Should Aid Reward Performance?

Evidence from a field experiment on health and education in Indonesia

Benjamin A. Olken, MIT
Junko Onishi, World Bank
Susan Wong, World Bank

February 2012

ABSTRACT

This paper reports an experiment in over 3,000 Indonesian villages designed to test the role of performance incentives in improving the efficacy of aid programs. Villages in a randomly-chosen one-third of subdistricts received a block grant to improve 12 maternal and child health and education indicators, with the size of the subsequent year's block grant depending on performance relative to other villages in the subdistrict. Villages in remaining subdistricts were randomly assigned to either an otherwise identical block grant program with no financial link to performance, or to a pure control group. We find that the incentivized villages performed better on health than the non-incentivized villages, particularly in less developed areas, but found no differential impact of incentives on education. We find no evidence of negative spillovers from the incentives to untargeted outcomes, and no evidence that villagers manipulated scores. The relative performance design was crucial in ensuring that incentives did not result in a net transfer of funds toward richer areas. Incentives led to what appear to be more efficient spending of block grants, and led to an increase in labor from health providers, who are partially paid fee-for-service, but not teachers. On net, between 50-75% of the total impact of the block grant program on health indicators can be attributed to the performance incentives.

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, and Lina Marliani 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.

1. Introduction

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 U.S. Millennium Challenge Corporation is explicitly conditioned on recipient countries meeting 17 indicators of good governance, ranging from civil liberties to immunization rates to girl's primary education rates to inflation (Birdsall and Savedoff 2009). The World Bank is similarly moving towards "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 U.S. has used a similar approach to encourage state-based education and local school performance reform through its Race To The Top and No Child Left Behind programs.

As in any principal-agent framework, linking aid to performance can be useful from the principal's perspective to the extent it creates incentives for lower tiers of government to improve effort and mobilize additional resources. But there are potential pitfalls as well. As with all incentive schemes, there can be multitasking problems, where effort allocated towards targeted indicators comes at the expense of other, non-incentivized indicators that the principal may also care about (Holmstrom and Milgrom 1991). There can also be direct attempts to manipulate indicators to increase payouts (Linden and Shastry forthcoming). And – particularly when government budgets are being allocated based on performance – there is a substantial risk

that performance-based aid will mean that budgets are directed to richer or otherwise better performing locations. If these richer or better performing regions have a lower marginal value of funds, this reallocation could offset some of the incentive effects.

To investigate these issues, we designed a randomized field experiment in the context of an aid program in Indonesia that seeks to improve maternal and child health and education and that incorporates explicit performance incentives. Under the program, known as *Generasi*, villages receive an annual block grant of approximately US $10,000, which each village can allocate to any activity that supported one 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). To give communities incentives to focus on the most effective policies, 20% of the subsequent year's block grant is allocated among villages in a subdistrict based on their relative performance on each of the 12 targeted health and education indicators. To test the impact of the performance incentives, the experiment randomized entire subdistricts to either receive the program with incentives, or to receive an otherwise identical program without the performance incentives. Other than the performance incentives, 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, etc). The experimental design thus precisely identifies the impact of the performance incentives.

A total of 264 subdistricts, each consisting of approximately 12 villages, were randomized into either a pure control group or one of two versions of the program. Surveys were conducted at baseline, 18 months after the program started, and 30 months after the program started. With over 2,100 villages randomized to receive either the incentivized or non-

incentivized version of the *Generasi* program (plus over 1,000 villages in control subdistricts), and over 1.8 million target beneficiaries of the program in treatment areas, to the best of our knowledge this represents one of the largest randomized social experiments conducted in the world to date.

We begin by examining the impact of the incentives on the 12 main indicators the program was designed to address. Using data from the household survey, we find that the incentives led to greater performance on the health indicators, but not on the education indicators. Over the two years of the program, the 8 targeted maternal and child indicators (e.g., prenatal visits, delivery by trained midwives, childhood immunizations, growth monitoring) were an average of 0.03 standard deviations higher in incentivized areas than in non-incentivized areas. This was driven by prenatal visits, which were 5% higher in incentivized areas than non-incentivized areas, and immunization rates, which were 3% higher in incentivized areas than non-incentivized areas. While these differences are modest, the impact of the incentives was more pronounced in areas with low baseline levels of service delivery: the incentives improved the 8 targeted maternal and child health indicators by an average of 0.06 standard deviations for a subdistrict at the 10^{th} percentile at baseline, and by as much as 0.10 – 0.15 standard deviations in the poorer, off Java provinces. On net, between 50-75% of the program's impact on health indicators came from the incentivized areas: for example, on average the 8 target health indicators were 0.052 standard deviations higher than pure controls in incentivized areas, compared to only 0.020 standard deviations higher than controls in non-incentivized areas. While the Generasi program overall improved enrollments after two years of implementation, there were no differences between incentivized and non-incentivized areas on the 4 education indicators examined (primary and junior secondary enrollment and attendance).

We find no evidence of adverse effects of the incentives. We find no evidence of a multi-tasking problem across the wide variety of non-incentivized metrics we examine: incentive areas were comparable or better than non-incentivized areas in terms of quality of health care services, use of adult health care services, good childcare and parenting practices, high school enrollment, enrollment in alternative forms of education, and child labor. We find no evidence that immunization recordkeeping, school attendance records, or program scores were manipulated in performance zones relative to non-performance incentive zones. And, we find that by making the incentive payments relative to other villages in the same subdistrict, the program prevented the incentives from resulting in a net transfer of funds to richer villages.

We investigate four potential mechanisms through which the incentives may have had an impact: changes in the composition of spending, worker effort, community effort, and targeting of benefits within the community. We find two main channels through which the incentives may have had an impact. First, the incentives led to a reallocation of the block grants away from school supplies and uniforms (4 percentage points lower, or about 16 percent) and towards health expenditures (3 percentage points higher, or about 6.5 percent). Despite the reallocation of funds away from school supplies and uniforms, households were in fact no less likely to receive these items and did not report that they were of lower quality, and were in fact more likely to receive education scholarships in the performance areas. This suggests that the changes in budgets may reflect more efficient use of funds rather than cutting down on quantity or quality. Second, we find that the performance incentives led to an increase in the labor of midwives, who are the major providers of the incentivized maternal and child health services (1.7 hours more over a 3 days recall window, or about 6 percent). 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

of the services they provide (e.g., pre and post natal care, deliveries), whereas teachers are not. We find no changes in community effort or the targeting of benefits within villages.

Interpreting the magnitudes we find is hard without some notion of the program's costs. To examine this, we perform a cost-effectiveness calculation. A general problem in cost-effectiveness calculations with multiple outcomes is that one needs to weight the various outcomes. In our case, we use the program's weights for each of the 12 indicators, which presumably reflect some notion of the government's relative weights among the indicators, to calculate a cost per "bonus point" achieved. Overall, the Generasi program cost about $8 - $11 per "bonus point." Translating "bonus points" into outcomes suggests, for example, that the cost of an additional child weight check is $16 - $22, the cost of preventing one malnourished child was $384 - $528, and the cost of enrolling one more child in primary school was $200 - $275. We show that these costs are similar to a traditional conditional cash transfer program implemented in Indonesia at the same time and evaluated using comparable methodologies, though are substantially more expensive than several recent interventions in Kenya and India aimed at improving school enrollment by improving health or providing school inputs (Dhaliwal et al. 2011). When we isolate the cost-effectiveness of the incentives, however, we obtain numbers on the order of $0.60 per point. This is because the incentives cost almost nothing: the total amount of money was held fixed between the incentivized and non-incentivized versions of the program, the marginal cost of collecting performance data was very low since program facilitators would have been there anyway administering the block grant, and the incentives just changed how the block grant was apportioned among communities. Thus, even though the impact of the performance incentives on outcomes was relatively modest, the results suggest that performance incentives can be a cost effective way of improving aid effectiveness.

This study is part of a newly expanding literature that seeks to identify the impact of performance incentives on health and education in developing countries, holding the amount of resources given constant. In the context of conditional cash transfers programs, Baird, McIntosh and Ozler (2011) find that adding conditions to a household-based CCT program 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. 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 quite substantial flexibility in budgetary responses to the aid (and indeed, we find evidence that changes in budget allocations is an important channel through which incentives worked), and maps most closely to the types of performance-based aid to central or regional governments being considered at the more macro level.

The remainder of the paper is organized as follows. Section 2 discusses the design of the program, the experimental design, and the econometric approach. Section 3 presents the main results of the impact of the incentives on the 12 targeted indicators. Section 4 examines the potential adverse effects of incentives, and Section 5 examines the mechanisms through which the incentives may have acted. Section 6 discusses the cost-effectiveness calculation, and Section 7 concludes.

2. Program and experimental design

2.1. The Generasi Program

To the best of our knowledge, the *Generasi* program is the first health and education program worldwide that combines community block grants with explicit performance bonuses for communities. The *Generasi* program, known formally as *Program National Pemberdayaan Masyarakat – Generasi Sehat dan Cerdas* (National Community Empowerment Program – Healthy and Smart Generation) began in mid-2007 in 129 subdistricts in rural areas of five 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 *Generasi* 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 of Indonesia to be as similar as possible to the conditions for a conditional cash transfer being piloted at the same time as *Generasi* (but in different locations), and, as such, they are in the same spirit as the conditions used by CCTs in other countries, such as Mexico's *Progresa* (Gertler 2004; Schultz 2004; Levy 2006). These 12 indicators represent health and educational behaviors that are within the direct control of villagers—such as, the number of children who receive immunizations, prenatal and postnatal care, and the number of children enrolled and attending school—rather than long-term outcomes, such as test scores or infant mortality. They also correspond to the Indonesian government's standard of service for maternal and child health and to the Indonesian government's stated goal of universal primary and junior secondary education.

