American Economic Journal: Applied Economics 2014, 6(4): 1–34
http://dx.doi.org/10.1257/app.6.4.1

Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia[†]

By Benjamin A. Olken, Junko Onishi, and Susan Wong*

We report an experiment in 3,000 villages that tested whether incentives improve aid efficacy. Villages received block grants for maternal and child health and education that incorporated relative performance incentives. Subdistricts were randomized into incentives, an otherwise identical program without incentives, or control. Incentives initially improved preventative health indicators, particularly in underdeveloped areas, and spending efficiency increased. While school enrollments improved overall, incentives had no differential impact on education, and incentive health effects diminished over time. Reductions in neonatal mortality in nonincentivized areas did not persist with incentives. We find no systematic scoring manipulation nor funding reallocation toward richer areas. (*JEL* F35, I18, I28, J13, J16, O15)

A recent movement throughout the world has sought to improve the links between development aid and performance. For example, the United Nations has sought to focus developing country governments on improving human development and poverty alleviation by defining and measuring progress against the Millennium Development Goals. Even more directly, foreign assistance given out by the

* Olken: Department of Economics, Massachusetts Institute of Technology, E17-212, 77 Massachusetts Avenue, Cambridge, MA 02139 (e-mail: bolken@mit.edu); Onishi: The World Bank, 1818 H Street, NW, Washington, DC 20433 (e-mail: jonishi@worldbank.org); Wong: The World Bank, 1818 H Street, NW, Washington, DC 20433 (e-mail: swong1@worldbank.org). We thank the members of the PNPM Generasi Team including: Sadwanto Purnomo, Gerda Gulo, Juliana Wilson, Scott Guggenheim, John Victor Bottini, and Sentot Surya Satria. Special thanks go to Yulia Herawati, Gregorius Pattinasarany, Gregorius Endarso, Joey Neggers, Lina Marliani, and Arianna Ornaghi for their outstanding support in survey preparation, oversight, and research assistance, and to Pascaline Dupas and Rema Hanna for very helpful comments and suggestions. We thank the Government of Indonesia through the Ministry of Planning (Bappenas), the Coordinating Ministry for Economy and Social Welfare (Menkokesra), and the Ministry of Home Affairs (Depdagri) for their support for the program and its evaluations. Special thanks to Sujana Royat (Menkokesra); Prasetijono Widjojo, Endah Murniningtyas, Pungky Sumadi, Vivi Yulaswati (Bappenas); and Ayip Muflich, Eko Sri Haryanto, and Bito Wikantosa (Ministry of Home Affairs). The University of Gadjah Mada (UGM), Center for Public Policy Studies, implemented the surveys used in this analysis. Financial support for the overall PNPM Generasi program and the evaluation surveys has come from the Government of Indonesia, the World Bank, the Decentralization Support Facility, the Netherlands Embassy, and the PNPM Support Facility, which consists of donors from Australia, the United Kingdom, the Netherlands, and Denmark, and the Spanish Impact Evaluation Fund; and funding for the analysis came in part from NIH under grant P01 HD061315. Olken was a consultant to the World Bank for part of the period under this evaluation (ending in 2008), Onishi consulted for the World Bank throughout the period under study, and Wong worked full time for the World Bank throughout the period under study. The views expressed in this paper are those of the authors alone and do not represent the views of the World Bank or any of the many individuals or organizations acknowledged here.

[†]Go to http://dx.doi.org/10.1257/app.6.4.1 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

US Millennium Challenge Corporation is explicitly conditioned on recipient countries meeting 17 indicators of good governance, ranging from civil liberties to immunization rates to girls' primary education rates to inflation, and a new movement has advocated that "Cash on Delivery" aid to countries that would explicitly give aid based on achieving specific outcome indicators (Birdsall and Savedoff 2009). The World Bank is similarly moving toward "Program for Results" loans, which would condition actual World Bank disbursements on results obtained. The idea of linking aid to performance is not limited to the developing world: the United States has used a similar approach to encourage state and local school reform through its Race To The Top and No Child Left Behind programs.

Yet despite the policy interest in linking aid to performance, there is little evidence on whether this approach works, and there are reasons it may not. For example, those individuals in charge of implementing aid programs may not directly reap the benefits of the performance incentives, most of which flow to program beneficiaries in the form of future aid programs, not direct payments to implementers. Even if implementers do respond, there can be multitasking problems, where effort allocated toward targeted indicators comes at the expense of other, nonincentivized indicators (Holmstrom and Milgrom 1991). There can also be attempts to manipulate indicators to increase payouts (Linden and Shastry 2012). And, if government budgets are allocated based on performance, there is a risk that performance-based aid will redirect budgets to richer areas that need aid less.

To investigate these issues, we designed a large-scale, randomized field experiment that tests the role of financial performance incentives for villages in improving maternal and child health and education. Villages received an annual block grant of approximately US$10,000, to be allocated to any activity that supported 1 of 12 indicators of health and education service delivery (such as prenatal and postnatal care, childbirth assisted by trained personnel, immunizations, school enrollment, and school attendance). In a randomly chosen subset of subdistricts, villages were given performance incentives, in that 20 percent of the subsequent year's block grant would be allocated among villages in a subdistrict based on their relative performance on each of the 12 targeted indicators. To test the impact of the incentives, in other randomly chosen subdistricts, villages received an identical block grant program with no financial performance incentives. Otherwise, the two versions of the program—with and without performance incentives—were identical down to the last detail (e.g., amounts of money, target indicators, facilitation manuals, monitoring tools, information presented to villagers, cross-village meetings to compare performance on targeted indicators, etc). The experimental design thus precisely identifies the impact of the performance incentives.

A total of 264 subdistricts, with approximately 12 villages each, were randomized into a pure control group or 1 of 2 versions of the program (incentivized or nonincentivized). Surveys were conducted at baseline, and then 18 and 30 months after the program started. With over 2,100 villages randomized to receive either the incentivized or nonincentivized program (plus over 1,000 control villages), and over 1.8 million target beneficiaries in treatment areas, to the best of our knowledge this represents one of the largest randomized social experiments conducted in the world to date, and, hence, a unique opportunity to study these issues at scale.

We begin by examining the impact of the incentives on the 12 main indicators. Given the large number of potential outcomes in a program of this type, we prespecified our analysis plan before looking at the outcome data, and we examine the average standardized effects across the 8 health and 4 education indicators. Using data from the household survey, we find that after 30 months, compared to controls the block grant program overall had a statistically significant, positive average impact on the 12 health and education indicators, such as weight checks, antenatal care, and school participation rates. Comparing the incentivized and nonincentivized treatments, we find the incentives led to greater initial performance (e.g., at 18 months) on health, but no differential performance on education. Specifically, the average standardized effect across the 8 health indicators was about 0.04 standard deviations higher in incentivized rather than nonincentivized areas. While this difference is modest, the incentives' impact was more pronounced in areas with low baseline levels of service delivery: the incentives improved the health indicators by an average of 0.07 standard deviations for a subdistrict at the tenth percentile at baseline. The estimates suggest the average increases we observe may have been particularly driven by preventative health (e.g., prenatal visits and weight checks) and reductions in malnutrition.

We find that the incentives primarily seem to be speeding up impacts on the targeted indicators rather than changing ultimate long-run outcomes. At 30 months, the differences between the incentivized and nonincentivized treatment areas are smaller and no longer statistically significant. This is not because the incentivized group ceased to perform, but rather because the nonincentivized group seems to have caught up with the incentivized group.

Other than the decline in malnutrition at 18 months, we find no evidence that ultimate health outcomes differentially improved with incentives. In fact, the evidence suggests that while neonatal mortality (mortality in 0–28 days) declined in the nonincentivized group relative to controls at both 18 and 30 months, the decline in the incentivized group that was present at 18 months did not persist at 30 months. The fact that reductions in neonatal mortality did not persist with incentives could be an indicator of multitasking problems (e.g., midwives in the incentivized group performed more prenatal care visits and weight checks, which were monitored, but perhaps lower quality prenatal care), or it could be because the improvements in prenatal care and maternal nutrition led some pregnancies that would have ended in miscarriage to survive through to birth, decreasing the health of those who survive to be born (Huang et al. 2013; Valente 2013). We cannot definitely distinguish between these hypotheses.

With respect to education, while the block grant program overall improved enrollments at 30 months, there were no differences between incentivized and nonincentivized areas on the 4 education indicators examined (primary and junior secondary enrollment and attendance) in either survey round. One reason for this may be that in the first year of the program, the program's funding became available after the school year had already started, so it was too late to affect enrollments.

We find evidence for two channels through which the incentives may have had an impact. First, we find that incentives led to an increase in the labor supply of midwives, who are the major providers of the preventative care services we saw increase (e.g., prenatal care, regular weight checks for children). By contrast, we found no change

in labor supplied by teachers. One possible explanation is that midwives are paid on a fee-for-service basis for many services they provide, whereas teachers are not.

Second, the incentives led to what looks like a more efficient use of funds. We find that the incentives led to a reallocation of funds away from education supplies (5 percentage points lower, or about 21 percent) and toward health expenditures (3 percentage points higher, or about 7 percent). Yet, despite the reallocation of funds away from school supplies and uniforms, households were no less likely to receive these items, and were, in fact, more likely to receive scholarships. We find no changes in community effort or the targeting of benefits within villages.

Explicit performance incentives have many potential disadvantages. As discussed above, we find that the incentives led to less of a reduction in neonatal mortality compared to the nonincentivized group, which could be indicative of multitasking problems. Otherwise, though, we find no evidence of a multitasking problem across a very wide array of measures we investigate. We also find no evidence that immunization or school attendance records were manipulated in performance zones relative to nonperformance incentive zones. In fact, we find more accurate record keeping in incentivized areas, where the records were actually being used. And, we find that the fact that incentive payments were relative to other villages in the same subdistrict prevented the incentives from resulting in a net transfer of funds to richer villages. Of course, the incentives studied here represented only 20 percent of the total funds available, and it is possible that these negative effects might only have emerged with even stronger incentives.

In sum, we find that providing incentives increased the speed with which impacts appeared on several targeted health indicators. We find no improvements on measured health and education outcomes due to the incentives through 30 months. An important mechanism appears to be the reallocation of budgets, suggesting that incentives may be more effective when implemented at a high enough geographic level to allow budgetary flexibility.

This study is part of a recent literature on performance incentives for health and education in developing countries.[1] The present study is unique in that incentives are provided to an entire community, and the performance incentives influenced the amount of future aid. This allows for flexibility in budgetary responses to the aid, which is an important channel for the type of performance-based aid to governments being considered at the more macro level.

The results are also related to the literature on the effectiveness of block grants (Musgrave 1997; Das et al. 2013). Most studies of conditional block grants motivate the conditionality concerns about interjurisdictional spillovers, where the conditionality or matching grant forces the local government to internalize the externalities (Oates 1999). In this case, instead, the idea of the incentives is more

[1] Baird, McIntosh, and Özler (2011) find that adding conditions to a household-based Conditional Cash Transfer program in Malawi reduced school dropouts and improved English comprehension. In health, Basinga et al. (2011), find that pay-for-performance for health clinics in Rwanda yields positive impacts of performance incentives on institutional deliveries, preventive health visits for young children, and quality of prenatal care, but not on the quantity of prenatal care or immunizations. In education, a recent series of papers studies the effects of incentives given to teachers and compares them to unincentivized block grants (Muralidharan and Sundararaman 2011; Das et al. 2013).

analogous to a principal-agent problem: the national government uses incentives in funding to incentivize local government in situations where the local government has control rights, much in the way the US federal government ties highway fund block grants to requirements about the minimum drinking age. While this approach is frequently used as a way of incentivizing local governments, there is relatively little rigorous evidence on its effectiveness (Baicker, Clemens, and Singhal 2012).

The remainder of the paper is organized as follows. Section I discusses the design of the program and incentives, the experimental design, and the econometric approach. Section II presents the main results of the impact of the incentives on the 12 targeted indicators. Section III examines the mechanisms through which the incentives may have acted, and Section IV examines the potential adverse effects of incentives. Section V concludes with a discussion of how the potential benefits of incentives compare with the costs of collecting and administering them.

