

Supplementary Robustness and Sensitivity Analyses

Gender, Income, and Financial Fragility Among Working-Age Adults in Kazakhstan · NHSJS final version (v3.1)

These analyses were specified after the primary results were obtained, in response to peer review. They are reported as sensitivity analyses rather than as part of the original design. All are reproducible from the supplied script. Version 3.1, prepared at the proof stage, computes the standard errors of every weighted Firth fit with the same linearized (sandwich) variance as the maximum-likelihood fits, and replaces trimming at the 1st and 99th percentiles, which altered no weight, with trimming at the 5th and 95th. Neither change alters a conclusion.

Table R1. Events per Variable by Outcome

Outcome	Events	n	Parameters	EPV	Assessment
Financial vulnerability	63	714	11 (full)	5.73	Below the conventional minimum of 10
Financial vulnerability	63	714	8 (reduced)	7.88	Improved; still below 10
Formal savings	239	722	10 (full)	23.90	No sparse-data concern

Note. The vulnerability models carry a genuine sparse-data risk; the savings models do not. This asymmetry determines which results the manuscript treats as confirmatory. The reduced specification collapses income from five quintiles to three categories and drops urbanicity, whose coefficient was null in every specification.

Table R2. Income Gradient Across Estimators

Maximum-likelihood logistic regression is biased away from the null with few events per variable. Firth penalization corrects that bias. The income coefficient is the manuscript's largest, so it is the one most exposed.

Model	Estimator	Term	OR	95% CI	p
Full	Weighted ML	Highest quintile vs lowest	0.084	[0.022, 0.320]	.0003
Full	Weighted Firth	Highest quintile vs lowest	0.102	[0.034, 0.310]	.0001
Full	Unweighted Firth	Highest quintile vs lowest	0.127	[0.045, 0.361]	.0001
Reduced	Weighted ML	Q4–5 vs Q1–2	0.202	[0.088, 0.465]	.0002
Reduced	Weighted Firth	Q4–5 vs Q1–2	0.214	[0.099, 0.462]	.0001
Reduced	Unweighted Firth	Q4–5 vs Q1–2	0.229	[0.114, 0.460]	< .0001

Note. The unpenalized full-model estimate of 0.084 is inflated. Penalization moves it to 0.102 and the reduced specification to roughly 0.21. The gradient is real and highly significant under every estimator, but the unpenalized estimate overstates its magnitude. The manuscript reports the reduced penalized estimate as primary and the full ML estimate only in this table. Weighted estimators use the linearized (sandwich) variance; the unweighted Firth fits use the model-based variance.

Table R3. Reduced Vulnerability Model: ML Compared with Penalized Estimation

Term	ML OR	ML p	Firth OR	Firth 95% CI	Firth p
Income Q3 (middle) vs Q1–2	0.798	.556	0.814	[0.397, 1.669]	.574
Income Q4–5 (high) vs Q1–2	0.202	.0002	0.214	[0.099, 0.462]	.0001
Female	2.108	.123	2.056	[0.834, 5.071]	.118
Age 36–55	0.943	.907	0.944	[0.369, 2.420]	.905
Female × Age 36–55	0.639	.486	0.649	[0.196, 2.155]	.481

Term	ML OR	ML p	Firth OR	Firth 95% CI	Firth p
Tertiary education	0.589	.086	0.606	[0.342, 1.075]	.087
Employed	0.644	.213	0.643	[0.331, 1.249]	.192

Note. $n = 714$. Both estimators use the linearized (sandwich) variance. Penalization leaves every conclusion unchanged. The Female $\times$ Age interaction (H1) remains non-significant under both estimators, so the null result for H1 is not an artefact of sparse-data bias.

Table R4. Education: Total Effect and Income-Adjusted Direct Effect

Income lies on the causal path from education to financial outcomes. Conditioning on it and observing attenuation indicates mediation, not that education is irrelevant. The two quantities are therefore reported separately.

Specification	ML OR	ML p	Firth OR	Firth 95% CI	Firth p
Total effect (unadjusted)	0.532	.034	0.545	[0.308, 0.964]	.037
Plus demographics, no income	0.515	.029	0.529	[0.298, 0.938]	.029
Plus income (direct effect)	0.589	.086	0.605	[0.340, 1.077]	.088

Note. About 20% of education's coefficient is transmitted through income. The total effect is significant under both estimators; the direct effect is not significant under either. The manuscript therefore does not claim that education fails to protect against financial vulnerability. It claims the narrower and better-supported point that women's substantial educational advantage does not appear as an advantage in either financial outcome — a descriptive comparison that does not depend on the regression.

Table R5. Formal Savings Gap: Every Specification Tested

This is the manuscript's single confirmatory finding, so it is reported against every challenge raised in review.

Specification	OR or estimate	95% CI	p
Unadjusted weighted proportion contrast	−7.95 pp	[−15.29, −0.62]	.034
Adjusted weighted ML, full	0.618	[0.425, 0.899]	.012
Adjusted weighted Firth, full	0.623	[0.430, 0.904]	.013
Adjusted weighted ML, reduced	0.596	[0.412, 0.863]	.0061
Adjusted weighted Firth, reduced	0.600	[0.415, 0.867]	.0065
Weights trimmed at 5th / 95th percentile	0.605	—	.0070
Clustering, $p = .03$ ($\bar{m} = 10$)	0.600	—	.016
Clustering, $p = .06$ ($\bar{m} = 10$)	0.600	—	.028
Manski bound: all 37 excluded coded as savers	0.672	[0.473, 0.954]	.026
Manski bound: all 37 excluded coded as non-savers	0.606	[0.422, 0.871]	.0067

Note. The estimate is stable between 0.60 and 0.67 across every specification, and remains significant at $\alpha = .05$ under penalization, parameter reduction, weight trimming, plausible unmeasured clustering, and both extreme assumptions about the excluded cases. The extreme bounds are the strongest of these: coding all 37 excluded respondents in the direction least favourable to the finding still leaves $p = .026$. Reporting the unadjusted proportion contrast alone would understate this, since that quantity does not survive the same bound ($p = .109$).

