Joseph George Caldwell, PhD (Statistics)
503 Chastine Drive, Spartanburg, SC 29301-5977 USA
Tel. (001)(864)541-7324, e-mail jcaldwell9@yahoo.com

15 February 2011

Memorandum

To: Dr. Ray Struyk
Subject: New power calculations for MiDA Ghana Roads Project

This memorandum presents new power calculations for the MiDA Ghana Roads Project. The new calculations differ from the earlier ones done for the project proposal and prior to construction of the sample design, in several ways. First, they examine situations ("cases") that apply to the mid-term ("Phase II") evaluation. The original calculations applied to estimating the impact between the baseline ("Phase I") and endline ("Phase III")) surveys, not to the analysis of impact from the baseline to the mid-term survey. There are two substantial differences for the mid-term estimates of impact: (1) the mid-term survey has half the sample size as the baseline and endline surveys (77 localities per design group (treatment and control), instead of 154); and (2) since the mid-term survey is being conducted shortly after completion of the project intervention (road improvements), it is expected that the project impact on prices at the time of the mid-term survey would be less than at the time of the endline survey to be conducted over a year after completion of the project intervention.

The second way in which the new power calculations differ from the earlier ones is that we now have baseline data from which the variation in prices among localities may be measured. For the original power calculations, usable data on price variation among localities was not available. In making power estimates, it is necessary to relate the minimum effect size to be detected to the level of price variation. Since data on the level of price variation were not available earlier, the power calculations were done under the assumption that it was desired to detect a price change equal to ten percent of the standard deviation of prices, whatever that (unknown) value was. This is not a very satisfactory approach, but it is often done in statistical power analysis when no data are available about the level of variation. Now that the baseline data are available, estimates are available of the level of price variation (measured by the coefficient of variation (COV), or ratio of the standard deviation to the mean). With these data, the minimum detectable effect size may be specified as a fractional change in the mean price level, and power calculations may be made by relating this value to the value of the COV.

The third way in which the new power calculations differ from before is with respect to the values assumed for the correlations between units in the four groups of the evaluation design. In the new calculations, these correlations have been reduced somewhat from the previous values (causing the power estimates to be a little more conservative).

The paragraphs that follow discuss these changes in greater detail, and present the new power calculations.

In addition to the parameters just mentioned (sample size, COV, and design correlations), power estimates depend on two other parameters, namely, the significance level of the hypothesis test and the design effect. The values of these parameters have not been changed from those used in the earlier

analysis (i.e., the value of the significance level = the probability of Type I error of rejecting the hypothesis when it is true = $\alpha = .05$, and the value of the design effect is 1.0.

(The design effect is the ratio of the variance of an estimate using a particular survey design to the variance of the estimate using simple random sampling. The design effect varies from 1.0 because of design features such as stratification, multi-stage (or cluster) sampling and selection of sample units with varying probabilities. For the statistical power analysis, the formula used to estimate the power refers to *localities*, not to markets or vendors within localities. With the "marginal stratification" approach used to construct the sample design, treatment localities were selected with varying probabilities. Under a "model-based" approach to survey design and analysis, there is no decrease in precision or power associated with the use of variable selection probabilities to select the first-stage sampling units (for two reasons: for analytical surveys, the first-stage sampling fraction may be assumed to be zero, so that the estimated variance depends only on the first-stage sampling unit means, not on the variance within units; under a model-based approach, the unweighted model-based estimates are unbiased if the model is correctly specified).)

(It is noted that the treatment and control groups contain localities that vary in distance from the project roads. In the data analysis, it is planned to develop regression models that estimate the relationship of impact to changes in travel time associated with the program intervention. The power analysis presented here relates to simple double-difference estimates of impact, not to estimates based on regression models (such as a covariate-adjusted single-difference model, in which the impact would be a coefficient on a treatment indicator variable). The power formulas for the impact estimates based on regression models are different from those for the double-difference estimate, but the results are similar (since the impact regression coefficient is similar to a difference estimate).)