In *Generasi,* each year all participating villages receive a block grant to improve maternal health, child health, and education. Block grants are usable for any purpose that the village can claim might help address one 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 for health care or school attendance, improving health or school buildings, or even building a road or path through the forest to improve access to health and educational facilities. The block grants averaged US$8,500 in the first 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 Generasi 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. Through social mapping and in-depth discussion groups, villagers identify problems and bottlenecks in reaching the 12 indicators. Inter-village meetings and consultation workshops with local health and education service providers allow community leaders to obtain information, technical assistance, and support from the local health and education offices as well as to coordinate the use of *Generasi* funds for multi-village projects. Following these discussions, the 11-member management team makes the final *Generasi* budget allocation.

2.2. Performance Incentives

In *Generasi,* the size of a village's block grant depends in part on its performance on the 12 targeted indicators in the previous year. The purpose of the performance bonus is to increase the village's effort at achieving the targeted indicators (Holmstrom 1979), both by encouraging a

more effective allocation of *Generasi* 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 villages in the subdistrict, 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 towards richer areas We discuss the impact of the relative bonus scheme on allocations, and compare it to a counter-factual with absolute performance bonuses, in Section 4.3 below.

The rule for allocating *Generasi* funds is as follows. The size of the overall *Generasi* allocation for the entire subdistrict is fixed by the subdistrict's population and province. Within a subdistrict, in year 1 of the program 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 2 of the program, 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, to be divided among villages based on their performance on the 12 *Generasi* indicators. The performance bonus pool is allocated to villages in proportion to a weighted sum of each village's

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

As discussed above, two versions of the *Generasi* program were implemented to separate the impact of the performance incentives *per se* from the overall impact of having additional financial resources available for health and education: the program with performance bonuses described above (referred to as "incentivized"), and an identical program without performance bonuses (referred to as "non-incentivized"). The non-incentivized version is identical to the incentivized version except that in the non-incentivized 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. 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

[1] 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 forthcoming). While more objective measures of monitoring attendance were considered, such as taking daily photos of students (as in Duflo, Hanna, and Ryan forthcoming) or installing fingerprint readers in all schools (Express India News Service 2008), *Generasi* 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 4.2 below.

versions, the same socialization 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 non-incentivized areas; the only difference is that in non-incentivized villages the points are used simply as an end-of-year monitoring and evaluation tool, and have no relationship to the allocation of funds. Within a given subdistrict, all villages participate in the same version of the program (i.e., either all villages received incentivized *Generasi* or all villages received non-incentivized *Generasi*).

2.3. Experimental design and data

Generasi locations were selected by lottery to form a randomized, controlled field experiment. The *Generasi* randomization was conducted at the subdistrict (*kecamatan*) level, so that all villages within the subdistrict either received the same version of *Generasi* (either all incentivized or all non-incentivized) or were in the control group.[2] A total of 264 eligible subdistricts were randomized into either one of the two treatment groups or the control group. Details on how the 264 subdistricts were selected and the randomization can be found in Appendix 2.

The program was phased in over two years, with 127 treatment subdistricts in year 1 and 174 treatment subdistricts in year 2. In year 1, for logistical reasons, the government prioritized those subdistricts who had previously received the regular PNPM village infrastructure program (denoted group P) to receive the program first. Since we observe group P status in treatment as well as control groups, we control for group P status (interacted with time fixed effects) in the experimental analysis below to ensure we use only the variation induced by the lottery. By year

[2] Randomizing at the subdistrict level is important since many health and education services, such as community health centers (Puskesmas) and junior secondary schools, provide services to multiple villages within a subdistricts. Randomizing at the subdistrict level ensures that we capture the total net effect of the program, since any within-subdistrict spillovers would also be captured in other treatment villages.

two of the program (2008) 96% of eligible subdistricts – 174 out of the 181 eligible subdistricts randomized to receive Generasi – were receiving the program. The remaining 7 eligible districts received the regular PNPM village infrastructure program instead of Generasi.[3] Conditional on receiving the program, compliance with the incentivized or non-incentivized randomization was 100%.

The phase-in and final allocation of Generasi 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 that shows that the randomization is balanced against baseline levels of covariates is discussed in Appendix 3 and shown in Appendix Table 1.

The main data we examine is a set of three waves of surveys of households, village officials, health service providers, and school officials. Wave I, the baseline round, was conducted from June to August 2007 prior to Generasi implementation.[4] Wave II, the first follow-up survey round, was conducted from October to December 2008, about 18 months after the program began. Wave III, a longer-term follow-up survey round, 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. 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, Yogyakarta, Indonesia. This survey data is unrelated to the data collected by the program

[3] We do not know why these 7 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. Appendix Table 2 shows that controlling for whether a subdistrict received traditional PNPM does not affect the results.

[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 Appendix Table 2 column 10.

for the purposes of calculating performance bonuses, and was not explicitly linked to the Generasi program in any way. Additional details about these data can be found in Appendix 4.

2.4. Estimation

Since the Generasi program was designed as a randomized experiment, the analysis is econometrically straightforward: we compare outcomes in those subdistricts randomized to be treatments with those subdistricts randomized to be control areas, controlling for the level of the outcome at baseline.

In implementing our analysis, we restrict attention to the 264 "eligible" subdistricts, as above, and use the randomization results combined with the government's prioritization rule to construct our treatment variables. Specifically, analyzing Wave II data (corresponding to the first treatment year), we define the *GENERASI* variable to be a dummy that takes value 1 if the subdistrict was randomized to receive *GENERASI* and either a) it was in the priority area (group P) or b) was in the non-priority area and selected in the additional lottery to receive the program in 2007. In analyzing Wave III data, we define the *GENERASI* variable to be a dummy that takes value 1 if the subdistrict was randomized to receive *Generasi*. We define the *GENERASIINCENTIVES* variable to be a dummy that takes value 1 if the GENERASI variable is 1 and if the subdistrict was randomized to be in the incentivized version of the program. *GENERASIINCENTIVES* thus captures the additional effect of the incentives above and beyond the main effect of having the program, and is the key variable of interest in the paper. Note that by defining the variables in this way, we are exploiting only the variation in program exposure due to the lottery. 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% of

subdistricts in 2007 and 96% 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 controlling for the average level of the outcome variable in the subdistrict in the baseline survey. Since we also have individual-specific panel data for half our sample, we include the pre-period value for those who have it, as well as a dummy variable that corresponds to having non-missing pre-period values. Since households came from one of three different samples (those with a child under age 2, those with a child age 2–15 but not in the first group, and all others; see Appendix 4 for more information), we include dummies for those three sample types, interacted with whether a household came from a panel or non-panel village. Finally, since many of the indicators for children vary naturally as the child ages, for all child-level variables we include age dummies.

We thus estimate the following regressions:

Wave II data:

$$y_{pdsi2} = \alpha_d + \beta_1 GENERASI_{pds2} + \beta_2 GENERASIINCENTIVES_{pds2} + \gamma_1 y_{pds1} + \gamma_2 1_{\{y_{pds1} \neq missing\}} + \gamma_3 \overline{y_{ds1}} + SAMPLE_{pdsi} + \alpha_p \times P_s + \varepsilon_{pdsi} \quad (1)$$

Wave III data:

$$y_{pdsi3} = \alpha_d + \beta_1 GENERASI_{pds3} + \beta_2 GENERASIINCENTIVES_{pds3} + \gamma_1 y_{pds1} + \gamma_2 1_{\{y_{pds1} \neq missing\}} + \gamma_3 \overline{y_{ds1}} + SAMPLE_{pdsi} + \alpha_p \times P_s + \varepsilon_{pdsi} \quad (2)$$

Wave II and III combined average effect:

$$y_{pdsit} = \alpha_{dt} + \beta_1 GENERASI_{pdst} + \beta_2 GENERASIINCENTIVES_{pdst} + \gamma_{1t} y_{pds1} + \gamma_{2t} 1_{\{y_{pds1} \neq missing\}} + \gamma_{3t} \overline{y_{ds1}} + \pi_t SAMPLE_{pdsi} + \alpha_{pt} \times P_s + \varepsilon_{pdsit} \quad (3)$$

where *i* is an individual respondent, *p* is a province, *d* is a district, *s* is a subdistrict, *t* is the survey wave (1 = baseline, 2 = interim survey, 3 = final survey), 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), $1_{ypdsi1\ missing}$ is a dummy for being a panel household, $\overline{y_{ds1}}$ is the average baseline value for the subdistrict, SAMPLE are dummies for how the household was sampled interacted with being a panel or cross-section household, and $\alpha_p \quad P_s$ are province-specific dummies for being in the sample areas having had prior community-driven development experience through the KDP program. Standard errors are clustered at the subdistrict level. Note that in the final equation for computing the average effect over Wave II and Wave III, all control variables (e.g., district FE, sample controls, baseline values, etc) are be fully interacted with wave dummies (shown in the equation with coefficients indexed by *t*), to capture the fact that there may be differential trends in different parts of the country.

The key coefficient of interest is β_2, which estimates the difference between the incentivized and non-incentivized version of the program. We can also calculate the total impact of the incentivized version of the program (vis-à-vis pure controls) by adding the coefficients on *GENERASIINCENTIVES* and *GENERASI*. We also examine a wide variety of additional specifications as robustness tests; these specifications are discussed in more detail in Section 3.

Since we have a large number of indicators, in order to calculate joint significance we will calculate average standardized effects for each family of indicators, following Kling, Liebman and Katz (2007). Specifically, for each indicator *i*, define σ^2_i to be the variance of *i.* We then 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 then define the average standardized effect as

$$\sum_i \frac{\beta_i}{\sigma_i} \quad (4)$$

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). This hypothesis document was registered with the Abdul Latif Jameel Poverty Action Lab at MIT and is available on request.

