Appendix A Detailed risk of bias assessment

The appendix discusses the critical appraisal domains relating to bias in the benchmark treatment effect, in turn, as well as any factors that are relevant in assessing bias in correspondence between benchmark and NRS.

Randomisation

Benchmark study data are typically from cluster-randomised field trials, five of which evaluated conditional cash transfers in Latin America. These programmes were typically randomised at public events with members of the government, media and field research teams present. Two benchmarks were assessed as being of 'low risk of bias' in the randomisation process, given the random assignment of clusters, and the similarity of cluster sizes and/or balance of household characteristics at pre-test; these included Barrera-Osorio et al. (2014), and Galiani and McEwan (2013) for the replications by Galiani and McEwan (2013) and Galiani et al. (2017).

In Chaplin et al. (2017), the difference in means and statistical tests did not suggest more frequent differences than would be expected by chance alone (9 out of 191 covariates at 5 percent significance). However, there were large differences in baseline variables relating to the outcomes (access to and spending on electricity, use of technologies requiring electricity (e.g., water pump, satellite television), suggesting 'some concerns' which were likely reflected in the small sample size for treatment clusters (27 communities) compared to controls (151 communities). In McKenzie et al. (2010), which compared fewer baseline characteristics, there was a difference in the baseline mean outcome, although that difference was not statistically significant. Nevertheless, it is notable that even small differences may appear significant in relatively large samples, or large differences appear non-significant in small samples. For example, the difference in baseline outcome amounted to 6 percent of the control mean in McKenzie et al. (2010).

In the case of the PROGRESA CCT replications (Buddelmeyer and Skoufias, 2004; Diaz and Handa, 2006), Behrman and Todd (1999) presented balance tables for several hundred baseline covariates at household level,

which suggested statistical differences between treatment and control may have arisen by chance.[1] In contrast, they did not find statistically significant differences in covariates measured at the locality level (where the total sample of treatment and control communities was 505), which suggests that the cluster randomisation led to balanced groups on average. However, no information was available on the randomisation process for PROGRESA – how it was implemented, e.g., with respect to a random number table, and by whom, whether done centrally by researchers – to assess the risk of subversion of randomisation, hence 'some concerns' were noted.

For the RPS CCT programme in Nicaragua, eligible clusters were randomised at a public event, but there appeared differences in group characteristics (the extreme poverty level was higher in controls), as reported in Maluccio and Flores (2005, for the replication by Handa and Maluccio, 2010). This may be due to restricted randomisation over a relatively small cluster sample size (42 clusters in total), which is common in RCT practice.

Recruitment of participants

This section assesses risk of selection bias into the study due to identification and recruitment of individual participants in relation to timing of randomisation. It appears the case that individuals were nearly always chosen after randomisation was done or communicated (Buddelmeyer and Skoufias, 2004; Diaz and Handa, 2006; Handa and Maluccio, 2010; McKenzie et al., 2010; Chaplin et al., 2017). In the case of PROGRESA in Mexico, "[t]he selection of households as PROGRESA beneficiaries was accomplished by first identifying the communities to be covered by the program (geographic targeting) and then selecting the beneficiary households within the chosen communities" (Buddelmeyer and Skoufias, 2004, p.6). Individual household selection was done in a two-part process where eligible households were selected if they fulfilled certain poverty criteria based on a household survey, and then the list presented to the community assembly for discussion, which Skoufias et al. (2001) note made very little difference to the final household choice. As discussed

[1] Randomisation leads to balanced samples in expectation over repeated trials, not in any specific draw (Deaton and Cartwright, 2018).

above, although there do not appear to be differences in treatment and control at cluster level, there are differences at the household level, which may go beyond that expected by chance. However, as the authors noted, the large sample size for the study (there were 24,000 households and 41,000 children aged under 17) suggested the study was powered to detect very small differences with statistical precision. A more appropriate approach would have been to analyse treatment group and control group differences using distance metrics (Bruhn and McKenzie, 2009), but this was not presented in Behrman and Todd (1999). The benchmark was therefore evaluated as having 'some concerns'.

