

Which incentives work?

An experimental analysis of incentives for trainers

OMAR AZFAR AND CLIFFORD ZINNES¹

The IRIS Center
Department of Economics, University of Maryland University at College Park

Abstract

One conjecture in the theory of incentives is that incentives based on broader outcomes may be better at motivating agents than incentives based on narrow measures. We designed an experiment to test these hypotheses using a “prospective randomized evaluation procedure” (PREP). We then apply PREP to training programs as typically funded by donors of economic development assistance. We randomly assigned 274 participating entrepreneurs in the Philippines to one of 26, simultaneous, one-day, training classes in marketing. Trainers were given cash incentives based on the average score of their “students” on a standardized test containing an alternative number of questions, which were randomly assigned to each class. We then examined outcomes based on student satisfaction ratings of the trainer. Our results suggest that incentives based on broad outcomes are more effective than incentives based on narrow outcomes. We conclude with ways to improve our approach as well as with a discussion of the implications for using prospective randomized evaluation for improving the evaluation of donor projects.

JEL Codes: H43, C93, I20, L33, O22

Key words: randomized trials, project evaluation, teacher incentives, aid effectiveness

¹ Omar@iris.econ.umd.edu, Zinnes@iris.econ.umd.edu. Corresponding author: Omar Azfar, The IRIS Center, 2100 Morrill Hall, University of Maryland, College Park, Maryland 20742, USA. Fax: 1-301-405-3020. We thank Peter Murrell for his comments on an earlier draft. All errors are our own. We gratefully acknowledge the generous support of USAID/EGAT and Fred Witthans for funding this research under Task Order 7, SEGIR/LIR PCE-I-00-97-00042-00 to the IRIS Center at the University of Maryland.

1 Introduction

It is widely conjectured (e.g., Laffont and Tirole 1993) that perverse incentives are the cause of underperformance at the level of the individual, the firm and the economy. The theoretical work on incentives has highlighted how incentives based on indicators, when not closely related to the objectives that matter most, can be distracting rather than motivating (Holmstrom and Milgrom 1991). These concerns about incentives for teachers were voiced even before the theoretical advances by Murnane and Cohen (1986). Our results suggest that, as theory predicts, incentives do have an effect on outcomes, but incentives based on narrow outcomes can be ineffective or even counterproductive.

The statistical study of the impact of incentives, however, is limited by the fact that incentives vary endogenously and most analyses of *ex post* data suffer from problems caused by omitted variable bias, simultaneity bias and selection bias. For example, individuals destined to succeed may adopt better incentives than individuals less likely to do well.² Also there may simply not be enough institutional variation across units to examine some interesting hypotheses.

While these observations are relevant in a wide variety of applications, the present authors are particularly interested in the implications for technical assistance provided by bilateral and multilateral donors. On the one hand, in recent years donors have become concerned that their assistance has not always been effective and, worse, that they don't have the tools to assess their impact.³ On the other hand, the amount of donor funding has been falling in real terms, making it more important than ever to have a way to evaluate aid performance for future improvement.

In recent years, artificial experimental methods have entered the mainstream of economic analysis, with experimental papers now appearing regularly in all the major journals (e.g., Glaeser *et al.* 2000). Indeed experimental economics has forced economists to rethink their cherished assumptions of rationality (e.g., Shiller 2004, Thaler 2000), and led to a *gestalt* shift in economic thought. In this paper, we present an experimental test in the context of donor training programs

² By "adopt" we mean they may choose sharper incentive schemes within their firms, if the opportunities exist, or they may choose firms with sharper incentives schemes.

³ There are signs that this situation may be changing. The Millennium Challenge Corp, set up by President Bush, has recently initiated a massive effort to introduce randomized trials as a means of evaluating its country programs (MCC 2005).

of the hypothesis that incentives only improve outcomes when based on broad measures of performance.

We begin by describing what we call a “Prospective, Randomized Evaluation Procedure” (PREP) approach to evaluating donor development assistance.⁴ We then present an application of PREP to technical assistance for the simple case of marketing workshops for small- and medium-scale enterprise (SME) entrepreneurs in the Philippines. The central question was the whether the effectiveness of trainers was increased by providing them cash incentives. Incentives are presumed to have an effect on behavior but there are several concerns about the effect of incentives on teachers and trainers, or for that matter any activity where the true value of outcomes is difficult to measure (Holmstrom and Milgrom 1991, Prendergast 1999; Dixit 2001, Azfar 2002). To address these concerns we offered trainers financial incentives based on an objective measure of outcomes and compared their performance across groups.

