



Highlights of [GAO-05-671](#), a report to the Committee on Finance, U.S. Senate

Why GAO Did This Study

In the past, the Internal Revenue Service (IRS) has experienced declines in enforcement productivity as measured by cases closed per Full Time Equivalent. Increasing enforcement productivity through a variety of enforcement improvement projects is one strategy being pursued by IRS. Evaluating the benefits of different projects requires good measures of productivity. In addition, IRS's ability to correctly measure its productivity has important budget implications.

GAO was asked to illustrate available methods to better measure productivity at IRS. Specifically, our objectives were to (1) describe challenges that IRS faces when measuring productivity, (2) describe alternative methods that IRS can use to improve its productivity measures, and (3) assess the feasibility of using these alternative methods by illustrating their use with existing IRS data.

What GAO Recommends

GAO recommends that the Commissioner of Internal Revenue put in place a plan for introducing wider use of alternative methods of measuring productivity, such as those illustrated in this report, taking account of the costs of implementing the new methods. The Commissioner of Internal Revenue agreed with our recommendation and assigned responsibility for considering alternative methods of measuring productivity.

www.gao.gov/cgi-bin/getrpt?GAO-05-671.

To view the full product, including the scope and methodology, click on the link above. For more information, contact James White at (202) 512-9110 or whitej@gao.gov.

TAX ADMINISTRATION

IRS Can Improve Its Productivity Measures by Using Alternative Methods

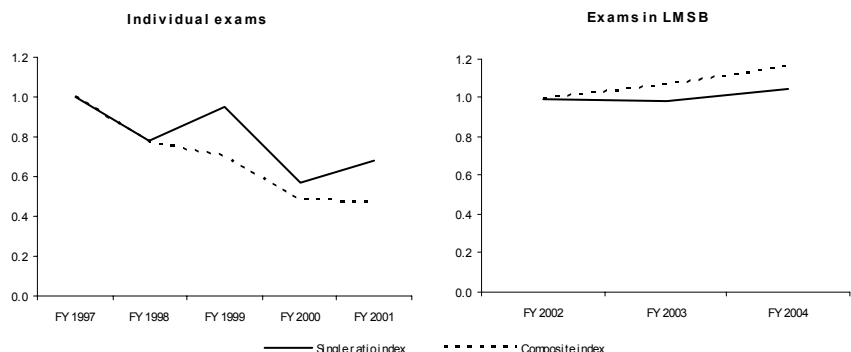
What GAO Found

Measuring IRS's productivity, the efficiency with which inputs are used to produce outputs, is challenging. IRS's output could be measured in terms of impact on taxpayers or the activities it performs. IRS's impacts on taxpayers, such as compliance and perceptions of fairness, are intangible and costly to measure. IRS's activities, such as exams or audits conducted, are easier to count but must be adjusted for complexity and quality. An increase in exams closed per employee would not indicate an increase in productivity if IRS had shifted to less complex exams or if quality declined.

IRS can improve its productivity measures by using a variety of methods for calculating productivity that adjust for complexity and quality. These methods range from ratios using a single output and input to methods that combine multiple outputs and inputs into composite indexes. Which method is appropriate depends on the purpose for which the productivity measure is being calculated. For example, a single ratio may be useful for examining the productivity of a single simple activity, while composite indexes can be used to measure the productivity of resources across an entire organization, where many different activities are being performed.

Two examples show that existing data, even though they have limitations, can be used to produce a more complete picture of productivity. For individual exams, composite indexes controlling for exam complexity show a larger productivity decline than the single ratio method. On the other hand, for exams performed in the Large and Mid-Size Business (LSMB) division, the single ratio understates the productivity increase shown, after again controlling for complexity. By using alternative methods for measuring productivity, managers would be better able to isolate sources of productivity change and manage resources more effectively. More complete productivity measures would provide better information about IRS effectiveness, budget needs, and efforts to improve efficiency.