In the case of RPS in Nicaragua, which used geographic targeting to identify treatment clusters, within which participation was voluntary but participation rates exceeded 90 percent due to the size of the transfer (Handa and Maluccio, 2010), households were chosen for data collection after cluster randomisation, using a random sample based on a household census conducted for the evaluation. Non-response was 10 percent in the first round, and similar in treatment and control groups (Maluccio and Flores, 2005). However, there were differences in baseline household characteristics for a few variables, warranting 'some concerns'.

Similar issues concerning imbalance occurred when assessing McKenzie et al. (2010) and Chaplin et al. (2017). McKenzie et al. (2010) noted difficulties in recruiting individuals into the study, the reasons for which were given for treated units (e.g., being located outside of the survey area) and weighted accordingly, but were less clear for controls. They also attempted to avoid bias in the recruitment strategy which was done by telephoning unsuccessful lottery participants from the same villages as successful participants by including "in the sample households from the Outer Islands of Vava'u and 'Eua" (p.919) that were less likely to have telephones. However, it is not clear how successful the strategy was at obtaining a representative sample of controls, while the reasons for missingness appeared different across treatment and control. In the case of Chaplin (2017), where it appears that sampling of households was done after cluster randomisation, the authors did make efforts to track whether there was migration from controls to treated communities before the household baseline was conducted. However, owing to the differences in

baseline characteristics noted above, the analysis suggested 'some concerns'.

In the case of Barrera-Osorio and Filmer (2016), which presented the benchmark RCT used in Barrera-Osorio et al. (2014), recruitment of students, who completed application forms for means-tested scholarships in treatment and control groups, was done before school-level stakeholders were aware of the school's randomised assignment. In the benchmark study in Galiani and McEwan (2013) and Galiani et al. (2017), all households living in treatment localities were eligible to receive benefits of the programme. In addition, the outcomes data were taken from an unrelated census, conducted 8 months after programme implementation had begun. So, while there could be threats to validity relating to deviation from intended intervention (e.g., due to migration), selection into the study is unlikely to be correlated with treatment status. There was therefore 'low risk of bias' in recruitment of participants for these benchmarks.

Departures from intended interventions

Departures from the intended interventions across the cluster-randomised studies is relevant for within-study comparisons using dependent design, when it affects the control group in the benchmark trial. Issues relating to intervention delays that would typically be of concern if the purpose of the analysis were to estimate treatment effectiveness, are not relevant. For example, referring to the experiment used in Handa and Maluccio (2010), Maluccio and Flores (2004, p.14) stated "it was not possible to design and implement all the components according to the original timelines. In particular, the health-care component was not initiated until June 2001… There were also delays in the payment of transfers to households due to a governmental audit that effectively froze RPS funds." Similarly, Buddelmeyer and Skoufias (2004, p.7) found "in the treatment localities 27% of the total eligible population had not received any benefits by March 2000." However, within-study comparisons based on dependent design estimate the same level of impact, regardless of whether that reflects a poorly implemented intervention. This is particularly relevant for Diaz and Handa (2006), Handa and Maluccio (2010), Galiani and McEwan (2013), Galiani et al. (2017), Chaplin et al. (2017), and two of the comparisons in

Buddelmeyer and Skoufias (2004), where the distance estimate is calculated solely from the comparison of means between randomised control and NRS comparison group. Hence, in these cases, the risk-of-bias rating was amended (upgraded) to capture the expectation that problems in implementation of the intervention would not cause bias between randomised and NRS estimators.