More recently, Eberts, Hollenbeck and Stone (2000) have shown that a school that introduced merit pay based on student retention, improved its performance as measured by retention rates, but may have undermined performance more broadly defined. Their paper, however, involves the comparison of only two schools, making any clear inference difficult. Our results, based on 26 training classes using an identical syllabus but trainer-varying incentives, suggest that, as theory predicts, incentives do have an effect on outcomes but, if inappropriately designed, can be ineffective or even counterproductive. Our results resonate with the concurrent work done by Glewwe *et al.* (2003) who find that in Kenya the provision of monetary incentives to teachers only led to short-term improvements in student performance.

Finally we offer some methodological suggestions on how to conduct better a study like this one in a way that produces clearer results in the future and then make some policy recommendations for funders of development assistance.

2 Evaluating the impact of trainer incentives

To test our hypothesis that incentives only improve outcomes when based on broad measures of performance we conducted a prospective random evaluation procedure (PREP) in The Philippines. As its name implies, PREP treats the evaluation of technical assistance in the same way

⁴ This approach was inspired by the work of Kremer (e.g., 2003; 2001, with Miguel, E).

that a pharmaceutical company might use clinical trials to assess drug effectiveness.⁵ There are several challenges for its application to donor interventions (ignoring the obvious need to convince a donor of the efficacy of the approach to begin with), however. Among these is, first, the difficulty of structuring the application's treatment design to be amenable to randomization. In the case of decentralization, freedom of information reform, or tax reform, for example, politics and equity⁶ if not practicality make it difficult to simply divide the country into a sufficient number of groups receiving varied treatments. Nonetheless, it is clear that the two insights of properly designed control groups and prospective randomization could be powerful techniques to evaluate the effectiveness of donor interventions.

As a concrete test of the PREP approach and one where the above concerns were easy to avoid, we picked the field of donor-funded training. Donors have been paying for training under the name of capacity building and human capital investment for decades and for a wide variety of topics, both economic and political. The training we selected was of SME managers in strategic marketing.⁷ Our evaluation was focused on the impact of trainer incentives on student outcomes. We offered different groups of trainers different incentives and compared outcomes across groups using a variety of alternative indicators.

2.1 Theoretical background and implementation design

We begin by providing some theoretical background for this exercise in terms of the theory of incentives, and the relevance of these ideas for donors of development assistance.

Theory and relevance of performance-based incentives. That monetary incentives matter and have an effect on behavior follows easily from the assumptions of rational choice and utility maximization. The early formal work on the theory of incentives showed that incentives could improve performance and empirical work on the provision of incentives in firms showed that incentives did improve performance.⁸ However, what these theoretical and empirical studies had in common was that the “principal” could observe something closely related to the principal's interest—in the present example the “principal” would be the aid donor of the trainings and the

⁵ For an introduction to randomized trials, see Greenberg *et al.* (2004) and on how to apply PREP, see Azfar and Zinnes (2003).

⁶ In fact, there is the analogous concern of withholding treatment of a potentially effective experimental drug from a sub-group where it is believed with high probability that the drug would be effective.

⁷ Espina and Zinnes (2003) find that USAID funded at least 600 entrepreneur training projects in the 1990s alone.

⁸ See Prendergast (1999) for a survey.

observable outcome would be the quality of the trainer's teaching. Changing the premise that outcomes are observable can call the value of incentives into question since one can no longer be certain what effect—positive or negative—the incentives are having.

Subsequent theoretical work has in fact shown that if the principal cannot observe whether his objectives are being maximized, then incentives can be distracting rather than motivating (Holmstrom and Milgrom 1991), i.e, lead to unwanted side-effects. The issue has also been highlighted in policy debates on subjects like education where observable variables like scores on standardized tests are thought by many experts to distract rather than motivate teachers. Teachers may “teach for the test” in response to such incentives, they may encourage students to concentrate on doing well on items the teachers’ incentives are based on, and, in extreme cases, may even lead teachers to turn a blind eye toward a cheating student (Jacob and Levitt 2003).