3. Main results on targeted indicators

3.1. Overall impact

Table 3 presents the results on the 12 targeted indicators. Each row in Table 3 presents three separate regressions. Column (1) shows the baseline mean of the variable. Columns (2) – (4) show the regression with the Wave II survey results (after one year of program implementation) estimated using equation (1); columns (5) – (7) show the regression with the Wave III survey results estimated using equation (2); columns (8) – (10) show the regression that average across both survey waves estimated using equation (3). For each specification, we show the total Generasi treatment effect in incentive areas (the sum of the coefficients on GENERASI and GENERASIINCENTIVES), the total Generasi treatment effect in non-incentive areas (the coefficient on GENERASI), and the additional treatment effect due to the incentives (the coefficient on GENERASIINCENTIVES). The first 12 rows present each of the main 12 indicators one by one. The next three rows present average standardized effects overall, for the 8 health indicators and for the 4 education indicators. The final three rows present our estimate of the 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 the household surveys.

Focusing first on the average standardized effects, column (10) shows that on average over the two years of the program, the 8 targeted maternal and child indicators (e.g., prenatal visits, delivery by trained midwives, immunizations, regular weight checks) were an average of 0.03 standard deviations higher in incentivized areas than in non-incentivized areas. The only one of the 12 indicators that is individually significant is prenatal visits, which were 5% higher in incentivized areas than non-incentivized areas. The effects appear somewhat larger in Wave II – column (4) shows that the average standardized effect of the incentives on health is 0.04 standard deviations, whereas column (7) shows that they were a (statistically insignificant) 0.026 standard deviations in Wave III. The comparison to pure controls suggest that the change over time is due to increases in effectiveness in the non-incentivized areas, rather than declining effectiveness of the incentivized areas, though these differences are not statistically significant.[5] In Wave II prenatal visits were 8% higher in incentivized areas than in non-incentivized areas; weight checks were 4% higher; and, most notably, malnutrition (defined as having weight-for-age more than 2 standard deviations below World Health Organization standards) was 15% lower in incentivized areas.

No effects of the incentives were seen in education in either wave. In both incentivized and non-incentivized areas, age 13-15 participation and gross attendance fell relative to controls in Wave II, and age 7-12 participation increased in Wave III. The average standardized effects for education for both incentivized and non-incentivized areas decrease in Wave II and increase

[5] To test whether the differences between waves are statistically significant, we conducted additional analysis where we restricted the sample to those subdistricts that were randomized to be treatment in both waves or control in both waves, so that differences over time could be separated from changing composition of participating districts over time. We then tested whether the impact of incentives in Wave II (after 1 year) was different from the impact in Wave III (after 2 years). Appendix Table 5 presents the results, which shows that none of the differences in average standardized effects for health (either separately by treatment or the additional effects of the incentives) are statistically significantly different between waves.

in Wave III.[6] Overall, this was a period where enrollments were increasing dramatically throughout Indonesia (including in our control areas).

It is worth noting that a substantial share of the overall effect of the Generasi program can be attributed to the performance incentives. In Wave II, the incentivized version of the program improved the 12 indicators by an average of 0.053 standard deviations compared to pure control (column 2), while the non-incentivized version improved the 12 indicators by only 0.012 standard deviations (column 3, statistically insignificant). This implies that 77% of the total impact of the program can be attributed to the incentives. In Wave III, even though the effect of the incentives is statistically insignificant, the point estimates suggest that 50% of the total impact of the program can be attributed to the incentives. Thus, when the incentive effect is scaled by the overall impact of the program, the incentives seem to have had a substantial effect.

An alternative approach to weighting the various individual effects is to use the weights used by the program in calculating bonus payments. This approach has the advantage 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 by each version of the program. The results, shown at the bottom of table, show a similar story to the average standardized effects. In Wave II, 93 percent of the program's impact on health (in terms of points) can be attributed to the incentives; in

[6] In particular, if we pool incentive and non-incentivized 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).

Wave III, 30 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.

Although we pre-specified equations (1) – (3) as the main regression specifications of interest, we have also considered the robustness of these results to a wide range of alternative specifications. Appendix Table 2 reports the coefficient on the additional effect of the incentives – the equivalent of columns (4), (7), and (10) – 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 whatsoever, estimate the regression in first-differences rather than including the baseline level as a control, and run the entire regression aggregated to the subdistrict level, rather than using individual level data. The results are very consistent with the main specification in Table 3.

3.2. Heterogeneity in impact

We test whether the incentives had a larger impact in areas where the baseline level was lower. The idea is that the marginal cost of improving achievement is higher if the baseline level is higher, e.g., moving from 98% to 98% enrollment rates is harder than from moving from 80% to 81%.[7] To examine this explicitly, we re-estimate equations (1) – (3), interacting the GENERASI and GENERASIINCENTIVES variables with the mean value of the indicator in the subdistrict at baseline. The results are shown in Table 4. A negative coefficient on the interaction implies that the program was more effective in areas with worse baseline levels. For ease of interpretation, we also present the implied impacts calculated at the 10^{th} percentile of the baseline distribution.

[7] Note that this is the main dimension of heterogeneity we specified in the pre-specified analysis plan.

The results confirm that the incentives were more effective in areas with lower baseline levels of service delivery – the standardized interaction term of GENERASIINCENTIVES * BASELINE_VALUE in columns (3), (7), and (11) are negative and, in both Wave II and overall, statistically significant. To interpret the magnitude of the heterogeneity, note that, in Wave II, the incentives added 0.074 standard deviations to the health indicators at the 10^{th} percentile of the baseline distribution. In Wave III, it was 0.06 standard deviations (not statistically significant), and across the two waves, it was 0.066 standard deviations. These effects are about double the average effect of the program shown in Table 3, and suggest that, indeed, incentives were more effective in areas with lower baseline levels.

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 non-incentivized areas in NTT province relative to Java, and 0.14 standard deviations higher in incentivized areas than non-incentivized areas in Sulawesi relative to Java (see Appendix Table 3). This is not surprising given the lower levels of baseline service delivery in these areas: malnutrition for under 3 year olds (more than 2 standard deviations below the age adjusted weight as defined by the WHO) is 12.6 percent in Java, but 24.7 percent in NTT and 23.4 percent in Sulawesi; similarly, 76.4 percent of births in our Java areas are attended by a trained medical professional (midwife or doctor), compared with only 42.7 percent are attended by a trained professional in NTT and only 55.9 percent in Sulawesi. These results confirm the idea that the program was substantially more effective in areas with lower levels of baseline service provision.

4. Potential pitfalls of incentives

The previous section shows that the performance incentives substantially increased the effectiveness of the program. 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 non-incentivized outcomes; manipulation of performance records; and reallocation of funds towards wealthier areas.

4.1. Spillovers on non-targeted indicators

Whether the incentives would increase or decrease performance on non-targeted indicators depends on the nature of the health and education production functions. For example, if there is a large fixed cost of getting 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 towards the types of service incentivized by Generasi 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 3 health domains: utilization of non-incentivized 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. In constructing these indicators, we erred on the side of including more rather than fewer indicators (once again, all indicators were pre-specified in the analysis plan prior to examining the data). We also examine potential impacts on family composition decisions (for example, does better maternal care induce people to have more children or migrate into the area). On the education side, we examine the impact on high school enrollment, hours spent in school, enrollment in

informal education (so-called Paket A, B, and C, which are the at-home equivalents of primary, junior secondary, and senior secondary schools), distance to school, and child labor.

Table 5 reports the average standardized effects for each of these domains; the detailed indicator-by-indicator results can be found in Appendix Table 4. In general, we find no differential negative spillover impacts of the performance incentives on any of these indicators, and if anything, find some slight evidence of positive spillovers. For example, we find that the performance incentives led to positive effects on reductions in child labor (.12 hours per child for age 7-15; 0.08 standard deviations in Wave II and 0.03 standard deviations overall). This suggests that negative spillovers on non-targeted indicators do not seem to be a substantial concern with the performance incentives in this context.

4.2. 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 (forthcoming) show that teachers in India inflate student attendance records to allow them to receive subsidized grain. Manipulation of recordkeeping can have substantial efficiency costs: for example, children could fail to get immunized properly if their immunization records were falsified.

We can check for two types of falsification of recordkeeping. First, for immunizations and school attendance, we can check for falsification of actual underlying records by comparing the official records to an independent measure observed directly by our survey team. Second, 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 using the household survey data.

For the first approach – checking underlying records against direct observation by the survey team – we use two measures that we can verify directly. The BCG vaccine leaves a distinctive scar on the arm, so we can compare a child's records on whether the BCG vaccine was administered to the presence of the BCG scar on the arm as measured by our surveyors (see Banerjee et al. 2008). Second, we compare attendance from random spot-checks of classrooms with attendance records from the same classroom on a specific day 1-2 months previously. Although the dates compared are not the identical (we could not obtain reliable records on the date of our survey, since the presence of our surveyors might affect the records), the difference between them should capture, on average, the markup in attendance.

The results are shown in Table 6. Panel A explores the differences between BCG scars and record keeping.[8] On average, 75 percent of children have the scar; 60% of children have a record of receiving the vaccine, and 85 percent of children either have a record of receiving the vaccine or have a parent who reports the child received a vaccine. 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 no statistically significant differences in false reports of the BCG scar based on the performance incentives, though the point estimates suggest the possibility of slight inflation in Wave II.

Panel B explores differences in attendance rates. On average, attendance is overstated: 88 percent of children were recorded as present by our random visits, whereas 95 percent were recorded present in the official attendance records. The discrepancy is unchanged by the performance incentives. In fact, recorded attendance appears lower in the incentive treatment

[8] Note that in 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 report that the child was vaccinated.

while actual attendance is unchanged, which suggests perhaps that the incentives led to better record keeping.