Illustrations of Exam Productivity Indexes before and after Controlling for Complexity



Source: GAO analysis of IRS data.

Contents

Letter	1
Results in Brief	3
Background	4
Measuring Productivity at IRS Is Challenging because Measuring the Output of Services Is Difficult	6
However Output Is Measured, IRS Can Improve Its Current Productivity Measures by Using Alternative Methods	9
Illustrations of Alternative Methods of Measuring Productivity	12
Conclusion	19
Recommendations for Executive Action	20
Agency Comments and Our Evaluation	20

Appendixes	
Appendix I: Methods for Calculating Productivity Indexes	21
Productivity Indexes	21
Estimation of Distance Functions	23

Table	Table 1: Summary of Output Measures	9
-------	-------------------------------------	---

Figures	Figure 1: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for All Individual Returns	14
	Figure 2: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for Individual Returns (without EIC)	15
	Figure 3: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for LMSB Exams	16
	Figure 4: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for LMSB Exams (Excluding Individual and Corporate Exams)	17

This is a work of the U.S. government and is not subject to copyright protection in the United States. It may be reproduced and distributed in its entirety without further permission from GAO. However, because this work may contain copyrighted images or other material, permission from the copyright holder may be necessary if you wish to reproduce this material separately.



United States Government Accountability Office
Washington, D.C. 20548

July 11, 2005

The Honorable Charles E. Grassley
Chairman
The Honorable Max Baucus
Ranking Minority Member
Committee on Finance
United States Senate

In the past, we have reported on declines in the Internal Revenue Service's (IRS) enforcement programs, including declining exam and collection efforts.¹ One factor we have cited as contributing to these declines is decreased enforcement productivity as measured by cases closed per staff time.² Increasing enforcement productivity through a variety of enforcement improvement projects is one strategy being pursued by IRS that could help reverse the declines. However, evaluating the benefits of these different projects requires good measures of productivity. IRS's ability to correctly measure its productivity has important budget implications. Productivity declines may indicate that IRS is not using its resources as efficiently as possible. Increasing the productivity of existing resources might lessen, to some extent, the need for budget increases.

Productivity is measured as a ratio of outputs to inputs. In a January 2004 report on IRS's enforcement improvement projects,³ we recommended that IRS invest in enforcement productivity data that better adjust for complexity and quality, taking into consideration the costs and benefits of doing so. More complete productivity data—data that adjust for complexity and quality—would give managers a clearer picture of how effectively resources are being used. In addition, Congress would have better information about IRS's performance and budget needs. To better understand productivity measurement at IRS, you asked us to illustrate methods available to better measure it. Specifically, our objectives were to

¹GAO, *Compliance and Collection: Challenges for IRS in Reversing Trends and Implementing New Initiatives*, [GAO-03-732T](#) (Washington, D.C.: May 7, 2003), and *IRS Modernization: Continued Progress Necessary for Improving Service to Taxpayers and Ensuring Compliance*, [GAO-03-796T](#) (Washington, D.C.: May 20, 2003).

²GAO, *Tax Administration: Impact of Compliance and Collection Program Declines on Taxpayers*, [GAO-02-674](#) (Washington, D.C.: May 22, 2002).

³GAO, *Tax Administration: Planning for IRS's Enforcement Process Changes Included Many Key Steps but Can Be Improved*, [GAO-04-287](#) (Washington, D.C.: Jan. 20, 2004).

(1) describe challenges that IRS faces when measuring productivity, (2) describe alternative methods that IRS can use to improve its productivity measures, and (3) assess the feasibility of using these alternative methods by illustrating their use with existing IRS data.

In the context of the productivity literature, output is a general concept representing what is produced. However, in the performance measurement literature, the term “output,” as defined in the Government Performance and Results Act of 1993 (GPRA)⁴ is limited to an activity or effort, while an outcome is the result of a program activity. Activities are typically easily measured, such as transactions completed. Results such as the difference an activity makes in the economy or people’s lives are usually less tangible. In this report, we use the general concept of “output” to define productivity but then distinguish between outputs that are results and those that are activities.

