.org/10.1177/096228029200100303>
11. Pratt, T. (2023). Affect in sociolinguistic style. *Language in Society*, 52(1), 1-26. <https://doi.org/10.1017/S0047404521000774>
12. Rickford, J. R., & McNair-Knox, F. (1994). Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 235-276). Oxford University Press. <https://doi.org/10.1093/oso/9780195083644.003.0011>
13. Sandow, R., Grieve, J., Stamp, R., & Schembri, A. (2026). Lexical variation and change: Integrating lexis into variationist sociolinguistics. *Language Variation and Change*, 37(3), 293-318. <https://doi.org/10.1017/S0954394525100628>
14. Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3), 278-295. <https://doi.org/10.1177/0962280210395740>

Appendix Estimation details

Index the four focal cells in the order Intimate-Familiar, Intimate-Unfamiliar, Language-Familiar and Language-Unfamiliar. Their contrast signs are (1, -1, -1, 1). Let the observed proportion in a retained cell be Y/m . For the two unweighted methods, each estimated cell mean is the average of Y/m over observed speakers in that cell. Its speaker influence contribution is R divided by the empirical retention proportion, multiplied by the deviation of Y/m from the estimated cell mean. Combining these cell contributions with the logit derivatives retains the cross-cell covariance.

For stratified weighting, let the observed mean in cell j and X stratum x be $\hat{m}_j(x)$, and let $\hat{\pi}_j(x)$ be the observed fraction among all intended speakers in that stratum. The weighted mean averages $\hat{m}_j(x)$ using the empirical proportions of X among all N intended speakers:

$$\hat{\mu}_j = \sum_{x \in \{-1,1\}} \frac{n_x}{N} \hat{m}_j(x)$$

This expression equals a normalised inverse probability weighted mean when the stratum retention probabilities are estimated by observed fractions. Its first-order speaker influence contribution accounts for both the covariate composition of the full sample and the retained outcomes:

$$IF_{ij} = \hat{m}_j(X_i) - \hat{\mu}_j + \frac{R_{ij}}{\hat{\pi}_j(X_i)} \left[\frac{Y_{ij}}{m} - \hat{m}_j(X_i) \right]$$

Writing d for the vector of contrast signs divided by the corresponding estimated cell mean times one minus that mean, the estimated interaction variance is:

$$SE(\hat{\gamma})^2 = \frac{1}{N(N-1)} \sum_{i=1}^N \left(\sum_{j=1}^4 d_j IF_{ij} \right)^2$$

The naive method instead sums the four independent binomial logit variances, each equal to the reciprocal of m times the observed speaker count times the estimated cell mean times one minus that mean. All calculations use the log-odds scale before interval endpoints are exponentiated where an odds-ratio display is required. The source code implements these expressions without optimisation. The independent validation script reconstructs them using scalar stratum calculations and a covariance-matrix calculation and checks that hidden or masked outcomes cannot affect the answers.