Panel C of Table 6 uses a different approach, and examines manipulation of the program scores used to calculate incentive payments. For each of the 12 indicators in the household survey, we calculate the difference between the between average level of performance on that indicator in the respondent's village according to the administrator and the corresponding values from the household survey.[9] We then regress the difference between the administrative and household data on a dummy for being in the incentivized version of the program. The average standardized effects are presented in Panel C of Table 6; the indicator-by-indicator results are available Appendix Table 8. Since there is no administrative data for control groups, the results show only the differences between the incentivized and non-incentivized groups. The results in Table 6 show that, for both Wave II and the pooled results, the difference between the administrative data and household survey is lower in the incentive than non-incentivized villages. This is the opposite of what one would expect if the incentives led villages to systematically inflate scores in the incentivized areas to increase their performance bonuses. Combined, these two pieces of evidence suggest that manipulation of recordkeeping is not a major problem of the performance incentives in this context. The fact that the performance bonuses were relative to other villages in the same subdistrict, and that those villages were allowed (indeed, encouraged) to regularly audit performance indicators in neighboring villages may have minimized the problems of over-reporting in this context.

[9] 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.

4.3. *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 towards 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 in Generasi attempted to mitigate this concern though relative incentives. Specifically, the performance pool was fixed for each subdistrict, and villages within a given subdistrict were competing only against one another for bonus funds, rather than against much wealthier or better performing areas in other parts of the country. This relative incentive system meant that unobserved, subdistrict specific common shocks would cancel out, and it mechanically prevents the performance bonus from resulting in funds migrating from poorer subdistricts to wealthier subdistricts.[10] Nevertheless, if most of the differences in productivity were within subdistricts, not between subdistricts, the same problem could still occur.

To explore whether relative performance measurement prevented funds from flowing to richer areas, in Table 7, in Panel A we regress the total amount of bonus funds each village received on village average per-capita consumption (measured in the household survey), village remoteness (measured in km from the district capital), and village poverty (measured as the share of households classified as poor by the eligibility criteria set by the national family planning board). In Panel B, we then repeat the same regressions for a counterfactual calculation for incentives without the relative performance component. Specifically, in the counterfactual we allocate bonus payments proportional to bonus points relative to all villages in the program, rather than relative only to other villages in the same subdistrict.

[10] Minimum performance levels (m_{vi}) were also adjusted based on coarse measures of access (distance to nearest midwife and distance to nearest junior secondary school). See Appendix 1.

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.

In sum, the results in this section document that there were little negative effects of incentives: we found no evidence of multitasking problems; we found no evidence of manipulation of records, and we found that the relative incentive scheme successfully prevented the incentives from resulting in funds flowing to richer areas.

5. Mechanisms

The results thus far showed that the incentives substantially improved the targeted health indicators with little obvious downside. 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.

5.1. Allocation of funds

Table 8 examines whether the incentives affected how the Generasi communities chose to allocate the block grants. Each row in Panels A and B shows the share of the village's block grant spent on the item.

The most notable finding that emerges is that the incentives led to a shift away from education supplies – uniforms, books, and other school supplies – and towards health expenditures. In particular, 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 types of 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 are likely to have relatively little impact on school attendance and enrollment. As shown in Table 3 above, the performance incentives improved health outcomes with no detrimental effect on education, so combined this suggests that the performance incentives may have led communities to reallocate funds away from potentially politically popular but ineffective education spending towards more effective health spending.

We also tested two other hypotheses that do not seem borne out in the data. First, we expected that, since performance incentives effectively increase the discount rate (since one places higher value on a return in the current year since it will affect bonuses), 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 overall (18 percent). Second, we expected that performance incentives would lead to a decrease in "capture" of the

funds to expenses benefitting providers (e.g., uniforms for health volunteers), but we see no impact on this dimension.

The evidence thus far was on how the money was spent. Table 9 shows the other side of the equation, namely, what households received from the block grants, using data from the household survey. Both incentive and non-incentive 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 than in the non-incentive treatments – in fact, if anything the point estimates suggest they were 1.0-2.7 percentage points (12-32 percent) more likely to receive a uniform in the groups with performance incentives and 1.0-1.7 percentage points (18-32 percent) more likely to receive other school supplies in the groups with performance incentives. Moreover, the self-reported Rupiah value of the uniform received is identical in both treatments. This suggests that the change in budget allocations away from uniforms and school supplies documented in Table 8 likely came from increased efficiency in procuring the uniforms rather than a reduction in the quality or quantity of uniforms. Likewise, there was also a substantial increase in scholarships (1 percentage point, or about 125 percent) and transport subsidies (0.6 percentage points, about 110 percent). Thus, on average more children received education subsidies, even though more money was being spent on health. Combined with the fact that the health outcomes improved and education did not suffer, the evidence here suggests that the performance incentives improved the efficiency of the Generasi funds.

5.2. Effort

A second dimension we examine is effort – both on the part of workers and on the part of communities. Table 10 begins by examining effort of 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 labor on the part of midwives. On average, midwives spent 1.7 hours (6 percent) more working over the 3 days prior to the survey in incentive areas than in non-incentive areas. There was no impact on teacher attendance or provider attendance at health care centers. Given that midwives are the main providers of maternal and child health services, the increase in midwife effort is consistent with the increase in these services we observed above.

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 10 with a dummy for whether the midwife is a tenured civil servant, we find that the Generasi incentive treatment led to a greater increase in private practice hours provided by tenured civil servant midwives (See Appendix Table 7), with no change in their public hours. This suggests that Generasi was able to leverage the fee-for-service component of midwives' lives to increase their service provision. Interestingly, the monetary compensation (e.g. value of subsidies per patient) Generasi provided to midwives did not differ between the incentivized and non-incentivized treatments (results not reported in table), so it was not the financial incentives alone that resulted in the difference. More likely, it was the combination of other, non-financial incentives to midwives (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 11 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 providers; community effort at outreach, such as door-to-door "sweepings" to get more kids into the posyandu net and school committee meetings with parents, and community effort at monitoring service providers, such as school committee membership and meetings with teachers. We find no evidence that the performance incentives had an impact on any of these margins, although the Generasi program as a whole increased community participation at monthly community health outreach activities (posyandu) where many maternal and child services are provided.

5.3. 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 their behavior. While we can't estimate each household's individual elasticity directly, we can examine whether incentivized communities targeted differently based on the household's per-capita consumption. The idea is that poorer households' behavior may be more elastic with respect to subsidies than richer households, who can afford the targeted services with or without subsidies. Incentives could therefore encourage e communities to targeted benefits to poorer households and resist the pressure from interest groups within the village to distribute benefits more evenly.[11]

The results in Table 12 show how the incentives affect how Generasi communities targeted the direct benefits they distribute. For each of the three specifications (Wave II, Wave

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

III, and Pooled), we re-estimate equations (1) – (3), allowing for subdistrict fixed effects, and interact the GENREASI variables with a dummy for the household being in the top 3 quintiles of the income distribution in the baseline survey. 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 12 shows estimated effects from the regression. We first present the difference between the top 3 quintiles and the bottom 2 quintiles for incentivized Generasi areas. A negative coefficient indicates that the poor received relatively more than the rich in Generasi areas relative to controls. The second column presents the difference between the top 3 quintiles and the bottom 2 quintiles for non-incentivized Generasi areas. The third column presents the difference between the first two columns. A negative coefficient in the third column indicates that the incentivized version of the program had more pro-poor targeting than the non-incentivized version. Panel A shows the average standardized effects for targeting of direct benefits (i.e. the subsidies and transfers examined in Table 9), 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 Appendix Tables 9 and 10. The results in Panel A show 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 non-incentivized 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 in this section 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 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.

6. Cost-effectiveness

It is difficult to interpret the magnitudes given above without some notion of costs. Conditional on implementing the Generasi program, adding the performance incentives was essentially free – the same monitoring of indicators was done in both the incentivized and non-incentivized versions of the program, no additional personnel were required to do monitoring (the program would have needed facilitators regardless), 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 non-incentivized areas. Thus the cost effectiveness of the incentives themselves for this program is easy to analyze: they improved outcomes, added virtually no costs, and therefore is likely to be cost effective.

The challenge in doing a cost-effectiveness calculation more formally is that there are many potential outcomes, and we do not necessarily know how to apportion the costs of the programs among the various outcomes. We therefore take the following approach: as in Section 3, we calculate the total number of "points" the program created, using the weighting scheme agreed upon in advance and shown in Table 1 and the point estimates for the impact of the program from Table 3.[12] We divide the total cost of the program by the total number of points created to generate a "cost per point", which can then be interpreted using the point values in

[12] Note that the number of points shown in this section is approximately half the total number of points reported in Table 3. The reason is that Table 3 is calculated using the total number of beneficiaries in the program, whereas in this section since we are mapping it to points we use number of beneficiaries in the particular treatment group only, which is half of the total.

Table 1. While naturally different weighting schemes could produce different answers, we use the points in the program since they presumably represent the government's relative weightings of the different interventions, i.e. we use a set of relative prices that should roughly correspond to the relative weights the government places on the various indicators.

To calculate the costs of the program, we divide the expenditures into transfers to households and real expenditures (i.e. real allocation of resources). For transfers, we assume that transfers are valued by recipients at cost, so the real social cost of transfers is the social deadweight loss of taxation to raise the funds for the transfers. For non-transfer costs (such as hiring a midwife), the social cost is the expenditure plus the deadweight loss of taxation. For the purposes of evaluating Generasi, we count school supplies, school fee subsidies, health care subsidies, and supplementary food as transfers, and all other expenditures as real expenditures. As shown in Table 10 above, about 75% of the block grant is spent on transfers by these definitions. We also include the cost of the facilitators who administer the program as real expenditures. We use the consensus estimate of the marginal cost of public funds of 0.3 (Ballard, Shoven, and Whalley 1985), though we note that there are not reliable estimates of this parameter for developing countries. We use the Wave III impact results (at the end of the program's second year), when the program was at full scale, for this calculation.

