Response to Jorgen Elklit's Rebuttal to KPTJ's Criticism of the IREC Report

Kenyans for Peace, Truth, and Justice

8 November 2008

"KPTJ claims that IREC should not infer population parameters from results based on analysis of a non-random sample. Theoretically, KPTJ has a point, but only theoretically. The evidence provided in the IREC analysis is so strong and so unequivocal that there is no reason to doubt that purposive sampling (the only possibility in the circumstances) allowed IREC to reach valid conclusions directly related to its mandate."[1]

Elklit argues in his "rebuttal" that purposive sampling was the only feasible option for IREC, given its mandate. KPTJ disagrees with that statement, and will provide a better way forward below. However, let's begin by *assuming* that Elklit is correct, and that purposive sampling was the only way to go. What implications does his assertion have on the credibility of the IREC report? First of all, section 6.5.1 (p. 129), entitled "The random nature of the errors affecting the presidential election in the eighteen constituencies analysed," should be struck from the report. In his response, Elklit states "KPTJ claims that IREC should not infer population parameters from results based on analysis of a non-random sample. Theoretically, KPTJ has a point, but only theoretically." Our "theoretical" point has a clear implication: we cannot reliably estimate population parameters from a non-random sample. Section 6.5.1 makes implicit claims about a specific population parameter: the average difference (for each candidate) between the ECK tallies and IREC's re-tallies. IREC appears to claim that, because the differences between ECK numbers and IREC's numbers do not systematically favor any one candidate, that we can conclude that discrepancies did not favor any one candidate.

This argument is simply invalid given the chosen sampling procedure. In order to demonstrate to interested readers exactly how inappropriate this claim is, we ran a simple experiment to answer this question: how likely is it that a random sample will provide a better estimate of a population parameter than the purposive sample chosen? To answer this question, we ran 10,000 simulations. In each simulation, we did the following:

1. Draw a random sample of 19 constituencies (without replacement) from the population of 210 constituencies.

[1] *Some Comments on the KPTJ piece: "Evaluation of IREC's Statistical Analysis and Claims" (dated 1 October 2008)* , p. 3.

2. Estimate a population parameter from this sample.[2]

3. Compare the population parameter estimate from the random sample and IREC's sample with the true population parameter, estimated from all 210 constituencies.

4. If the difference between the estimate from the random sample and the true population parameter is less than the difference between the estimate from IREC's sample and the the true population parameter, report that the random sample beat IREC's sample. If otherwise, report that IREC's sample beat the random sample.

We repeated these steps 10,000 times, recording each time whether the random sample or IREC's sample did a better job recovering the population parameters of interest. Table 1 reports the results from this exercise. The implications of this exercise are clear: the purposive sample IREC chose does not do a good job of estimating population parameters. Therefore, generalizations regarding characteristics of the entire population of constituencies should not be made from their sample. This means that we cannot take IREC's *ad hoc cui bono?* analysis in section 6.5.1 seriously, since it, in all likelihood, does not represent the population of constituencies.

Variable	% Success
Reg. Voters, Nov. 2007	90%
Pop. Density	31%
% Urban Population	46%
Poverty Incidence	93%
Inequality	82%
% Chg. in Reg. Voters	81%
% Kibaki Vote, 2007	99%
% Diff., Pres. and Parl. Vote	100%
# Polling Stations	57%
# Parl. Candidates	99%
# Civic Candidates	99%

Table 1: A random sample performs much better than IREC's purposive sample in most cases. As a result, claims about the population of constituencies should not be drawn from IREC's sample.

Moreover, Elklit states "Non-probability sampling had to be used because random sampling was not an option, even though the KPTJ paper seems to consider that a realistic

[2]The parameter we choose to estimate is the population mean. We do so for two constituency-level characteristics, (a) the difference between presidential and parliamentary vote counts as a percentage of registered voters, and (b) percentage urban population. We chose these two variables because they represent the characteristics least and most similar (respectively) between IREC's sample and the population of all constituencies. This approach will give us a sense of the best and worst case scenarios in terms of how frequently a random sample beats IREC's purposive sample in terms of getting close to the true population average.

possibility."[3] The tautological nature of Elklit's justification hides the basic fact: probability sampling was an option, but IREC did not choose it. Instead, IREC chose non-probability sampling. As the above simulation demonstrates, random sampling does a much better job than purposive sampling in recovering population parameter estimates. Rather than make the tenuous assumption that the purposive sample provided a sufficient picture of what the rest of the data would look like, IREC could have used a stratified probability sample, using disputed/undisputed constituencies as the strata. This would have provided a picture of *both* likely problematic and likely unproblematic constituencies, and enabled IREC to make a credible claim regarding the quality of the ECK data.

The KPTJ paper goes on to argue that IREC too easily dismisses the possibility of using one or more of the available methods to deal with messy, problematic data like those produced by ECK.