The theory of incentives also suggests that incentives based on broader measures of performance are more likely to motivate “agents” (the trainers in the present case) toward the objectives of their principal than incentives based on narrower measures of performance. To test this theory we offered trainers incentives based on outcomes of different breadth. How this was done is described in the next section. From a policy perspective, we are also interested in what type of trainer incentives should donors use to maximize the effectiveness of their aid.

Experimental design of the PREP application to trainers. The trainings were conducted by the Institute for Small Scale Industries (ISSI) at the University of The Philippines. ISSI routinely conducts trainings for SME managers and proved to be a good partner. We chose strategic marketing as the subject to be taught because ISSI indicated it was a popular topic and would make it easier to attract course participants.

The training was done in two stages. First, ISSI recruited 30 trainers from a large applicant pool. These trainers were subsequently trained to train the SME managers.⁹ All trainers took an exam at the end of this training. Over the subsequent week, trainers were asked to give trial presentations of how they would teach from a common syllabus and their performance as

⁹ We were fortunate that ISSI was able to hire Dr. Felix Lao, the author of a widely used textbook on strategic marketing in The Philippines, to train the trainers.

teachers was graded by ISSI staff.¹⁰ We would use later this exam score and trial presentation grade as two controls when assessing incentive effects on their performance. Finally, though we began with 30 trainers, one trainer left during the training and another did not show up for the trial training. Both were taken out of the sample. Of the remaining 28 trainers, we removed the one who got the lowest score, leaving us with 27 trainers.

The 27 trainers were randomly placed into one of six incentive groups according to Table 1. Eighteen trainers were provided incentives to teach a one-day class where the incentive was based on the percentage of responses the students in their class would answer correctly on an exam to be administered at the end of the training. Of these eighteen trainers, the performance of six would be based on an exam of 20 questions, six would be based on an exam of 40 questions, and six would be based on an exam of 80 questions. The possible monetary payoffs to trainers depended on the average exam score of the students in their class as listed in Appendix B and varied from 0 to 10,000 Pesos (approximately 0 to 200 U.S. dollars). Eight trainers did not have incentives based on their students' performance. These trainers got a fixed, extra payment of 3,000 Pesos, which we guessed would be the average amount of the incentive payment.¹¹ One of these eighteen trainers failed to show up on the day of the training, reducing the number of trainers in Group 1 by one trainer.

Table 1: Randomization matrix for PREP Training Study

<i>Group</i>	<i>Incentives</i>	<i>Number of questions on test</i>	<i>Number of classrooms/trainers</i>
1	No	20	2*
2	No	40	3
3	No	80	3
4	Yes	20	6
5	Yes	40	6
6	Yes	80	6

*In fact, 3 had been planned, but one trainer did not show up.

One week before the training, we mailed each trainer their incentive package and the exam questions their students had to answer. Having this material in advance allowed trainers time to take the incentive scheme into account while preparing to teach their sessions. Thus, our

¹⁰ A set of specific criteria, each graded from 1 to 5, was used to minimize the subjectivity of these teaching performance evaluations.

¹¹ It turned out we underestimated the average incentive payment by around 25 percent.

conjecture was that trainers who were given incentives based on the answers to 20 questions would likely plan to teach in a way that communicated the answer to the twenty questions, but probably without creating a broader understanding of the material. Trainers who were assigned 80 questions, on the other hand, faced the choice of either communicating the answers to 80 questions or conveying a broad understanding of the material. Given the greater difficulty of making their students learn and remember 80 answers (as compared to 20 answers), they may have preferred developing an understanding of the material. Trainers were asked not to discuss any information regarding their incentive package or the exams they were to teach to with other trainers to minimize spillover effects (see Appendix B). The likelihood of this was low to begin with since trainer selection was such that trainers would have never met prior to the day of the training.