Nevertheless, several cluster-RCTs were considered to have biases in this domain due to potential contamination, spillover effects or performance bias. For PROGRESA (Buddelmeyer and Skoufias, 2004; Diaz and Handa, 2006), Behrman and Todd (1999) explained that individuals may migrate between control and treatment clusters in order to receive the benefits of the intervention and that the incidence of such issues should be tracked. This source of bias may have existed because participating households were not fixed at the start of the study, and may have occurred even where treated localities had not received benefits (perhaps being a factor that might have encouraged migration out of underperforming treatment clusters). One of the treatment effect estimates made in Buddelmeyer and Skoufias (2004) suggests potential issues with comparability of the control. In addition, controls within clusters may be affected, where they change their behaviour in response to 'peer effects' from observing treatment participants (spillovers), or possibly with the expectation of becoming eligible for the benefits (John Henry effects) (as also assessed for RPS by Maluccio and Flores, 2005). Buddelmeyer and Skoufias (2003) tested for this by comparing groups where spillovers were unlikely due to geographical separation – i.e., ineligible households in treatment communities were compared with eligible but untreated control households (group C versus group B) and ineligible control households (group C versus group D). They did not find significant differences with the estimates that could have been compromised by spillovers.

However, in general the studies did not indicate the extent that departures from intended interventions may have occurred. An exception was Maluccio and Flores (2005), which examined the presence of substitution effects in control groups (differential contamination by other interventions) for the RPS CCT programme, finding that there may have been reduced access in control communities for school supplies, but not other interventions. They

also reported that a small number of controls who received treatment were dropped from analysis to avoid bias in the estimate. Galiani et al. (2017) highlighted contamination of controls as an unlikely issue in the benchmark experiment, since the value of the cash transfer was small relative to average income (and there were also severe delays in distribution of the cash transfers beyond the follow-up data collection period). Therefore, the transfers were unlikely to provide incentives or liquidity for poor people to move to treated localities to obtain them, in the benchmark study, meriting 'low risk of bias'. Similarly, the scholarship benchmark experiment used by Barrera-Osorio et al. (2014) was assigned 'low risk of bias' on deviation from intended interventions. The analysis used ITT and there were no opportunities for controls to cross over to treatment, since "[i]f a student had dropped out and could not collect the scholarship, the funds could not be reassigned to another student but would be returned to a central fund for use in a subsequent distribution round" (p.473).

Finally, in the case of the natural experiment of the effects of migration on income (McKenzie et al., 2010), there was considerable non-compliance due to no-shows in the treatment group (i.e., a large proportion of participants randomised into the treatment group did not emigrate by the time of the follow-up survey). Two types of experimental estimates were provided by the authors to accommodate deviations from intended interventions. These were ITT, which estimates the effect of assignment, and the effect of starting and adhering to treatment, correcting for non-random deviations from the intended intervention, measured in instrumental variables.[2] Because the instrument was randomisation, and the correlation with treatment status migration was high (F-statistic=60), this domain was assessed as being of 'low risk of bias'.[3]

[2] The CACE estimate (where the randomised outcome of the random ballot is an instrument for the variable of interest – the migration decision) was the one that was incorporated in subsequent analysis and hence is presented in this analysis.

[3] This is therefore an override to the decision tree used in the Cochrane tool (Higgins et al., 2016) which indicates that even appropriate analysis using IV to correct for non-compliance cannot score more highly than having 'some concerns'. McKenzie et al. (2010: 923) noted: "One could conceive of stories such as that winning the ballot and not being able to migrate causes frustration and leads individuals to work less, or conversely, that winning the ballot acts as a spur to work harder in order to afford the costs of trying to find a job in New Zealand. However, we did not encounter any evidence of such changes in behaviour in our field work, lending support to this identification assumption."