To describe the challenges IRS faces when measuring productivity and alternative methods IRS can use to improve its productivity measures, we reviewed the literature on the methods used to measure productivity in the public and private sectors. We also consulted IRS officials and reviewed IRS documentation on IRS’s methods for measuring productivity.

To assess the feasibility of using these alternative methods by illustrating their use with existing IRS data, we used currently available IRS data to calculate alternative exam, or audit, productivity measures. These methods included calculating unweighted productivity indexes and weighted productivity indexes. We compared these indexes to show how implementing different methods can provide IRS with better measures of productivity and better ways to identify the causes of productivity change.

For this report existing IRS examination data were used to illustrate the feasibility of using alternative methods of productivity. The data are from IRS’s Tax Compliance Report and Automated Inventory Management System.⁵ In prior reports we recognized that IRS’s existing examination data have limitations. For example, direct measures of complexity were not available. We use type of exam as a proxy for complexity. We have also

⁴P. L. No. 103-62 (1993).

⁵IRS, *Tax Compliance Activities Report*, June 24, 2002, prepared in response to a directive in the House Report accompanying the legislation (P.L. 107-67).

recommended that IRS improve its input data by implementing a cost accounting system. While there are reliability issues related to the data, we are using the available IRS data for illustrative purposes, and we will not be representing these illustrations as complete measures of IRS productivity. Therefore, we determined that the information contained in IRS's Tax Compliance Report and Automated Inventory Management System databases were sufficiently reliable for illustrative purposes.

We initiated our review in September 2003 but conducted most of our review from August 2004 through April 2005 in accordance with generally accepted government auditing standards.

Results in Brief

Because IRS provides services, such as providing information to taxpayers and enforcing the tax laws, that are intangible and complex, measuring output—and therefore productivity—is challenging. Productivity is the efficiency with which inputs are used to produce outputs. IRS can use its activities or the results of its activities or services as measures of output. IRS's results are the impacts on the condition or behavior of taxpayers, such as compliance and compliance burden. IRS's activities are what IRS does to achieve those results, such as phone calls answered and exams conducted. Generally, information about results is preferred, but measuring such results is often difficult. Activities may be used instead to provide information about internal efficiency—how effectively IRS is using resources to perform a specific function—or as proxies for ultimate results to which the activities are closely related.

IRS can improve its productivity measures by using alternative methods for calculating productivity that adjust for complexity and intangibles such as quality. The methods range from computing ratios of single outputs to inputs—exams closed per Full Time Equivalent (FTE)—to using statistical methods to combine multiple indicators of outputs and inputs. Which method is appropriate depends on the purpose for which the productivity measure is being calculated. For example, a single ratio may be useful for examining the productivity of a single simple activity, while composite indexes can be used to measure the productivity of resources across an entire organization, where many different activities are being performed.

Existing IRS data can be used to illustrate alternative exam productivity measures that adjust for complexity and quality. For example, the single ratio index, unadjusted for complexity or quality, shows a decline in individual exam productivity (as measured by exams closed per FTE) of

32 percent from 1997 to 2001. A composite index, controlling for complexity, shows a larger decrease of 53 percent. The composite measure shows a greater decline because it accounts for IRS's shift to less complex Earned Income Credit (EIC) exams. On the other hand, for examinations conducted by IRS's Large and Mid-Sized Business (LMSB) division from 2002 to 2004, a single ratio index understates productivity improvements. The single ratio index shows a productivity gain of 4 percent. After adjusting for changes in the complexity of exams over those years, productivity increased by 16 percent. Consistent with our 2004 report on IRS's enforcement improvement projects, IRS officials said they generally use single ratios as measures of productivity. More complete productivity measures would provide better information about the effectiveness of IRS resources, IRS's budget needs, and IRS's efforts to improve efficiency.

