To benchmark is to compare performance against a standard. As part of an effort to improve the effectiveness of Monitoring and Evaluation (M&E) in the policy cycle, benchmarking can be useful in three ways. First, benchmarking can help place an outcome in context. Was the achievement good, bad, or indifferent? Second, benchmarking can help assess the reasonableness of targets that may be set. Third, benchmarking can help identify specific regions or subgroups whose exceptionally good or poor results hint at what factors drive performance. In the third case, when there is very little variation in performance, it can be hard to extract information about potential determinants. Therefore, the process of analyzing performance at the extremes can make it easier to extract information and to introduce more feedback into the policy cycle.

Setting targets is often an important part of initiatives to induce a greater focus on results and may or may not be linked explicitly to the budget. The Millennium Development Goals are targets that are not explicitly linked to a budgetary process. However, a performance-based or performance-informed budget process is often accompanied by targets as part of an effort to set clear objectives. The targets are usually set through negotiation between a principal (typically a finance or planning ministry) and an agent (typically a sector ministry). The outcome of this negotiation is critical to the effectiveness of the performance-informed budget process. If the negotiation results in an unreasonably high target, which is subsequently not met, the principal has only two choices – either waive the requirement that the agent meet the target (which damages the system’s credibility) or hold the agent to the target (which could breed resentment). There are, perhaps, fewer threats to the operation of the performance-informed budgeting system from setting a target that is unreasonably low, one that could be easily reached by the agents. While the problems of setting an unreasonably low target are not as visible as those associated with setting a target that is unreasonably high, there is still a cost. If nothing different is done—if there is no improvement in performance—then society is actually worse off because it takes real resources to mount the superstructure to support the performance-informed budgeting process.

A first approach to judging the reasonableness of a target is to compare the implied change between the current value and the target to the empirical distribution of changes for a reference population. Would the performance improvement implied by meeting the target place the change within the top 10 or 20 percent of all recorded changes in the indicator for the reference population? If so, then the target should certainly be considered a very
ambitious target. Carrying out such a comparison is a benchmarking exercise. Figure 1 shows the result of such a comparison for a target to reduce homicides in the state of Minas Gerais, Brazil (Peixoto, Cruz, and Azevedo, 2010). The empirical distribution is of changes in the number of homicides for every year between 2004 and 2008 for all the police regions in the state. Figure 1 suggests that the target for 2009 would require a significant effort in terms of average annual change because the state would have to perform, on average, better than 80 percent of the observed recorded changes to date.

Figure 1 – Cumulative distribution on changes of homicide between 2004-2008, by Police Regions in the State of Minas Gerais

Today, the online availability of large internationally comparable data sets (such as those in the World Bank’s Data Development Platform, Edstats or HNPstats) and household survey data makes it considerably easier to use international data for benchmarking exercises. However, perhaps even more useful than internationally comparable data sets are national data sets that provide information on the indicator of interest for different subgroups. For example, homicide data may be available for all police regions in a state (as was the case for Minas Gerais). The target that the principal and the agent may negotiate may be for improvement in homicides at the state level, but if the target is to be met, it will be up to individual police departments to improve the outcomes in each one of their regions. It is often the case that a sector, ideally before, but often after, concluding negotiations on targets at a state level will then set targets for the “sub-state” administrative units. Recognizing that different police regions are starting from different initial conditions, what is the implied performance improvement that each police district would have to meet if the same state-level target was set for each police region? What is being asked of them and is it reasonable to expect that they could improve? Comparing the targets to some of the changes that have taken place in the recent past may help provide answers to those questions.

Figure 2 represents a simple way to present the comparison made in Figure 1 for many police regions simultaneously. Rather than presenting the entire empirical distribution for the indicator of interest and showing where the target falls in that distribution, Figure 2 presents the percentiles for recent changes, changes over a longer period, and changes that would be implied by meeting the targets—for all the police regions. Figure 2 uses only the empirical distribution of changes in homicides. That is, the analysis does not try to control for differences in the initial conditions of the different police regions.