For the second step, ISSI advertisements and direct marketing succeeded in registering several hundred course participants. In the event, a total of 274 participants comprising SME owners or senior managers showed up for the training. They were randomly placed into 26 classes of approximately equal size. All of these “students” provided some basic information and took a pretest of 20 questions. Students then had a five-hour training on strategic marketing (with a lunch break). This was followed by a multiple-choice test of varying length. Students got time proportional to the number of questions they had to answer. All students then answered two essay questions. Finally, we asked several questions about their rating of the trainer as a teacher, whether the training would change how they conducted their business, and whether they would be willing to pay for subsequent trainings.

2.2 Measures of outcomes

The success of an evaluation depends on the construction of the appropriate indicators of outcomes and performance. The trainers were graded prior to giving their class on their answers to the multiple choice test, essay questions, and graded on teaching ability by ISSI experts. A composite measure of these scores was used as the trainer quality variable.¹² We used several different indicators to measure training outcomes. These included student scores on: multiple choice tests, essay questions, satisfaction ratings of the trainer, whether the training changed the way the students would do business, and the willingness to pay for subsequent trainings.

Our experience as students and educators has been that the quality of teaching significantly affects learning and the value of the lesson. An obvious way to check for the validity of trainer performance variables is to correlate the performance variables with the trainer score. We calculated the bivariate correlations of the performance variables with the trainer quality variable, and also ran robust, univariate regressions to examine if these correlations were robust (Table 2). In each case only the satisfaction rating variable showed a correlation with the trainer quality variable. We also found that the satisfaction ratings were more correlated with components of the trainer score variable than were the other performance variables.¹³ For this reason, we only used satisfaction ratings as a performance variable when investigating the determinants of trainer performance. The other performance variables seem to contain sufficient noise so that even variation in trainer quality does not have a statistically measurable effect on them.¹⁴ This makes it *a priori* less likely that a variable, like the incentives we offer the trainers, would have a statistically measurable impact on these indicators of performance.

The satisfaction training variable is the composite score of the questions we asked the students to determine their satisfaction of their trainer. The questions themselves and the composite indicator are presented in Appendix A. These questions are similar to satisfaction ratings used in American universities. We had some initial concern about the relevance of the questions to the Philippines context but found that ISSI actually uses a similar approach. Questions

¹² The variable, *Trainer Quality* is a standardized, composite indicator, created by first standardizing and then averaging the trainer's scores in the multiple-choice and essay tests and the mini-lecture. Scores were collected before the teaching sessions.

¹³ These results are not shown in to save space but are available from the authors.

¹⁴ As discussed below, had this level of noise been anticipated, it could have been partly corrected by setting a larger sample size. Normally, such problems can be avoided by administering a pre-test to establish *inter alia* the adequate sample size. Unfortunately, time limitations for the application prevented this step from being implemented.

addressed how clearly the lecturer spoke, whether she was well-prepared, etc. and a general question asking for the overall rating of the lecturer. As expected, all the indicator components were significantly correlated with each other. An aggregate score based on an average of all these ratings is used as one measure of trainer performance.

Table 2: Correlation of performance variable with trainer quality.

<i>Performance indicator</i>	<i>Correlation coefficient and P-value</i>	<i>Significant*</i>	<i>Robust regression t-stat and P-value</i>	<i>Significant*</i>
Satisfaction rating	0.13 0.09	Yes	1.87 0.06	Yes
Average multiple-choice score	0.04 0.46	No	-0.62 0.58	No
Average essay score	0.07 0.21	No	1.08 0.28	No
Training will change business practices	0.11 0.21	No	0.11 0.91	No
Willingness to pay (fixed value)	0.07 0.35	No	Did not converge ¹⁵	No
Willingness to pay (maximum value)	0.01 0.82	No	-0.30 0.76	No

Regressions are all univariate. *At the 90-percent level of confidence.

2.3 The determinants of trainer performance

We now begin our examination of the determinants of the performance variables that seem to capture the quality of teaching: satisfaction ratings. But first let us recap the theoretical predictions.

Theory predicts that incentives can improve performance but may also encourage trainers to try to “game the outcome” by teaching to the tests, which, recall, were provided to the trainers in advance and contained from 20 to 80 questions. Students may also find they are not learning useful material for their businesses and give poor satisfaction ratings for trainers. This would be particularly true for trainers whose incentives were based on 20 questions. Trainers, whose incentives were based on their students’ answers to 80 questions, may find it best to teach well and impart useful knowledge, rather than focus on the answers to specific questions. This is

¹⁵ A robust regression down-weights outliers and re-estimates regressions based on the new weights. It then identifies outliers from the new weighted regression and re-estimates the weights. This iterative process sometimes does not converge, as in this instance.

because the best strategy to impart the knowledge on 80 topics may be to provide broader-based knowledge, rather than attempting to cram 80 “facts” into a student’s head in a short period.