We are making a recommendation to investigate the use of alternative methods of measuring productivity.

Background

Productivity is defined as the efficiency with which inputs are used to produce outputs. It is measured as the ratio of outputs to inputs. Productivity and cost are inversely related—as productivity increases, average costs decrease. Consequently, information about productivity can inform budget debates as a factor that explains the level or changes in the cost of carrying out different types of activities. Improvements in productivity either allow more of an activity to be carried out at the same cost or the same level of activity to be carried out at a lower cost.

IRS currently relies on output-to-input ratios such as cases closed per FTE to measure productivity and productivity indexes. A productivity change is measured as an index which compares productivity in a given year to productivity in a base year. Measuring productivity trends requires choosing both output and input measures, and the methods for calculating productivity indexes.

In the past we have reported on declining enforcement trends, finding in 2002 that there were large and pervasive declines in six of eight major compliance and collection programs we reviewed. In addition to reporting these declines, we reported on the large and growing gap between collection workload and collection work completed and the resultant increase in the number of cases where IRS has deferred collection action

on delinquent accounts.⁶ In 2003, we reported on the declining percentage of individual income tax returns that IRS was able to examine or audit each year, with this rate falling from 0.92 percent to 0.57 percent between 1993 and 2002.⁷ Since 2000, the audit rate has increased slightly but not returned to previous levels. IRS conducts two types of audits: field exams that involve complex tax issues and usually face-to-face contact with the taxpayer, and, correspondence exams that cover simpler issues and are done through the mail. We also reported on enforcement productivity measured by cases closed per FTE employee, finding that IRS's telephone and field collection productivity declined by about 25 percent from 1996 through 2001 and productivity in IRS's three exam programs—individual, corporate, and other audit—declined by 31 to 48 percent.⁸

In January 2004 we reported on the extent to which IRS's Small Business and Self-Employed (SB/SE) division followed steps consistent with both GAO guidance and the experience of private sector and government organizations when planning its enforcement process improvement projects. We reported on how the use of a framework would increase the likelihood that projects target the right processes for improvement and lead to the most fruitful improvements. In that report, we also reported that more complete productivity data—input and output measures adjusted for the complexity and quality of cases worked—would give SB/SE managers a more informed basis for decisions on how to identify processes that need improvement, improve processes, and assess the success of process improvement efforts. This report elaborates on that recommendation, providing more information about the challenges of obtaining complete productivity data.

Improving productivity by changing processes is a strategy SB/SE is using to address these declining trends. However, the data available to SB/SE managers to assess the productivity of their enforcement activities, identify processes that need improvement, and assess the success of their process improvement efforts are only partially adjusted for complexity and quality of cases worked. This problem of adjusting for quality and complexity is not unique to SB/SE process improvement projects—the data available to

⁶GAO-02-674.

⁷GAO, *Tax Administration: IRS Should Continue to Expand Reporting on Its Enforcement Efforts*, GAO-03-378 (Washington, D.C.: Jan. 31, 2003).

⁸GAO-02-674.

process improvement project managers are the same data used throughout SB/SE to measure productivity and otherwise manage enforcement operations.

Measuring Productivity at IRS Is Challenging because Measuring the Output of Services Is Difficult

Because IRS provides services, such as providing information to taxpayers and enforcing the tax laws, that are intangible and complex, measuring output—and therefore productivity—is challenging. Like other providers of intangible and complex services, IRS has a choice of measuring activities or the results of its services. Generally, information about results is preferred, but measuring results is often difficult. In the absence of direct measures of results, activities that are closely related to the results of the service can be used as proxies.

Measuring productivity in services is difficult. Unlike manufacturing, which lends itself to objective measurement because output can be measured in terms of units produced, services, which involve changes in the condition of people receiving the service, often have intangible characteristics. Thus, the output of an assembly line is easier to measure than the output of a teacher, doctor, or lawyer. Services may also be complex bundles of individual services, making it difficult to specify the different elements of the service. For example, financial services provide a range of individual services, such as financial advice, accounts management and processing, and facilitating financial transactions.

