

Amended Pre-Analysis Plan: Child Development Status in Communities within a Randomized Evaluation of a Social Franchising Intervention

Amendment to “Improving Childcare Quality through Social Franchising”

Pre-registration: AEARCTR-0011747 · JDE Pre-Results Review: DEVEC-D-24-01906R2

Anne Fitzpatrick, Emily Beam, Maira Reimao

This date: October 2, 2026

1. Purpose of this amendment

Our original pre-analysis plan (PAP) for “Improving Childcare Quality through Social Franchising” (DEVEC-D-24-01906R2) listed child development as a secondary household outcome (HB6) but deferred specification of the measurement approach:

“We will file an amendment to the pre-analysis plan once the post-intervention household survey scheduled for June 2026, which will include caregiver-reported child development questions, is appropriately piloted.”

This document is that amendment. It specifies the child development measures that we plan to collect alongside household data during endlines of the primary study, reports the power calculations that determine what these measures can and cannot detect, and sets out how we will analyze and report the results.

We are filing this amendment before the completion of endline data collection of households for two reasons. First, it lets us be transparent about a limitation of the study design: child development measures were added to the household survey after baseline, and as such, we lack a baseline measure of child development. Separately, through our midline data collection efforts, we learned that only a minority of children in our household sample (less than 30%) are likely to be enrolled at a center supported by our partner organization. Second, pre-specifying that we expect these analyses to be underpowered protects against the risk that an imprecise null result on child development is later read as evidence that the social franchising program does not benefit children, when in fact the study was not designed or powered to answer that specific question.

2. Status of this analysis: exploratory, not confirmatory

Given these empirical limitations, and prior to any analysis of endline data related to child development, we note the following:

- **All child development results in this study are exploratory.** None of the four measures described below are likely powered to detect the community-level effects we would consider a policy-relevant benchmark for this intervention (Section 4). We are reporting them to build descriptive evidence on the state of child development in these communities and its correlates,

and to be transparent about what the data do and do not show — not to test a confirmatory hypothesis about the causal effect of the program on child development.

- **We will not apply a multiple-hypothesis-testing (MHT) correction** to the child development measures. The Westfall-Young step-down procedure we use for our confirmatory secondary hypotheses (HB1–HB5, HB7) is designed to control for the family-wise error rate across a set of hypothesis tests we intend to interpret individually as evidence for or against the program's effect. Because we do not intend to interpret the child development estimates this way — and because MHT adjustment would misleadingly suggest a level of inferential rigor these underpowered comparisons do not have — we will report unadjusted point estimates and confidence intervals for all four measures and their sub-domains, explicitly labeled as exploratory.
- **We will not draw a policy conclusion from a null result.** Given the power calculations in Section 4, a null or noisy estimate on any of these measures is uninformative about whether the program affects child development — it is the expected result even if the program has a substantively large true effect among children actually exposed to a center supported by our implementing partner. We will state this explicitly alongside any null finding rather than let it be read as evidence of no effect.
- **Where a measure is not individually valid, we will say so.** ECDI2030 in particular is designed and validated as a population-level indicator (it feeds SDG indicator 4.2.1) and has limited ability to detect individual- or program-level changes; we treat it accordingly (Section 3). The GSED short-form has similar limitations. Finally, SDQ has not been validated in our context, with potentially important differences in clinical cut-offs.

3. Measures and the age range each one serves

Households in our sample have a child born 2021–2024, so by the June 2026 endline these children will be roughly 2 to 5 years old, with some children outside of that range. No single publicly-available instrument in this space is validated across that entire span, so each tool covers a different, overlapping slice of our sample — which is the main reason the number of children assessed varies so much across measures in Section 4.

3.1 IDELA — International Development and Early Learning Assessment (direct assessment)

IDELA is administered directly to the child by a trained enumerator. It is validated for children aged approximately 3.5 to 6 years, and produces domain scores for motor development, emergent literacy, emergent numeracy, and socio-emotional development, in addition to an overall score. Because it requires an in-person, one-on-one assessment with the child, it is the most resource-intensive of the four measures. While this instrument has been widely used to measure child development in low-income settings, one major limitation of our study sample is that some of the children are not age-eligible for it.

3.2 ECDI2030 — Early Childhood Development Index 2030 (caregiver-reported)

ECDI2030 is a 20-item caregiver-report module developed by UNICEF to track SDG indicator 4.2.1 (the share of children developmentally on track in health, learning, and psychosocial well-being). It is validated for children aged 24–59 months (2 to just under 5 years), which is why, of the four instruments, it covers the largest share of our sample. It is important to note that ECDI2030 was explicitly designed and validated as a population-level indicator for national and sub-national monitoring, and its developers state that it is not intended for individual-level measurement. We report this measure because it is the most widely used cross-country ECD benchmark, which makes our sample's baseline levels informative for comparison with other populations — not because we expect it to be a sensitive tool for detecting program impact.

3.3 SDQ — Strengths and Difficulties Questionnaire (caregiver-reported)

The SDQ is a caregiver-reported screening instrument covering emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behavior, combined into a total difficulties score. It is validated across ages 2–17, with a dedicated preschool-adapted version for ages 2–4; because our sample spans roughly 2–5 years, we can use it across nearly the full study cohort. Like ECDI2030, the SDQ is a screening tool designed to flag population-level patterns and risk, not a diagnostic or high-sensitivity instrument for detecting incremental program effects on socio-emotional development.