Missing outcomes data

Benchmarks were assessed as having 'low risk of bias' where attrition was at a similar level across treatment and control and where missingness of observations was not differentially correlated with covariates. Studies were of 'some concern' where information was not available. Chaplin et al. (2017) reported data collection in all target communities and 20 percent overall household attrition between baseline and follow-up, evenly split between treatment (19.9%) and control (16.9%), suggesting 'low risk'. The benchmark underlying Galiani and McEwan (2013) and Galiani et al. (2017) was assessed as being of 'low risk of bias' as the analysis was based on census data. McKenzie et al. (2010) performed purposeful sampling of the control group during the follow-up survey because of concerns that the method of follow-up (using a telephone directory) may have led to bias in selection into the study (for those that did not have telephones). They elected to include a sample of participants from the outer islands of Tonga deliberately, in order to correct for the possible bias introduced. However, we remained unclear as to the effect that this purposeful sampling may have had on the composition of the control group and their outcome data during the follow-up. Robustness checks and further details are not available, and therefore the study was rated as having 'some concerns'.

No information was available about differential attrition from the benchmark study for PROGRESA in available published reports. Rubalcava et al. (2009) noted that "one-third of households left the sample during the study period" and "no attempt was made to follow movers" (p.515). No information was reported on differential attrition across groups. PROGRESA was coded as having 'high risk of bias' due to high overall attrition and lack of information about differential attrition. In the case of RPS in Nicaragua, there was 5 percent attrition between baseline and follow-up, which was approximately equal in both groups (Maluccio and Flores, 2004). Analysis suggested that attrition may have been correlated with treatment status, but the differences were small, warranting 'some concerns'. In Barrera-Osorio et al. (2014), overall attrition was 23 percent, comprising 20 percent of treated students and 28 percent of controls.

Barrera-Osorio and Filmer (2016) presented significance tests of differences in characteristics between attritors and non-attritors in treatment and control, which they argue are consistent with 'pure chance'. However, due to the differential attrition between groups, the category was classified as having 'some concerns'. Galiani and McEwan (2013) and Galiani et al. (2017) analysed census data for two outcomes – school enrolment, which was available for all households, and child labour, available for 82 percent of households. Although attrition was large for child labour, the data were collected from the census which would not have been linked to the CCT programme by participants or enumerators. Therefore, 'low of risk of bias' was given for attrition.

Outcomes measurement

While assessment of confounding and differential selection bias (into and out of study) are important in determining bias in the benchmark estimate itself, as well as bias in relation to the NRS estimate, it is not immediately clear whether risk of bias in the method of collecting outcomes data is an important source of bias in correspondence between them. For example, if outcomes data were collected using identical methods (whether observed or reported) in an open benchmark control and NRS comparison study, these potential biases might be expected to 'cancel out' in the calculation of the distance estimate. Whether this is the case would depend on the motivations of participant or outcome assessors in unblinded studies, which may vary between trials where data are clearly linked to an intervention (due to informed consent), and studies where data are not. Hence, in the case of McKenzie et al. (2010), an individually randomised lottery, where benchmark and NRS outcomes data were collected using the same tools by the same enumerators, it is possible that migrants were incentivised to over-report income (e.g., due to the 'false success' narratives that are known to exist), which could have upwardly biased the benchmark estimate.[4]

Across all but three benchmarks, outcome measurements were considered to have 'some concerns'. This was largely due to the issue of lack of blinding

[4] McKenzie et al. (2010) also collected pre-test income using one-year recall, which might be expected to be less reliable for international migrants than non-migrants. However, the treatment effect estimates calculated for benchmark and NRS in this study do not use the pre-test outcome.

of assessors in trials, where participants and outcome assessors may have had incentives affecting how they report outcomes (e.g., relating to social desirability). It is also unknown (there was insufficient evidence) to confidently state whether outcomes were likely to be influenced by knowledge of intervention received, since outcomes data were usually collected from household surveys through self-report, rather than more rigorous methods such as formal tests.[5] However, these are also cluster-RCTs where informed consent for the outcomes survey does need not refer to a specific intervention. Unfortunately, no information was reported about the process of consent in any of the studies, so it was unclear whether consent for the benchmark studies informed participants that the purpose of data collection was to evaluate the intervention of interest.