It is noted that the power calculations done earlier and those presented here assume that one-sided tests of hypothesis are used. It is not known how price levels will change after the road improvements. They may go up, down, or stay the same. What is important is how they change *compared to what they would have been had the road improvements not been made*. Because of the road improvements, however, it is expected that they will be *somewhat lower than what they would have otherwise been*. Since the *direction of the impact* (i.e., the direction of the price change after the project intervention relative to a counterfactual) is specified, one-sided tests of hypothesis are appropriate. (This is important because a one-sided test is more powerful than a two-sided test, for detecting a change in a certain direction.)

<u>Minimum Detectable Effect; Power Curve Estimation</u>

The power of a test of hypothesis is the probability of rejecting the null hypothesis, when it is false. In the present application, it is the probability of detecting an effect (impact) of a specified size, D, called the minimum detectable effect (or minimum detectable impact). The power depends on the effect size, D, and the standard deviation, σ, (of the variable of interest) only through the ratio, D/σ. There are two approaches to statistical power analysis. One is to specify a minimum detectable effect size and determine the power of the sample design (sample structure and sample size) to detect an effect of that size. The second is to specify a power level (i.e., the probability of detecting an effect of a specified size) and determine the minimum detectable effect that can be detected with that power.

In the present project, the approach to assessing impact is called the "Rubin causal model," the "potential outcomes model," or the "counterfactuals model." With this approach, the impact is the difference between the price changes (for an item or group of items) that occur when the project is implemented and the price changes that would have occurred had the project not been implemented (i.e., the counterfactual). The major problem facing the evaluator is that the counterfactual cannot be observed, and the impact must be estimated from data obtained from an appropriate evaluation design (in this case, using a double-difference estimator based on a pretest-posttest-comparison-group design).

If it is desired to estimate the power of the mid-term survey data to detect a minimum detectable effect, then the size of that effect must be specified. The problem that arises here is that little is known about the expected effect of the roads-improvement project intervention on price levels. For many evaluation projects, such as those intended to increase income, the expected effect may be estimated from previous similar projects, or from policy guidelines such as the Millennium Challenge Corporation *Guidelines for Economic and Beneficiary Analysis* (Revised April 2009). The MCC guidelines are not helpful in the present case – they specify effect sizes in terms of economic rate of return (ERR), and little is known about the relationship of price changes to ERR. Because little is known about the expected impact of the roads improvement on affecting price levels, the approach of estimating the power associated with a minimum detectable effect is not very productive. As mentioned, this problem was overcome in the power analysis done for the proposal by specifying the minimum detectable effect as a proportion of the standard deviation of prices. Now that the baseline data are available, we have estimates of the coefficient of variation (standard deviation divided by the mean) for the items priced in the survey. Using the approach of determining power from a specification of the minimum detectable effect, this still doesn't solve the problem, since we still don't know the size of the minimum detectable effect.

In view of the lack of information about the size of the minimum detectable effect, it is necessary in this project to use the second approach to power analysis, viz., to estimate the size of the minimum detectable effect *given* a value for the power. That is the approach that will be used here. Specifically, we shall conduct a "sensitivity analysis," in which we vary the coefficient of variation (now known for the items included in the survey) and determine the "power curve," i.e., the power as a function of the effect size, D. From the power curve, we may determine the power corresponding to any specified effect size, or the minimum detectable effect size corresponding to any specified power. Given the lack of information on the anticipated impact of the project intervention on price levels, this is the best that can be done. Although we cannot identify a particular minimum detectable effect that is of particular interest, this approach is nevertheless much more informative than the approach used in our proposal, in which neither the minimum detectable effect size nor the coefficient of variation was known, so that all that could be done was to estimate the power as a function of the ratio of these two unknown quantities. The proposed sensitivity analysis makes full use of the baseline data (on COVs), *to estimate the power curves* associated with the evaluation design (pretest-posttest-comparison-group design) and impact estimator (double-difference estimator (or perhaps a covariate-adjusted single-difference estimator)). While the inability to specify a particular minimum detectable effect may be frustrating, the ability to estimate the power curves as a function of COV is very useful, and a substantial improvement over what could be done in the absence of the baseline data.

In reviewing the power curves and assessing the power associated with impacts of various sizes, it is important to consider the length of time between the completion of the project intervention and the follow-up survey on which the impact estimates are being made. Some time may be required before

the full impact of a road improvement on price levels is manifest. If the mid-term follow-up survey is conducted only a few months after the completion of the project intervention, it may be that the effect is very small, in which case the power to detect it will also be very small.