I. Program and Experimental Design

A. *The* Generasi *Program*

The program we study is, to the best of our knowledge, the first health and education program worldwide that combines community block grants with explicit performance bonuses for communities. The program, known formally as *Program Nasional Pemberdayaan Masyarakat—Generasi Sehat dan Cerdas* (National Community Empowerment Program—Healthy and Smart Generation; henceforth Generasi) began in mid-2007 in 129 subdistricts in rural areas of 5 Indonesian provinces: West Java, East Java, North Sulawesi, Gorontalo, and Nusa Tenggara Timur. In the program's second year, which began in mid-2008, the program expanded to cover a total of 2,120 villages in a total of 176 subdistricts, with a total annual budget of US$44 million, funded through a mix of Indonesian government budget appropriations, World Bank, and donor country support.

The program is oriented around the 12 indicators of maternal and child health behavior and educational behavior shown in column 1 of Table 1. These indicators were chosen by the government to be similar to the conditions for a conditional cash transfer being piloted at the same time (but in different locations), and are in the same spirit as those used by other CCTs, such as Mexico's *Progresa* (Gertler 2004; Schultz 2004; Levy 2006). These 12 indicators represent behaviors that are within the direct control of villagers, such as immunizations, prenatal and postnatal care, and school enrollment and attendance, rather than long-term outcomes, such as test scores or infant mortality.

Each year all participating villages receive a block grant. Block grants are usable for any purpose that the village can claim might help address 1 of the 12 indicators shown in Table 1, including, but not limited to, hiring extra midwives for the village, subsidizing the costs of prenatal and postnatal care, providing supplementary feeding, hiring extra teachers, opening a branch school in the village, providing scholarships, providing school uniforms, providing transportation funds, or improving health or school buildings. The block grants averaged US$8,500 in the first

Table 1—Generasi Program Target Indicators and Weights

Performance metric	Weight per measured achievement	Potential times per person per year	Potential points per person per year
1. Prenatal care visit	12	4	48
2. Iron tablets (30 pill packet)	7	3	21
3. Childbirth assisted by trained professional	100	1	100
4. Postnatal care visit	25	2	50
5. Immunizations	4	12	48
6. Monthly weight increases	4	12	48
7. Weight check	2	12	24
8. Vitamin A pill	10	2	20
9. Primary enrollment	25	1	25
10. Monthly primary attendance >= 85%	2	12	24
11. Middle school enrollment	50	1	50
12. Monthly middle school attendance >= 85%	5	12	60

Note: This table shows the 12 indicators used in the *Generasi* program, along with the weights assigned by the program in calculating bonus points.

year of the program and US$13,500 in the second year of the program, or about US$2.70–US$4.30 per person living in treatment villages in the target age ranges.

To decide on the allocation of the funds, trained facilitators help each village elect an 11-member village management team, as well as select local facilitators and volunteers. This management team usually consists of villagers active in health and education issues, such as volunteers from monthly neighborhood child and maternal health meetings. Through social mapping and in-depth discussion groups, villagers identify problems and bottlenecks in reaching the 12 indicators. Inter-village meetings and consultation with local health and education service providers allow the team to obtain information, technical assistance, and support. Following these discussions, the 11-member management team makes the final budget allocation.

B. *Performance Incentives*

The size of a village's block grant depends on its performance on the 12 targeted indicators in the previous year. The purpose is to increase the village's effort toward achieving the targeted indicators (Holmstrom 1979), both by encouraging a more effective allocation of funds and by stimulating village outreach efforts to encourage mothers and children to obtain appropriate health care and increase educational enrollment and attendance. The performance bonus is structured as relative competition between villages within the same subdistrict (*kecamatan*). By making the performance bonuses relative to other local villages, the government sought to minimize the impact of unobserved differences in the capabilities of different areas on the performance bonuses (Lazear and Rosen 1981; Mookherjee 1984; Gibbons and Murphy 1990) and to avoid funds flowing toward richer areas. We discuss the impact of the relative bonus scheme in Section IVC below.

The rule for allocating funds is as follows. The size of the overall block grant allocation for the entire subdistrict is fixed by the subdistrict's population and province.

Within a subdistrict, in year one, funds are divided among villages in proportion to the number of target beneficiaries in each village (i.e., the number of children of varying ages and the expected number of pregnant women). Starting in year two, 80 percent of the subdistrict's funds continue to be divided among villages in proportion to the number of target beneficiaries. The remaining 20 percent of the subdistrict's funds form a performance bonus pool, divided among villages based on performance on the 12 indicators. The bonus pool is allocated in proportion to a weighted sum of each village's performance above a predicted minimum achievement level, i.e.,

$$ShareOfBonus_v = \frac{P_v}{\sum_{j=1}^{N} P_j} \quad \text{where} \quad P_v = \sum_{i=1}^{I} w_i \max[y_{vi} - m_{vi}, 0],$$

where y_{vi} represents village v's performance on indicator i, w_i represents the weight for indicator i, m_{vi} represents the predicted minimum achievement level for village v and indicator i, and P_v is the total number of bonus "points" earned by village v. The minimums (m_{vi}) were set at 70 percent of the predicted level, so that virtually all villages would be "in the money" and face linear incentives on all 12 indicators. The weights, w_i, were set by the government to be approximately proportional to the marginal cost of having an additional individual complete indicator i, and are shown in Table 1. Simple spreadsheets were created to help villagers understand the formulas. Additional details can be found in online Appendix 1.

To monitor achievement of the health indicators, facilitators collect data from health providers and community health workers on the amount of each type of service provided. School enrollment and attendance data are obtained from the official school register.[2]

C. *The Nonincentivized Group*

As discussed above, two versions of the program were implemented to separate the impact of the performance incentives *per se* from the overall impact of the block grant program: the program with performance bonuses (referred to as "incentivized"), and an identical program without performance bonuses (referred to as "nonincentivized"). The nonincentivized version is absolutely identical to the incentivized version except that in the nonincentivized version, there is no performance bonus pool; instead, in all years, 100 percent of funds are divided among villages in proportion to the number of target beneficiaries in each village. Since each entire subdistrict is either entirely incentivized or entirely nonincentivized, and since the total amount of funds per subdistrict is fixed in advance and is the same regardless

[2] Obtaining attendance data from the official school register is not a perfect measure, since it is possible that teachers could manipulate student attendance records to ensure they cross the 85 percent threshold (Linden and Shastry 2012). While more objective measures of monitoring attendance were considered, such as taking daily photos of students (as in Duflo, Hanna, and Ryan 2012) or installing fingerprint readers in all schools (Express India News Service 2008), the program decided not to adopt these more objective measures due to their cost and logistical complexity. We test for this type of differential manipulation in Section IVB.

of whether the subdistrict is incentivized, the expected amount of resources a village obtains is unaffected by incentives.

In all other respects, the two versions of the program are identical: the total amount of funds allocated to each subdistrict is the same in both versions, the same communication materials and indicators are used, the same procedures are used to pick village budget allocations, and the same monitoring tools and scoring system are used. Even the annual point score of villages P_v is also calculated in nonincentivized areas and discussed in comparison to other villages in the community, but as an end-of-year monitoring and evaluation tool, not to allocate funds. The fact that monitoring is identical was an experimental design choice made to precisely isolate the impact of financial performance incentives, holding monitoring constant.

D. *Experimental Design and Data*

Project locations were selected by lottery to form a randomized, controlled field experiment. The randomization was conducted at the subdistrict (*kecamatan*) level, so all villages within a subdistrict either received the same version (either all incentivized or all nonincentivized) or were in the control group. Since some services (e.g., health services, junior secondary schools) service multiple villages within the same subdistrict, but rarely serve people from other subdistricts, randomizing at the subdistrict level and treating all villages within the subdistrict estimates the program's true net impact, rather than possible reallocations among villages. A total of 264 eligible subdistricts were randomized into either 1 of the 2 treatment groups or the control group. Details can be found in online Appendix 2.

The program was phased in over 2 years, with 127 treatment subdistricts in year 1 and 174 treatment subdistricts in year 2. In year one, for logistical reasons, the government prioritized those subdistricts that had previously received the regular village infrastructure program (denoted group P). Since we observe group P status in treatment as well as control, we control for group P status (interacted with time fixed effects) in the experimental analysis to ensure we use only the variation induced by the lottery. By year 2 (2008), 96 percent of eligible subdistricts—174 out of the 181 eligible subdistricts randomized to receive the block grants—were receiving the program. The remaining seven eligible districts received the regular PNPM village infrastructure program instead.[3] Conditional on receiving the program, compliance with the incentivized or nonincentivized randomization was 100 percent.

The phase-in and allocation is shown in Table 2. In all analysis, we report intent-to-treat estimates based on the computer randomization we conducted among the 264 eligible subdistricts and the prioritization rule specified by the government. A balance check against baseline variables is discussed in online Appendix 3 and shown in online Appendix Table 1.

The main dataset we examine is a set of three waves of surveys of households, village officials, health service providers, and school officials. Wave I, the baseline

[3] We do not know why these seven districts received regular PNPM rather than Generasi. We therefore include them in the treatment group as if they had received the program, and interpret the resulting estimates as intent-to-treat estimates. Online Appendix Table 2 shows that controlling receipt of traditional PNPM does not affect the results.

TABLE 2—*GENERASI* RANDOMIZATION AND IMPLEMENTATION

	Incentivized Generasi		Nonincentivized Generasi		Control		
	P	NP	P	NP	P	NP	Total
Total subdistricts in initial randomization	61	39	55	45	55	45	300
Total eligible subdistricts	57	36	48	40	46	37	264
Eligible and received Generasi in:							
2007	57	10	48	12	0	0	127
2008	57	33	48	36	0	0	174

Notes: This table shows the randomization and actual program implementation. P indicates the subdistricts that were ex ante prioritized to receive Generasi in 2007 should they be randomly selected for the program; after the priority areas were given the program, a second lottery was held to select which NP subdistricts randomly selected to receive the program should receive it starting in 2007. The randomization results are shown in the columns (Incentivized Generasi, Nonincentivized Generasi, and Control). Actual implementation status is shown in the rows. Note that conditional in receiving the program, the randomization into the incentivized or nonincentivized version of the program was always perfectly followed.

round, was conducted from June to August 2007, prior to implementation.[4] Wave II, the first follow-up survey, was conducted from October to December 2008, about 18 months after the program began. Wave III was conducted from October 2009 to January 2010, about 30 months after the program began. Approximately 12,000 households were interviewed in each survey wave, as well as more than 8,000 village officials and health and education providers. Within each subdistrict we sampled 5 households from each of 8 villages, for a total of 40 households per subdistrict. Households were selected from a stratified random sample, with the strata consisting of those households with a pregnant woman or mother who had given birth within the past 24 months; households with children under age 15 but not in the first group, and all other households. In the second and third waves, in 50 percent of villages, all households were followed up to form an individual panel, and in the remaining villages new households were selected. These surveys were designed by the authors and were conducted by the Center for Population and Policy Studies (CPPS) of the University of Gadjah Mada, Indonesia. This survey data is unrelated to the data collected by the program for calculating performance bonuses, and was not explicitly linked to the program. Additional details can be found in online Appendix 4.

E. *Estimation*

Since the program was designed as a randomized experiment, the analysis is econometrically straightforward. We compare outcomes in subdistricts randomized to be treatments with subdistricts randomized to be controls, controlling for outcomes at baseline.

We restrict attention to the 264 "eligible" subdistricts, as above, and use the randomization results combined with the government's prioritization rule to construct

[4] Note that in a very small number of villages, the Generasi program field preparations may have begun prior to the baseline survey being completed. We have verified that the main results are unaltered if we do not use the baseline data in these villages. See online Appendix Table 2, column 10.

our treatment variables. Specifically, analyzing Wave II data (corresponding to the first treatment year), we define *BLOCKGRANTS* to be a dummy with value 1 if the subdistrict was randomized to receive either version of the block grants, and either it was in the priority area (group P) or was in the nonpriority area and selected in an additional lottery to receive the program in 2007. In analyzing Wave III data, we define *BLOCKGRANTS* to be a dummy that takes value 1 if the subdistrict was randomized to receive either version of the block grants. We define *INCENTIVES* to be a dummy with value 1 if *BLOCKGRANTS* is 1 and if the subdistrict was randomized to be in the incentivized version. *INCENTIVES* captures the additional effect of the incentives beyond the main effect of having the program, and is the key variable of interest in the paper. These variables capture the intent-to-treat effect of the program, and since the lottery results were very closely followed—they predict true program implementation in 99 percent of subdistricts in 2007 and 96 percent of subdistricts in 2008—they will be very close to the true effect of the treatment on the treated (Imbens and Angrist 1994).