Outcomes were observed for Barrera-Osorio et al. (2014), where enumerators determined grade completion and administered mathematics tests. The data were collected using the same survey instrument at the same time in benchmark and NRS comparison. Therefore, even though outcome assessors were not blinded, any effect that enumerator incentives may have had was likely to be equivalent in benchmark control and NRS comparison. In the benchmark study in Galiani and McEwan (2013) and Galiani et al. (2017), there was effectively blinding of outcome assessment, since outcomes data used the national census which had been collected shortly after implementation of the cash transfer programme. Participants and outcome assessors would therefore not have been able to associate the data collection with the programme or household treatment status. Both studies were therefore rated as having 'low risk of bias' in outcomes measurement.

In the case of Chaplin et al. (2017), outcomes data were collected through self-report in benchmark and NRS using the same survey instruments at the same time of year by the authors. Furthermore, the NRS comparison group was selected from the comparison group of a concurrent non-randomised evaluation of electrification being done by the authors for the same project at the same time as the RCT. Therefore, since both benchmark control and NRS comparison data were from communities taking part in

[5] In some instances, outcomes were collected at community level (e.g., household electricity grid connections in Chaplin et al., 2017) but these were not used in estimation of within-study comparisons.

evaluations, any effect that knowledge of treatment status by participants or enumerators may have had on responses may be expected to 'cancel out'. Hence, this study was also assigned 'low risk of bias'.

Selection of the reported result

The purpose of the within-study comparisons was usually to test for differences across multiple outcomes and specifications of benchmark and NRS comparison. Selective reporting was assessed as being of 'low risk bias' across all benchmark studies, due to the large number of effects usually reported for different outcomes and samples. For example, all studies reported results of RCTs across multiple outcome domains, which were subsequently used in comparison with non-randomised replications. Some studies also reported findings for particular sub-groups, such as boys and girls in Buddelmeyer and Skoufias (2004), which was judged as common practice in the evaluation of school programmes, and non-selectively reported since all findings were reported by sex for all specifications. However, there is potentially a problem with multiple hypotheses, suggesting that statistical significance thresholds should be more conservative when comparing differences between RCT and NRS.

Bias in the within-study comparison estimate

The final source of bias relates to confounding of the correspondence between benchmark and NRS due to differences in measurement and differences in target population (sampling bias). This section discusses these sources of bias, as well as threats to validity due to design and conduct of the NRS (Table A2 provides a detailed summary, which is presented as an overall rating in Table 4 in the main text). Regarding measurement, McKenzie et al. (2010) reported NRS findings for two surveys, one done by the authors identical to that done for the randomised benchmark, comprising a relatively small sample of 60 non-applicant households living in the same village as lottery applicants. The second survey drew on nationally representative survey containing 3,000 households in the relevant target population. The findings reported below are therefore taken

from the author survey to ensure identical survey instruments.[6] However, Diaz and Handa (2006) reported differences in sampling frame and season of data collection between benchmark and NRS for all outcomes, as well as specific differences in detail of questions and recall period for expenditure data, stating that "differences in expenditure outcomes may be entirely due to questionnaire design rather than evaluation technique" (p.327). These differences were noted and explored in the meta-analysis below.

Handa and Maluccio (2010) used Living Standards Measurement Survey (LSMS) data, which were collected at the start of the rainy season in April to July 2001, to generate the NRS comparison for RCT data collected in October 2001, at the end of the rainy season. Given likely seasonal variation in the outcomes measured (food expenditure, preventive health care behaviour, child health), it was useful that both surveys were done in the same season, although it is possible the RCT data were collected at the time when infectious diseases (e.g., diarrhoea and ARIs) were more prevalent, which would tend to cause the mean in the RCT control to exceed the NRS comparison.[7] Furthermore, for one of the 12 outcomes collected, use of preventive health check-up for children aged 0-36 months, there were slight differences in the reference period being recalled and specific type of check-up. However, the authors made refinements to the LSMS sample used in the NRS to foster comparability with the RCT sample. Firstly, they excluded localities where the programme was operating from the NRS sample, to avoid possible contamination from treated households (since the programme began the previous year). From this sample, they calculated three NRS treatment estimates: the full sample estimate; a sub-sample estimate including only those localities that would have been eligible for treatment using the marginality index that determined eligibility for treatment; and a second sub-sample limiting eligible localities to the same geographical zone as treated households. Differences in findings for these sub-samples were explored in the meta-analysis.

