This survey, written at the request of the experts committee, is being published in its original form, as written by the author.

On The Use of Educational Indicators in Policy- and Decision-Making: The GEMS Tests

Henry Braun
Boston College
Chestnut Hill, MA, USA

October 23, 2008.

Introduction

Competent administrative management requires information to make appropriate decisions in order to achieve its short-term and long-term goals. This information should possess at least three key characteristics: relevance, credibility and usability. One would like to add another characteristic; namely, completeness. Unfortunately, it is rarely – if ever – the case that completeness is achievable. However, the degree of completeness should contribute to the evaluation of usability. In practice, the massive amounts of data usually available are summarized in a small number of statistics that are often termed "indicators" and it is these indicators that are considered in policy-making and decision-making. Thus, it is these indicators that should be judged with respect to the three key characteristics.

In educational settings, indicators are constructed from both administrative data and pedagogical results. The percentage of students promoted from grade 4 to grade 5 is an example of the former, and the mean score on the grade 5 end-of-year mathematics test is an example of the latter. With respect to educational indicators, relevance and credibility depend in a fundamental way on the data sources. Thus, indicators derived from test scores should be judged, in part, on the relevance and credibility of those scores! But they should also be evaluated on the basis of their psychometric and statistical properties. A thorough analysis of these properties will reveal their strengths and weaknesses with respect to the intended inferences.

Beyond its obvious dependence on relevance, usability is determined by considering the policy issues at stake (i.e., what purpose will be served by the indicator system), the intended audiences (i.e., who will use it and what degree of specialized knowledge or training is required), the type of application (i.e., how it will be used), and the modes of presentation (how the indicators are communicated).

Indicator systems do not exist in isolation. They are a design artifact and, as such, are created and developed under real-world constraints that include, but are not limited to, time and money. Consequently, every existing indicator system is the result of a series of "design tradeoffs" that lead to a compromise among desired features and (usually) imposed constraints. Thus, the critical question is not whether the system achieves the ideal but, rather, whether it represents a satisfactory compromise that serves adequately

well for the purposes at hand and is commensurate with the costs incurred. The costs involve both resources expended and unintended consequences that can subvert the hoped for results.

The answer to the question is not obtained by calculation but by the weighing of evidence. That process inevitably involves value judgments and, thus, should be entered into with the understanding that various observers can reach different conclusions. It should also be recognized at the outset that there are inherent limitations to an indicator system that relies on externally mandated assessments. One is that the data underlying the indicator are limited in what they can reveal about the system. Another is that policy uses of indicators implicitly or explicitly rely on causal interpretations of the indicators and that such attributions of causality are highly problematic, given the nature of the data and how it is collected.

GEMS

GEMS was initiated by the Ministry of Education (MoE) in 2002 with the principal goal of providing useful information to the principals and teaching staffs of elementary schools and junior high schools regarding the achievement of their students in four core subjects: Mathematics, science and technology, Hebrew or Arabic (as a first language) and English (as a second language). A second goal was to provide the MoE with an aggregate picture of achievement, as well as facilitating comparisons among subsets of schools categorized by sector, region, etc. Since 2006, GEMS has been under the supervision of RAMA and is administered in grades 2 (first language only), 5 and 8.

Each year, one-quarter of the schools in each grade are administered examinations in two subjects as an external assessment; this means that the test papers are graded outside of the school in a process that is (now) implemented by RAMA. The results are combined in the aggregate report. The remaining three-quarters of the schools may administer the same two examinations, but the papers are graded internally and the results remain within the school. Thus, a particular school experiences an external assessment every two years, with a particular subject tested every four years. In addition to the aggregate reports that are produced each year by RAMA, school staff can access their school's data at the RAMA website, along with interpretive information.

Purpose of the present note

With the recognition of the increasing importance of human capital and the key role that public education plays in its development at a national level, the Committee on Revising Educational Indicators for the State of Israel asked Professor Cahan to write a report on the use of educational indicators. Subsequently, he suggested, and the committee agreed, that he focus on the GEMS as a particular case. A draft report was submitted in July 2008 and made available to me in mid-August 2008 (in English translation).

It should be said at the outset that every indicator system should be subjected periodically to critical evaluation. Such an evaluation should inform policy makers whether, and to what degree, the system is accomplishing its purpose, as well as provide insights on how the system might be improved. Where possible, one should glean insights that can contribute to the development of improved educational indicator systems generally.