The graph suggests that (1) for some regions, the given target is not very different from the medium-term and recent tendencies (Juiz de Fora), (2) for other regions, the targets are quite a bit more demanding than the medium-term tendencies, but not from recent tendencies (Vespasiano, Belo Horizonte and Contagem), and,
finally (3) for certain regions, meeting the targets would require a considerable shift from both the medium-term and recent tendencies (Teofilio Otoni and Montes Claros).

Identifying where a district may have to improve its performance considerably should alert stakeholders to examine what improvements in policy or implementation are in place to lead to the improvement.

However, it is quite likely that stakeholders involved in a target setting process may wish to define and control for some specific characteristics to generate a more relevant comparison than a comparison to all other regions. With many more observations than what is normally available, the researcher might form a comparator group by selecting some units to be included in the comparator group, dropping others. However, constructing a comparison group by dropping observations is not efficient and would greatly reduce the sample size. Instead, it is possible to retain all observations and carry out some multivariate statistical analysis to take into account differences in characteristics.

One approach to controlling for differences in characteristics used in the Minas Gerais benchmarking exercise was to use quantile regression to estimate the relation between the change in the indicator and a set of characteristics (for more details, see Newman, Azevedo, Saavedra, and Molina, 2008). This has some advantages over controlling for the characteristics using OLS regressions. An OLS regression estimates the relation at the mean of the distribution. However, a benchmarking exercise is not about performance at the mean. It is about performance that is considerably better or worse than average. Using quantile regression allows the researcher to estimate the relationship between changes in the indicator of interest average annual changes in poverty and country characteristics at different percentiles in the distribution. In the analysis carried out in this paper, we estimate 99 different quantile regressions, corresponding to percentiles 1 through 99. This generates 99 sets of coefficients, which when combined with a particular police region’s set of characteristics, yields a set of 99 predicted percentiles.

When working with the empirical distribution, the outcomes for a police region were judged to be good or poor simply by noting where the particular police region’s observed change fell in the empirical distribution of all observed annual changes. When controlling for characteristics of the police regions, the observed change can be compared with a counterfactual distribution of predicted percentiles that is specific to each police region.

The final consideration is the choice of the characteristics to control for. The idea is not to explain the change in homicides, but to generate a standard for comparing performance that could be more relevant for the police region in question. In the comparison, each government or interested party could pick the specific characteristics that it would like to control for. The only requirement is that the characteristic be observable and that data exist for the units of analysis of interest. In this particular case, only two controls were used, population and level of homicide at the beginning of the period.

---

1 The data used in the quantile regressions is for each observed change over adjacent years for each police district.
2 A Taylor series expansion to calculate the 95 percent confidence interval of the predicted values. This technique is sometimes called the delta method (for more details, see Newman, et al., 2008).
A comparison of Figures 3 and 2 suggests that taking into account the characteristics does make a slight difference to the comparisons of the targets to recent tendencies. In general, taking into account the particular and current characteristics of the police regions, the targets appear somewhat less demanding relative to past performance.

In most benchmarking exercises, it is useful to consider not only the nature of the changes in the indicator of interest but also the level. Focusing only on the relative performance in the change can cause the researcher to be overly optimistic. A district, state or country may be advancing comparatively rapidly, but it may have very far to go. Focusing only on the relative performance on the level can cause the researcher to be overly pessimistic, as it may not be sufficiently sensitive to pick up recent changes in efforts to improve.

Conclusions

The approach described in this note is of general applicability. The key points are:

1. Benchmarking can be useful to help place an observed change in an indicator in context, to help assess the reasonableness of targets and to help inform the negotiations over the setting of targets.

2. With the ready availability of online comparable data, it is useful and increasingly easy to compare a change in an indicator of interest to the empirical distribution of all changes in the indicator for a reference population.

3. If controlling for observed characteristics is needed to make a valid comparison, quantile regressions can be used to generate a relevant counterfactual distribution of predicted percentiles for the particular district, state or country in question.

References


About the Authors

João Pedro Azevedo is an Economist with LCSPP, John L. Newman is the Lead Poverty Specialist with SASEP and Juliana Pungiluppi is a Consultant with LCSPP.