We control for the subdistrict baseline average level of the outcome variable, and the preperiod outcome variable for those who have it, as well as a dummy variable for having nonmissing preperiod values. Since households came from 1 of 3 different samples (those with a child under 2, those with a child age 2–15 but not in the first group, and all others; see online Appendix 4), we sample type dummies, interacted with whether it is a panel village, and for all child-level variables, we include age dummies. We, thus, estimate the following regressions.

Wave II data:

$$(1)\quad y_{pdsi2} = \alpha_d + \beta_1 BLOCKGRANTS_{pds2} + \beta_2 INCENTIVES_{pds2} + \gamma_1 y_{pdsi1} + \gamma_2 1_{\{y_{pdsi1} \neq missing\}} + \gamma_3 \overline{y_{pds1}} + SAMPLE_{pdsi} + \alpha_p \times P_{pds} + \varepsilon_{pdsi}$$

Wave III data:

$$(2)\quad y_{pdsi3} = \alpha_d + \beta_1 BLOCKGRANTS_{pds3} + \beta_2 INCENTIVES_{pds3} + \gamma_1 y_{pdsi1} + \gamma_2 1_{\{y_{pdsi1} \neq missing\}} + \gamma_3 \overline{y_{pds1}} + SAMPLE_{pdsi} + \alpha_p \times P_{pds} + \varepsilon_{pdsi},$$

where i is an individual respondent, p is a province, d is a district, s is a subdistrict, t is the survey wave, y_{pdsit} is the outcome in Wave t, α_d is a district fixed effect, y_{pdsi1} is the baseline value for individual i (assuming that this is a panel household, and 0 if it is not a panel household), $1_{\{y_{pdsi1} \neq missing\}}$ is panel household dummy, $\overline{y_{pds1}}$ is the average baseline value for the subdistrict, *SAMPLE* are sample type dummies interacted with being a panel household, and $\alpha_p \times P_s$ are province-specific dummies

for having had prior community-driven development experience through the PNPM program. We also report pooled results across the two waves in the online Appendix. Standard errors are clustered at the subdistrict level.

The key coefficient of interest is β_2, which estimates the difference between the incentivized and nonincentivized program. We also calculate the total impact of the incentivized version of the program (vis-à-vis pure controls) by adding the coefficients on *INCENTIVES* and *BLOCKGRANTS*. We discuss additional specifications for robustness in Section II.

Since we have many indicators, to estimate joint significance we calculate average standardized effects for each family of indicators, following Kling, Liebman, and Katz (2007). For each indicator *i*, define σ_i^2 to be the variance of *i*. We estimate (1) for each indicator, but run the regressions jointly, clustering the standard errors by subdistrict to allow for arbitrary correlation among the errors across equations within subdistricts both between and across indicators. We define the average standardized effect as $\frac{1}{N}\sum_i \frac{\beta_i}{\sigma_i}$. Following our preanalysis plan, these average standardized effects are the main way we handle multiple inference problems.

Since we also are interested in which individual indicators drive effects, in addition to reporting standard *p*-values for each indicator, we implemented family-wise errors rates (FWER) using the stepwise procedure of Romano and Wolf (2005) and report the results in the table notes. The FWER uses a bootstrap-based method to calculate which hypotheses would be rejected, taking into account the fact that multiple hypotheses are being tested within a given family. We report, in the notes to each table, which individual hypotheses are still rejected once family-wise error rates are taken into account within each family of indicators.

Note that all of the analysis presented here (regression specifications including control variables, outcome variables, and aggregate effects) follows an analysis plan that was finalized in April 2009 for the Wave II data (before we examined any of the Wave II data) and in January 2010 (before we examined any of the Wave III data). These hypothesis documents were registered with the Abdul Latif Jameel Poverty Action Lab at MIT.[5]

II. Main Results on Targeted Outcomes

A. *Overall Impact on Targeted Indicators*

Table 3 presents the results on the 12 targeted indicators. Each row presents three separate regressions. Column 1 shows the baseline mean of the variable. Columns 2–4 show the Wave II survey results (after 18 months of program implementation) from equation (1); columns 5–7 show the Wave III results estimated using equation (2). For each specification, we show the total treatment effect in incentive areas (the sum of

[5] The hypotheses documents, showing the date they were archived, are publicly available at http://www.povertyactionlab.org/Hypothesis-Registry. A full set of tables that correspond to the Wave III analysis plan can be found in the full Generasi impact evaluation report (Olken, Onishi, and Wong 2011), available at http://goo.gl/TudhZ. Other recent economics papers using these types of prespecified analysis plans include Alatas et al. (2012); Finkelstein et al. (2012); and Casey, Glennerster, and Miguel (2012).

Table 3—Impact on Targeted Outcomes

	Wave II				Wave III		
Indicator	Baseline mean (1)	Incentive treatment effect (2)	Non-incentive treatment effect (3)	Incentive additional effect (4)	Incentive treatment effect (5)	Non-incentive treatment effect (6)	Incentive additional effect (7)
Panel A. Health							
Number prenatal visits	7.451	0.333	−0.274	0.608***	0.162	−0.018	0.180
	[4.292]	(0.234)	(0.201)	(0.220)	(0.192)	(0.188)	(0.173)
Delivery by trained midwife	0.673	0.037	0.040	−0.004	0.012	−0.008	0.019
	[0.469]	(0.027)	(0.027)	(0.025)	(0.021)	(0.023)	(0.021)
Number of postnatal visits	1.734	−0.160	−0.056	−0.104	−0.028	−0.024	−0.004
	[2.465]	(0.140)	(0.120)	(0.140)	(0.129)	(0.124)	(0.129)
Iron tablet sachets	1.587	0.130	0.051	0.078	0.076	0.045	0.031
	[1.255]	(0.084)	(0.081)	(0.081)	(0.058)	(0.065)	(0.063)
Percent of immunization	0.654	0.027	0.012	0.015	0.010	−0.006	0.016
	[0.366]	(0.018)	(0.018)	(0.018)	(0.015)	(0.015)	(0.014)
Number of weight checks	2.127	0.164***	0.069	0.096*	0.176***	0.199***	−0.024
	[1.189]	(0.052)	(0.049)	(0.054)	(0.055)	(0.052)	(0.051)
Number Vitamin A	1.528	−0.008	0.005	−0.013	0.085*	0.002	0.082
Supplements	[1.136]	(0.052)	(0.055)	(0.058)	(0.048)	(0.054)	(0.053)
Percent malnourished	0.168	−0.016	0.011	−0.027*	−0.017	−0.026*	0.009
	[0.374]	(0.016)	(0.015)	(0.016)	(0.014)	(0.015)	(0.016)
Average standardized effect health		0.055**	0.014	0.041*	0.0523**	0.027	0.026
		(0.024)	(0.023)	(0.024)	(0.023)	(0.022)	(0.022)
Panel B. Education							
Age 7–12 participation rate	0.948	−0.001	0.003	−0.004	0.005	0.011***	−0.006
	[0.222]	(0.005)	(0.006)	(0.006)	(0.005)	(0.004)	(0.005)
Age 13–15 participation rate	0.823	−0.034*	−0.050**	0.016	0.020	0.013	0.007
	[0.382]	(0.020)	(0.023)	(0.024)	(0.017)	(0.016)	(0.014)
Age 7–12 gross attendance	0.904	0.001	0.002	−0.001	0.003	0.004	−0.001
	[0.277]	(0.005)	(0.005)	(0.006)	(0.007)	(0.006)	(0.006)
Age 13–15 gross attendance	0.769	−0.040*	−0.065***	0.025	0.026	0.016	0.010
	[0.412]	(0.021)	(0.024)	(0.025)	(0.018)	(0.017)	(0.015)
Average standardized effect education		−0.062	−0.090**	0.027	0.048	0.045*	0.003
		(0.039)	(0.045)	(0.045)	(0.029)	(0.027)	(0.027)

(*Continued*)

the coefficients on *BLOCKGRANTS* and *INCENTIVES*), the total treatment effect in nonincentive areas (the coefficient on *BLOCKGRANTS*), and the additional treatment effect due to the incentives (the coefficient on *INCENTIVES*). We first present the eight health indicators, along with the average standardized effect for those indicators. We then present the 4 education indicators with standardized effect, and then the overall standardized effect for all 12 indicators. The final three rows show the impact on total "bonus points," where the 12 indicators are weighted using the weights in Table 1 and an estimate for the number of affected households (using the same estimated number of households in both treatment groups). All data is from household surveys.

We begin by examining the average standardized effects. Focusing first on the Wave II (18 month) results, the average standardized effect among the 8 health

TABLE 3—IMPACT ON TARGETED OUTCOMES (*Continued*)

		Wave II			Wave III		
Indicator	Baseline mean (1)	Incentive treatment effect (2)	Non-incentive treatment effect (3)	Incentive additional effect (4)	Incentive treatment effect (5)	Non-incentive treatment effect (6)	Incentive additional effect (7)
Panel C. Overall							
Average standardized effect		0.016	−0.021	0.036	0.051**	0.033*	0.018
Overall		(0.023)	(0.022)	(0.024)	(0.020)	(0.018)	(0.019)
Panel D. Calculation of total points							
Total points		0.698	−1.638	2.336*	2.889**	2.150*	0.738
(millions)		(1.376)	(1.263)	(1.414)	(1.249)	(1.152)	(1.176)
Total points health		1.941**	0.206	1.735*	1.962**	1.388	0.574
(millions)		(0.987)	(0.933)	(1.005)	(0.985)	(0.982)	(0.969)
Total points education		−1.243*	−1.844**	0.601	0.927	0.762	0.165
(millions)		(0.710)	(0.822)	(0.836)	(0.585)	(0.553)	(0.516)

Notes: Data is from the household survey. Column 1 shows the baseline mean of the variable shown, with standard deviations in brackets. Each row of columns 2–4 and 5–7 shows coefficients from a regression of the variable shown on an incentive treatment dummy, a nonincentive treatment dummy, district fixed effects, province × group P fixed effects, and baseline means, as described in the text. Robust standard errors in parentheses, adjusted for clustering at the subdistrict level. In columns 2–4 the treatment variable is defined based on year one program placement, and in columns 5–7 it is defined based on year two program placement. All treatment variables are defined using the original randomizations combined with eligibility rules, rather than actual program implementation, and so are interpretable as intent-to-treat estimates. Columns 4 and 7 are the calculated difference between the previous two columns. Average standardized effects and total points reported in the bottom rows are calculated using the estimated coefficients from the 12 individual regressions above using the formula shown in the text, adjusted for arbitrary cross-equation clustering of standard errors within subdistricts. Applying family-wise error rates (Romano and Wolf 2005) to the incentive additional effect (columns 4, 7, and 10), the only coefficient where the null is rejected, taking into account multiple comparisons, is prenatal visits in Wave II, which is rejected at the 10 percent level.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

indicators is 0.04 standard deviations higher in the incentivized group than in the nonincentivized control group (statistically significant at the 10 percent level). There are no statistically detectable impacts on education or overall (though the overall effect has a p-value of 0.12).

To understand what may be driving the impact on health indicators, we examine the indicators one by one, and find differential effects of the incentives on 3 of 12 indicators (column 4). Two of the 3 indicators that respond appear to be preventative care: prenatal care (increased by 0.61 visits, or 8.2 percent of the baseline mean) and regular monthly weight checks for under-5 year olds (increased by 0.096 over the previous 3 months, or about 4.5 percent of the baseline mean). When we apply the FWER multiple-hypothesis testing correction, the only coefficient that is still statistically significant is prenatal visits, which is statistically significant at the 10 percent level even taking into account the multiple hypothesis testing.

One reason preventative care may be responsive is that it is largely conducted at *posyandus*, the neighborhood village health posts where mothers and children gather monthly to get their children weighed and receive preventative care (usually provided by village midwives). These meetings are organized by community

volunteers. Since many of these volunteers may have been involved in managing the block grant program, they may have been particularly responsive to the incentives.

The other indicator that responds is malnutrition, defined as being more than 2 standard deviations below the weight-for-age normal *z*-score for children under three. This is measured directly by the survey teams, who brought scales and independently weighed children at home. Malnutrition is 2.6 percentage points (15 percent) lower in the incentivized group than in the nonincentivized group, though this coefficient is not statistically significant once we take into account multiple hypothesis testing. Since the purpose of regular weight checks for children is precisely to identify those who are not growing properly so that they can receive supplemental nutrition, it is not surprising that an impact on improving regular weight monthly checks for children in turn leads to fewer children who are severely underweight.