In his report, Professor Cahan discusses a number of issues regarding the design, development and implementation of educational indicators. However, in the main he focuses on the GEMS, addressing several different topics. Some are technical, some are policy-related and some are political in nature. His conclusions, as I understand them are: (1) The current GEMS is so deeply flawed (despite some small improvements under the auspices of RAMA) as to be nearly worthless; (2) The necessary flexibility and resources to respond at the school level to any indicator system is essentially absent; and, consequently, (3) The annual cost of the system is not commensurate with its utility. Professor Cahan's recommendation is to scrap the current system entirely and begin a broad consultative process that would likely lead to a very different set of indicators. He posits that the new system will be characterized by: (1) Emphasis on measuring the teaching process in all schools, grades and subjects; (2) Greater focus on "opportunity to learn"; (3) Testing samples of students with appropriate exposure to the material; (4) Reporting of results by question, rather than by scale scores that combine data from different questions.

Based on my reading of the report and on an admittedly primitive understanding of GEMS, I come to a somewhat different set of conclusions. Although a number of the technical issues raised by Professor Cahan have merit, others are problematic or even beyond present day technology. Accordingly, I would argue that the current system should be considered as a stage along a path of development. Responsible staff should be engaged in a process of continuous improvement, informed both by technical evaluations and by extensive consultations with users, as well as lessons learned in other countries with various assessment systems. Thus, there should be a long-term design effort, leading to improved credibility and utility. The measurement of other features of the education system should certainly be undertaken, but as a complement to the current set of outcomes-focused indicators.

The following two sections discuss many of the technical and policy related topics raised in Professor Cahan's report.

Technical issues

[Information quality] This issue addresses some aspects of (test) construct validity. The two main threats to construct validity are (i) construct under-representation and (ii) construct-irrelevant variance. Threat (i) is not mentioned in the report, although it is arguably a critical one. In brief, one should examine the alignment between the curriculum goals and the test specifications, as well as congruence between the test specifications and the actual test instrument. There is much recent work on the methodologies for such alignment studies.

Threat (ii) is treated in terms of three concerns: (a) opportunity to learn, (b) pupil motivation and (c) score inflation through guessing and copying. With respect to (a), I am puzzled by the expectation that an external test be used to disentangle the contributions of exposure to the relevant material and pupils' mastery of the material to which they were exposed. Variability among schools in coverage and quality of instruction are certainly factors in average score differences across schools. The test can tell us (within measurement error) whether a cohort of students has achieved a certain level of proficiency. If not, then further investigation is called for at the school level. This seems to me a proper use of an indicator. With respect to (b), pupil motivation is a perennial problem in low-stakes (for pupils) assessments. Certainly, strategies to encourage maximal engagement and effort should be investigated and there is some relevant literature on the matter. The possibility of gathering student self-reports on engagement and effort, as is often done in large-scale assessment surveys, should be explored. Nonetheless, the likelihood of differential engagement across schools will continue to be a challenge. With respect to (c), guessing is a problem inherent to multiple choice items. The problem can be partially addressed through the development of objectively scored items that minimize guessing, the formulation of attractive distractors, and the use of item response models that incorporate a guessing parameter. Although none of these offer a complete solution, they can mitigate the contribution of guessing to construct-irrelevant variance. Incorporating more items that require a student produced response is also a viable strategy – although it has implications for both cost and reliability. Copying can be controlled through careful proctoring. If it is considered a serious problem, there are a number of statistical techniques that can be used to detect response patterns in a classroom that are suggestive of cheating.

Another issue raised under this heading concerns the proportion of students in a class that is assessed. Apparently, many students can be excluded (e.g., students with special needs, new immigrants). Exclusion rules are acceptable, as long as they are reasonable and are uniformly applied. In the U.S., the National Assessment of Educational Progress employs such rules and makes clear that the inferences from the assessed sample employ to the subpopulation of students in the grade that can meaningfully participate in the assessment. Over the years, NAEP has worked hard to clarify the rules for exclusion and to help school officials apply them consistently. It must be admitted, however, that NAEP has had limited success in this respect. Certainly, this issue must be considered when interpreting the results of GEMS. Examining the distribution of school-level exclusion rates by subject can suggest where further investigation may be fruitful.

More worrisome, perhaps, is the fact that students who respond to fewer than 20% of the items are also excluded from the calculations. The rationale for this rule should be explicated and the consequences for school-level inferences examined. At the same time, since any change in the rule can have unintended consequences (e.g., for the number of students excluded), the implications of such changes should be considered before implementation, with some monitoring mechanism put into place.