We now examine the determinants of trainer performance as measured by the satisfaction rating. The results are presented in Table 3. Ordinary least squares would mis-estimate the standard errors on the group level variables if, as is likely, the error terms were correlated across groups. The estimation method we use is regression with the cluster command in STATA. This method controls for correlated error terms within groups and calculates robust standard errors. As student’s ability may affect their enjoyment of the course, all regressions control for the student’s pretest score. The variable is generally insignificant. Omitting it does not substantially change the results.

Model 1 in Table 3 is a simple regression of the satisfaction rating on whether or not the trainer had incentives. The coefficient is positive but insignificant, allowing no clear inference on the effect of incentives on the quality of teaching. This result remains valid in Model 2 when we control for gender of the student and whether the student owns his firm.¹⁶ This result seems to give credence to the frequently aired concern that giving incentives based on multiple choice questions would simply induce trainers to “teach to the test”. The students—the SME managers who had come to take the course to improve their business practices—may, therefore, be dissatisfied with such narrowly focused instruction. We had anticipated this issue and planned to test for this concern. Our strategy was to offer incentives on outcomes of different breadths and we, therefore, gave different trainers incentives based on 20, 40 or 80 questions (see Table 1). This allows us to test whether broader-based incentives are more motivating.

In Models 3 and 4 we test the role of the broadness of the outcome measure on the effectiveness of incentives by varying the number of questions on the trainer’s students’ exam. These models include a dummy variable for the presence or absence of trainer incentives, the number of questions, their interaction term, and trainer quality. Variables are transformed to allow an easier interpretation of the coefficients.¹⁷

¹⁶ Other variables like age were tried but turned out to be statistically insignificant and led to a large reduction in the number of observations.

¹⁷ Variables are transformed in the following way. *Incentives* is a dummy variable equal to 1 if incentives were given and 0 if not. *Questions* is created using the linear transform, $(Examsize-40)/20$, where *Examsize* indicates the number of questions (20, 40 or 80) in the multiple-choice exam the student received. This makes interpreting the coefficient easier—because it has round number values of -1 at 20 questions, 0 at 40 questions, and 2 at 80 questions. The variable, *Trainer Quality* is described in Footnote 12.

Here, we see that increasing the number of questions has a clear negative effect. One possible interpretation of this is that students do not like answering 80 questions and let their frustrations out when giving their satisfaction ratings of the trainer. Since educators or policy makers are typically not concerned about these short-term frustrations, if this was the reason for the question effect, then it can be discounted. We proceed by making this “frustration” assumption and will turn to other interpretations later.

Table 3: The determinants of trainer performance, as measured by satisfaction ratings

Dependant variable ¹ : Composite satisfaction rating (Mean= 24.62, Standard deviation= 3.20)				
Name of explanatory variable	Model ¹			
	1	2	3	4
Constant	24.1	23.7	26.0	26.4
Incentives	0.775 (1.03)	0.581 (0.81)	0.321 (0.79)	0.157 (0.40)
Questions			-1.28** (8.77)	-1.27** (7.91)
Incentives*Questions			0.586* (2.25)	0.597* (2.26)
Additional student controls ²	No	Yes	No	Yes
Number of observations	172	162	172	162
Adjusted R-squared	0.02	-0.01	0.13	0.10

Notes: Estimation of regression has standard errors adjusted for within-group correlations. Absolute value of robust *t*-statistics are below coefficients. * Significant at 5%; ** Significant at 1%. (No coefficient is significant at 10% level).

¹ All regressions include the student’s pre-training score and control for trainer quality as described in Footnote 12.

² Additional student controls are gender and whether the student is the firm owner.

³ Higher values are better.