The estimates are presented in Table 13. The key results are shown in the first two columns of Panel A: Generasi with incentives had a real cost per point of about $8, and Generasi without incentives had a real cost per point of about $11. Since the estimates in Table 3 show that, for year 3, the difference in the total number of points between incentivized and non-incentivized versions of the program is not significantly different, we should treat the difference between $8 and $11 as also not statistically significant. Panel B separately estimates the cost

effectiveness for the health and education components of the program, allocating facilitation costs equally between the two portions of the program and allocating expenditures based on how communities actually allocated block grants. For health, this yields estimates of $7 per point for the incentivized version and $9 for the non-incentivized version. For education, this yields estimates of $13 per point for the incentivized version and $16 for the non-incentivized version.

How do we interpret the $8 - $11 per point average cost effectiveness of the program? One approach is to back out what this implies to move a given indicator. Applying the weights from Table 1, for example, suggests that the cost of additional child weight check was $16 - $22, the cost of preventing one malnourished child was $384 - $528, the cost of getting one additional child fully covered with Vitamin A was $160 - $220, and the cost of enrolling one more child in primary school was $200 - $275.

Are these numbers large or small? While that is ultimately a judgment question for the reader, we provide two benchmarks. First, the closest comparison is Indonesia's conditional cash transfer program (PKH). The PKH program was conducted at the same time and evaluated using a randomized evaluation using the same survey instruments as Generasi, though it was conducted in somewhat different areas of the country (more urban and with better supply of services), and was targeted at the same set of indicators. We use the randomized evaluation results from Alatas (2011) of the PKH program, combined with the same weights in Table 1.[13] Alatas reports an estimated effect just for those households receiving PKH, as well as a "placement effect" on all poor households in the subdistricts regardless of whether they received PKH or not. We report

[13] We calculate the number of households of different age ranges from the PKH survey data on all PKH respondents. For spillovers, we assume that there are 2 non-PKH households for every one PKH household. This is consistent with the data used in the PKH evaluation, which consisted of a survey of previous cash transfer recipients and is the population over which they estimate spillovers. If there are spillovers to other parts of the population , this may be an underestimate.

cost-effectiveness numbers based on both calculations. The results suggest that, if one focuses only on the benefits enjoyed by PKH households, Generasi is more cost effective – with the $8 – $11 per point in Generasi comparing to about $22 per point for PKH. If one includes estimated spillover effects from PKH to non-recipient households in the same subdistricts, then the $8 - $11 per point for Generasi is comparable to the $11 per point estimate for PKH.[14] Thus, the Generasi program looks roughly comparable to an alternative program tried in Indonesia at the same time.

An alternative benchmark is to look at international comparisons. For example, school-based deworming in Kenya costs $3.50 per additional year of school attendance, and iron and deworming tablets in India cost $29 per additional year of school attendance. School meals in Kenya cost $35 per additional year of attendance, and school uniforms cost about $100 per additional year of attendance (JPAL 2011). By comparison, Generasi would cost between $125 - $400 per additional year schooling. [15] By this metric, Generasi as a whole is substantially less cost-effective than these other interventions, although it is worth noting that Generasi affects enrollment rates, whereas the international comparisons affect attendance, and that there is already high baseline enrollment in Generasi areas, which makes the marginal cost higher.

While these numbers suggest that Generasi, as a whole, may be more expensive than these other programs, the performance incentives themselves – at $0.62 per point, which translates into $16 per additional child enrolled in school and $30 per case of malnutrition avoided – compares favorably with all of the above interventions except deworming. The reason

[14] It is also worth noting that neither calculation includes the benefits from redistribution.

[15] Note that these estimates are not strictly comparable to the estimates in Table 13, since they count all program expenditures at cost regardless of whether they were transfers or not, and do not include the deadweight loss of taxation. When we redo the Generasi cost-effectiveness in this way, we obtain a cost per point of around $12. See Appendix Table 11.

the incentives themselves are cost effective is that in our case they are essentially free – the block grant was the same with and without incentives, and collecting the data used to validate the incentives was done in both the incentivized and non-incentivized versions of the program, so the only "costs" come from the fact that there were slightly more real expenditures and slightly fewer transfers in the incentivized version of the program. This suggests that while the Generasi program as a whole was not as cost effective as some other international comparators, adding incentives to existing programs may be a cost-effective way to improve performance.

7. Conclusion

In sum, the evidence presented here suggests that properly designed, performance based incentives can be a useful addition to aid programs. In Indonesia, we found that adding a relative performance-based incentive to a community-based health and education program increased performance. This was particularly true in areas with the lowest levels of performance before the program began. Incentives worked through increasing the efficiency with which funds were spent and through increasing providers' hours worked. Though the gains from incentives were relatively modest, we found little downside from the incentives – there was no evidence of multitasking problems, 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 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 suggest that one may not want to make the incentives too narrow – instead, to the extent the multitasking issue can be controlled (and it was not an issue here), 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 which where it may have lower 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.

References

Alatas, V. (2011). Program Keluarga Harapan: Main Findings from the Impact Evaluation of Indonesia's Pilot Household Conditional Cash Transfer Program, World Bank.

Baird, S., C. McIntosh, et al. (2011). "Cash or Condition? Evidence from a Cash Transfer Experiment." Quarterly Journal of Economics 126(4): 1709-1753.

Ballard, C. L., J. B. Shoven, et al. (1985). "General Equilibrium Computations of the Marginal Welfare Cost of Taxes in the United States." American Economic Review 75(1): 128-135.

Banerjee, A. V., E. Duflo, et al. (2008). Improving Immunization Coverage in Rural India: A Clustered Randomized Controlled Evaluation of Immunization Campaigns with and without Incentives, MIT.

Basinga, P., P. Gertler, et al. (2011). "Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation." The Lancet 377(9775): 1421-1428

Birdsall, N. and W. D. Savedoff (2009). Cash on Delivery: A New Approach to Foreign Aid With an Application to Primary Schooling. Washington, DC, Center for Global Development.

Dhaliwal, I., E. Duflo, et al. (2011). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education, MIT.

Duflo, E., R. Hanna, et al. (forthcoming). "Monitoring Works: Getting Teachers to Come to School " American Economic Review.

Express India News Service (2008). Biometric attendance to keep track of students, teachers in primary schools. Express India.

Gertler, P. (2004). "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA' s Control Randomized Experiment." American Economic Review 94(2): 336-341.

Gibbons, R. and K. J. Murphy (1990). "Relative Performance Evaluation for Chief Executive Officers." Industrial and Labor Relations Review 43(3): 30-51.

Holmstrom, B. (1979). "Moral Hazard and Observability." Bell Journal of Economics 10(1): 74-91.
Holmstrom, B. and P. Milgrom (1991). "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." Journal of Law, Economics and Organizations 7: 24.
Imbens, G. W. and J. D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects." Econometrica 62(2): 467-475.
Kling, J. R., J. B. Liebman, et al. (2007). "Experimental Analysis of Neighborhood Effects." Econometrica 75(1): 83-119.
Lazear, E. P. and S. Rosen (1981). "Rank-Order Tournaments as Optimum Labor Contracts." The Journal of Political Economy 89(5): 841.
Levy, S. (2006). Progress against poverty: sustaining Mexico's Progresa-Oportunidades program, Brookings Institution Press, Washington, DC.
Linden, L. and G. K. Shastry (forthcoming). "Identifying Agent Discretion: Exaggerating Student Attendance in Response to a Conditional School Nutrition Program." Journal of Development Economics.
Mookherjee, D. (1984). "Optimal Incentive Schemes with Many Agents." Review of Economic Studies 51(3): 433-446.
Schultz, T. P. (2004). "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." Journal of Development Economics 74(1): 199-250.
Weitzman, M. (1980). "The ratchet principle and performance incentives." Bell Journal of Economics 11(1): 302-308.

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

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

Table 2: *Generasi* randomization and implementation

	Incentivized *Generasi*		Non-incentivized *Generasi*		Control		Total
	P	NP	P	NP	P	NP	
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, Non-incentivized Generasi, and Control). Actual implementation status is shown in the rows. Note that conditional in receiving the program, the randomization into the incentivized or non-incentivized version of the program was always perfectly followed.

Table 3: Impact on targeted outcomes

Notes: Column 1 shows the baseline mean of the variable shown, with standard deviations in brackets. Each row of columns (2) – (4), (5) – (7), and (8) – (10) show coefficients from a regression of the variable shown on an incentive treatment dummy, a non-incentive 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 1 program placement, and in columns (5) – (7) it is defined based on year 2 program placement, and in columns (8) – (10), which uses pooled data from both waves, it is defined as year 1 placement for the Wave II data and as year 2 placement for the Wave III data. 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), (7), and (10) 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. * = 10% significance, ** = 5% significance, *** = 1% significance.