"However, the KPTJ paper also admits that this kind of analysis remains stymied as polling centre [must be polling station – JE] data have not yet been released by the ECK, so this suggestion is also not something which might have been of much use. The value of the claim would also have been increased if it had been demonstrated how the various methods KPTJ refers to would have allowed IREC to reach other conclusions, especially in view of the many problems identified prior to vote counting at polling stations (in particular vote-buying, zoning, intimidation and ballot-stuffing)."[4]

While it is true that statistical modeling of polling centre data would be the ideal, such models could have been plied on constituency level data. They were not. [Add in example here? Could take a day or two.] Moreover, if IREC had chosen a stratified probability sample, IREC's analysts could have modeled the subset of polling stations from the relatively "clean" constituencies, and then used that model to estimate counter-factual outcomes for the polling stations in suspect constituencies. Discrepancies between the actual outcomes in suspect polling stations and these counter-factual outcomes generated from the "clean" model would be a strong indication of vote-stuffing or other kinds of localized fraud at the polling station level. But, due to the poor methodological choices made by IREC, such modeling was stymied.

"The issue was not whether and how constituency counting centre personnel might have generated numbers, but whether they had transferred numbers correctly from forms 16A to form 17A – there is usually a "correct" number (in form 16A) to use as a benchmark for the number found in 17A. ...[T]he usefulness of tests based on Benson's [sic] law on digit patterns is repeated but that does not, of course, increase its applicability in this context[.]"[5]

The issue at hand with respect to polling centre level numbers is whether and where fraud might have been committed by ECK employees. Unfortunately, as Elklit's statement lays bare, IREC was not interested in "how constituency counting centre personnel might have generated numbers" – that is, by fumble or by fraud. Rather, IREC made two assump-

[3] ibid., p. 2.

[4] ibid, p. 3 and 4.

[5] ibid, p. 4 and 5.

tions. First, they assumed that the numbers on form 16A's were correct. This seems like a reasonable assumption, since re-counting ballots to get the "true" counts would have been a difficult endeavor.(More on this assumption below.) Second, they assumed that errors arising from transcription to form 17A's were simply mistakes, and not indications of fraud. This assumption is not reasonable, since it assumes away the question of interest.

A more reasonable approach would have been to remain agnostic about how polling station level numbers were generated and develop a test to detect suspect returns. Benford's law and related pattern-based benchmarks (see the previous KPTJ critique) provide a way to develop such tests. For example, suppose that constituency i has N polling stations. Given ECK procedures, we have 2 sets of numbers for those N polling stations. First, we have N form 16A's, which report the turnout and candidate vote counts from that polling station. Second, we have 1 form 17A, which is filled out by constituency counting-centre personnel from the form 16A's. In an ideal world, the numbers on the form 16A's and the form 17A would match. However, in many cases, they do not. The task at hand is to set up a test that detects whether the differences between the forms are accidental or otherwise. As we demonstrated above, IREC's *ad hoc cui bono?* analysis fails since it is derived from a non-random sample.

Does such a test exist? Benford's law based χ^2-tests will allow such an analysis, and allow us to detect suspicious returns and locate them at the polling station or constituency levels.[6] We will develop this idea in the context of total votes cast, though similar tests could be carried out using the returns from each candidate. Let's call the N turnout numbers from the form 16A's N_p, and the N turnout numbers on the form 17A N_c. Now, let's explore two cases. First, when constituency-level counting personnel make no changes (intentional or by accident) during transcription from the form 16A's, then the numbers in N_p are the same as those in N_c. Thus, we face one χ^2-test, which the digits will either pass or fail. Since the polling station is the source of the form 16A and these is no difference between N_p and N_c, we can reasonably assume that a passed or failed test is an indicator of polling-station level behavior. The second case occurs when N_p and N_c are different; as a result, we can run two χ^2-tests, one for each set of numbers. This leaves us with four possible outcomes:

1. N_p passes and N_c passes: This would seem to indicate that changes occurring at the constituency level were random, and did not affect the distribution of digits.

2. N_p passes and N_c fails: This would seem to indicate that changes occurring at the constituency level may be indicative of fraud.

3. N_p fails and N_c passes: N_p appears manipulated, but changes in N_c at the constituency level do not. Constituency-level manipulation "corrects" polling centre manipulation. Suggestive of fraud at the polling station level; indeterminate at the constituency level.

[6]Such tests could examine the second, third, or any later digit other than the first. Related tests, discussed in our previous critique could also be applied. Given the need to detect intentional manipulation, the Beber-Scacco test would probably be the best choice. Since these tests are easy to implement, an optimal approach would be to carry out a standard Benford's law χ^2-test, followed by a more discerning analysis focusing on the specific ways in which the expected distribution is violated.

4. N_p fails and N_c fails: N_p appears manipulated, as does N_c numbers.

What are the short-comings of such tests? They do not allow us to pinpoint the exact polling station in which the fraud occurred, only that the patterns displayed in the form 16A's as a whole do not conform to our expectations. While this indeed a weakness, the test still provides an important diagnostic in locating potential fraud at the polling station versus constituency level.

While statistical tests like these are imperfect vehicles, they are perhaps our best ally in the absence of a time machine and some omniscient observer of all polling stations. Perhaps most importantly, these tests do not assume that the form 16A numbers are "correct," but rather set reasonable external benchmarks with which to compare sets of numbers which we do not know *ex ante* to be generated by error or by fraud. These tests, rather than assumptions about cause or a flawed *cui bono?* analysis, proceed with a measure of objectivity that human beings alone do not possess. The choice not to use such tests – especially given that IREC had access to polling centre level data for at least 19 constituencies – amounts to leaving evidence on the table, unexplored.