# A Report on "Impact of Guaranteed Income on Health, Finances, and Agency: Findings from the Stockton Randomized Controlled Trial" by West and Castro (2023)

Reviewer 2

February 11, 2026

v1

isitcredible.com

# Disclaimer

This report was generated by large language models, overseen by a human editor. It represents the honest opinion of The Catalogue of Errors Ltd, but its accuracy should be verified by a qualified expert. Comments can be made here. Any errors in the report will be corrected in future revisions.

> I am wiser than this person; for it is likely that neither of us knows anything fine and good, but he thinks he knows something when he does not know it, whereas I, just as I do not know, do not think I know, either. I seem, then, to be wiser than him in this small way, at least: that what I do not know, I do not think I know, either.
>
> Plato, *The Apology of Socrates*, 21d

> To err is human. All human knowledge is fallible and therefore uncertain. It follows that we must distinguish sharply between truth and certainty. That to err is human means not only that we must constantly struggle against error, but also that, even when we have taken the greatest care, we cannot be completely certain that we have not made a mistake.
>
> Karl Popper, 'Knowledge and the Shaping of Reality'

# Overview

**Citation:** West, S., and Castro, A. (2023). Impact of Guaranteed Income on Health, Finances, and Agency: Findings from the Stockton Randomized Controlled Trial. *Journal of Urban Health*, Vol. 100, No. 2, pp. 227–244.

**URL:** https://doi.org/10.1007/s11524-023-00723-0

**Abstract Summary:** This mixed-methods randomized controlled trial tested the effects of a $500 per month guaranteed income for two years on health and financial outcomes in Stockton, CA, finding that the treatment group reported lower income volatility, lower mental distress, better energy and physical functioning, and greater agency.

**Key Methodology:** Mixed-methods randomized controlled trial (RCT) with a treatment group ($n = 131$) receiving 500/*monthfor2yearsandacontrolgroup*(n=200$), using quantitative data (surveys, SMS) and qualitative data (interviews) analyzed via ANCOVA and thematic analysis.

**Research Question:** How does guaranteed income impact monthly income volatility, psychological distress, physical functioning, agency over one's future, and how were financial wellbeing and agency attenuated by the pandemic?

# Summary

## Is It Credible?

The West and Castro article presents findings from the Stockton Economic Empowerment Demonstration (SEED), a mixed-methods randomized controlled trial (RCT) designed to test the effects of a $500 monthly guaranteed income. The authors argue that the study provides "causal evidence of positive health and financial outcomes" (p. 227), specifically highlighting reductions in income volatility, improvements in psychological distress and physical functioning, and an increase in agency. Furthermore, they posit that the intervention did not negatively impact labor market participation, a finding they describe as "particularly important" given historical speculation about unconditional cash transfers (p. 242). While the study offers a detailed narrative of the recipients' experiences, the credibility of its central causal claims is severely compromised by critical flaws in design execution and statistical analysis.

The most significant threat to the study's validity is the catastrophic and differential rate of attrition. By the study's endline (Wave 7), retention rates had plummeted to approximately 35% for the control group and 55% for the treatment group (p. 229). While retention was higher during earlier waves where primary outcomes were measured—roughly 48% for control and 79% for treatment at the pre-pandemic endpoint—the differential loss of participants remains problematic (p. 232). When nearly half of the control group is lost even at earlier stages, the initial randomization is effectively nullified, and the groups can no longer be assumed to be comparable. The authors acknowledge that attrition "could have been differential by outcome variables" but admit that "there is no possibility to test this" (p. 243). This creates a high probability of selection bias; for instance, the remaining control participants might be those with more stable lives, or conversely, those in more desperate need of the compensation for completing surveys. Without a valid counterfactual, the observed differences in health and financial well-being cannot be confidently attributed to the guaranteed income.

The claim regarding labor supply is also methodologically weak. The conclusion that there was "no significant difference between the treatment and control group on labor" (p. 242) is likely a result of the study being underpowered, a problem exacerbated by the small sample size and attrition. An absence of evidence is not evidence of absence, particularly when the study was designed with a minimum detectable effect size that might miss subtle but meaningful labor market shifts (p. 228). Moreover, the primary statistical analysis operationalized employment status by grouping "stay-at-home parent or caregiver" with full-time and part-time employment (p. 229). While the authors do provide a descriptive breakdown of these categories elsewhere (p. 231), the main analytical model prevents the detection of potential withdrawals from the formal labor market. If a recipient quit a job to care for a relative, this model would register no change in "employment status," thereby failing to capture the very labor supply reduction policymakers often seek to measure.