Table 4: Interactions with baseline level of service delivery

	Wave II				Wave III				AVERAGE			
Indicator	Generasi Incentive Total Effect * Pre-Period Level	Generasi Non-Incentive Total Effect * Pre-Period Level	Generasi Incentive Additional Effect * Pre-Period Level	Incentive Additional Effect at 10th Percentile	Generasi Incentive Total Effect * Pre-Period Level	Generasi Non-Incentive Total Effect * Pre-Period Level	Generasi Incentive Additional Effect * Pre-Period Level	Incentive Additional Effect at 10th Percentile	Generasi Incentive Total Effect * Pre-Period Level	Generasi Non-Incentive Total Effect * Pre-Period Level	Generasi Incentive Additional Effect * Pre-Period Level	Incentive Additional Effect at 10th Percentile
Main 12 indicators	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Number prenatal visits	0.167	0.115	0.052	0.490	-0.045	-0.1456*	0.101	-0.095	0.065	-0.015	0.080	0.158
	(0.131)	(0.117)	(0.133)	(0.365)	(0.080)	(0.082)	(0.083)	(0.284)	(0.080)	(0.071)	(0.082)	(0.253)
Delivery by trained midwife	-0.088	0.042	-0.1296*	0.050	-0.056	0.049	-0.105	0.065	-0.074	0.040	-0.1143**	0.0588*
	(0.074)	(0.070)	(0.071)	(0.046)	(0.071)	(0.076)	(0.068)	(0.042)	(0.055)	(0.056)	(0.053)	(0.034)
Number of postnatal visits	-0.141	-0.039	-0.102	0.012	-0.2502*	-0.061	-0.189	0.221	-0.1904**	-0.043	-0.147	0.123
	(0.126)	(0.130)	(0.142)	(0.193)	(0.137)	(0.128)	(0.136)	(0.182)	(0.094)	(0.101)	(0.103)	(0.142)
Iron tablet sachets	-0.142	-0.206	0.064	0.044	0.009	0.160	-0.151	0.116	-0.082	-0.020	-0.061	0.086
	(0.126)	(0.130)	(0.152)	(0.111)	(0.124)	(0.143)	(0.153)	(0.097)	(0.093)	(0.093)	(0.111)	(0.074)
Percent of immunization	-0.1884**	-0.086	-0.102	0.041	0.016	0.074	-0.057	0.033	-0.102	-0.021	-0.080	0.0376*
	(0.085)	(0.078)	(0.087)	(0.029)	(0.079)	(0.073)	(0.066)	(0.025)	(0.062)	(0.057)	(0.056)	(0.020)
Number of weight checks	-0.071	-0.086	0.016	0.083	-0.065	0.000	-0.065	0.022	-0.069	-0.043	-0.025	0.045
	(0.098)	(0.099)	(0.106)	(0.108)	(0.110)	(0.115)	(0.129)	(0.120)	(0.069)	(0.086)	(0.094)	(0.093)
Number Vitamin A supplements	-0.030	-0.024	-0.007	-0.008	-0.001	-0.044	0.043	0.060	-0.013	-0.037	0.024	0.026
	(0.128)	(0.115)	(0.154)	(0.096)	(0.115)	(0.130)	(0.133)	(0.085)	(0.085)	(0.093)	(0.106)	(0.065)
Percent malnourished	-0.2564**	-0.100	-0.156	-0.0481*	-0.2677**	-0.2400**	-0.028	0.006	-0.2591***	-0.1657**	-0.093	-0.020
	(0.129)	(0.113)	(0.138)	(0.027)	(0.132)	(0.116)	(0.128)	(0.027)	(0.095)	(0.078)	(0.100)	(0.021)
Age 7–12 gross enrollment	-0.042	-0.087	0.045	-0.007	-0.114	-0.1800**	0.066	-0.011	-0.074	-0.129	0.055	-0.009
	(0.090)	(0.106)	(0.127)	(0.012)	(0.094)	(0.081)	(0.098)	(0.010)	(0.066)	(0.080)	(0.091)	(0.009)
Age 13–15 gross enrollment	-0.063	-0.079	0.016	0.013	-0.006	-0.115	0.110	-0.021	-0.036	-0.100	0.065	-0.006
	(0.120)	(0.121)	(0.149)	(0.044)	(0.109)	(0.098)	(0.101)	(0.028)	(0.085)	(0.090)	(0.098)	(0.028)
Age 7–12 gross attendance	-0.051	-0.045	-0.006	0.000	-0.1064**	-0.1085**	0.002	-0.001	-0.0738**	-0.0721**	-0.002	-0.001
	(0.039)	(0.039)	(0.033)	(0.006)	(0.050)	(0.049)	(0.034)	(0.007)	(0.032)	(0.031)	(0.027)	(0.005)
Age 13–15 gross attendance	-0.052	-0.033	-0.019	0.031	-0.022	-0.110	0.087	-0.016	-0.037	-0.077	0.040	0.004
	(0.111)	(0.109)	(0.133)	(0.048)	(0.108)	(0.078)	(0.099)	(0.032)	(0.080)	(0.073)	(0.087)	(0.029)
Average standardized effect	-0.2140**	-0.157	-0.057	0.056	-0.1919**	-0.2225***	0.031	0.025	-0.2055***	-0.1957**	-0.010	0.037
	(0.096)	(0.112)	(0.133)	(0.042)	(0.091)	(0.085)	(0.088)	(0.033)	(0.065)	(0.079)	(0.087)	(0.029)
Average standardized effect health	-0.1975***	-0.078	-0.1193*	0.0706*	-0.106	-0.020	-0.086	0.061	-0.1615***	-0.059	-0.1022**	0.0638**
	(0.061)	(0.056)	(0.068)	(0.037)	(0.066)	(0.066)	(0.064)	(0.039)	(0.047)	(0.045)	(0.051)	(0.031)
Average standardized effect educ.	-0.247	-0.314	0.067	0.026	-0.363	-0.6274***	0.264	-0.048	-0.2934*	-0.4685**	0.175	-0.017
	(0.253)	(0.312)	(0.369)	(0.086)	(0.244)	(0.219)	(0.235)	(0.050)	(0.174)	(0.221)	(0.241)	(0.051)

Notes: See Notes to Table 3. Columns (1), (5), and (9) interact the incentive treatment dummy with the baseline subdistrict mean of the variable shown, and columns (2), (5), and (10) interact the non-incentive treatment dummy with the baseline subdistrict mean of the variable shown. Columns (3), (7), and (11) are the difference between the two previous columns. Columns (4), (8), and (12) show the estimated additional impact of incentives evaluated at the 10th percentile of the indicator at baseline.

Table 5: Spillovers on non-targeted indicators, average standardized effects by indicator family.

	Wave II			Wave III			AVERAGE		
Family of indicators	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Average Treatment Effect	Non-Incentive Average Treatment Effect	Incentive Average Additional Effect
Health									
Utilization of non-incentivized	0.019	-0.010	0.029	0.029	0.011	0.018	0.023	0.001	0.022
health services	(0.020)	(0.021)	(0.022)	(0.021)	(0.020)	(0.019)	(0.016)	(0.015)	(0.016)
Health services quality	0.0901**	0.0752*	0.015	0.041	0.040	0.001	0.0646**	0.0567**	0.008
	(0.039)	(0.039)	(0.040)	(0.036)	(0.038)	(0.036)	(0.029)	(0.029)	(0.028)
Maternal knowledge and	0.026	0.024	0.002	0.033	0.043	-0.011	0.029	0.034	-0.005
practices	(0.029)	(0.028)	(0.030)	(0.029)	(0.027)	(0.026)	(0.022)	(0.022)	(0.021)
Family composition decisions	0.014	-0.012	0.026	0.023	-0.007	0.029	0.025	-0.014	0.0381*
	(0.019)	(0.021)	(0.022)	(0.022)	(0.026)	(0.023)	(0.020)	(0.024)	(0.022)
Education									
Other enrollment metrics	-0.070	-0.051	-0.019	-0.013	0.006	-0.019	-0.022	-0.011	-0.012
	(0.049)	(0.046)	(0.049)	(0.021)	(0.020)	(0.018)	(0.017)	(0.018)	(0.018)
Transportation to school (cost	-0.077	-0.034	-0.043	0.004	0.022	-0.018	-0.025	0.002	-0.027
and distance	(0.058)	(0.050)	(0.060)	(0.042)	(0.041)	(0.042)	(0.036)	(0.035)	(0.039)
Avoiding child labor	-0.025	-0.1074***	0.0825**	0.012	0.007	0.005	-0.006	-0.0391*	0.0335*
(higher #s = less child labor)	(0.022)	(0.038)	(0.034)	(0.025)	(0.020)	(0.022)	(0.018)	(0.021)	(0.020)
Overall									
Average overall standardized	-0.003	-0.017	0.013	0.012	0.007	0.005	0.013	0.004	0.008
effect	(0.016)	(0.017)	(0.019)	(0.025)	(0.020)	(0.022)	(0.010)	(0.011)	(0.011)
Average standardized effect	0.0373**	0.019	0.018	0.012	0.007	0.005	0.0354***	0.020	0.016
health	(0.016)	(0.016)	(0.017)	(0.025)	(0.020)	(0.022)	(0.013)	(0.013)	(0.013)
Average standardized effect	-0.0574**	-0.0643**	0.007	0.012	0.007	0.005	-0.018	-0.016	-0.002
educ.	(0.029)	(0.030)	(0.032)	(0.025)	(0.020)	(0.022)	(0.016)	(0.016)	(0.017)

Notes: See Notes to Table 3. Each row presents average standardized effects from a family of indicators, with the detailed indicator-by-indicator results shown in Appendix Table 4. 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 4 per pregnancy, vitamin A drops beyond 2 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 fraction initiating breastfeeding within the first hour after birth, share with exclusive breastfeeding, maternal knowledge about how to proper treatment of several child health conditions, and a questions about a women'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 work for a wage, hours spend working for a wage, a dummy for doing any wage work, and a dummy for doing any household work.