IRS provides a service. IRS's mission, to help taxpayers understand and meet their tax responsibilities and to apply the tax law with integrity and fairness, requires IRS to provide a variety of services ranging from collecting taxes to taxpayer education. IRS, like other service providers, could measure output in terms of its results—its impact on taxpayers—or in terms of activities. The results of IRS's service are the impacts on the condition or behavior of taxpayers. These taxpayer conditions or behaviors include their compliance with the tax laws, their compliance burden (the time and money cost of complying with tax laws), and their perception of how fairly taxpayers are treated. IRS's activities are what IRS does to achieve those results. These activities include phone calls answered, notices sent to taxpayers, and exams conducted.

Generally, information about results is preferred, but measuring such results is often difficult. In the case of the public sector, this preference is reflected in GPRA, which requires that federal agencies measure performance, whenever possible, in terms of results or outcomes for

people receiving the agencies' services. However, results such as compliance and fairness have intangible characteristics that are difficult to measure. In addition, results are produced in complicated and interrelated ways. For example, a transaction or activity may affect a number of results: IRS's exams may affect taxpayers' compliance, compliance burden, and perceptions of the fairness of the tax system. In addition, a result may be influenced by a number of transactions or activities: A taxpayer's compliance may be influenced by all IRS exams (through their effect on the probability of an exam) as well as by other IRS activities such as taxpayer assistance services.

IRS's activities are easier to measure than results but still present challenges. Activities are easier to measure because they are often service transactions such as exams, levies issued, or calls answered that can be easily counted. However, unlike measures of results, more informative measurement of activities requires that they be adjusted for quality and complexity, as we noted in our report on IRS's enforcement and improvement projects.⁹ A productivity measure based on activities such as cases closed per FTE may be misleading if such adjustments are not made. For example, an increase in exam cases closed per FTE would not indicate an increase in true productivity if the increase occurred because FTEs were shifted to less complex cases or the examiner allowed the quality of the case review to decline to close cases more quickly.

Activities-based productivity measures can provide IRS with useful information on the efficiency of IRS operations. Measuring output, and therefore productivity, in terms of activities provides IRS with measures of how efficiently it is using resources to perform specific functions or transactions. However, activities do not constitute—and should not be mistaken for—measures of IRS's productivity in terms of ultimate results.

While the productivity measures we have examined are based on activities, IRS has efforts under way to measure results such as compliance and compliance burden. Recently, we reported on IRS's National Research Program (NRP) to measure voluntary compliance and efforts to measure

⁹By measuring the actual impact on taxpayers, measures of results incorporate the quality and complexity of the service.

compliance burden.¹⁰ As we mentioned previously, measuring these results is difficult. For some results, such as compliance, measurement is also costly and intrusive because taxpayers must be contacted and questioned in detail. Despite these difficulties, IRS can improve its productivity measurement by continuing its efforts to get measures of results. These efforts would give Congress and the general public a better idea of what is being achieved by the resources invested in IRS.

In the absence of direct measures of results, activities that are closely related to the results of the service are used as proxies. The value of these proxies depends on the extent to which they are correlated with results. By carefully choosing these measures, IRS may gain some information about the effect of its activities on ultimate results. Because activities may affect a number of results and a single result may be affected by a number of activities, a single activity likely will not be a sufficient proxy for the results of the service. Therefore, a variety of activities would likely be necessary as proxies for the results of the service.