The term of interest in these models is the interaction term of incentives and the number of questions. This term is positive and significant ($P=0.033$), suggesting that, regardless of the interpretation of the direct effect of the trainer’s performance maximand, incentives based on more questions are in fact more effective at raising student’s satisfaction ratings than incentives based on fewer questions. The transformation of these variables allows us to easily interpret the coefficient on the interaction term. At 40 questions, the impact of incentives is simply the coefficient on incentives (an insignificant 0.32). At 20 questions, the impact is $0.32-0.59=-0.27$ (also insignificant). At 80 questions, the impact is $0.32+2*0.59=1.50$ (significant $P=0.01$). It is worth

re-emphasizing the point that trainers knew the 20, 40, and 80 questions, respectively, a week before the teaching session, and so could have planned to teach for the test.

We also ran another set of regressions of satisfaction ratings on incentives separately for groups which answered 20, 40 and 80 questions. These regressions are presented in Table 4. There was no significant effect of incentives on performance when measured by the student's answers to 20 or 40 questions. For groups with 80 questions, incentives have a positive and significant effect on satisfaction ratings. The magnitude of the effect is not small. Incentives improve performance by 40 percent of the standard deviation of the satisfaction rating measure.

Table 4: The determinants of trainer performance, as measured by satisfaction ratings.

Dependant variable: ³ Composite satisfaction rating (Mean= 24.62, Standard deviation= 3.20)						
Explanatory variable	Model ²					
	1	2	3	4	5	6
Number of questions	20	40	80	20	40	80
Constant	29.266	27.743	21.333	29.283	25.185	24.106
Incentives	-2.041 (1.27)	-0.5 (0.99)	1.428 (3.26)*	-1.851 (1.15)	-0.775 (1.33)	1.376 (2.79)*
Additional student controls ¹	No	No	No	Yes	Yes	Yes
Number of observations	58	54	60	55	52	55
Adjusted R ²	0.04	0.11	0.09	0.01	0.07	0.01

Notes: Regression estimated with standard errors adjusted for within-group correlations. Absolute values of robust *t*-statistics are below coefficients. * Significant at 5%; ** Significant at 1%. (No coefficient is significant at 10% level)

¹ All regressions include student's pre-training score. Additional student controls are gender and whether the student is the firm owner.

² All regressions control for trainer quality as assessed in the mini lecture and the trainer's score on exam questions.

³ Higher values are better.

How can we interpret these results? It appears that incentives do improve performance but only if based on 80 questions. This accords with the modern view that incentives can be useful, but only if based on a broad measure of success.

As always, some caveats bear making. First, it is important to remember that *if* the negative coefficient on the number of student exam questions is interpreted differently, for instance, as “teachers teach badly if they feel they have to cover too much material”, then the results presented here would have to be reinterpreted. In this case, an alternative trainer incentive maximand would be more appropriate. For these and other reasons, it is evident with the benefit of hindsight that there is scope for refining our methodology in order to get clearer results. Second,

a larger sample size at the level of each group would have allowed for more subtle testing and to detect weaker incentive effects. The next section explores the lessons we have learned from this PREP application that will help refine future such studies.

3 Lessons learned and conclusion

The first predictable lesson is “collect more data”. While this is always true, we seem to have erred on the side of collecting too little data in the sense that the marginal usefulness of more data would have quite high. We did in fact try to collect more data. We had specified a larger number of course participants than eventually appeared on the day of the training (450 rather than 274). Our local collaborator, ISSI, tried hard with phone calls and advertisements in newspapers and even television to increase enrollments but with limited success. Likewise, budgetary and time restrictions prevented our conducting another set of trainings in other major Philippine cities.

We also erred by assuming that there would be less individual-level “noise” than there actually was in all the variables other than the multiple choice scores. Normally pre-tests could have alleviated this risk by helping to definitively establish the necessary sample size. The presence of these individual-level factors, together with the small number of observations in each classroom, made all the performance variables, other than satisfaction ratings, noisy indicators of the quality of teaching. Had we been able to anticipate this problem—say, from a pre-test, we would have conducted one or two trainings where the entire population of students was taught by the same trainer. Responses on the willingness to pay for more trainings from this trainer, satisfaction ratings of this trainer’s performance, and statements about whether these trainings would affect the way business would be conducted could be used to control for (purge) individual student-level effects. However, we should not overstate the expected effectiveness of such a method. For example, we did anticipate this problem for the test-scores variable and gave the students an exam prior to conduct the trainings. Cleaning the post-training score using the pre-training exam score did not significantly improve the performance of the test-score variable.