The results in Wave III, after 30 months of the program, are more muted, showing no statistically significant differences between incentivized and nonincentivized. Closer inspection suggests that most of the changes from Wave II are driven by the nonincentivized group improving, rather than the incentivized group declining. For example, columns 5 and 6, which show the incentivized and nonincentivized groups relative to pure controls, show that in Wave III, the nonincentivized group saw improvements in weight checks and malnutrition of similar magnitude to the incentivized group.[6] This suggests that the main impact of the incentives was to speed up the impacts of the program on preventative care and malnutrition, rather than on changing the ultimate long-run impacts of the programs on the targeted indicators.

No effects of the incentives were seen on education in either wave. In both incentivized and nonincentivized areas, age 13–15 participation and attendance fell relative to controls in Wave II, and age 7–12 participation increased in Wave III. Consistent with this, the average standardized effects for education for both incentivized and nonincentivized areas decreased in Wave II and increased in Wave III.[7] One reason enrollments did not increase until the second year of program implementation (i.e., Wave III) is that the program did not disburse money until after the school year had begun, so it was structurally very unlikely to generate effects until the subsequent school year (though the fact that enrollments actually declined in Wave II in both incentivized and nonincentivized groups relative to pure control is something of a mystery). Overall, this was a period when enrollments were increasing dramatically throughout Indonesia, so enrollments increased everywhere relative to baseline.

The average standardized effects weight the indicators by the control groups' standard deviations of the relevant variables. An alternative approach is to use the weights used by the program in calculating bonus payments. This approach has the advantage

[6] To test the changes over time more directly, online Appendix Table 6 restricts the sample to those subdistricts that were either treated both years or control both years (i.e., drops those subdistricts where treatment started in the second year). Columns 7 and 8 show the differences between the impact in Wave II and the impact in Wave III relative to control, and column 9 shows the difference between the incentive effect in Wave II and Wave III. Online Appendix Table 6 shows that the decline in the incentive effect of weight checks (due to the increase in the nonincentivized group) is statistically significant, while the decline in malnutrition is not statistically significant.

[7] In particular, if we pool incentive and nonincentivized treatments, the change in 7–12 participation and the education average standardized effects become statistically significant. We also find a statistically significant 4 percentage point (6 percent) improvement in the percentage of people age 13–15 enrolled in middle school. These results are in Olken, Onishi, and Wong (2011).

that it weights each indicator by the weight assigned to it by the government. For each indicator, we use the weights in Table 1, multiplied by the number of potential beneficiaries of each indicator (garnered from population data in different age ranges from the program's internal management system, and using the same numbers for both treatment groups), and aggregate to determine the total number of "points" created. The results show a similar story to the average standardized effects. In Wave II, 89 percent of the program's impact on health (in terms of points) can be attributed to the incentives, and the incentives had a statistically significant increase on both points from health and total points overall. In Wave III, 29 percent of the program's impact on health (in terms of points) can be attributed to the incentives, though the Wave III difference is not statistically significant either for health or overall.

Although we prespecified equations (1) and (2) as the main regression specifications of interest, we have also considered a wide range of alternative specifications. Online Appendix Table 2 reports the coefficient on *INCENTIVES*—the equivalent of columns 4 and 7, as well as average effects across both waves—for specifications where we control for the baseline level of all 12 indicators instead of just the indicator in question, control only for subdistrict averages at baseline rather than also using individual baseline controls, include no controls, estimate using first-differences rather than controlling for the baseline level, and run everything aggregated to the subdistrict, rather than using individual-level data. The results are very consistent with the main specification in Table 3.

B. *Heterogeneity in Impact on Targeted Indicators*

We test whether incentives had a larger impact in areas with low baseline levels. The idea is that the marginal cost of improving achievement is higher if the baseline level is higher, e.g., moving from 98 percent to 99 percent enrollment rates is harder than moving from 80 percent to 81 percent.[8] We re-estimate equations (1) and (2), interacting *BLOCKGRANTS* and *INCENTIVES* with the mean value of the indicator in the subdistrict at baseline. The results are shown in Table 4 (indicator-by-indicator results are in online Appendix Table 4). A negative interaction coefficient implies that the program was more effective in areas with worse baseline levels. For ease of interpretation, we also calculate the implied impacts at the tenth percentile of the baseline distribution.

The results confirm that the incentives were more effective in areas with lower baseline levels—the standardized interaction term of *INCENTIVES* $\times$ *BASELINE_VALUE* in columns 3 and 7 are negative and, in both Wave II and overall, statistically significant. To interpret the magnitude, note that, in Wave II, the incentives added 0.072 standard deviations to the health indicators at the tenth percentile of the baseline distribution. In Wave III, it was 0.061 standard deviations (not statistically significant). Pooled across the two waves, it was 0.065 standard deviations (statistically significant at 5 percent; results not shown). These effects are about double the average effect of the program shown in Table 3.

[8] Note that this is the main dimension of heterogeneity we specified in the prespecified analysis plan.

Table 4—Interactions with Baseline Level of Service Delivery, Average Standardized Effects

	Wave II				Wave III			
Indicator	Generasi incentive total effect × preperiod level (1)	Generasi nonincentive total effect × preperiod level (2)	Generasi incentive additional effect × preperiod level (3)	Incentive additional effect at 10th percentile (4)	Generasi incentive total effect × preperiod level (5)	Generasi nonincentive total effect × preperiod level (6)	Generasi incentive additional effect × preperiod level (7)	Incentive additional effect at 10th percentile (8)
Average standardized effect	−0.211** (0.096)	−0.154 (0.112)	−0.057 (0.133)	0.057 (0.042)	−0.187** (0.092)	−0.218** (0.085)	0.031 (0.088)	0.025 (0.033)
Average standardized effect health	−0.187*** (0.062)	−0.065 (0.057)	−0.122* (0.068)	0.072* (0.037)	−0.091 (0.066)	−0.004 (0.066)	−0.088 (0.064)	0.061 (0.039)
Average standardized effect education	−0.259 (0.253)	−0.333 (0.313)	0.074 (0.369)	0.025 (0.086)	−0.378 (0.245)	−0.647*** (0.219)	0.269 (0.235)	−0.049 (0.050)

Notes: See notes to Table 3. Data is from the household survey. Columns 1 and 5 interact the incentive treatment dummy with the baseline subdistrict mean of the variable shown, and columns 2 and 5 interact the nonincentive treatment dummy with the baseline subdistrict mean of the variable shown. Columns 3 and 7 are the difference between the two previous columns. Columns 4 and 8 show the estimated additional impact of incentives evaluated at the tenth percentile of the indicator at baseline. The indicator-by-indicator regressions corresponding to these average standardized effects are shown in online Appendix Table 4.

Consistent with the results in Table 4, we find that the incentives were more effective in the poorer, off-Java locations: on average across all waves, the total standardized effect for health was 0.11 standard deviations higher in incentivized areas than nonincentivized areas in NTT province relative to Java, and 0.14 standard deviations higher in incentivized areas than nonincentivized areas in Sulawesi relative to Java (see online Appendix Table 3). This is not surprising given the lower levels of baseline service delivery in these areas: malnutrition for under 3-year olds is 12.6 percent in Java, but 24.7 percent in NTT and 23.4 percent in Sulawesi. These results confirm the idea that the incentives were substantially more effective in areas with lower levels of baseline service provision.

C. *Impacts on Health and Education Outcomes*

The 12 targeted outcomes of the program are, other than malnutrition, inputs to health and education—things like health-seeking behavior and educational enrollment and attendance—rather than actual metrics of health and education. To examine impacts on health, we measure anthropometrics in the household survey (malnourishment, measured being 2 or 3 standard deviations below normal in weight-for-age; wasting, measured as being 2 or 3 standard deviations below normal in weight-for-height; and stunting, measured as being 2 or 3 standard deviations below normal in height-for-age), acute illness (prevalence of diarrhea or acute respiratory infections in the previous month), and mortality (neonatal and infant). To measure learning, we conducted at-home tests of children on both reading in Bahasa Indonesia and in math, using test questions drawn from the standard Ministry of National Education test databank.

The results are presented in Table 5, and generally show no systematic improvements in these indicators between the incentivized and nonincentivized group. In fact, neonatal mortality actually appears worse in Wave III in incentivized areas relative to nonincentivized areas.

Table 5—Impacts on Nutrition, Mortality, and Test Scores

Indicator	Baseline mean (1)	Wave II Incentive treatment effect (2)	Wave II Non-incentive treatment effect (3)	Wave II Incentive additional effect (4)	Wave III Incentive treatment effect (5)	Wave III Non-incentive treatment effect (6)	Wave III Incentive additional effect (7)
Panel A. Health							
Malnourished (0–3 years)	0.168 [0.006]	−0.016 (0.016)	0.011 (0.015)	−0.027* (0.016)	−0.017 (0.014)	−0.026* (0.015)	0.009 (0.016)
Severely malnourished (0–3 years)	0.046 [0.003]	−0.007 (0.009)	−0.005 (0.008)	−0.003 (0.009)	−0.016 (0.010)	−0.014 (0.010)	−0.002 (0.009)
Weight for age *z*-score	−0.841 [0.020]	−0.016 (0.050)	−0.017 (0.046)	0.001 (0.052)	0.056 (0.052)	0.067 (0.050)	−0.010 (0.048)
Wasting (0–3 years)	0.124 [0.006]				−0.005 (0.017)	0.003 (0.016)	−0.008 (0.015)
Severe wasting (0–3 years)	0.048 [0.004]				0.000 (0.011)	0.006 (0.012)	−0.006 (0.012)
Weight for height *z*-score	−0.066 [0.030]				0.032 (0.081)	0.135 (0.084)	−0.103 (0.089)
Stunting (0–3 years)	0.383 [0.008]				0.034* (0.020)	0.027 (0.020)	0.006 (0.021)
Severe stunting (0–3 years)	0.206 [0.007]				−0.007 (0.019)	0.019 (0.019)	−0.026 (0.018)
Height for age *z*-score	−1.369 [0.035]				0.052 (0.096)	−0.013 (0.093)	0.066 (0.098)
Diarrhea or ARI	0.356 [0.008]	−0.026 (0.023)	0.012 (0.020)	−0.038 (0.024)	0.003 (0.023)	−0.003 (0.022)	0.006 (0.019)
Neonatal mortality (0–28 days) (births in past 18 months)	0.013 [0.002]	−0.006* (0.003)	−0.006 (0.004)	0.000 (0.003)	0.006 (0.005)	−0.008* (0.004)	0.014*** (0.004)
Infant mortality (1–12 months) (births in past 24 months)	0.012 [0.002]	−0.004 (0.004)	−0.005 (0.004)	0.001 (0.004)	0.005 (0.005)	0.000 (0.004)	0.005 (0.004)
Mortality 0–12 months (births in past 24 months)	0.024 [0.003]	−0.006 (0.005)	−0.011** (0.005)	0.005 (0.005)	0.012* (0.006)	−0.004 (0.005)	0.016*** (0.006)
Average standardized effect health		0.048** (0.021)	0.029 (0.019)	0.019 (0.021)	−0.029 (0.027)	0.025 (0.025)	−0.054** (0.026)

(*Continued*)

As previously discussed, malnutrition is lower in the incentivized group in Wave II (though this is not statistically significant once one takes into account family-wise error rates). In Wave III, after 30 months, the nonincentivized group also shows improvements in malnutrition, so there is no longer a difference between them. Height-based anthropometrics, which were measured in Wave III only, show no systematic differences. It is also worth noting that the weight-for-age *z*-score is not statistically significantly different, suggesting that the malnutrition result is being driven by changes at the very bottom of the distribution (consistent with a program that targets highly malnourished children).