Both types of output measures, those that reflect the results of IRS's service and those that use activities to measure internal efficiency, should be accurate and consistent over time. In addition, both output measures should be reliably linked to inputs. Linking the results of IRS's service to inputs may be difficult because of outside factors that may also affect measured results. For example, an increase in compliance could result both from IRS actions such as exams and from changes in tax laws. Another challenge is that IRS currently has difficulties linking inputs to activities, as we note in a previous report, where we reported IRS's lack of a cost accounting system. In particular, IRS only recently implemented a cost accounting system, and has not yet determined the full range of its cost information needs. Table 1 summarizes some of the key differences between activities and results measures. Table 1 also indicates some general criteria that apply to both types of measures.

¹⁰GAO, *Tax Administration: IRS Is Implementing the National Research Program as Planned*, [GAO-03-614](#) (Washington, D.C.: June 16, 2003), and *Tax Administration: New Compliance Research Effort Is on Track, but Important Work Remains*, [GAO-02-769](#) (Washington, D.C.: June 27, 2002) look at IRS's research on compliance, and *Tax Administration: IRS Is Working to Improve Its Estimates of Compliance Burden*, [GAO/GGD-00-11](#) (Washington, D.C.: May 22, 2000) reported on IRS's measures of compliance burden.

Table 1: Summary of Output Measures

Type of measure	Purpose	Criteria
Activities	• Measure internal efficiency	Activities measures should
	• Serve as a proxy for results	• reflect the work performed • adjust for quality and complexity • be accurate and consistent over time and reliably linked to inputs
Results	• Measure impact on taxpayers	Results measures should • reflect the effects of the service • be accurate and consistent over time and reliably linked to inputs

Source: GAO analysis.

Because inputs are more easily measured and identifiable than outputs, measuring them is more straightforward. IRS, as a government agency, may be able more often to use labor costs or hours as a single input in its productivity measures because it relies heavily on labor. However, it may be particularly important for IRS to use a multifactor measure that includes capital along with labor during periods of modernization that involve increased or high levels of capital investment. As with outputs, inputs should be measured accurately and consistently over time. Measuring inputs consistently over time may require adjusting for changes in the quality of the labor, which has been done using proxies such as education level or years of experience. Also, as mentioned previously, inputs should be reliably linked to outputs.

However Output Is Measured, IRS Can Improve Its Current Productivity Measures by Using Alternative Methods

The appropriate method for calculating productivity depends on the purpose for which the productivity measure is being calculated. The alternative methods that can be used for calculating productivity range from computing single ratios—exams closed per FTE—to using complex statistical methods to form composite indexes that combine multiple indicators of outputs and inputs. While single ratios may be adequate for certain purposes, the composite indexes based on statistical methods may

be more useful because they provide information about the sources of productivity change.¹¹

Comparing the ratios of outputs to inputs at different points in time defines a productivity index that measures the percentage increase or decrease in productivity over time. The ratios that form the index may be single, comparing a single output to a single input or composite, where multiple outputs and inputs are compared. The single ratios may be useful for evaluating the efficiency of a single noncomplex activity. Composite indexes can measure the productivity of more complicated activities, controlling for complexity and quality. Composite indexes can also be used to measure productivity of resources across an entire organization, where many different activities are being performed.¹²

One method of producing composite indexes is to use weights to combine such disparate activities as telephone calls answered and exams closed. One common weighting method, used by the Bureau of Labor Statistics (BLS), is a labor weight. Weighting outputs by their share of labor in a baseline period controls for how resources are allocated between different types of outputs. If the productivity of two activities is unchanged but resources are reallocated between the activities, the composite measure of productivity would change unless these weights are employed. For example, if IRS reallocates exam resources so that it does more simple exams and fewer complex exams, the number of total exams might increase. Consequently, a single productivity ratio comparing total exams to inputs would show an increase. Labor weighting deals with this issue. The weights allow any gains from resource reallocation to be distinguished from gains in the productivity of the underlying activities. When types of activities can be distinguished by their quality of complexity, labor weighting can also be used to control for quality and complexity differences when resources are shifted between types of outputs.

¹¹For a more technical description of these methods, see app. I.