Another lesson relates to the appropriate incentives to motivate trainers. We found that some of the trainers were motivated by the desire to work for ISSI or USAID in the future—even though we had explicitly stated this was a one-off project, and there were no future plans for col-

laboration with ISSI.¹⁸ Such motivations could easily drown out the monetary incentives we had provided. Several possible solutions could be considered to remedy this problem. First, one could provide much larger financial incentives. Second, one could employ an experimental design that is semi-blind so as to prevent (in the present case) the experimenter (i.e., the authors) and ISSI from learning the trainers' true performance. Finally, one could have required ISSI to sign an agreement not to hire any of the trainers for say three years. In the latter two cases, the strategy is to convince the trainers *ex ante* that the authors and ISSI have credibly pre-committed to having no future plans to collaborate with them.

We should, perhaps, also have had a group of trainers rewarded on the basis of scores on essay questions and multiple choice tests. This would have substantially broadened the basis for incentives. A comparison between this group and the other groups would have been instructive. Unfortunately, funding did not permit this option. There are other methodological improvements we would have liked to use. The essay question should have been temporally administered randomly, either before or after the multiple-choice questions to control for the "test fatigue" effect. Likewise, several student controls measured before the trainings would have been helpful, such as whether students had previously participated in training programs (with or without donor sponsorship), years in business (which would also proxy for age), and expected willingness to pay for the course.

In terms of policy advice on offering incentives, our results suggest that only incentives based on broadly defined outcomes are likely to motivate better outcomes. This corresponds with theoretical conjectures, concerns expressed in the education literature, and concurrent empirical work. However, we must end on a note of calling for more research on this subject before definitive policy-reform conclusions can be drawn. More studies structured with the improvements we have suggested here are needed in more countries to develop a clear understanding on how explicit monetary incentives motivate teachers and trainers.

On a more general note, there is an enormous push in the donor community to use indicators of performance and more rigorous techniques of monitoring and evaluation. Surprisingly, many of these laudable efforts appear to overlook the need for properly controlled experimental design. The apparent excuse is often either that the nature of the intervention only permits an

¹⁸ USAID was the funder and has a long-term presence in The Philippines. Perhaps, it was not credible for the trainers to believe that ISSI would not wish to contract them in the future, if they were good.

inadequately small number of treatment groups or a belief that it is unethical to “withhold” treatment (in the case of the control groups). The results of the PREP application illustrated herein suggest PREP to be a rather promising approach for those donors who are serious about project evaluation. At the very least, PREP encourages donors to engage in better project design and to build data collection into project execution *ex ante*.

References

- Azfar, O. (2002).** “The NIE approach to economic development: An analytical primer”, *IRIS Discussion Papers on Institutions and Development*, College Park, Maryland, USAID SEGIR/LIR Task Order 7.
- _____ **and C. Zinnes (2003).** “Success by Design: The Science of Self-Evaluating Projects”, *IRIS Discussion Papers on Institutions and Development*, College Park, Maryland, USAID SEGIR/LIR Task Order 7.
- Davidson and Mckinnon (1993).** *Estimation and Inference in Econometrics*, Oxford University Press.
- Dixit, A. (2001).** “Incentives and organizations in the public sector: An interpretative review”, *mimeo*, Princeton University.
- Ebert, R., K. Hollenbeck and J. Stone (2002).** “Teacher performance incentives and student outcomes”, *mimeo*, Department of Economics, University of Oregon.
- Espina, C. and C. Zinnes (2003).** “Incentive Characteristics of Institutional Development,” *IRIS Discussion Papers on Institutions and Development*, College Park, Maryland, F5-2, USAID SEGIR/LIR Task Order 7.
- Glaeser, E., D. Laibson, J. Scheinkman, and C. Soutter (2000).** “Measuring Trust”, *Quarterly Journal of Economics*, 115(III).
- Glewwe, P., N. Ilias and M. Kremer (2003).** “Teacher incentives”, *mimeo Harvard*.
- Greenberg, D., and M. Shroder (2004).** *The Digest of Social Experiments*. 3rd ed. Washington D.C.: The Urban Institute.
- Holmstrom, B. and P. Milgrom (1991).** “Multi-task Principal Agent Analysis: Incentive Contracts, Asset Ownership and Job Design”, *Journal of Law, Economics and Organization* 7(0), pp. 24-52.
- Jacob, Brian A. and Steven D. Levitt (2003).** “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating,” *NBER Working Papers* 9413, National Bureau of Economic Research, Incentive.
- MCC (2005).** “Impact Evaluation Services”, solicitation number MCC-05-RFP-0029, Washington, DC: Millennium Challenge Corporation.
- Miguel, E. and M. Kremer (2001).** “Worms: Education and Health Externalities in Kenya”, *NBER Working Paper* No. 8481, September.
- Kremer, M. (2003).** “Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons”, *American Economic Review*, 93(2), May, pp. 102-106.