[Absolute interpretation of test scores] I find the discussion here somewhat confusing, at least at first. Professor Sorel asserts that "Performance … must be expressed via an absolute scale, one that is uniform across grades and subjects. This is not possible in educational measurement, except in the most trivial cases. However, in Section 3.5 he explains that this "demand" is driven by the type of score interpretations favored by the MoE in the first years of the program. He then quotes more recent statements that provide cautions on the absolute interpretation of test scores, and includes a relevant excerpt from the classic monograph by Angoff that discusses this very point.

I agree that official interpretations of test scores, especially those communicated to educators and the public-at-large, should be both clear and correct. This appears not to have been the case earlier on. Certainly, the adoption of a 0-100 scale or the use of terms such as "relatively low" and "performed well" without being explicit about the nature of the comparisons can invite misinterpretation.

Normative comparisons are meaningful within the context of a single examination. For example, it is possible to compare the performance of two distinct groups of students (e.g. secular Jewish students and Arab students) on the grade 5 mathematics test given in a particular year. Comparisons within a given subject but in different years are possible to the extent that the examinations are very similar in both content and difficulty. If such comparisons are desired, then the test development process likely needs to be made more stringent and some sort of test equating procedure adopted.

To approach the ideal of an absolute interpretation, test specialists look to criterion-referenced (CR) examinations. This involves setting standards for performance at different levels (e.g., Basic, Proficient, Advanced) and then developing a test that can provide sufficient evidence to reliably classify students into the appropriate category. Although this is an attractive, and increasingly widespread, strategy, it is not without its difficulties. These include devising and implementing a consensus process to establish credible standards, as well as building a test with the requisite psychometric characteristics, while fully representing the content standards for the particular subject-grade combination. Thus, I would argue, contra Professor Sorel, that psychometric theory does provide the education system with the tools to measure the performance levels of pupils vis-à-vis the curriculum. It is challenging to do this well and, however well it is done, it cannot support interpretations tied to a percentage-mastery scale.

[Meaningful comparisons of scores between subjects and grade levels] Again, this demand is occasioned by statements in various reports that make explicit comparisons between subjects in a given year and grade or between grades in a given subject. Certainly, such comparisons would be valuable for policy-makers and educators who are faced with decisions on how to allocate scarce resources. Nonetheless, Professor Sorel is undoubtedly correct that these comparisons are not supported by the current testing system. A CR testing system would provide a more credible foundation for making comparisons between subjects. The validity of such comparisons would rest heavily on the quality and meta-equivalence of the standards set for the different subjects. Although this can be accomplished, it is difficult to do.

[Meaningful comparisons of scores across years within a subject/grade combination] If there is interest, for example, in comparing the performance (at the national level) in mathematics in grade 5 from one year to the next, then the sine qua non is that the tests given in the two years be parallel in content and psychometric characteristics; moreover, an equating procedure should be carried out to ensure that the reporting scales for the two tests are equivalent. Based on my reading of Professor Sorel's critique, these conditions are not yet satisfied by the GEMS.

A different, but equally important, issue raised by Professor Sorel is the proper interpretation of an observed trend in scores. He argues correctly that to interpret a score change as indicating a change in "pedagogical effectiveness" requires a careful evaluation of alternative explanations of how such a change may have occurred. He cites such factors as the degree of standardization in test administration and scoring, and the stability of the student cohort with respect to characteristics related to academic performance. The first factor can be controlled with sufficient effort. The second is dependent, in part, on the equivalence of the one-quarter samples employed each year. The degree of equivalence can be estimated by comparing the samples with respect to the distributions of relevant background characteristics. Major demographic shifts are unlikely to be a problem in adjacent years but could be an issue for the analysis of longer term trends. I would add that trends observed at the aggregate level are likely to be of limited utility. Observing trends for comparable subgroups of schools (determined by a small set of school-level characteristics such as sector, location, etc.) should be more informative.

Trend comparisons for an individual school are more problematic since the external examination occurs only on a four year cycle. It appears that differences in the administration and scoring between the external and internal examinations are sufficiently great as to preclude meaningful comparisons. Comparisons over four or eight years may be confounded with changes in curriculum, school personnel, etc. In addition, small sample sizes cause substantial uncertainty in the annual estimates and even greater uncertainty in the estimate of the difference. Thus, school-level trends should be approached with extreme caution.