Table R6. Clustering Sensitivity for All Key Tests

Test	$p = 0$	$p = .01$	$p = .03$	$p = .06$
Savings gap, pooled adjusted (Firth, reduced)	.0065	.0092	.0158	.0284
Savings gap, pooled unadjusted	.034	.042	.059	.087

Test	$\rho = 0$	$\rho = .01$	$\rho = .03$	$\rho = .06$
Vulnerability gap, pooled	.076	.089	.115	.153
H1 gender $\times$ age interaction	.307	.328	.365	.411
H3 digital $\times$ age (weighted Firth)	.0090	.0123	.0204	.0353

Note. The design effect from clustering is $1 + \rho(\bar{m} - 1)$, assuming an average cluster size of 10. The adjusted savings gap and the H3 interaction on the primary outcome clear $\alpha = .05$ across the full band. On the composite outcome the H3 interaction does not ($p = .068, .105$ and $.141$ at $\rho = 0, .03$ and $.06$; Table R7).

Table R7. Digital Account $\times$ Age Among Women: Four Estimators

The four estimators disagree, and the source of the disagreement is the survey weights rather than the penalty. Penalization moves the estimate very little within either weighting scheme; dropping the weights moves it a great deal. Unweighted estimators do not produce population quantities, so they are reported for diagnosis rather than as competing answers.

Estimator	Design-based?	Corrects sparse-data bias?	OR	p
Weighted Firth penalized	Yes	Yes	0.109	.009
Weighted maximum likelihood	Yes	No	0.092	.009
Unweighted Firth penalized	No	Yes	0.264	.113
Unweighted maximum likelihood	No	No	0.241	.100

Note. The decisive cell contains 6 vulnerability events among 26 women. Both design-based estimators support the interaction and both unweighted estimators do not, while penalization changes the p -value by at most .013 within either pair. Attributing the divergence to sparse-data correction would therefore be incorrect.

Additional checks on the same interaction:

Check	Result	Assessment
Holm–Bonferroni within the three pre-specified hypotheses	$p = .020$	Clears $\alpha = .05$
Clustering, $\rho = .06, \bar{m} = 10$	$p = .035$	Clears $\alpha = .05$
Composite outcome (174 events among women), weighted Firth	OR = 0.276, 95% CI [0.069, 1.099], $p = .068$	Does not clear $\alpha = .05$
Composite outcome, weighted ML	OR = 0.265, $p = .060$	Does not clear $\alpha = .05$
Leave-one-out across all 411 observations	p ranges .004 to .022; no deletion exceeds .05	Not driven by individual respondents
Bootstrap difference-in-differences	27.65 pp, 95% CI [2.10, 52.74]	Upper bound not substantively attainable; interval unstable

Note. The composite outcome combines the source and difficulty items and uses the full sample, so it carries 174 events among women rather than 6 in the decisive cell and addresses the design and the sparsity simultaneously. It does not reach conventional significance. The manuscript therefore reports H3 as supported on the primary outcome and not supported on the better-powered composite, and does not resolve the disagreement.

Table R8. Weight Diagnostics

Statistic	Value
Minimum weight	0.2855
Maximum weight	2.2796
Mean weight	0.9263
Coefficient of variation	0.486

Statistic	Value
Ratio of maximum to minimum	7.99
Kish design effect	1.236
Savings gap after trimming at 5th / 95th percentile	OR = 0.605, p = .0070
H1 interaction after trimming at 5th / 95th percentile	p = .326

Note. Statistics refer to the vulnerability model sample (n = 714). The released weights are already capped: the 1st and 99th percentiles equal the minimum and maximum, so trimming at those points alters no weight. Trimming at the 5th and 95th percentiles changes no conclusion in the manuscript.

Summary

The revision leaves one confirmatory finding, one strong descriptive contrast, one hypothesis supported only on the primary outcome, and one hypothesis reported as unsupported.

The pooled adjusted gender gap in formal savings survives every challenge raised in review: penalized estimation, parameter reduction, weight trimming, plausible unmeasured clustering, and both extreme assumptions about excluded cases. It is estimated on 239 events, so the sparse-data problems affecting the vulnerability models do not apply to it.

Women hold tertiary education 13.34 pp more often than men (95% CI [6.36, 20.33]), a contrast large enough to survive any plausible design correction, and this advantage appears in neither financial outcome.

The income gradient is real under every estimator, but the unpenalized full-model odds ratio of 0.084 is inflated by sparse-data bias. The manuscript reports the penalized reduced-model estimate of approximately 0.21.

H1 was tested and is not supported; H3 is supported on the primary outcome but not on the better-powered composite. With 63 vulnerability events, an interaction test needs roughly four times the sample of a main-effect test to detect an effect of the same magnitude, so the design was underpowered for interaction inference from the outset. The manuscript states this rather than presenting the H1 null as informative evidence of absence: the H1 interaction interval runs from 0.16 to 1.79 and does not exclude a substantial difference.