With respect to mortality, neonatal mortality fell in both incentivized and non-incentive areas relative to control in Wave II, by about six deaths per thousand. In Wave III, however, it was lower only in nonincentive areas (by about eight deaths per thousand), and there was no decline in mortality in incentivized areas compared

Table 5—Impacts on Nutrition, Mortality, and Test Scores (*Continued*)

		Wave II			Wave III		
Indicator	Baseline mean (1)	Incentive treatment effect (2)	Non-incentive treatment effect (3)	Incentive additional effect (4)	Incentive treatment effect (5)	Non-incentive treatment effect (6)	Incentive additional effect (7)
Panel B. Education							
Home-based Bahasa test 7–12 years (age-adjusted *z*-score)	−0.037 [0.019]				−0.048 (0.048)	−0.001 (0.044)	−0.046 (0.044)
Home-based math test 7–12 years (age-adjusted *z*-score)	−0.036 [0.019]				−0.026 (0.049)	0.002 (0.049)	−0.027 (0.048)
Home-based total test 7–12 years (age-adjusted *z*-score)	−0.046 [0.019]				−0.042 (0.049)	0.010 (0.047)	−0.052 (0.046)
Home-based Bahasa test 13–15 years (age-adjusted *z*-score)	−0.010 [0.032]				0.034 (0.071)	0.093 (0.078)	−0.059 (0.061)
Home-based math test 13–15 years (age-adjusted *z*-score)	−0.002 [0.032]				−0.002 (0.068)	0.085 (0.071)	−0.087 (0.063)
Home-based total test 13–15 years (age-adjusted *z*-score)	−0.006 [0.032]				0.012 (0.071)	0.088 (0.076)	−0.076 (0.064)
Average standardized effect on education					−0.012 (0.039)	0.043 (0.042)	−0.055 (0.037)
Panel C. Overall							
Average standardized effect overall		0.048** (0.021)	0.029 (0.019)	0.019 (0.021)	−0.026 (0.020)	0.032* (0.019)	−0.058*** (0.019)

Notes: See notes to Table 3. Data is from the household survey. Test scores were conducted at home as part of the household survey. Note that for computing average standardized effects, we multiply the health variables by −1, so that all coefficients are defined so that improvements in health or education are positive numbers. Average standardized effects do not include infant mortality (1–12 months), weight for age *z*-score, weight for height *z*-score, and height for age *z*-score, as these variables were not specified in the preanalysis plan. Applying family-wise error rates (Romano and Wolf 2005) to the incentive additional effect (columns 4 and 7), in Wave III, 0–28 day mortality is rejected at the 5 percent level in the family of all indicators, and 0–28 day and 0–12 month mortality are rejected in the health comparison at the 5 and 10 percent levels, respectively.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

to control. The results in column 7 therefore suggest greater mortality in the incentivized areas relative to the nonincentivized areas. The difference in neonatal mortality between incentivized and nonincentivized areas survives multiple hypothesis testing corrections. The results suggest that the difference in Wave III is entirely in neonatal mortality (mortality during the first 28 days), as there is no difference in infant mortality (mortality from 1 to 12 months). In fact, of the 14 neonatal deaths that occur in the incentivized group, 10 of them occur within the 1 day after birth, suggesting that the increase is being driven almost entirely by these early deaths within 24 hours of birth.

The fact that the decline in infant mortality in Wave III only occurs in nonincentivized areas is a puzzle. There are two possible interpretations. One interpretation

is that this is evidence of a multitasking problem. For example, perhaps quantity of prenatal services increased but quality of prenatal services decreased. We know, for example, that midwives performed many more weight checks and prenatal visits in the incentivized areas relative to the nonincentivized areas, so it is possible that this extra effort on incentivized indicators crowded out other important dimensions of prenatal care. The results on quality of prenatal care presented below suggest this is not an issue so far as we can measure quality, but it is possible there is an unobserved dimension of quality that we cannot measure.

A second interpretation is that this increase in neonatal mortality is a consequence of the fact that the increase in prenatal care in incentive areas, which occurred in Wave II, led to an increase in marginal pregnancies actually surviving to become live births, which, in turn, counteracted the initial improvement in mortality.[9] Unfortunately, data on miscarriages are unreliable, and many of the potentially vulnerable early-stage pregnancies are not even detected, so one cannot directly test this hypothesis.[10] One therefore cannot know for sure whether these additional early births represent a decline in miscarriages and, hence, an improvement in health (births carried to term that would otherwise have miscarried), or instead represent births that are somehow being delivered earlier than they would have (and, hence, a deterioration of health), so it is not possible to fully distinguish between these two alternative interpretations of these results.[11]

D. *Discussion*

These results suggest that the incentives' main impact was to accelerate performance improvements on preventative care (e.g., prenatal care and regular weight

[9] This latter hypothesis, that improvements in prenatal care can negatively affect the health of the born population because marginal pregnancies are carried to term rather than resulting in miscarriage, is related to Bozzoli, Deaton, and Quintana-Domeque (2009), who investigate the link between adult height and childhood disease, and Gørgens, Meng, and Vaithianathan (2012), who study stunting and selection effects of the 1959–1961 Chinese famine. The closest papers that pay particular attention to selection effects occurring through early miscarriage (i.e., in utero selection versus selection via early childhood mortality) are Huang et al. (2012), which studies this issue in the context of the Chinese famine, and Valente (2013), which studies the impact of civil conflict in Nepal.

[10] We do ask about miscarriage rates in our survey, and find that there is no statistically significant difference in stillbirth rates in our survey between incentivized and nonincentivized treatments. However, it is important to note that most of the change in marginal births may be coming from very early (e.g., first trimester) miscarriages, which are associated with maternal nutrition and stress (Almond and Mazumder 2011). These types of early miscarriages appear to be grossly underreported in our data (and many may not even be detected), so it is not surprising we do not find an effect in the data.

[11] While it is not possible to definitively determine which hypothesis is behind the results, there are several pieces of evidence that suggests that this latter hypothesis is at least plausible in this context. First, as noted above, it is important to note that all of the mortality effects we find are driven by neonatal mortality—0 to 28 days—and virtually all are driven by the first day after birth. This is consistent with the idea that these increased deaths are related to fragile newborns. Second, there is a statistically significant decline in gestational age associated with the incentive treatment, of about 0.4 weeks. (See online Appendix Table 14.) This is driven by premature births: live births less than 37 weeks are 3.1 percentage points more likely and live births less than 36 weeks are 2.1 percentage points more likely in the incentive treatment. These early live births, in turn, drive the neonatal (under 28 day) mortality—80 percent of the mortality increase is associated with births less than 37 weeks and 50 percent of the mortality increase is associated with births less than 36 weeks. Combined, this suggests a link between earlier births and the increase in mortality. We also find that mothers in the incentivized areas reported being more likely to receive prenatal information about maternal nutrition. (See online Appendix Table 14.) Since maternal nutrition is a key link in the inutero selection effects documented elsewhere (e.g., Almond and Mazumder 2011; Huang et al. 2013), all these facts are consistent with the idea that there was a change in the selection margin in the incentivized areas. Nevertheless, since we do not observe these "missing births" directly in the control group, it is difficult to know for sure.

checks for young children). One reason both prenatal care and weight checks may be particularly responsive is that they are organized by community members at monthly neighborhood primary care sessions, and many of these same community members were involved in managing the block grants and may have been particularly attuned to the incentives. While some effects are substantial (16 percent reduction in malnutrition rates from baseline in just 18 months), when we consider all 8 health indicators together, the average standardized effect is a modest 0.04 standard deviations, and there was no impact on education. The effects of the incentives seem to be about accelerating improvements rather than changing long-run outcomes—30 months after the program started, the nonincentivized group had improved, and was showing the same impacts as the incentivized group compared to pure controls.

An interesting question is why health indicators appear to have been more responsive than education. One possibility, explored below, is that the health providers are more responsive than education providers. Another possible explanation is costs: it may be that simple preventative care is less costly to provide and easier for the community to mobilize than getting the few remaining children who are not yet in school enrolled. Indeed, the government set the scores in Table 1 with high weights on education indicators precisely because it believed these were more difficult to achieve, but it is possible that the incentives were not sufficient to cover the differential costs, and communities were optimizing. The subsequent sections explore, to the extent we can in the data, how the incentives may have worked and test for potential downsides.

III. Mechanisms

In this section, we explore three potential mechanisms through which the incentives may have had an impact: by inducing a change in the allocation of funds, by changing provider or community effort, and by changing the targeting of funds and benefits.

A. *Allocation of Funds*

Table 6 examines whether the incentives had impacts on communities' allocation of the grants. Each row shows the share of the village's grant spent on the item.

The most notable finding is that the incentives led to a shift away from education supplies—uniforms, books, and other school supplies—and toward health expenditures. Spending on education supplies is about 4 percentage points (15 percent) lower in incentivized villages, and health spending is about 3 percentage points (7 percent) higher. One interpretation is that these education supplies are essentially a transfer—when distributed, they tend to be distributed quite broadly to the entire population, the vast majority of whose children are already in school, and therefore may have little impact on school attendance and enrollment. As shown in Table 3, the incentives improved health outcomes with no detrimental effect on education, so combined this suggests that the incentives may have led communities to reallocate funds away from potentially politically popular but ineffective education spending towards more effective health spending.

Table 6—Change in Budget Allocations

Indicator	Wave II			Wave III		
	Incentive Mean (1)	Nonincentive Mean (2)	Incentive additional effect (3)	Incentive Mean (4)	Nonincentive Mean (5)	Incentive additional effect (6)
Panel A. Health versus education						
All health expenditures	0.470	0.432	0.033**	0.490	0.470	0.029**
			(0.015)			(0.012)
Health durables	0.099	0.085	0.011	0.126	0.110	0.011
			(0.012)			(0.015)
Health benefiting providers	0.017	0.014	0.003	0.022	0.023	0.001
			(0.005)			(0.004)
Panel B. Transfers						
All transfers	0.731	0.756	−0.028	0.728	0.743	−0.004
			(0.025)			(0.024)
Education supplies	0.236	0.274	−0.049**	0.236	0.270	−0.028
			(0.025)			(0.018)
Supplementary feeding	0.217	0.177	0.022	0.212	0.215	0.009
			(0.014)			(0.013)
Subsidies	0.279	0.305	−0.001	0.280	0.258	0.015
			(0.024)			(0.020)
Uniform unit values	146,132	158,407	−45,230	108,789	99,881	12,517
			(54,467)			(12,128)

Note: See notes to Table 3. Data from administrative records, one observation per village. Since budgets are only available for treatment areas, columns 3 and 6 regress the variable on an incentive subdistrict dummy.

Communities often use the grants to provide a small amount of food at the monthly weighing sessions, mostly to encourage poor mothers to bring at-risk children to the weighing sessions to be checked by the midwife. Table 6 suggests that expenditures on supplementary feeding activities—which work both as a show-up incentive and are used intensively for underweight children—appear higher in the incentivized group in Wave II, although the difference is not significant. By Wave III, this effect reversed, which may explain why the initial differential impacts on weighings and malnutrition are reversed subsequently.

We also tested two hypotheses that were not borne out in the data. First, we expected that, since incentives effectively increase the discount rate (since a return in the current year will affect bonuses next year), we would expect a shift away from durable investments – if anything, the opposite appears to have occurred, with spending on health durables increasing by about 1.7 percentage points (15 percent). Second, we expected that incentives would lead to a decrease in "capture" for expenses benefitting providers (e.g., uniforms for health volunteers), but we see no impact on this dimension.

This evidence was on how the money was spent. Table 7 examines what households actually received from the block grants, using data from the household survey. Both incentivized and nonincentivized versions show substantial increases in virtually all items, confirming that the block grant did indeed result in noticeable transfers of many types to households.