Furthermore, the statistical evidence for positive outcomes appears fragile. The primary finding that guaranteed income reduced income volatility in year one relies on a one-tailed $t$-test with a $p$-value of 0.039 (p. 230). Had a standard two-tailed test been used, this result would likely not have reached statistical significance ($p \approx 0.078$). Similarly, the article reports numerous positive effects on the Kessler 10 and SF-36 health scales without applying corrections for multiple comparisons (pp. 232, 236–237). Given the large number of tests conducted across various domains and time points, the likelihood of finding significant results purely by chance is substantial. Additionally, the exclusion of 14 participants from the treatment group after randomization—though pre-specified in the analysis plan—deviates from a strict intention-to-treat principle, introducing further potential for bias (p. 229).

Finally, the qualitative findings, while rich, present a largely positive narrative of increased agency and risk capacity (p. 238). While the authors do note negative contextual factors such as pandemic fatigue (p. 230), the reporting on the intervention's effects is uniformly positive. The absence of negative cases or disconfirming evidence regarding the cash transfer itself raises concerns about confirmation bias, particularly given the authors' note about the "politically purposive cohort" (p. 229). While the study succeeds in generating hypotheses and illustrating the lived experience of financial stability, the quantitative flaws—specifically the broken randomization and low statistical power—render the causal claims largely inconclusive.

## The Bottom Line

The SEED trial provides valuable descriptive and qualitative insights into the potential benefits of guaranteed income, but its claims of causal impact on health, financial volatility, and employment are not credible. The study is fundamentally undermined by extreme and differential attrition that likely invalidated the randomized design, as well as by low statistical power and questionable analytical choices such as one-tailed testing and non-standard definitions of employment in primary models. Policymakers should view these findings as exploratory and illustrative rather than as definitive evidence of the program's efficacy.

# Potential Issues

**High and differential attrition:** The study's causal claims are significantly threatened by high and differential rates of attrition. By the study's endline, retention was approximately 35% in the control group and 55% in the treatment group (p. 229). Even at earlier measurement points used for primary outcomes, retention showed significant disparity (e.g., roughly 48% control vs. 79% treatment at Wave 4) (p. 232). A 20- to 30-percentage-point difference in retention between groups strongly suggests that the remaining participants are no longer comparable and that the initial randomization was compromised by selection bias. The authors acknowledge this as a limitation, stating that "attrition in the study could have been differential by outcome variables" but that "there is no possibility to test this" (p. 243). While they report that attrition was not predicted by baseline characteristics, this check cannot account for unobservable differences that may have driven the differential dropout. The severity of this attrition represents a fundamental challenge to the validity of the study's findings.

**Deviation from strict intention-to-treat principle:** The study claims to use an "intention to treat" (ITT) analysis but deviates from a strict application of this principle by excluding 14 participants from the treatment group after randomization. The ITT principle requires that all randomized participants be analyzed in the group to which they were assigned, regardless of post-randomization events. The text states that 131 individuals were allocated to treatment, but analyses were conducted on only 110, with the difference accounted for by 7 withdrawals and the 14 post-randomization exclusions (p. 230). The authors justify this by noting the exclusion of this "politically purposive cohort" was specified in the pre-analysis plan due to the potential for "differential outcomes" (p. 229). While pre-specification mitigates concerns of ad-hoc data manipulation, the act of removing participants from an analysis after they have been randomized introduces a risk of selection bias and is inconsistent with a strict ITT approach.

**Underpowered design and interpretation of null findings:** The study was likely underpowered to detect small or moderate effects, which complicates the interpretation of its null findings, particularly regarding employment. The power analysis was designed to detect a medium-to-large effect size (MDE was $f = 0.30$), meaning the study had a high probability of failing to detect a true but smaller effect (p. 228). This issue was exacerbated by attrition, which further reduced the sample size available for analysis. Consequently, the conclusion that "There was no significant difference between the treatment and control group on labor" (p. 242) should be interpreted with caution. It represents an absence of evidence for an effect, not necessarily evidence of its absence. The authors acknowledge that "Statistical power was limited by attrition" (p. 229), but the implications of this for interpreting key null results are substantial.