Table 6: Manipulation of performance records

		Wave II			Wave III			Average		
Indicator	Baseline mean	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Average Treatment Effect	Non-Incentive Average Treatment Effect	Incentive Average Additional Effect
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: BCG Scar										
False "yes" in recorded BCG	0.100	0.0304**	0.006	0.025	-0.002	0.002	-0.003	0.009	0.001	0.008
vaccine	[0.2995]	(0.015)	(0.014)	(0.015)	(0.013)	(0.014)	(0.014)	(0.010)	(0.011)	(0.011)
False "yes" in declared BCG	0.122	0.0297*	0.019	0.010	0.007	0.000	0.008	0.014	0.005	0.009
vaccine	[0.3274]	(0.015)	(0.015)	(0.016)	(0.014)	(0.015)	(0.014)	(0.011)	(0.011)	(0.011)
Children with no record card	0.195	-0.0481**	-0.0349*	-0.013	-0.020	-0.0515***	0.0314*	-0.0289*	-0.0423***	0.013
	[0.3963]	(0.021)	(0.021)	(0.019)	(0.019)	(0.018)	(0.017)	(0.015)	(0.015)	(0.014)
Panel B: Attendance										
Attend. Rate – difference between	7.299	-2.088	-2.6623*	0.574	0.496	2.158	-1.663	-0.589	0.245	-0.834
recorded and observed		(1.660)	(1.500)	(1.686)	(2.018)	(2.119)	(1.971)	(1.324)	(1.266)	(1.295)
Attend. rate observed	88.366	1.348	2.9302**	-1.582	-0.795	-1.975	1.180	0.070	-0.036	0.106
		(1.577)	(1.466)	(1.617)	(1.885)	(2.024)	(1.900)	(1.254)	(1.233)	(1.266)
Attend. rate recorded	95.726	-0.7465**	0.164	-0.9106**	-0.253	0.160	-0.414	-0.472	0.168	-0.6395*
		(0.368)	(0.399)	(0.459)	(0.441)	(0.435)	(0.441)	(0.306)	(0.308)	(0.333)
Panel C: Difference between admin. and household data										
Average standardized effect				-0.0952**			-0.084			-0.0772**
				(0.045)			(0.070)			(0.036)
Average standardized effect				-0.068			-0.068			-0.064
health				(0.049)			(0.066)			(0.040)
Average standardized effect				-0.1433***			-0.111			-0.1007***
educ.				(0.053)			(0.096)			(0.039)

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.

Table 7: Do relative payments prevent money from flowing to richer areas?

	Wave II				Wave III				Pooled			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Actual incentive payments												
	-1.325			-1.749	13.48			15.09				

Table 9: Direct benefits received, incentivized vs. non-incentivized

		Wave II			Wave III			AVERAGE		
Indicator	Control Mean	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Average Treatment Effect	Non-Incentive Average Treatment Effect	Incentive Average Additional Effect
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Received scholarship	0.025	0.0162**	0.008	0.009	0.0208**	0.009	0.012	0.0187***	0.008	0.0108*
	[0.0048]	(0.007)	(0.006)	(0.008)	(0.009)	(0.007)	(0.009)	(0.005)	(0.005)	(0.006)
Received uniform	0.013	0.1095***	0.0828***	0.027	0.0824***	0.0723***	0.010	0.0948***	0.0777***	0.017
	[0.0036]	(0.019)	(0.012)	(0.018)	(0.013)	(0.010)	(0.015)	(0.013)	(0.008)	(0.012)
Value of uniforms (Rp.)	731	7,845***	6,098***	1,746	7,122***	5,936***	1,186	7,451***	6,028***	1,423
	[271]	(1,569)	(1,034)	(1,447)	(1,312)	(1,117)	(1,521)	(1,264)	(845)	(1,225)
Received other school supplies	0.008	0.0634***	0.0535***	0.010	0.0701***	0.0534***	0.017	0.0670***	0.0533***	0.014
	[0.0027]	(0.012)	(0.009)	(0.012)	(0.012)	(0.010)	(0.015)	(0.010)	(0.007)	(0.011)
Received transport subsidy	0.000	0.0143***	0.0049*	0.009	0.0078***	0.0050***	0.003	0.0108***	0.0051***	0.0056*
	[0.0000]	(0.005)	(0.003)	(0.006)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)
Received other school support	0.000	0.000	0.000	0.001	0.0072**	0.0063*	0.001	0.0039**	0.0033*	0.001
	[0.0000]	(0.000)	(0.000)	(0.000)	(0.003)	(0.003)	(0.004)	(0.002)	(0.002)	(0.002)
Received supp. feeding at school	0.000	0.005	0.0041**	0.001	0.006	0.003	0.003	0.005	0.004	0.002
	[0.0000]	(0.003)	(0.002)	(0.004)	(0.006)	(0.005)	(0.007)	(0.004)	(0.003)	(0.004)
Received supp. feeding at posyandu	0.469	0.1533***	0.1563***	-0.003	0.1745***	0.2044***	-0.030	0.1647***	0.1843***	-0.020
	[0.0171]	(0.028)	(0.027)	(0.028)	(0.025)	(0.022)	(0.023)	(0.022)	(0.019)	(0.019)
Received intensive supp. feeding at school	0.027	0.008	0.0252**	-0.018	0.0242**	0.0191**	0.005	0.0173***	0.0212***	-0.004
	[0.0055]	(0.007)	(0.011)	(0.011)	(0.010)	(0.009)	(0.010)	(0.007)	(0.007)	(0.007)
Received health subsidy for pre/postnatal care	0.005	0.0343***	0.0270***	0.007	0.0273***	0.0364***	-0.009	0.0304***	0.0323***	-0.002
	[0.0023]	(0.008)	(0.007)	(0.009)	(0.006)	(0.007)	(0.009)	(0.006)	(0.006)	(0.007)
Received health subsidy for childbirth	0.039	0.1010***	0.1273***	-0.026	0.0974***	0.1249***	-0.028	0.0991***	0.1260***	-0.027
	[0.0078]	(0.017)	(0.017)	(0.019)	(0.016)	(0.020)	(0.023)	(0.012)	(0.015)	(0.016)
Average standardized effect		0.3394***	0.2995***	0.040	0.3076***	0.2950***	0.013	0.3526***	0.3140***	0.039
		(0.041)	(0.030)	(0.040)	(0.031)	(0.028)	(0.039)	(0.032)	(0.026)	(0.035)
Average standardized effect health		0.2847***	0.3122***	-0.028	0.2657***	0.3136***	-0.048	0.3179***	0.3620***	-0.044
		(0.037)	(0.031)	(0.039)	(0.031)	(0.035)	(0.042)	(0.039)	(0.040)	(0.047)
Average standardized effect educ.		0.3940***	0.2867***	0.1073*	0.3495***	0.2764***	0.073	0.3734***	0.2852***	0.0883*
		(0.063)	(0.041)	(0.060)	(0.049)	(0.041)	(0.059)	(0.045)	(0.032)	(0.046)

Note: See Notes to Table 3. 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 avg. standardized effects do not include value of uniforms since this variable wasn't pre-specified in the analysis plan. Value of uniforms is coded as 0 if the HH doesn't receive the uniforms.

Table 10: Worker Behavior

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	AVERAGE: Incentive Average Treatment Effect	AVERAGE: Non-Incentive Average Treatment Effect	AVERAGE: Incentive Average Additional Effect
Midwives:										
Hours spent in outreach over past 3 days	3.165 [4.4875]	0.7961* (0.410)	-0.074 (0.337)	0.8700** (0.425)	0.076 (0.389)	0.038 (0.419)	0.038 (0.400)	0.391 (0.299)	0.007 (0.305)	0.383 (0.327)
Hours spent providing public services over past 3 days	13.548 [10.0559]	0.536 (0.608)	-1.1020* (0.594)	1.6380** (0.721)	0.675 (0.619)	0.417 (0.567)	0.257 (0.585)	0.579 (0.460)	-0.248 (0.419)	0.8272* (0.487)
Hours spent providing private services over past 3 days	10.805 [12.5048]	0.212 (0.832)	-0.469 (0.826)	0.681 (0.886)	0.894 (0.674)	0.591 (0.669)	0.304 (0.644)	0.570 (0.525)	0.112 (0.524)	0.458 (0.524)
Total hours spent working over past 3 days	27.518 [15.7132]	1.477 (1.047)	-1.7182* (1.039)	3.1956*** (1.154)	1.6276* (0.951)	0.936 (0.932)	0.692 (0.884)	1.5004** (0.712)	-0.224 (0.728)	1.7246** (0.723)
Number of posyandus attended in past Month	4.166 [3.3213]	0.189 (0.332)	0.059 (0.227)	0.130 (0.348)	-0.162 (0.247)	0.053 (0.268)	-0.215 (0.324)	-0.009 (0.241)	0.064 (0.195)	-0.073 (0.294)
Number of hours midwife per posyandu	3.039 [1.6932]	0.137 (0.130)	0.181 (0.120)	-0.044 (0.127)	0.110 (0.152)	-0.083 (0.133)	0.192 (0.153)	0.127 (0.111)	0.032 (0.095)	0.095 (0.111)
Teachers:										
Percent present at time of interview (primary)	. [.]	0.006 (0.016)	0.016 (0.015)	-0.010 (0.017)	0.000 (0.011)	0.008 (0.011)	-0.009 (0.012)	0.006 (0.016)	0.016 (0.015)	-0.010 (0.017)
Percent present at time of interview (junior secondary)	. [.]	0.001 (0.015)	-0.010 (0.014)	0.011 (0.014)	-0.008 (0.012)	-0.015 (0.012)	0.007 (0.013)	-0.004 (0.010)	-0.013 (0.010)	0.009 (0.010)
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)	-0.005 (0.028)	-0.028 (0.029)	0.023 (0.028)
Percent observed teaching (j. sec.)	. [.]	-0.069 (0.044)	-0.052 (0.047)	-0.018 (0.049)	0.039 (0.049)	0.024 (0.048)	0.015 (0.044)	-0.010 (0.033)	-0.011 (0.033)	0.002 (0.032)
Puskesmas:										
Minutes wait at recent health visits	25.201 [23.7360]	0.778 (3.637)	6.035 (4.685)	-5.257 (3.953)	2.409 (4.269)	1.281 (4.224)	1.128 (4.400)	1.696 (3.033)	3.042 (3.302)	-1.345 (3.320)
Percent of providers present at time of observation	. [.]	0.0714** (0.036)	0.1090*** (0.039)	-0.038 (0.035)	-0.009 (0.029)	-0.0757** (0.030)	0.0667** (0.030)	0.030 (0.022)	0.006 (0.023)	0.024 (0.024)
Average standardized effect		0.045 (0.029)	-0.040 (0.028)	0.0846*** (0.032)	0.043 (0.028)	0.021 (0.028)	0.021 (0.030)	0.0409* (0.022)	-0.005 (0.020)	0.0463* (0.024)
Average standardized effect health		0.0892** (0.044)	-0.036 (0.038)	0.1250*** (0.048)	0.056 (0.040)	0.030 (0.039)	0.026 (0.040)	0.0665** (0.031)	0.000 (0.028)	0.0662** (0.034)
Average standardized effect educ.		-0.022 (0.043)	-0.046 (0.044)	0.024 (0.047)	0.023 (0.041)	0.009 (0.042)	0.014 (0.042)	0.002 (0.030)	-0.014 (0.029)	0.016 (0.030)

Note: See Notes to Table 3.