With respect to the incentives, there are two notable results. First, households were no less likely to receive a uniform or school supplies in the incentive treatments

Table 7—Direct Benefits Received, Incentivized versus Nonincentivized

		Wave II			Wave III		
Indicator	Control mean	Incentive treatment effect (1)	Non-incentive treatment effect (2)	Incentive additional effect (3)	Incentive treatment effect (4)	Non-incentive treatment effect (5)	Incentive additional effect (6)
Panel A. Health							
Received supp. feeding at school	0.005 [0.001]	0.005 (0.003)	0.004** (0.002)	0.001 (0.004)	0.006 (0.006)	0.003 (0.005)	0.003 (0.007)
Received supp. feeding at posyandu	0.464 [0.017]	0.153*** (0.028)	0.156*** (0.027)	−0.003 (0.028)	0.175*** (0.025)	0.204*** (0.022)	−0.030 (0.023)
Received intensive supp. feeding at school	0.026 [0.005]	0.008 (0.007)	0.025** (0.011)	−0.018 (0.011)	0.024** (0.010)	0.019** (0.009)	0.005 (0.010)
Received health subsidy for pre/postnatal care	0.005 [0.002]	0.034*** (0.008)	0.027*** (0.007)	0.007 (0.009)	0.027*** (0.006)	0.036*** (0.007)	−0.009 (0.009)
Received health subsidy for childbirth	0.038 [0.008]	0.101*** (0.017)	0.127*** (0.017)	−0.026 (0.019)	0.097*** (0.016)	0.125*** (0.020)	−0.028 (0.023)
Average standardized effect health		0.287*** (0.037)	0.315*** (0.031)	−0.028 (0.039)	0.267*** (0.031)	0.315*** (0.035)	−0.048 (0.042)
Panel B. Education							
Received scholarship	0.024 [0.005]	0.016** (0.007)	0.008 (0.006)	0.009 (0.008)	0.021** (0.009)	0.009 (0.007)	0.012 (0.009)
Received uniform	0.013 [0.004]	0.110*** (0.019)	0.083*** (0.012)	0.027 (0.018)	0.082*** (0.013)	0.072*** (0.010)	0.010 (0.015)
Value of uniforms (Rp.)	712 [264]	7,845*** (1,569)	6,099*** (1,035)	1,746 (1,447)	7,123*** (1,313)	5,936*** (1,118)	1,187 (1,521)
Received other school supplies	0.007 [0.003]	0.063*** (0.012)	0.054*** (0.009)	0.010 (0.012)	0.070*** (0.012)	0.053*** (0.010)	0.017 (0.015)
Received transport subsidy	0.007 [0.002]	0.014*** (0.005)	0.005* (0.003)	0.009 (0.006)	0.008*** (0.002)	0.005*** (0.002)	0.003 (0.003)
Received other school support	0.000 [0.000]	0.000 (0.000)	0.000 (0.000)	0.001 (0.000)	0.007** (0.003)	0.006* (0.003)	0.001 (0.004)
Average standardized effect education		0.399*** (0.064)	0.290*** (0.042)	0.109* (0.061)	0.351*** (0.050)	0.278*** (0.041)	0.073 (0.059)
Panel C. Overall							
Average standardized effect overall		0.343*** (0.041)	0.303*** (0.030)	0.040 (0.041)	0.309*** (0.031)	0.296*** (0.028)	0.013 (0.039)

Notes: See notes to Table 3. Data is from the household survey. Note that instead of showing a baseline mean, we show the Wave II control group mean because there is no data available for these categories in Wave I. These regressions also therefore do not control for baseline values. Note that average standardized effects do not include value of uniforms since this variable wasn't prespecified in the analysis plan. Value of uniforms is coded as zero if the HH doesn't receive the uniforms. Applying family-wise error rates (Romano and Wolf 2005) to the incentive additional effect (columns 4 and 7), none of the coefficients are rejected.

than in the nonincentive treatments—in fact, the point estimates suggest they were 1.0–2.7 percentage points (14–32 percent) more likely to receive a uniform with incentives and 1.0–1.7 percentage points (18–32 percent) more likely to receive other school supplies with incentives. Moreover, the self-reported monetary value of the uniform received is identical in both treatments. This suggests that the change in budgets away from uniforms and school supplies documented in Table 6 likely came from increased efficiency in procuring the uniforms rather than a reduction in quality or quantity. In fact, the average standardized effect suggests more direct benefits

for education were received in incentivized areas, not less. Thus, on net more children received education subsidies, even though more money was spent on health. Combined with the improvements in health outcomes and the fact that education did not suffer, the evidence suggests that the incentives improved the efficiency of the block grant funds.

B. *Effort*

A second dimension we examine is effort—both on the part of workers and on the part of communities. Table 8 begins by examining labor supplied by midwives, who are the primary health workers at the village level; teachers; and subdistrict level health center workers. The main impact is an increase in hours worked by midwives, particularly in Wave II, where midwives spent 3.2 hours (12 percent) more working over the 3 days prior to the survey in incentive areas compared to in nonincentive areas. This effect is statistically significant even when we apply the family-wise error rates among all health indicators. Since midwives are primary main providers of maternal and child health services, the increase in midwife hours is consistent with the increase in these services we observed above. Likewise, in Wave III, there was no statistically significant difference in midwife hours worked between incentivized and nonincentivized treatments, as hours also appear to increase in the nonincentivized groups, consistent with the improvements in weight checks observed in the household survey in the nonincentivized group in Wave III. Teacher attendance showed no clear pattern.

Virtually all of the midwives in our area have a mix of both public and private practice, but they vary in whether their government practice is as a full-fledged, tenured civil servant (*PNS*) or is instead on a temporary or contract basis. When we interact the variables in Table 8 with a dummy for whether the midwife is a tenured civil servant, we find that the incentive treatment led to a greater increase in private-practice hours provided by tenured civil servant midwives (see online Appendix Table 8), with no change in their public hours. This suggests that the fee-for-service component of midwives' practices may have been a reason why they increased their service provision. Interestingly, the monetary compensation (e.g., value of subsidies per patient) provided to midwives did not differ between the incentivized and nonincentivized treatments (results not reported in table), so it was not the financial incentives per patient seen that resulted in the difference. More likely, it was the combination of other efforts to increase demand (e.g., effort from the community to bring people to health posts), combined with the fact that midwives were indeed paid for additional services they provided, that resulted in the midwives' increase in effort.

Table 9 examines the effort of communities. We examine three types of community effort: holding more posyandus, the monthly village health meetings where most maternal and child health care is provided; community effort at outreach, such as door-to-door "sweepings" to get more kids into the posyandu and school committee meetings with parents, and community effort at monitoring, such as school committee membership and teacher meetings. We find no evidence that the incentives had an impact on these margins, although the program as a whole increased community participation at monthly community health outreach activities (posyandu).

Table 8—Worker Behavior

Indicator	Wave II				Wave III		
	Baseline mean	Incentive treatment effect	Non-incentive treatment effect	Incentive additional effect	Incentive treatment effect	Non-incentive treatment effect	Incentive additional effect
Panel A. Health							
Midwives							
Hours spent in outreach over past 3 days	3.165 [4.488]	0.796* (0.410)	−0.074 (0.337)	0.870** (0.425)	0.073 (0.389)	0.036 (0.419)	0.038 (0.400)
Hours spent providing public services over past 3 days	13.548 [10.056]	0.534 (0.608)	−1.104* (0.594)	1.638** (0.721)	0.672 (0.618)	0.414 (0.566)	0.258 (0.586)
Hours spent providing private services over past 3 days	10.805 [12.505]	0.211 (0.832)	−0.470 (0.826)	0.681 (0.886)	0.892 (0.674)	0.588 (0.669)	0.304 (0.644)
Total hours spent working over past 3 days	27.518 [15.713]	1.474 (1.046)	−1.722* (1.039)	3.195*** (1.154)	1.621* (0.950)	0.930 (0.931)	0.692 (0.884)
Number of posyandus attended in past month	4.166 [3.321]	0.202 (0.334)	0.071 (0.225)	0.131 (0.348)	−0.155 (0.248)	0.060 (0.267)	−0.215 (0.324)
Number of hours midwife per posyandu	3.039 [1.693]	0.137 (0.130)	0.180 (0.120)	−0.044 (0.127)	0.109 (0.152)	−0.083 (0.133)	0.192 (0.153)
Health centers							
Minutes wait at recent health visits	25.201 [23.736]	0.435 (3.695)	5.693 (4.690)	−5.258 (3.935)	3.361 (4.345)	2.234 (4.342)	1.127 (4.336)
Percent of providers present at time of observation	. [.]	0.071** (0.036)	0.109*** (0.039)	−0.038 (0.035)	−0.009 (0.029)	−0.076** (0.030)	0.067** (0.030)
Average standardized effect health		0.107** (0.043)	0.057 (0.044)	0.050 (0.047)	0.055 (0.040)	−0.012 (0.039)	0.066 (0.041)
Panel B. Education–Teachers							
Percent present at time of interview (primary)	. [.]	0.013 (0.014)	−0.009 (0.013)	0.021 (0.015)	0.000 (0.012)	0.022** (0.011)	−0.023** (0.011)
Percent present at time of interview (junior secondary)	. [.]	−0.002 (0.027)	0.020 (0.024)	−0.022 (0.026)	0.005 (0.020)	−0.026 (0.020)	0.031 (0.022)
Percent observed teaching (primary)	. [.]	−0.006 (0.038)	−0.050 (0.042)	0.044 (0.042)	−0.003 (0.040)	−0.012 (0.041)	0.009 (0.038)
Percent observed teaching (junior secondary)	. [.]	−0.069 (0.044)	−0.052 (0.047)	−0.018 (0.049)	0.039 (0.049)	0.024 (0.048)	0.015 (0.044)
Average standardized effect education		−0.022 (0.043)	−0.046 (0.044)	0.024 (0.047)	0.023 (0.041)	0.009 (0.042)	0.014 (0.042)
Panel C. Overall							
Average standardized effect overall		0.064** (0.031)	0.023 (0.032)	0.041 (0.034)	0.044 (0.030)	−0.005 (0.028)	0.049 (0.031)

Notes: Data is from the survey of midwives (top panel); direct observation of schools (middle panel), household survey (bottom panel; wait times), and direct observation of health centers (bottom panel, provider presence). See also notes to Table 3. Applying family-wise error rates (Romano and Wolf 2005) to the incentive additional effect (columns 4 and 7), the only coefficient where the null is rejected, taking into account multiple comparisons, is total hours spent working over the past three days in Wave II, where health is the family.

C. *Targeting*

A third mechanism through which incentives could matter is by encouraging communities to target resources to those individuals who are the most elastic—i.e., those individuals for whom a given dollar is most likely to influence behavior. While we cannot estimate each household's elasticity directly, we can examine whether

TABLE 9—COMMUNITY EFFORT

Indicator	Baseline mean	Wave II Incentive treatment effect	Wave II Non-incentive treatment effect	Wave II Incentive additional effect	Wave III Incentive treatment effect	Wave III Non-incentive treatment effect	Wave III Incentive additional effect
Community effort at direct service provision							
Number of posyandus in village	4.519 [3.504]	−0.092 (0.124)	0.004 (0.147)	−0.096 (0.126)	0.128 (0.178)	0.196 (0.176)	−0.068 (0.148)
Number of posyandu meetings in past year at selected posyandu	. [.]	−0.003 (0.102)	0.082 (0.111)	−0.084 (0.102)	−0.113 (0.112)	−0.061 (0.091)	−0.052 (0.100)
Number of cadres at posyandu	. [.]	0.174 (0.113)	0.197 (0.153)	−0.023 (0.138)	0.294** (0.139)	0.358** (0.171)	−0.064 (0.165)
Community effort at outreach							
Number of sweepings at selected posyandu in last year	. [.]	−0.296 (0.394)	0.042 (0.377)	−0.338 (0.389)	−0.140 (0.341)	−0.628* (0.344)	0.488* (0.294)
Number of primary school comm. meetings with parents in past year	. [.]	0.066 (0.133)	−0.070 (0.133)	0.136 (0.121)	0.002 (0.181)	−0.125 (0.182)	0.126 (0.137)
Number of junior sec. school committee meetings w/parents	2.309 [1.973]	−0.121 (0.112)	0.032 (0.118)	−0.153 (0.126)	0.214 (0.147)	0.209 (0.222)	0.005 (0.206)
Community effort at monitoring							
Number of primary school committee members	. [.]	0.761* (0.392)	−0.503 (0.410)	1.264*** (0.478)	−0.003 (0.334)	0.195 (0.402)	−0.198 (0.344)
Number of junior sec. school committee members	8.259 [4.763]	−0.844 (0.992)	−1.421 (0.933)	0.577 (0.539)	0.199 (0.331)	0.216 (0.332)	−0.017 (0.291)
Number of prim. school committee meetings with teachers in past year	. [.]	−0.124 (0.358)	−0.367 (0.357)	0.243 (0.354)	−0.121 (0.316)	−0.096 (0.319)	−0.025 (0.268)
Number of j. sec. school committee meetings with teachers in year	4.476 [5.465]	0.471 (0.424)	0.125 (0.394)	0.346 (0.456)	0.532 (0.342)	0.567 (0.346)	−0.035 (0.365)
Average standardized effect		0.013 (0.022)	−0.009 (0.025)	0.023 (0.023)	0.043* (0.025)	0.047 (0.031)	−0.004 (0.029)

Notes: Data is from survey of the head of the posyandu and the head of schools. See also notes to Table 3. Applying family-wise error rates (Romano and Wolf 2005) to the incentive additional effect, no coefficients are individually rejected.

incentivized communities targeted differently based on per capita consumption. The idea is that poorer households' behavior may be more elastic with respect to subsidies than that of richer households, who can afford the targeted services with or without subsidies. Incentives could therefore encourage communities to target benefits to poorer households and resist pressure to distribute benefits more evenly.[12]

The results in Table 10 show how the incentives affect the targeting of direct benefits from the grants. For each specification, we re-estimate equations (1) and (2) with subdistrict fixed effects and interact the Generasi variables with a dummy for

[12] Of course, this prediction is theoretically ambiguous—one might also imagine that very poor households cannot afford services with very large subsidies, so incentives would encourage targeting of middle-income households that are closest to the margin.