**Failure to adjust for multiple comparisons:** The study conducts a large number of statistical tests across multiple outcomes and time points without reporting an adjustment to the significance threshold. This practice increases the probability of finding a statistically significant result purely by chance (a Type I error). Several of the study's key positive findings in the pre-pandemic period have $p$-values that are close to the conventional 0.05 threshold, such as for psychological distress ($p = 0.027$), bodily pain ($p = 0.031$), and physical functioning ($p = 0.037$) (pp. 232, 234, 237). These results might not remain statistically significant if a standard correction for multiple comparisons, such as a Bonferroni correction, were applied. While providing exact $p$-values allows readers to make their own assessments, the lack of a formal correction means some of the reported positive effects may be spurious.

**Operationalization of the employment variable:** The study's conclusion of no negative impact on labor supply is based on a primary statistical model that uses a non-standard definition of employment. The authors coded their employment variable as a binary where "stay-at-home parent or caregiver" was grouped with full-time and part-time employment, in opposition to being unemployed (p. 229). While the authors do provide descriptive statistics elsewhere that break these categories down (p. 231), the main statistical test collapses them. This methodological choice makes it impossible for the primary model to detect a shift from the formal labor market to unpaid caregiving, a key hypothesized effect of guaranteed income programs. This limits the study's ability to speak directly to the common policy question of whether guaranteed income reduces market labor participation.

**Use of a one-tailed test for income volatility:** The statistical significance of the primary finding on income volatility in year one rests on the use of a one-tailed $t$-test ($t = 1.76$, $p = 0.039$) (p. 230). A standard two-tailed test would yield a $p$-value of approximately 0.078, which is not significant at the conventional $\alpha = 0.05$ level. While the authors state in their methods that "One-tailed t-tests... were conducted" (p. 229), they also note in the power analysis section that they assumed a "non-directional hypothesis" (p. 228). This inconsistency, combined with the marginal nature of the result, suggests the finding on income volatility is less robust than presented.

**Absence of negative cases in qualitative reporting:** The qualitative analysis presents a largely positive narrative of the treatment experience, focusing on themes of reduced anxiety and increased agency. While the authors do acknowledge negative contextual factors like "fatigue from shifting one's entire life online" during the pandemic (p. 230), the reporting on the intervention's effects is uniformly positive. Rigorous qualitative reporting typically involves a transparent accounting of data that may not fit the dominant pattern, and the absence of any such discussion regarding the cash transfer itself raises the possibility of selection bias in the presentation of the qualitative data.

**Transparency of statistical methods:** The article relies on Analysis of Covariance (ANCOVA) for its quantitative results but does not report whether the statistical assumptions underlying this method were tested. This is particularly relevant for the analysis of the employment outcome, which was a binary variable (p. 229). Using ANCOVA, a linear model, for a binary outcome is a non-standard approach, as its assumptions regarding normality of residuals and homoscedasticity are likely to be violated. Without confirmation that these assumptions were met or that the method is robust to their violation in this context, the validity of the resulting *p*-values and estimates is uncertain.

**Clerical and presentation issues:** The article contains several minor clerical errors and inconsistencies that affect clarity and replicability. First, baseline sample sizes vary substantially across different measures without explanation; for example, 294 participants completed the Kessler 10 scale while only 195 completed the SF-36 General Health scale at baseline (pp. 232–233). Second, a sentence regarding the use of multiple imputation is ambiguous and appears to contain a typo, stating that "it was employed due to the conditions of the pre-analysis plan" immediately after suggesting it was not the preferred method (p. 229). Context suggests the authors likely meant it was *not* employed, but the text is confusing. Finally, there is a transcription error in the text, which reports an *F*-statistic of 0.491 for physical functioning, while the corresponding table correctly shows an *F*-statistic of 4.49 (pp. 232, 237).

# Future Research

**Robust retention strategies:** Future trials must prioritize retention to preserve the integrity of randomization, as high attrition rates render control groups useless. This could involve significantly higher incentives for control group participation or, more reliably, the use of administrative data linkages (e.g., tax filings or unemployment insurance records) that allow for outcome tracking even when participants do not respond to survey instruments.

**Granular labor market analysis:** To accurately assess the impact on labor supply, future research should operationalize employment variables to distinguish clearly between formal market employment and unpaid caregiving in primary statistical models. Studies should be sufficiently powered to detect shifts between these specific categories, rather than collapsing them into a single "productive" status, to provide a clearer picture of how guaranteed income affects workforce participation versus care work.

**Strict statistical controls:** Future analyses should adhere to a rigorous pre-analysis plan that includes standard corrections for multiple comparisons (such as Bonferroni or Benjamini-Hochberg) to prevent spurious significant findings arising from the testing of numerous health and financial outcomes. Furthermore, primary outcomes should be tested using standard two-tailed hypotheses unless there is overwhelming, pre-specified justification for directional testing.

isitcredible.com