Sample Sizes

The earlier power calculations, along with budget considerations, led to the sample sizes being used in the study, i.e., 308 localities in the baseline and endline surveys, evenly split between treatment and control, and half that number in the mid-term survey. Prior to completing the mid-term survey, it is desired to estimate the power associated with the lower sample sizes of the mid-term survey (i.e., 154 split between treatment and control groups, 77 in each). There are two reasons for doing this. First, not all of the planned road improvements have been completed, so that the effect of the program intervention may be realized for fewer than 77 of the treatment localities. In this case, the treatment sample size will be less than 77. Second, some items are not available in all markets, so that the sample size for estimating their price change is less than the full sample size. For the mid-term evaluation, we shall examine the power associated with sample sizes of 77 (the full planned mid-term sample size per design group), 67, 60 and 55.

Correlations Associated with Matching and Panel Sampling

For estimating changes, it is advantageous to construct the sample design so that there are correlations between the treatment and control groups, and between the before and after groups. Correlations are introduced between the treatment and control groups by matching control units (localities) to treatment units. Correlations are introduced between the before and after groups by conducting a panel survey in which the same localities are surveyed in each round of the panel survey. A number of the original power calculations assumed rather high values for these correlations. The statistical power analysis presented here assumes somewhat more conservative values (i.e., up to .5 for panel sampling and .3 for matching).

Cases Analyzed

Here follows a summary of the power analysis conducted under the preceding assumptions. (The formulas used in the analysis are included in the Appendix.) The table presents a number of "power curves," which specify the power for a range of values of the effect size (D), given selected values for the locality sample size (n) for each design group and selected values of the coefficient of variation, COV (holding other parameters, such as α and the various design correlations fixed). From the baseline data, it was observed that the COV generally varied over the range .1 to 1.00 for commodities (fresh food, packaged foods and non-food items) and was approximately 1.5 for transport costs (passenger and freight). The table presents power estimates for the following values of the COV: .1, .5, 1.0, and 1.5. The locality sample sizes (n) used are (as discussed) 55, 60, 67, and 77 (for each of the four design groups (treatment before, treatment after, control before, control after)). (The table also includes the case n=154, corresponding to the endline survey.) The power is calculated for the following values of the effect size, D: 0, .05, .1, .15, .2, .25, .3, .35, .4, .45, .5, and .55. Note that the power associated with D=0 is the significance level of the test, α=.05. The entry in each cell of the table is the power corresponding to the coefficient of variation (COV), sample size (n) and effect size (D) specified in the table margins. Each row of the table is a power "curve" (function specifying power as a function of effect size (D)).

Power Estimates for Various Values of Design-Group Locality Sample Size (n) and Priced-Item Coefficient of Variation (COV)													
		Minimum Detectable Effect Size (D)											
n	COV	0	.05	.1	.15	.2	.25	.3	.35	.4	.45	.5	.55
55	.1	.05	.96	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	.5	.05	.17	.39	.65	.85	.96	.99	1.0	1.0	1.0	1.0	1.0
	1.0	.05	.10	.17	.27	.39	.52	.65	.76	.86	.92	.96	.98
	1.5	.05	.08	.12	.17	.23	.30	.39	.47	.56	.65	.73	.80
60	.1	.05	.97	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	.5	.05	.18	.41	.68	.88	.97	.99	1.0	1.0	1.0	1.0	1.0
	1.0	.05	.10	.18	.28	.41	.55	.68	.80	.88	.94	.97	.99
	1.5	.05	.08	.12	.18	.24	.32	.41	.50	.60	.68	.76	.83
67	.1	.05	.98	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	.5	.05	.19	.44	.72	.91	.98	1.0	1.0	1.0	1.0	1.0	1.0
	1.0	.05	.10	.19	.30	.44	.59	.72	.83	.91	.96	.98	.99
	1.5	.05	.08	.13	.19	.26	.34	.44	.54	.64	.72	.80	.86
77	.1	.05	.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	.5	.05	.20	.48	.78	.94	.99	1.0	1.0	1.0	1.0	1.0	1.0
	1.0	.05	.11	.20	.33	.48	.64	.78	.87	.94	.97	.99	1.0
	1.5	.05	.09	.13	.20	.28	.38	.48	.59	.69	.78	.85	.90
154	.1	.05	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	.5	.05	.30	.73	.96	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	1.0	.05	.14	.30	.52	.73	.88	.96	.99	1.0	1.0	1.0	1.0
	1.5	.05	.10	.19	.30	.45	.60	.73	.84	.91	.96	.98	.99