Table 10—Within-Subdistrict Targeting

Indicator	Baseline mean	Wave II: Generasi incentive top 3 quintiles additional effect	Wave II: Generasi nonincentive top 3 quintiles additional effect	Wave II: Generasi incentive additional effect top 3 quintiles additional effect	Wave III: Generasi incentive top 3 quintiles additional effect	Wave III: Generasi nonincentive top 3 quintiles additional effect	Wave III: Generasi incentive additional effect top 3 quintiles additional effect
Panel A. Targeting of direct benefits							
Average standardized effect health		−0.073	0.093	−0.165	−0.124	−0.109	−0.014
		(0.169)	(0.117)	(0.202)	(0.126)	(0.102)	(0.147)
Average standardized effect education		−0.058	−0.067	0.009	−0.170**	−0.085	−0.085
		(0.147)	(0.163)	(0.210)	(0.078)	(0.073)	(0.096)
Average standardized effect overall		−0.066	0.022	−0.088	−0.147*	−0.097	−0.050
		(0.112)	(0.094)	(0.143)	(0.087)	(0.059)	(0.096)
Panel B.							
Average standardized effect health		−0.072	0.047	−0.119	0.063	0.000	0.063
		(0.064)	(0.067)	(0.077)	(0.069)	(0.063)	(0.065)
Average standardized effect education		−0.044	−0.073	0.029	−0.076	0.057	−0.133*
		(0.087)	(0.104)	(0.120)	(0.073)	(0.077)	(0.071)
Average standardized effect overall		−0.062	0.007	−0.070	0.017	0.019	−0.002
		(0.056)	(0.060)	(0.070)	(0.057)	(0.050)	(0.050)

Notes: Data is from the household survey. For each indicator in Table 3, the regression interacts the Generasi treatment variables for a dummy for a household being in the top three quintiles of the baseline per-capita consumption distribution. Average standardized effects for the interaction with the top three quintiles variable are shown in the table. Panel A examines the indicators of direct benefits shown in Table 7 and panel B examines the 12 main program indicators examined in Table 3.

the household being in the top three quintiles of the income distribution at baseline. The subdistrict fixed effects mean that this is controlling for the overall level of the outcome variable in the subdistrict, and thus picks up changes in the targeting of the outcomes among the rich and poor only.

Table 10 shows the results. We first present the difference between the top three quintiles and the bottom two quintiles for incentivized areas. A negative coefficient indicates that the poor received relatively more than the rich in treatment areas relative to controls. The second column presents the difference between the top three quintiles and the bottom two quintiles for nonincentivized treatment areas. The third column presents the difference between the first two columns. A negative coefficient indicates that the incentivized version of the program had more pro-poor targeting than the nonincentivized version. Panel A shows the average standardized effects for targeting of direct benefits (i.e., the subsidies and transfers examined in Table 7), and panel B shows the average standardized effects for targeting of improvements in actual outcomes (i.e., the main indicators examined in Table 3). Detailed indicator-by-indicator results are shown in online Appendix Tables 10 and 11. The results in panel A suggest there is somewhat more targeting of direct benefits to the poor in the incentivized version of the program, but the difference between the incentivized versions and nonincentivized versions is not statistically significant overall. Likewise in panel B there is mild suggestive evidence that incentives improve targeting of improvements in outcomes, but this is generally not statistically significant.

In sum, the results point to two main channels through which incentives mattered. Incentives led to a more efficient allocation of block grants, reducing expenditure on uniforms and other school supplies while not affecting a household's receipt of these items, and using the savings to increase expenditures on health. And, incentives led to an increase in midwife hours worked, particularly from tenured, civil servant midwives working in their private capacity.[13] The fact that the budget impacts persist over time, whereas the timing of the effort impacts more directly match the timing of the impact on indicators shown in Table 3, suggests that the effort impacts may be the more important channel.

IV. Potential Pitfalls of Incentives

In this section, we test for three types of negative consequences from the incentives: multi-tasking problems (Holmstrom and Milgrom 1991), where performance incentives encourage substitution away from nonincentivized outcomes; manipulation of performance records; and reallocation of funds toward wealthier areas.

A. *Spillovers on Nontargeted Indicators*

Whether the incentives would increase or decrease performance on nontargeted indicators depends on the nature of the health and education production functions. For example, if there is a large fixed cost for a midwife to show up in a village, but a small marginal cost of seeing additional patients once she is there, one might expect that other midwife-provided health services would increase. Alternatively, if the major cost is her time, she may substitute toward the types of service incentivized in the performance bonuses and away from things outside the incentive scheme, such as family planning, or might spend less time with each patient.

We test for spillover effects on three health domains: utilization of nonincentivized health services (e.g., adult health, prenatal visits beyond the number of visits that qualify for incentives), quality of health service provided by midwives (as measured by the share of the total required services they provide in a typical meeting), and maternal knowledge and practices. We also examine potential impacts on family composition decisions. On the education side, we examine the impact on high school enrollment, hours spent in school, enrollment in informal education, distance to school, and child labor.

Table 11 reports average standardized effects for each of these domains; the detailed indicator-by-indicator results can be found in online Appendix Table 5. In general, we find no differential negative spillover impacts of the incentives on any of these indicators, and, if anything, find some slight evidence of positive spillovers. For example, we find that the incentives led to positive effects on reductions in child labor (0.12 hours per child for age 7–15 in Wave II; this translates to 0.08 standard deviations across all child labor meaures). With regard to the neonatal mortality

[13] A final area we examined was prices for health services and school fees. While we found that the Generasi program did lead to increases in prices for some health services, we did not find any differential impact on prices between the incentivized and nonincentivized treatments. See Olken, Onishi, and Wong (2011) for more information.

Table 11—Spillovers on Nontargeted Indicators, Average Standardized Effects by Indicator Family

	Wave II			Wave III		
Family of indicators	Incentive treatment effect	Non-incentive treatment effect	Incentive additional effect	Incentive treatment effect	Non-incentive treatment effect	Incentive additional effect
Panel A. Health						
Utilization of nonincentivized health services	0.019	−0.009	0.029	0.038*	0.017	0.021
	(0.020)	(0.021)	(0.022)	(0.020)	(0.020)	(0.019)
Health services quality	0.079**	0.064*	0.015	0.041	0.040	0.001
	(0.038)	(0.039)	(0.040)	(0.036)	(0.038)	(0.036)
Maternal knowledge and practices	0.026	0.025	0.002	0.033	0.043	−0.011
	(0.029)	(0.028)	(0.030)	(0.029)	(0.027)	(0.026)
Family composition decisions	0.014	−0.012	0.026	0.023	−0.007	0.029
	(0.019)	(0.021)	(0.022)	(0.022)	(0.026)	(0.023)
Average standardized effect health	0.035**	0.017	0.018	0.034**	0.023	0.010
	(0.016)	(0.016)	(0.017)	(0.016)	(0.016)	(0.014)
Panel B. Education						
Other enrollment metrics	−0.071	−0.051	−0.019	−0.013	0.006	−0.019
	(0.049)	(0.046)	(0.049)	(0.021)	(0.020)	(0.018)
Transportation to school (cost and distance)	−0.077	−0.034	−0.043	0.004	0.022	−0.018
	(0.058)	(0.050)	(0.060)	(0.042)	(0.041)	(0.042)
Avoiding child labor (higher #s = less child labor)	−0.025	−0.107***	0.083**	0.012	0.007	0.005
	(0.022)	(0.038)	(0.034)	(0.025)	(0.020)	(0.022)
Average standardized effect education	−0.057**	−0.064**	0.007	0.001	0.012	−0.011
	(0.029)	(0.030)	(0.032)	(0.018)	(0.017)	(0.017)
Panel C. Overall						
Average overall standardized effect	−0.005	−0.018	0.013	0.020	0.018	0.001
	(0.015)	(0.017)	(0.019)	(0.012)	(0.012)	(0.011)

Notes: See notes to Table 3. Data is from the household survey. Each row presents average standardized effects from a family of indicators, with the detailed indicator-by-indicator results shown in online Appendix Table 5. The individual indicators consist of the following: Health utilization consists of deliveries based in facilities (as opposed to at home), use of family planning, use of curative health services, prenatal visits beyond four per pregnancy, vitamin A drops beyond two per child. Health services quality consists of quality of prenatal care services and quality of posyandu services, where quality is measured as the share of services that are supposed to be provided that are actually provided during a typical visit. Maternal knowledge and practices are the fraction initiating breastfeeding within the first hour after birth, share with exclusive breastfeeding, maternal knowledge about proper treatment of several child health conditions, and questions about a woman's role in decisions about children. Family composition is the fertility rate and out migration. Other enrollment metrics are gross high school enrollment, dropout rates, primary to junior secondary transition rates, number of hours children attend school, and the numbers attending primary, junior secondary, and senior secondary informal education (Paket A, B, and C). Transportation to school is the distance to junior secondary school, time spent traveling one-way to junior secondary school, and transportation cost each way to school. Child labor is the fraction age 7–15 who works for a wage, hours spent working for a wage, a dummy for doing any wage work, and a dummy for doing any household work.

result, we find no evidence that the quality of health services (defined as the share of activities midwives were supposed to do during various types of visits that were actually performed) declined in the incentivized relative to the incentivized treatment; in fact, it appeared to improve equally in both the incentivized and nonincentivized treatments relative to control. The results here suggest that, with the possible important exception of neonatal mortality discussed above, negative spillovers on nontargeted indicators do not seem to be a substantial concern with the incentives in this context.

B. *Manipulation of Performance Records*

A second potential downside of performance incentives is that communities or providers may manipulate records to inflate scores. For example, Linden and Shastry (2012) show that teachers in India inflate student attendance records to allow them to receive subsidized grain. Manipulation of record keeping can have substantial efficiency costs: for example, children could fail to get immunized properly if their immunization records were falsified.

For immunizations and school attendance, we can check for this by comparing the official immunization records to an independent measure observed directly by our survey team. For immunization, we compare official records to the scar left by the BCG vaccine on the arm where it was administered (see Banerjee et al. 2008), and for attendance, we compare official records to random spot-checks of classrooms. We can check for general manipulation of the administrative data used to calculate the incentives by checking whether the administrative data is systematically higher or lower than the corresponding estimates from the survey data.

The results are shown in Table 12. Panel A explores the differences between BCG scars and record keeping.[14] We defined a false "yes" if the child is recorded/declared as having had the vaccine but has no scar, and likewise for a false "no." We find some differences in false reports of the BCG scar based on the performance incentives in Wave II, though only when we compare the scar to the official immunizations in the immunization record book. It is also worth noting that the number of children without record cards also decreased in Generasi areas, which makes this comparison hard to conclusively interpret as manipulation as opposed to being a consequence of a change in record keeping.

Panel B explores differences in attendance rates, and finds that the discrepancy is unchanged by the performance incentives. In fact, recorded attendance appears lower in the incentive treatment while actual attendance is unchanged, which suggests perhaps that the incentives led to better record keeping. Panel C examines the difference between administrative data on performance and the corresponding values from the household survey.[15] Average standardized effects across all 12 indicators are presented in panel C of Table 12; the indicator-by-indicator results are available online Appendix Table 9. The results show that, for Wave II, the difference between the administrative data and household survey is lower in the incentive than nonincentivized villages, which is the opposite of what one would expect if the incentives led villages to systematically inflate scores in the incentivized areas.

Combined, these two pieces of evidence suggest that manipulation of record-keeping is not a major problem of the performance incentives in this context; in fact,

[14] Note that if the child did not have a record card, we asked the mother if the child was immunized. The "declared" vaccinated variable is 1 if either the record book or the mother reports that the child was vaccinated.

[15] For each indicator, the administrative data contains the total number of achievements per year. We divide by the number of people eligible to achieve the indicator (e.g., number of children age 13–15) to determine the average rate of achievement, which is comparable to what we observe in the household survey. Since there is no administrative data for control groups, the results show only the differences between the incentivized and nonincentivized groups.