[Normative interpretations of test scores] The normative interpretation of a test score involves the comparison of the score of an individual unit (e.g. a school) to the distribution of scores for some reference group of units. The choice of the reference group will depend on the question the comparison is intended to answer. Different questions will suggest different reference groups. For example, we may want to know the standing of a school relative to all schools in the country or, alternatively, relative to all schools with similar student demographics. Thus, a school with a relatively advantaged student population may stand at the 80^{th} percentile in the distribution of scores for all schools but at the 50^{th} percentile in the distribution of scores for all schools with similar demographics. Both results may be of interest and useful in different ways.

It is certainly the case that normative results are no substitute for absolute results, were the latter available. As is argued above, in educational measurement absolute scales are rarely found and the criterion-referenced standards are the best one can do. Nonetheless, with criterion-referenced standards it is still meaningful to track trends in performance over time. That is, one can say that, in comparison to last year, this year a higher proportion of schools obtained a school average that was above the proficient level. To be sure, that difference may be due in part to differences in the cohort of students examined so that the causal attribution of the trend to improved school effectiveness must be made carefully, if at all.

Certainly, a school's percentile relative to a particular reference group can also be tracked over time. However, as Professor Cahan points out, there are many pitfalls in interpretation. Since the reference group remains constant, one school's improvement must be matched by another school's decline. It is also the case that one cannot infer trends with respect to criterion-referenced standards from normative trends. For example, it is possible for all schools to improve with respect to the standards but for there to be changes in their relative positions in the distribution of scores of the reference group. Moreover, a school's change in position can be different depending on the choice of the reference group since the change depends crucially on the performance of all the other schools in the particular reference group.

For this reason and others, it is important that users of normative results should be educated on the proper interpretations -- and warned of the most common misinterpretations. Unfortunately, as Professor Cahan warns, there are no easy fixes and the issues can be subtle, particularly for those with little or no background in measurement. Moreover, some of the issues discussed above, such as the lack of comparability in standards across subjects, preclude certain kinds of inferences. Thus, a reasonable question is whether normative results can be of any use at all.

The answer, I believe, is a qualified "yes". For example, if the reference distribution is determined at one point in time and then fixed, in subsequent years, schools can compare their performance to the historical results. In that case, all schools could record increases in their percentile rankings and, perhaps, eventually be above (the historical) average! Of course, such results depend on the year-to-year comparability of tests and their reporting scales.

Comparisons to a selected norm group can be useful to school personnel, even in the absence of the possibility of a direct causal inference. For example, if the principal learns that her school's percentile ranking relative to a reference group of similar schools fell from the 50^{th} percentile to the 30^{th} percentile, then that is at least a signal that she and her staff should take a careful look at what has transpired in the school between test administrations. The point is that school staff need not – and should not --rely solely on the test results to reach a conclusion as to whether there is a problem and, if so, what to do about it. Unlike the central authorities, they have a much richer knowledge base for evaluating the situation. Changes in demographics, student mobility, flux in the teaching staff, and the economy of the area can all contribute to changes in performance. To be

sure, not every school will have the drive and the capacity to make such an evaluation, and there will be a natural tendency to "explain away" uncomfortable results. It is exactly here that targeted support in the form of expert advisor(s) and additional resources can play a key role.

Policy issues

[Completeness of the information provided to the school] One way to frame Professor Cahan's critique is to imagine a three-dimensional matrix whose dimensions are schools, grades and subjects. A cell in this matrix corresponds to a combination of a particular school, a specific grade and a specific subject. In any one year, the proportion of cells representing the data collected by the external examination system is relatively small; that is, the data matrix is quite sparse. Indeed, it is so sparse that Professor Cahan argues that it is nearly useless with respect to its intended purposes. Further, he asserts that, in order to meet that purpose, a full matrix is needed each year.

The current GEMS and Professor Cahan's preference can be thought of as standing at two ends of a continuum. Policy makers must decide what point along the continuum represents the best trade-off between utility and (generalized) costs. Obviously, the expense of a testing system that annually generates a full matrix is enormous. The costs include development, administration, scoring and reporting, as well as the burden placed on school officials and teachers. A comparable system introduced in England was eventually abandoned because of protests by school personnel motivated by their unhappiness with the system as implemented.

One strategy that might lead to some resolution of this impasse would be to engage in extensive consultations with various stakeholders (especially principals and other instructional leaders) about what changes could significantly enhance GEMS' value to school-level decision-making. It may be that a modest expansion of the current system would offer a reasonable cost/benefit trade-off. This would correspond to a point on the continuum between the two ends.