¹²For example, in GAO, *Tax Administration: IRS Needs to Further Refine Its Tax Filing Season Performance Measures*, [GAO-03-143](#) (Washington, D.C.: Nov. 22, 2002), we distinguished between the information provided by a productivity measure of individual returns processing functions and IRS's submission processing composite productivity measure of several different functions, including processing returns, remittances and refunds, and issuing notices and letters.

More complicated statistical methods can be used for calculating composite indexes that allow for greater flexibility in how weights are chosen to combine different outputs and for a wider range of output measures that include both qualitative and quantitative outputs. Data Envelopment Analysis (DEA), which has been widely used to measure the productivity of private industries and public sector services, is an example of such methods. DEA estimates an efficiency score for each producing unit, such as the firms in an industry or the schools in a school district, or for IRS, the separately managed areas and territories composing its business units. DEA estimates the relative efficiency of each producing unit by identifying those units with the best practice—those making the most efficient use of inputs, under current technology, to produce outputs—and measuring how far other units are from this best practice combination of inputs used to produce outputs. DEA estimates provide managers with information on how efficient they are relative to other units and the costs of making individual units more efficient.

These efficiency scores are used to form a composite productivity index called a Malmquist index. An advantage of the Malmquist index is that IRS managers can restrict the weights to adjust for managerial or congressional preferences to investigate the effect on productivity of a shift, for example, from an organization that emphasizes enforcement to one that emphasizes service. IRS can also include many different types of outputs and inputs, control for complexity and quality, and isolate the effects of certain historical changes, such as the IRS Restructuring and Reform Act of 1998 (RRA98).¹³

Another advantage of the Malmquist index is that productivity changes can be separated into their components, such as efficiency and technology changes. In this context, efficiency can be measured holding technology constant, and technology can be measured holding efficiency constant. Holding technology constant, IRS might improve productivity by improving the management of its existing resources. On the other hand, technology changes might improve productivity even if the management of resources has not changed. Thus, the productivity change of a given IRS unit is determined by both changes in its efficiency relative to the current best-practice IRS units and changes in the best practices or technology.

¹³P. L. No. 105-206 (1998).

Illustrations of Alternative Methods of Measuring Productivity

Currently available IRS data can be used to produce productivity indexes that control for complexity and quality. The examples that follow focus on productivity indexes that use exams closed as outputs and FTEs as inputs. The data on examinations cover individual returns across IRS and IRS's LMSB division. For both individuals and LMSB, the complexity and quality of exams can vary over time. For example, the proportion of exams that are correspondence versus field, business versus nonbusiness, and EIC versus non-EIC can vary over time. As already discussed, failing to take account of such variation can give a misleading picture of productivity change.

While these examples do not encompass all the methods, data, and adjustments that may be used, they illustrate the benefits of the additional analysis that IRS can perform using current data. In addition, as we pointed out in our 2004 report, IRS can improve its productivity measurement by investing in better data, taking into account the costs and benefits of doing so. These better data include measures of complexity, improved measures of quality, and additional measures of output.

Figures 1 through 4 illustrate, using currently available data between fiscal years 1997 and 2004, the difference between weighted indexes that make an adjustment for complexity and unweighted indexes that make no adjustments.¹⁴ In the illustrations, a labor-weighted composite index, which can control for complexity, is contrasted with a single unweighted index to show how the simpler method may be misleading. (See app. I for a fuller description of the labor-weighted index.) In each case, complexity is proxied by type of exam. Although the data were limited (for example, the measure of complexity was crude), the illustrations show that making the adjustments that are possible provides a different picture of productivity than would otherwise be available.¹⁵

¹⁴In addition to using labor weighting and similar methods for adjusting for complexity and quality, IRS may be able to use Malmquist indexes estimated using statistical methods such as DEA.