The table shows that for items having low variability (low values of the coefficient of variation, COV), the power is high for detecting impacts that are small relative to the mean price level, such as D=.1 or .2. Out of the 82 priced commodities, about half (42) had COVs less than .3 (10 out of 39 priced fresh foods, 19 out of 24 packaged foods, and 13 out of 19 non-food items). For transport, unlike the commodities, there were not a lot of different items considered. Transport cost was measured for just two items, passengers and freight tariffs, and both of these had COVs on the order of 1.5.

Summary

From the preceding analysis, it is seen that the power to detect minimum detectable effects from the mid-term survey data varies considerably by item. For a substantial number of products, the variation in prices (COV) is sufficiently small that it will be possible to detect effects (price changes, as measured by the double-difference estimator) that are small relative to the mean price level with high probability (power). For products with high price variation, the likelihood of detecting changes of this magnitude will be small. For transport costs, the COV is so large (about 1.5) that only very high price changes would be likely to be detected (transport costs are highly variable because the destinations are specified for each locality, and their costs vary widely).

It is important to keep in mind that it may take some time for the program intervention effects to become manifest. If the mid-term survey is conducted just a few months following the completion of the road improvements, the effect may be very small, simply because it takes some time for prices to adapt to the new situation. Before proceeding on the mid-term survey, it should be recognized that if it is done shortly after completion of the road improvements, few or no price changes may be detected, even for products with small COVs. Given this fact, it may be desirable to consider cancellation of the mid-term survey, and allocating the funds to some other purpose. Little value would be realized from

simply postponing the mid-term survey, since then it would show about the same results as the endline survey. It would be of very little advantage to reallocate the funds used for the mid-term survey to increase the sample size for the endline survey, since the evaluation design is a panel survey in which the same localities are observed in the follow-up surveys as were observed in the baseline survey (the addition of new localities for which no baseline data is available are of little value in increasing power).

Assessment of Validity of Power Estimates

The power calculations are made assuming values for a number of parameters. This memorandum has presented a "sensitivity analysis" to quantify the effect of the assumptions on the power estimate. For the most important parameters – price variation (as measured by the COV, estimated from the baseline survey data) – the parameter values are well founded (i.e., based on baseline survey data), and it is simply a matter of deciding which case (COV and sample size) is of interest. For some design parameters – the correlation associated with panel sampling and with matching – the assumed values are somewhat speculative, but considered reasonable based on experience. The remaining parameter is the significance level (α, the probability of a Type I error) used for tests of hypotheses. This has been set at .05, which is a "conservative" (and standard) value (i.e., the chance of deciding that an effect is significant when it is not, i.e., .05).

Analysis of the baseline survey data included regression analysis to estimate the relationship of prices to a number of explanatory variables, and these regressions were seen to have low explanatory power (value of R^2, the coefficient of determination, of .343 for fresh food items, .254 for package food items, and .178 for non-food items). Furthermore, the relationships are expected to be weaker for double-difference estimates of impact than for the raw (undifferenced) price data. Hence the power will not be appreciably increased through the use of covariate adjustment.

In summary, it is believed that the power estimates presented above are sound.