On the other hand, it may be that a solution "off the continuum" would be superior. For example, one can imagine enhancing the credibility of the internally administered exams by instituting an audit system for both administration and scoring. The scoring component of such an audit system could be incorporated into a program of teacher professional development. Research has shown that participation in carefully designed training based on evaluation of student work is highly valued by teachers – and when combined with appropriate pedagogical support – leads to improved student learning. Such a strategy would certainly not be cost free, but could provide useful information for school-level decision-making on an annual basis, even if the results were not incorporated into the system-level reports.

[Timing of test results] There is always a trade-off between the timing of test administration and the provision of the score reports to schools. If the tests are meant to

evaluate student attainment through the end of a particular grade, then they must be administered sometime near the conclusion of the school year. Unless the system is fully computer-based (for administration, scoring and reporting), it is difficult to make the results available before the end of the school year. Even if that were possible, it would still not be of use for that year. On the other hand, if the results are delivered some time during the summer, they could be useful for planning for the coming academic year – particularly if they were considered in conjunction with the results of the internally administered examinations. One must be resigned to the reality that the outcomes of large-scale assessment surveys such as GEMS, standing on their own, will generally be of greater value at the aggregate level than at the pupil level or even the school level.

[Limited school-level administrative flexibility] This is an important issue in the context of overall strategy and resource allocation because it speaks to the ultimate utility of the testing system. Professor Cahan asserts that principals have relatively limited discretionary funds to employ in response to GEMS data. In view of his critique of the current system, the essential point is that if the scope of GEMS were to be expanded in some way then concomitant steps should be taken to provide support (financial and otherwise) for schools where significant improvement is needed. Of course, this would add to the cost burden for the state.

With respect to resource allocation, then, the alternatives may come down to: (i) Major expansion of GEMS but no increase in school-level funding; (ii) Modest increases in the testing budget (e.g. investing in the enhancement of the internal exam administration) along with modest increases in targeted school-level funding; (iii) Minor increases in the testing budget but significant increases in targeted school-level funding; (iv) Some other combination, yet to be defined. As Professor Cahan argues elsewhere, that decision should be informed both by extensive consultations with stakeholders and by rigorous technical (psychometric, economic and logistical) analyses of the various options.

Final thoughts

In his wide-ranging critique of GEMS, Professor Cahan raises a number of important issues and also provides some suggestions on how the MoE might proceed in reconsidering its investment in this area, as well as possible strategies for enhancing the relevance, credibility and utility of the information generated.

As stated at the outset, my view is not as pessimistic as Professor Cahan's. In its present incarnation, GEMS may well be more useful to the central authorities than to individual schools. Modest investments in improving test quality, strengthening comparability across years and setting performance standards should yield substantial dividends in credibility and utility at both levels. In addition, special consideration should be given to providing more guidance and support to school officials who want to make use of GEMS results. In addition to traditional channels, new approaches such as web seminars (webinars) can be employed to good effect.

Beyond these short-term enhancements, revisiting the purposes of a national testing program in the context of current and future realities is called for. A number of alternatives were suggested by Professor Cahan and by the writer. Undoubtedly, others will emerge through further consultations both with experts and with various stakeholders. The lessons learned with the design, development, implementation and evolution of GEMS should be invaluable.
Ultimately, a decision will be the result of a political compromise and the key is to ensure that the recommended design can plausibly lead to improved learning for all students.

There is a large literature on the use of indicators to monitor the performance of governmental and quasi-governmental organizations. By their nature, indicators provide summary descriptions of one or more aspects of how the organization functions. However, they are usually intended to stimulate action by one or more stakeholders. Thus, the working assumption is that the indicator outcomes can be correctly interpreted and that appropriate actions can then be taken.

Reality is generally much more complex and rarely do indicator systems work as promised. To cite but one example, when the stakes attached to the indicators are negligible or low, then the responses to the reported outcomes tend to be quite variable; that is, some will take strong action, some will respond to a degree and others will ignore the matter. This can be frustrating. However, if, to counter the frustration, meaningful consequences are attached to the indicators, then the result over time can be severe distortions in both the indicators and the processes they are meant to monitor. This is particularly the case if there is a substantial discrepancy between what the indicator captures and what is truly valued – a situation all too common in education.

The point of the above example is simply to urge caution in designing an indicator system. The challenge is an enormous one: In addition to the various issues discussed in this note in the context of GEMS, one must try to anticipate the dynamic human and institutional responses to any externally imposed data collection system. One must be humble in the face of human frailty, ingenuity, and perversity!