[6] The findings from the nationally representative survey data were reported for OLS and PSM specifications in McKenzie et al. (2010). These yielded distance metrics larger than for the survey data collected by the authors (reported in Table 5 in the main text). The mean distance for OLS specifications is 0.104 (95%CI=0.013, 0.194); for PSM it is 0.096 (95%CI=-0.021, 0.214).

[7] The rainy season in Nicaragua is from May to October, with the wettest months being September and October.

A second question is whether there are differences in the NRS treatment estimand (e.g., ATET or LATE) with the benchmark (estimating ATE) that would lead to differences in treatment quantity over and above any bias or sampling error. In nearly all cases, the authors ensured NRS target populations were as similar as possible to RCTs, or the bias estimates were able to incorporate the differences. For example, in all RDD within-study comparisons, the RCT results were estimated at the same bandwidth around the treatment threshold. In the matched NRS comparisons, the bias was calculated with reference to the RCT control group only (Diaz and Handa, 2006; Handa and Maluccio, 2010), hence adjustments based on non-compliance were not necessary.[8] However, in McKenzie (2010) the bias estimates relied on the treatment mean. There was substantial non-compliance with the migrant lottery (mainly due to delays in migration). Therefore, the estimate from instrumental variables was taken for the benchmark estimate, rather than the ITT estimate.

The studies reported sensitivity analysis using different NRS estimators. For example, the studies of matching assessed the differences with nearest-neighbour, caliper, kernel and local-linear algorithms (e.g., Diaz and Handa, 2006), the inclusion of baseline outcome (McKenzie et al., 2010; Chaplin et al., 2017), use of 'rich covariates' and geographically proximate observations (Handa and Maluccio, 2010; Chaplin et al., 2017). The study of instrumental variables examined sensitivity to alternative instruments (McKenzie et al., 2010). Studies of regression discontinuity compared different bandwidth estimates (Buddelmeyer and Skoufias, 2004; Barrera-Osorio et al., 2014). Whether these differences are correlated with bias is an empirical question that was explored in the meta-analysis below.

Table A2 Bias in NRS-RCT comparisons

Within study comparison (outcome)	*Risk rating*	*Cause of confounding in NRS-RCT bias estimate*
Buddelmeyer and Skoufias (2004)	Low risk	NRS and RCT use same survey and bandwidth around eligibility threshold

[8] For example, in Handa and Maluccio (2010), the benchmark effect estimand was the intention-to-treat. The intervention participation rate was 90 percent, however, suggesting that NRS estimates of treatment effect, using ATET, would need to be rescaled by dividing ATET by (1-0.9) = 0.1, in order to equivalise the denominator and ensure comparability.

Within study comparison (outcome)	*Risk rating*	*Cause of confounding in NRS-RCT bias estimate*
Diaz and Handa (2006) (education)	Some concerns	Difference in season and sampling frame between NRS and benchmark surveys
Diaz and Handa (2006) (expenditure and child labour)	High risk	Measurement of expenditure and child labour differ between NRS and RCT surveys
Handa and Maluccio (2010) (expenditure, child feeding practices, immunisation)	Low risk	NRS and benchmark use same survey questions and target populations, during same season
Handa and Maluccio (2010) (child illness in previous month)	Some concerns	NRS and benchmark surveys conducted at opposite ends of the rainy season
Handa and Maluccio (2010) (preventive health)	High risk	NRS and benchmark questions are different for preventive health check-ups.
McKenzie et al. (2010)	Low risk	NRS and RCT use same survey and bandwidth around eligibility threshold
Barrera-Osorio et al. (2014)	Low risk	NRS and RCT use same survey and bandwidth around eligibility threshold
Galiani and McEwan (2013)	Some concerns	NRS and RCT use same survey and bandwidth around eligibility threshold; some concerns about the method used to identify the NRS comparison group.
Galiani et al. (2017)	Some concerns	NRS and RCT use same survey and bandwidth around eligibility threshold; some concerns about the comparability of the NRS population.
Chaplin et al. (2017)	Low risk	NRS and RCT use same survey conducted during same time of year