- Laffont, J.-J. and J. Tirole (1993).** *A Theory of Incentives in Procurement and Regulation*, Cambridge, MA: The MIT Press.
- Locke, E.A. and G.P. Latham (1990).** *A theory of goal setting and task performance*, Englewood Cliffs, NJ: Prentice Hall.
- Murnane, R. and D. Cohen (1986).** “Merit pay and the valuation problem: Why most merit pay plans fail and a few survive”, *Harvard Educational Review*, **56**(1), pp. 1-17.
- Prendergast, C. (1999).** “The provision of incentives in firms”, *Journal of Economic Literature*, **37**(1), pp. 7-63.
- Shiller, R. (2004).** *The New Financial Order: Risk in the 21st Century*, Princeton, NJ: Princeton University Press.
- Thaler, R. (2000).** “From Homo Economicus to Homo Sapiens”, *Journal of Economic Perspectives*, **14**(1), pp. 133-141.

Appendix A

Table A.1 Means of performance variables (individual questions on student satisfaction ratings of trainer)

<i>Variable</i>	<i>Obs.</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Min</i>	<i>Max</i>
G3. Could you understand the lecture	262	3.59	0.563	1	4
G4. How loudly did the lecturer speak? 1=too softly 4=loud	259	3.65	0.523	2	4
G5. How easy was it to understand what the lecturer spoke about? explanations? 1=difficult...4=easy	191	3.61	0.548	2	4
G6. How well did the lecturer present his/her topics	260	3.42	0.638	1	4
G7. How well did the lecturer conduct class discussion	257	3.40	0.649	1	4
G8. How well did the lecturer satisfy inquiries from participants	260	3.38	0.643	2	4
G13. What is your overall rating of the lecture	257	3.48	0.593	1	4

Appendix B: Trainer incentive payment schedule

1\$=approximately 50 pesos

Per-Capita GDP = 45,490 Pesos in 2001. (Trainers can reasonably be assumed to be in top decile of income earners. In 1997 the top decile earned twice as much as the country average).

For 6 trainers:

Dear Trainer (PERSONALIZE)

We have determined on the basis of your performance in the training on the 10th of November and after examining your resume to offer you some additional compensation. PLEASE DO NOT DISCUSS THIS ADDITIONAL COMPENSATION WITH ANYONE. We have tried to ensure that all trainers will get the same additional compensation in expected terms.

You will be given 3000 Pesos in addition to the \$400 that was agreed between you and ISSI. Of course all compensation will be taxed.

For 18 trainers (3 groups of 6). Some get 20 questions, some 40, and some 80:

Dear Trainer (PERSONALIZE)

We have determined on the basis of your performance in the training on the 10th of November and after examining your resume to offer you some additional compensation. PLEASE DO NOT DISCUSS THIS ADDITIONAL COMPENSATION WITH ANYONE. We have tried to ensure that all trainers will get the same additional compensation in expected terms.

In addition to the \$400 you agreed with ISSI your compensation will depend on the average performance of your class on the enclosed multiple choice questions according to the table below. Of course all compensation will be taxed.

Class performance	Your additional compensation
50% or below	0
51-55%	1000
56-60%	2000
61-65%	3000
66-70%	4000
71-75%	5000
76-80%	6000
81-85%	7000
86-90%	8000
91-95%	9000
96-100%	10000