Appendix. Formulas Used to Estimate Power

The formula for the power of a test of hypothesis about a mean double difference is as follows:

$$\Pr\left(\frac{\hat{\mu}_1 - \hat{\mu}_2 - \hat{\mu}_3 + \hat{\mu}_4}{(deff\ \mathrm{var}(\hat{\mu}_1 - \hat{\mu}_2 - \hat{\mu}_3 + \hat{\mu}_4))^{1/2}}\right) > z_{1-\alpha} \,\middle|\, \mu_1 - \mu_2 - \mu_3 + \mu_4 = D) = 1 - \beta$$

where

$$\mathrm{var}(\hat{\mu}_1 - \hat{\mu}_2 - \hat{\mu}_3 + \hat{\mu}_4) = \frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2} + \frac{\sigma_3^{\ 2}}{n_3} + \frac{\sigma_4^{\ 2}}{n_4} - \frac{2\rho_{12}\sigma_1\sigma_2}{\sqrt{n_1 n_2}} - \frac{2\rho_{13}\sigma_1\sigma_3}{\sqrt{n_1 n_3}} + \frac{2\rho_{14}\sigma_1\sigma_4}{\sqrt{n_1 n_4}}$$

$$+ \frac{2\rho_{23}\sigma_1\sigma_3}{\sqrt{n_1 n_3}} - \frac{2\rho_{24}\sigma_2\sigma_4}{\sqrt{n_2 n_4}} - \frac{2\rho_{34}\sigma_3\sigma_4}{\sqrt{n_3 n_4}}$$

where

μ_1 = mean for group 1 (treatment, time 1)
μ_2 = mean for group 2 (treatment, time 2)
μ_3 = mean for group 3 (control, time 1)
μ_4 = mean for group 4 (control, time 2)
n_1 = sample size for group 1
n_2 = sample size for group 2
n_3 = sample size for group 3
n_4 = sample size for group 4
σ_1 = standard deviation for group 1
σ_2 = standard deviation for group 2
σ_3 = standard deviation for group 3
σ_4 = standard deviation for group 4
ρ_{12} = correlation between items of groups 1 and 2
ρ_{13} = correlation between items of groups 1 and 3
ρ_{14} = correlation between items of groups 1 and 4
ρ_{23} = correlation between items of groups 2 and 3
ρ_{24} = correlation between items of groups 2 and 4
ρ_{34} = correlation between items of groups 3 and 4
(The correlation matrix should be positive definite.)
α = significance level of one-sided test of hypothesis of equality of group means (the probability of Type I error, i.e., the probability of rejecting the hypothesis of equality of group means, when it is in fact true) (e.g., .05)
β = the probability of making a Type II error, i.e., the probability of accepting the hypothesis of equality of the group means, when it is in fact false) (e.g., .1)
$1 - \beta$ = power of the test (e.g., .9)
$z_{1-\alpha}$ = $1-\alpha$ percentile point of normal distribution (e.g., 1.6449 for α=.05, or 1.2816 for α=.1)
deff = design effect (The design effect is the ratio of the variance of an estimate for a specified survey design, compared to the variance using simple random sampling.)
D = (true) size of the mean double difference
a caret (ˆ) over a symbol denotes a sample estimate
and Pr(.) denotes the normal probability distribution function.

Note that the preceding formula does not contain a finite population correction (FPC). The FPC is not relevant for analytical surveys, where the objective is to make inferences about a process, not about the particular finite population at hand.

For the present application, the sample sizes of the four design groups are the same (say, n), and the values of the standard deviations (σ's) or the coefficients of variation COV = σ/μ are the same for the four design groups. In this case, the following formulas (derived from the formula given above) may be used to calculate the following quantities:

Power, $1-\beta$: $1-\beta = 1 - Pr(z_k)$

where $z_k = z_{1-\alpha} - (D/\sigma)\ sqrt(n)\ /(deff\ c) = z_{1-\alpha} - ((D/\mu)/(\sigma/\mu))\ sqrt(n)\ /(deff\ c)$.

Minimum detectable effect, D: $D = (z_{1-\alpha} + z_{1-\beta})\ deff\ c\ \sigma\ /\ sqrt(n)$.

Minimum detectable effect as a ratio to σ, D/σ: $(z_{1-\alpha} + z_{1-\beta})$ deff c / sqrt(n).

Sample size, n: $n = (z_{1-\alpha} + z_{1-\beta})^2 \text{ deff}^2 c^2 / (D/\sigma)^2 = (z_{1-\alpha} + z_{1-\beta})^2 \text{ deff}^2 c^2 / ((D/\mu)/(\sigma/\mu))^2$

where $c^2 = 4 - 2(\rho_{12} + \rho_{13} - \rho_{14} - \rho_{23} + \rho_{24} + \rho_{34})$.