3.4 GSED — Global Scales for Early Development, Short Form (caregiver-reported)

The GSED Short Form is a caregiver-reported instrument developed by the WHO to measure development from birth to 36 months across motor, language, cognitive, socio-emotional, and adaptive domains, summarized in a single D-score. It is explicitly designed as a population- and programmatic-level measure. Due to time and budget constraints, we are unable to field the long form which may be more sensitive to detecting programmatic impact. Because the GSED is only validated through age 3, it has the smallest expected sample size of the four measures, as shown below. As a result, it is not powered to detect a program effect for this cohort.

4. Power calculations

Table 1 summarizes power for each measure, using data-informed parameters. Two figures matter most: the number of children assessed reflects the age overlap between our sample and the tool's validated group described in Section 3; and the estimated share of the treatment group actually enrolled at a center supported by our implementing partner (25%) reflects the take-up information we learned through midline data.

Measure	Children assessed	Share of treatment group at a center affiliated with our	Smallest community-level effect detectable	Effect required among children who attended

		implementing partner		
IDELA (direct assessment)	~1,000	25%	0.24 SD	~0.95 SD
ECDI2030 (caregiver-reported)	~2,100	25%	0.16 SD	~0.63 SD
SDQ (caregiver-reported)	~2,900	25%	0.26 SD	~1.06 SD
GSED (caregiver-reported)	~380	25%	—	not powered at this cohort's age

Table 1. Power calculations for child development measures at endline. “Smallest community-level effect detectable” is the minimum detectable effect (MDE), in standard deviations, on the intention-to-treat estimate at the community level. “Effect required among children who attended” rescales that MDE by the estimated 25% share of the treatment group enrolled at a center supported by our partner organization, to show the effect size the program would need to produce among children actually exposed to it in order for our community-level ITT estimate to be statistically distinguishable from zero.

The last column is the key number for interpreting this exercise. Because only about a quarter of children in treatment communities are estimated to attend a center affiliated with our implementing partner, our community-level intention-to-treat design can only detect an effect if the true effect among exposed children is roughly 4 times the community-level MDE (or 2 times if the effect on children in non-affiliated centers is equally large as in affiliated centers). The implied required effects among those that attend centers affiliated with our partner organization — approximately 0.95 SD for IDELA, 0.63 SD for ECDI2030, and 1.06 SD for SDQ — are far larger than effect sizes typically documented for early childhood interventions of this kind and duration in the literature (0.2-0.4 SD), meaning we are unlikely to detect even a genuinely meaningful effect of the program on these outcomes.

5. Empirical specification

Where we report treatment-control comparisons on these measures, we will use the same intention-to-treat specification as our confirmatory analyses:

$$Y_{ic} = \alpha + \beta_1 \text{Treatment}_c + \varphi_c + \varepsilon_{ic}$$

where Y_{ic} is the child development outcome for child i in community c , Treatment_c is the randomly assigned community-level treatment indicator, and φ_c is a set of strata (county) fixed effects. We do not have a baseline measure of child development to include as a control (these instruments were not part of the baseline survey), which is one further reason to treat these estimates as exploratory rather than

as precision-matched to our confirmatory ANCOVA specifications elsewhere in the study. Standard errors will be clustered at the community level, although as we explore patterns and correlates of development scores we may also relax the level of clustering. To examine the robustness of our findings, and explore how the tools work more broadly, we will also conduct sensitivity analysis. This sensitivity analysis may include but is not limited to confirmatory factor analysis, exploratory factor analysis, or correlations with key variables of interest.

We will report:

- The overall score and (where relevant) each domain sub-score for each instrument, by treatment arm, with unadjusted point estimates and 95% confidence intervals. For ECDI we will report the measure of children “on track”.
- The same comparison restricted to children whose household reports current attendance at a center that receives support from our partner organization, alongside the full intention-to-treat sample, since the two provide different (and differently limited) information: the ITT estimate is well-identified but diluted by low take-up, while the attender comparison has a more direct interpretation but is not protected by randomization.
- Descriptive statistics on levels of child development in our sample overall (pooling treatment and control), to contribute comparable evidence on child development in this population regardless of what the treatment comparison shows.

6. Planned Use of Results

To keep this amendment useful as a check on our own reporting once we have the data, we commit to the following when writing up these results to share. Note that writing up does not imply that the child development results will appear in working or published versions of papers, which are typically more specific and centered on well-powered analyses:

- State the power calculations in Table 1 alongside any child development result we report, in the same section or figure, not relegated to an appendix.
- Describe every estimate on these four measures as exploratory, and avoid language (e.g., “we find no effect,” “the program did not improve...”) that implies a well-powered null result.
- Note explicitly, wherever ECDI2030 or GSED results are presented, that these instruments are validated for population-level monitoring rather than individual-level or program-evaluation use.
- Use the collected information to describe the children in our setting, as well as differences between those who attended/ are attending daycare and those who are not, along with the type of daycare. Not use these results, on their own, to make claims about our partner organization’s program’s effect on child development to funders, media, or policy audiences without the same caveats included here.

8. Administrative information

This amendment covers only the child development measures described above; all other elements of the original pre-analysis plan (randomization, firm-level outcomes, household labor-supply outcomes, MHT procedure for confirmatory secondary hypotheses, ethical approvals, and funding) remain as specified in DEVEC-D-24-01906R2 and are not restated here.