However, it was also important to evaluate conduct in the NRS. For example, matching should be done using covariates that are likely to be correlated with treatment and outcome, preferably using higher-order polynomials and interactions with the treatment variable (Handa and Maluccio, 2010), but importantly the covariates must not be affected by the treatment. Handa and Maluccio (2010) used locality variables measured five years prior to treatment (which could not have been affected by treatment), and household variables measured one year after treatment commenced, some of which were fixed (e.g., age and parental education) but others may have been affected (e.g., working patterns). By contrasting the findings with NRS matches made using a survey from the previous year,

they interpreted the findings as presenting evidence of bias in some of the household level matching variables.

Matches should also not be geographically proximate so as to lead to possible bias in the treatment effect due to contamination or spillovers. In the case of McKenzie et al. (2010) where bias is calculated using the treatment mean, and NRS comparisons are taken from the same communities where treated observations used to live, there is little risk of contamination owing to the nature of the intervention (international migration). It is also worth noting that 'geographical proximity' is fairly loosely defined, as coming from the same central part of the country in Handa and Maluccio (2010). In Tanzania, Chaplin et al. (2017, p.G.7) stated they "initially concluded that 30 km would be a reasonable radius based on the following criteria: that 30 kilometers is an upper bound for the distance most adults would reasonably walk in a day and used it as one measure of how much two communities would be subject to similar influences."[9]

In the scholarship RDD (Barrera-Osorio et al., 2014), assignment was based on one of two indexes – a merit threshold based on a student test score, and a poverty threshold based on students' reported household and family socioeconomic factors. The tests were scored centrally by an independent firm employed specifically to reduce manipulation of eligibility. The authors noted that the official list of scholarship recipients provided by the government was identical to the list provided by the firm. Furthermore, "spot checks at a number of schools yielded no cases of the manipulation of the selection process" (Barrera-Osorio and Filmer, 2014, p. 486).

In Galiani and McEwan (2013), precise HAZ-score programme eligibility data were only available for the benchmark localities. However, a report on the height census conducted four years previously gave the proportion of children with severe and moderate stunting (HAZ-scores below -3 and -2, respectively) for all localities nationally. Eligibility for the RDD comparison localities was then predicted by a regression of the mean HAZ-score from the censored data on the stunting proportions from the previous height

[9] However, Chaplin et al. (2017) also discussed the potential limitations of local matching on reducing the availability, and therefore quality, of potential matches, and settled on a radius of 40 kilometres.

census. The authors found a high correlation between predicted HAZ-score and actual HAZ-score for treatment communities (r=0.96), although it should be borne in mind that eligibility for the RDD comparison is therefore estimated and 'fuzzy'. In Galiani et al. (2017), there were also concerns in the design of the NRS replication due to the "persistent imbalance in one covariate (Lenca) that is plausibly correlated with unobserved determinants of child outcomes" (p.207) between treated and control municipalities. As the authors argued, it was therefore not possible to assume continuity in potential outcomes at municipal borders, suggesting some threats to the internal validity of the replication. Therefore, despite the benchmark in Galiani and McEwan (2013) and Galiani et al. (2017) being assessed as of 'low risk of bias', concerns about conduct of the NRS suggested 'some concerns' about confounding of the difference estimator.

Additional references

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: randomization in practice in development field experiments. American Economic Journal: Applied Economics, 1 (4), 200-232.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine 210: 2-21.