Table 11: Community effort

		Wave II			Wave III			AVERAGE		
Indicator	Baseline Mean	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Treatment Effect	Non-Incentive Treatment Effect	Incentive Additional Effect	Incentive Average Treatment Effect	Non-Incentive Average Treatment Effect	Incentive Average Additional Effect
Community effort at direct service provision:										
Number of posyandus in village	4.5191	-0.092	0.004	-0.096	0.128	0.196	-0.068	0.027	0.107	-0.080
	[3.5043]	(0.124)	(0.147)	(0.126)	(0.178)	(0.176)	(0.148)	(0.140)	(0.151)	(0.120)
Number of posyandu meetings in	.	-0.003	0.082	-0.084	-0.112	-0.063	-0.049	-0.061	0.002	-0.063
past year at selected posyandu	[.]	(0.102)	(0.111)	(0.102)	(0.112)	(0.091)	(0.100)	(0.079)	(0.076)	(0.078)
Number of cadres at posyandu	.	0.174	0.197	-0.023	0.2890**	0.3577**	-0.069	0.2349**	0.2854**	-0.051
	[.]	(0.113)	(0.153)	(0.138)	(0.139)	(0.171)	(0.165)	(0.105)	(0.139)	(0.133)
Community effort at outreach										
Number of sweepings at selected	.	-0.296	0.042	-0.338	-0.127	-0.6155*	0.4888*	-0.186	-0.337	0.150
posyandu in last year	[.]	(0.394)	(0.377)	(0.389)	(0.342)	(0.346)	(0.295)	(0.266)	(0.257)	(0.257)
Number of primary school comm..	.	0.066	-0.070	0.136	0.002	-0.125	0.126	0.031	-0.099	0.130
meetings with parents in past year	[.]	(0.133)	(0.133)	(0.121)	(0.181)	(0.182)	(0.137)	(0.117)	(0.119)	(0.093)
Number of junior sec. school	2.3093	-0.121	0.032	-0.153	0.213	0.210	0.003	0.066	0.125	-0.060
committee meetings w parents	[1.9728]	(0.113)	(0.118)	(0.126)	(0.147)	(0.223)	(0.207)	(0.103)	(0.147)	(0.140)
Community effort at monitoring										
Number of primary school	.	0.7613*	-0.503	1.2638***	-0.003	0.195	-0.198	0.317	-0.085	0.401
committee members	[.]	(0.392)	(0.410)	(0.478)	(0.334)	(0.402)	(0.344)	(0.287)	(0.314)	(0.297)
Number of junior sec school	8.2592	-0.845	-1.421	0.577	0.199	0.231	-0.032	-0.296	-0.511	0.215
committee members	[4.7625]	(0.993)	(0.934)	(0.539)	(0.332)	(0.332)	(0.291)	(0.498)	(0.475)	(0.297)
Number of prim. school committee	.	-0.124	-0.367	0.243	-0.121	-0.096	-0.025	-0.129	-0.213	0.084
meetings with teachers in past year	[.]	(0.358)	(0.357)	(0.354)	(0.316)	(0.319)	(0.268)	(0.255)	(0.252)	(0.211)
Number of j. sec. school committee	4.4761	0.477	0.132	0.345	0.530	0.5755*	-0.045	0.4957*	0.381	0.115
meetings with teachers in year	[5.4650]	(0.424)	(0.394)	(0.455)	(0.342)	(0.346)	(0.364)	(0.262)	(0.258)	(0.269)
Average standardized effect		0.014	-0.009	0.023	0.0431*	0.048	-0.004	0.026	0.017	0.010
		(0.022)	(0.025)	(0.023)	(0.025)	(0.031)	(0.029)	(0.018)	(0.022)	(0.019)

Note: See Notes to Table 3.

Table 12: Within-subdistrict targeting

		Wave II			Wave III			AVERAGE		
Indicator	Baseline Mean	Generasi Incentive Top 3 Quintiles Additional Effect	Generasi Non-Incentive Top 3 Quintiles Additional Effect	Generasi Incentive Additional effect Top 3 Quintiles Additional Effect	Generasi Incentive Top 3 Quintiles Additional Effect	Generasi Non-Incentive Top 3 Quintiles Additional Effect	Generasi Incentive Additional effect Top 3 Quintiles Additional Effect	Generasi Incentive Top 3 Quintiles Additional Effect	Generasi Non-Incentive Top 3 Quintiles Additional Effect	Generasi Incentive Additional effect Top 3 Quintiles Additional Effect
Panel A: Targeting of direct benefits										
Average standardized effect		-0.036	0.032	-0.068	-0.076	-0.059	-0.017	-0.065	-0.029	-0.036
		(0.212)	(0.167)	(0.268)	(0.098)	(0.130)	(0.155)	(0.087)	(0.088)	(0.110)
Average standardized effect health		-0.050	0.070	-0.120	-0.060	-0.061	0.002	-0.071	-0.022	-0.049
		(0.296)	(0.256)	(0.390)	(0.166)	(0.234)	(0.273)	(0.131)	(0.153)	(0.179)
Average standardized effect educ.		-0.019	-0.016	-0.003	-0.093	-0.057	-0.036	-0.059	-0.036	-0.022
		(0.220)	(0.103)	(0.231)	(0.084)	(0.079)	(0.103)	(0.090)	(0.065)	(0.101)
Panel B: Heterogeneity in improvements in main indicators										
Average standardized effect		-0.075	0.007	-0.082	0.062	0.054	0.008	-0.017	0.031	-0.048
		(0.065)	(0.074)	(0.085)	(0.075)	(0.066)	(0.068)	(0.847)	(0.773)	(0.922)
Average standardized effect health		-0.078	0.057	-0.135	0.133	0.038	0.095	0.020	0.043	-0.024
		(0.078)	(0.080)	(0.092)	(0.097)	(0.087)	(0.091)	(0.731)	(0.632)	(0.752)
Average standardized effect educ.		-0.068	-0.093	0.025	-0.078	0.087	-0.1651*	-0.090	0.006	-0.097
		(0.096)	(0.129)	(0.143)	(0.084)	(0.091)	(0.084)	(1.234)	(1.193)	(1.397)

Notes: For each indicator, the regression interacts the Generasi treatment variables for a dummy for a household being in the top 3 quintiles of the baseline per-capita consumption distribution. Average standardized effects for the interaction with the top 3 quintiles variable are shown in the table. Panel A examines the indicators of direct benefits shown in Table 9 and Panel B examines the 12 main program indicators examined in Table 3.

Table 13: Cost-effectiveness Calculation

	Generasi with Incentives	Generasi without Incentives	Additional effect of incentives	Conditional Cash Transfer (PKH) (no spillover)	Conditional Cash Transfer (PKH) (w. spillover)	Generasi with Incentives	Generasi without Incentives	Additional effect of incentives	Conditional Cash Transfer (PKH) (no spillover)	Conditional Cash Transfer (PKH) (w. spillover)
Panel A: Social Cost Effectiveness	Entire program									
Transfers	0.00	0.00	0.00	0.00	0.00					
Non-transfers	3.91	3.68	0.23	0.00	0.00					
Facilitation	2.54	2.54	0.00	18.40	18.40					
Marginal cost public funds	5.07	5.07	0.00	32.04	32.04					
Total costs (millions USD)	11.51	11.28	0.23	50.44	50.44					
Millions of points	1.42	1.04	0.373	2.24	4.46					
Dollars per point	8.13	10.81	0.62	22.43	11.30					
Panel B: Social Cost Effectiveness by Area	Health					Education				
Transfers	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Non-transfers	2.24	1.97	0.28	0.00	0.00	1.63	1.68	-0.05	0.00	0.00
Facilitation	1.27	1.27	0.00	21.72	21.72	1.27	1.27	0.00	21.72	21.72
Marginal cost public funds	2.87	2.77	0.10	16.02	16.02	2.95	3.05	-0.10	16.02	16.02
Total costs (millions USD)	6.39	6.01	0.38	37.74	37.74	5.85	6.00	-0.15	37.74	37.74
Millions of points	0.96	0.67	0.291	2.25	4.47	0.46	0.38	0.081	0.00	0.00
Dollars per point	6.66	9.01	1.30	16.78	8.45	12.78	15.93	N/A	N/A	N/A

Notes: Note that the costs and points for Generasi have been divided by 2, so that in this calculation exactly half the benefits and costs have been allocated to the program with and without incentives. The estimated points are therefore 50% of the estimated numbers in Table 3 above. PKH calculations are authors calculations based on the coefficients given in Alatas et. al (2010), as well as authors' calculations of the average number of beneficiaries of different age ranges per PKH household based on the PKH wave 3 survey. For health and education, we allocate the facilitation costs and PKH transfers 50-50 between health and education, and allocate actual Generasi expenditures based on the actual distribution of expenditures between health and education in the MIS data.