Table 12—Manipulation of Performance Records

Indicator	Baseline mean	Wave II: Incentive treatment effect (1)	Wave II: Non-incentive treatment effect (2)	Wave II: Incentive additional effect (3)	Wave III: Incentive treatment effect (4)	Wave III: Non-incentive treatment effect (5)	Wave III: Incentive additional effect (6)
Panel A. BCG scar							
False "yes" in recorded BCG vaccine	0.079 [0.270]	0.032** (0.015)	0.006 (0.014)	0.026* (0.015)	0.004 (0.013)	0.003 (0.014)	0.001 (0.014)
False "yes" in declared BCG vaccine	0.111 [0.314]	0.033** (0.015)	0.021 (0.015)	0.012 (0.016)	0.013 (0.013)	0.000 (0.014)	0.013 (0.013)
Children with no record card	0.246 [0.431]	−0.054*** (0.019)	−0.038** (0.019)	−0.016 (0.018)	−0.023 (0.019)	−0.053*** (0.018)	0.030* (0.017)
Panel B. Attendance							
Attend. rate—difference between recorded and observed	8.178 [26.000]	−1.925 (1.696)	−2.593* (1.506)	0.668 (1.736)	0.740 (2.021)	2.360 (2.125)	−1.620 (1.910)
Attend. rate observed	87.496 [25.577]	1.350 (1.632)	2.890* (1.469)	−1.540 (1.669)	−0.970 (1.874)	−2.157 (2.017)	1.187 (1.839)
Attend. rate recorded	95.795 [7.438]	−0.609* (0.356)	0.186 (0.367)	−0.794* (0.434)	−0.201 (0.441)	0.142 (0.423)	−0.343 (0.437)
Panel C. Difference between admin. and household data							
Average standardized effect health				−0.074 (0.047)			−0.058 (0.067)
Average standardized effect education				−0.137*** (0.052)			−0.115 (0.097)
Average standardized effect				−0.097** (0.044)			−0.079 (0.071)

Notes: See notes to Table 3. Data from panel A comes from the household survey. False "yes" is defined as 1 if the child has no observed BCG scar on his/her arm but the records say that the child received the BCG immunization. For panel B, the observed attendance is the percent of students attending on the day of the survey, and the recorded attendance rate is the attendance in the record book on a fixed day prior to the survey taking place. For panel C, the dependent variable is the difference between what is recorded in MIS data for each of the 12 indicators and the corresponding number from the household survey, with average standardized effects shown in the table. A positive coefficient would indicate inflation of the program statistics (i.e., MIS is systematically higher than household). Note that since MIS data is available only for Generasi areas, panel C only compares the incentivized with non-incentivized areas. Applying family-wise error rates (Romano and Wolf 2005) to the incentive additional effect (columns 4 and 7), no individual coefficients are rejected.

if anything, the fact that records were being used for funding decisions in incentivized areas seems to have led to more accurate record keeping, not less.

C. *Allocation of Bonus Money to Wealthier Areas*

A third potential pitfall of incentive schemes in an aid context is that they can result in a transfer of funds toward areas that need aid less. Poorer or more remote areas, for example, might have lower performance levels, yet might actually have the highest marginal return from funds. The incentives attempted to mitigate this by creating relative incentives, with a fixed performance bonus pool for each subdistrict. The idea was that unobserved, subdistrict-specific common shocks would cancel out. Nevertheless, if most of the differences in productivity were within subdistricts, not between subdistricts, the same problem could still occur.

Table 13—Do Relative Payments Prevent Money from Flowing to Richer Areas?

	Wave II				Wave III			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Actual incentive payments								
Avg. pc exp.	−1.325			−1.749	13.48			15.09
	(7.078)			(6.769)	(12.28)			(12.45)
Distance to district		79,237**		82,873**		83,353**		78,305**
		(33,578)		(34,038)		(39,741)		(36,635)
Village poverty rate			976,885	1,806,000			−2,413,000	−766,739
			(2,980,000)	(2,752,000)			(6,102,000)	(5,976,000)
Observations	453	453	441	441	388	388	377	377
Panel B. Counter-factual incentive payments without relative performance within subdistricts								
Avg. pc exp.	4.330			4.405	−2.190			−1.052
	(3.172)			(2.826)	(5.832)			(5.758)
Distance to district		9,335		9,249		3,932		3,076
		(9,646)		(10,100)		(20,136)		(20,298)
Village poverty rate			−6,301,000***	−6,060,000***			−694,408	−702,532
			(1,945,000)	(1,952,000)			(4,014,000)	(4,025,000)
Observations	453	453	441	441	388	388	377	377

Notes: Data is from program administrative records. Dependent variable is the amount of bonus money given to a village, in Rupiah. Each column reports the result from a separate regression. Each observation is a village. The sample is the eight sampled villages within each of the incentivized subdistricts. Note that MIS data on total points is incomplete for Wave III (second year of program). Standard errors adjusted for clustering by subdistrict.

To investigate this, in Table 13, panel A, we regress the total amount of bonus funds each village received on village average per capita consumption, village remoteness (km from the district capital), and village poverty (share of households classified as poor by the national family planning board). In panel B, we repeat the same regressions for a counterfactual calculation for incentives without the relative performance component, where we hypothetically allocate bonus payments proportionally to bonus points relative to all villages in the program, rather than relative only to other villages in the same subdistrict.

The results show that, in the actual allocation shown in panel A, villages that were more remote (further from the district capital) received more bonus funds. The allocation of bonus funds was unrelated to average village consumption or to village poverty levels. By contrast, in the counterfactual calculation shown in panel B where incentives were based just on points earned rather than points earned relative to other villages in the same subdistrict, poor villages received substantially less, and more remote villages no longer received more. The calculation thus shows that the relative performance scheme was successful in preventing funds from migrating from poorer villages to richer villages. The counterfactual shows that had the program not awarded incentives relative to other villages in the same subdistrict, richer villages would have ended up receiving more bonus funds.

V. Conclusion

We found that adding a relative performance-based incentive to a community-based health and education program accelerated performance improvements in preventative health and malnutrition, particularly in areas with

the lowest levels of performance before the program began. We found that while the block grant program overall improved enrollments after 30 months, the incentives had no differential impact on education. Incentives worked through increasing the efficiency with which funds were spent and through increasing health providers' hours worked, particularly initially. There was no evidence of manipulation of records, and no evidence that performance incentives led to funds systematically flowing to richer or otherwise more advantaged areas. The main potential concern with the incentives was that the decline in neonatal mortality in the nonincentivized group was not observed in the incentivized areas. Though this finding is difficult to conclusively interpret, it is important that in implementing incentivized schemes care be taken to avoid multitasking problems.

It is difficult to interpret the magnitudes given above without some notion of costs. Conditional on implementing the program, adding the performance incentives added very few additional costs—the same monitoring of indicators was done in both the incentivized and nonincentivized versions of the program, no additional personnel were required to do monitoring (the program would have needed facilitators regardless, and the additional amount of time spent on calculating performance bonuses was small), and since the performance bonuses were relative within a subdistrict and the amount of money was fixed, there was no difference in the total size of block grants in incentivized and nonincentivized areas. In this case, the incentives thus accelerated outcomes, while adding few monetary costs to the program.[16] The degree to which this applies to other contexts depends, of course, on the degree to which there are additional real costs associated with collecting outcome data for monitoring.

The results have several implications for design of performance-based aid schemes. First, the fact that an important channel through which incentives appeared to work was the reallocation of budgets suggests that one may not want to make the incentives too narrow—instead, to the extent the multitasking issue can be controlled, it may be better to give broad incentives and let the recipients have sufficient power to shuffle resources to achieve them. Second, the results suggest that while performance-based aid can be effective, care must be taken to ensure that it does not result in aid money flowing to richer areas, where it may have less benefit. Indeed, we show that in this case, the fact that performance incentives were relative to a small set of close geographical neighbors meant that performance bonus money did not accrue to richer areas, but it would have in the absence of this relative competition. Incorporating these types of features into performance-based aid schemes may help obtain the promise of incentives while mitigating many of their risks.

[16] A more formal cost-effectiveness calculation can be found in online Appendix V.

REFERENCES

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review* 102 (4): 1206–40.

Almond, Douglas, and Bhashkar Mazumder. 2011. "Health Capital and the Prenatal Environment: The Effect of Ramadan Observance during Pregnancy." *American Economic Journal: Applied Economics* 3 (4): 56–85.

Baicker, Katherine, Jeffrey Clemens, and Monica Singhal. 2012. "The Rise of the States: U.S. Fiscal Decentralization in the Postwar Period." *Journal of Public Economics* 96 (11–12): 1079–91.

Baird, Sarah, Craig McIntosh, and Berk Özler. 2011. "Cash or Condition? Evidence from a Cash Transfer Experiment." *Quarterly Journal of Economics* 126 (4): 1709–53.

Banerjee, Abhijit Vinayak, Esther Duflo, Rachel Glennerster, and Dhruva Kothari. 2008. "Improving Immunization Coverage in Rural India: A Clustered Randomized Controlled Evaluation of Immunization Campaigns with and without Incentives." Unpublished.

Basinga, Paulin, Paul J. Gertler, Agnes Binagwaho, Agnes L. B. Soucat, Jennifer Sturdy, and Christel M. J. Vermeersch. 2011. "Effect on Maternal and Child Health Services in Rwanda of Payment to Primary Health-Care Providers for Performance: An Impact Evaluation." *Lancet* 377 (9775): 1421–28.

Birdsall, Nancy, and William D. Savedoff. 2009. *Cash on Delivery: A New Approach to Foreign Aid*. Washington, DC: Center for Global Development.

Bozzoli, Carlos, Angus Deaton, and Climent Quintana-Domeque. 2009. "Adult Height and Childhood Disease." *Demography* 46 (4): 647–69.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *Quarterly Journal of Economics* 127 (4): 1755–1812.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. 2013. "School Inputs, Household Substitution, and Test Scores." *American Economic Journal: Applied Economics* 5 (2): 29–57.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

Express India News Service. 2008. "Biometric Attendance to Keep Track of Students, Teachers in Primary Schools." http://expressindia.indianexpress.com/story_print.php?storyId=340201.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, and Katherine Baicker. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127 (3): 1057–1106.

Gertler, Paul. 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment." *American Economic Review* 94 (2): 336–41.

Gibbons, Robert, and Kevin J. Murphy. 1990. "Relative Performance Evaluation for Chief Executive Officers." *Industrial and Labor Relations Review* 43 (3): 30–51.

Gørgens, Tue, Xin Meng, and Rhema Vaithianathan. 2012. "Stunting and Selection Effects of Famine: A Case Study of the Great Chinese Famine." *Journal of Development Economics* 97 (1): 99–111.

Holmstrom, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10 (1): 74–91.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics and Organization* 7 (Special Issue): 24–52.

Huang, Cheng, Michael R. Phillips, Yali Zhang, Jingxuan Zhang, Qichang Shi, Zhiqiang Song, Zhijie Ding, Shutao Pang, and Reynaldo Martorell. 2013. "Malnutrition in Early Life and Adult Mental Health: Evidence from a Natural Experiment." *Social Science & Medicine* 97: 259–66.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119.

Lazear, Edward P., and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89 (5): 841–64.

Levy, Santiago. 2006. *Progress Against Poverty: Sustaining Mexico's Progresa-Oportunidades Program*. Washington, DC: Brookings Institution Press.

Linden, Leigh L., and Gauri Kartini Shastry. 2012. "Grain Inflation: Identifying Agent Discretion in Response to a Conditional School Nutrition Program." *Journal of Development Economics* 99 (1): 128–38.

Mookherjee, Dilip. 1984. "Optimal Incentive Schemes with Many Agents." *Review of Economic Studies* 51 (3): 433–46.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1): 39–77.

Musgrave, Richard A. 1997. "Devolution, Grants, and Fiscal Competition." *Journal of Economic Perspectives* 11 (4): 65–72.

Oates, Wallace E. 1999. "An Essay on Fiscal Federalism." *Journal of Economic Literature* 37 (3): 1120–49.

Olken, Benjamin A., Junko Onishi, and Susan Wong. 2011. *Indonesia's PNPM* Generasi *Program: Final Impact Evaluation Report*. World Bank. Jakarta, March.

Olken, Benjamin A., Junko Onishi, and Susan Wong. 2014. "Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia: Dataset." *American Economic Journal: Applied Economics*. http://dx.doi.org/10.1257/app.6.4.1.

Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–82.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1): 199–250.

Valente, Christine. 2013. "Civil Conflict, Gender-Specific Fetal Loss, and Selection: A New Test of the Trivers-Willard Hypothesis." Unpublished.