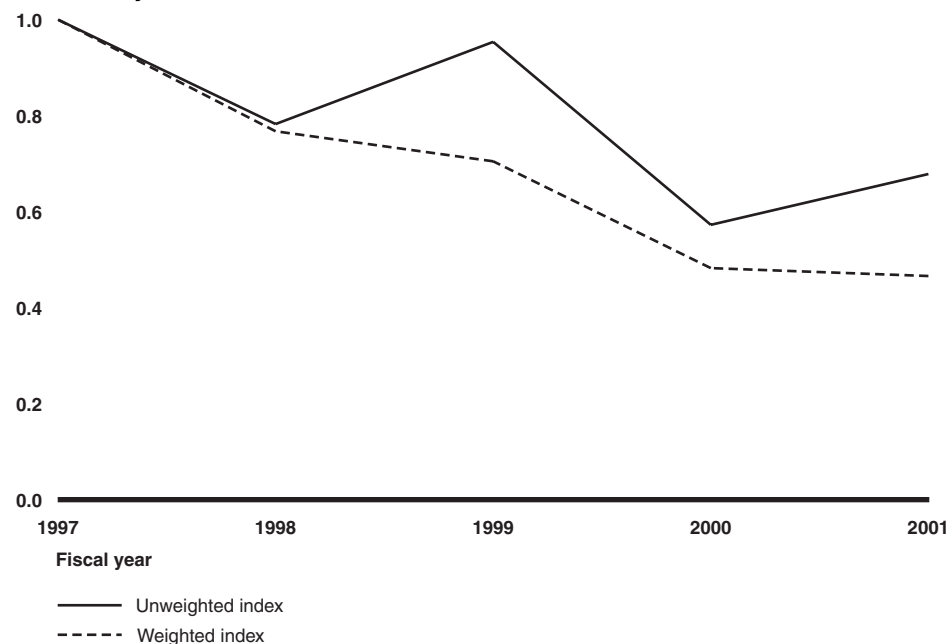
¹⁵We used the type of exam as a proxy for complexity based on the availability of data. Other proxies or direct measures might be used, although direct measures might be difficult to define and calculate. We included limited quality adjustments for the LMSB illustration only because, given that the purpose of the analysis is to illustrate methods, we determined it was not worthwhile to fully investigate the extent to which quality data currently available at IRS could be integrated with the exam-level data that we used for our analysis. Due to a lack of readily available data, capital inputs were not included.

The advantage of weighted indexes is that they allow changes in the mix of exams to be separated from changes in the productivity of performing those exams. In the examples that follow, an unweighted measure could be picking up two effects. One effect is the change in the number of exams that an auditor can complete if the complexity or quality of the exam changes. The second effect is the change in the number of exams an auditor can complete if the time an auditor requires to complete an exam changes, holding the quality and complexity of exams constant. By isolating the latter effect, the weighted index more closely measures productivity, or the efficiency with which the auditor is working the exams.

For individual exams, the comparison of productivity indexes shows that the unweighted index understates the decline in productivity. As figure 1 shows, between fiscal years 1997 and 2001, the unweighted productivity index declined by 32 percent while the weighted index declined by 53 percent. The difference is due largely to the increase in EIC exams during the period. Over the period between fiscal years 1997 and 2001, exams were declining. However, the mix of exams was changing, with increases in the number of EIC exams. EIC exams are disproportionately correspondence exams, and IRS can do these exams faster than field exams. IRS shifted to “easier” exams, and that shift caused the unweighted index to give an incomplete picture of productivity. The shift masked the larger productivity decline shown by the weighted index.¹⁶

¹⁶In figures 1 and 2, the exam types are correspondence and field exams, business and individual exams, and EIC exams. More specifically, the types for the weighted index are combinations of the following return categories: EIC and non-EIC; business and nonbusiness; low, medium, and high income; and correspondence and field exams. An example of an output type would be correspondence exams of non-EIC, nonbusiness high-income filers. The output types are meant to reflect differences in degrees of audit difficulty. Altogether, there are 13 output types used in the BLS index for individual returns.

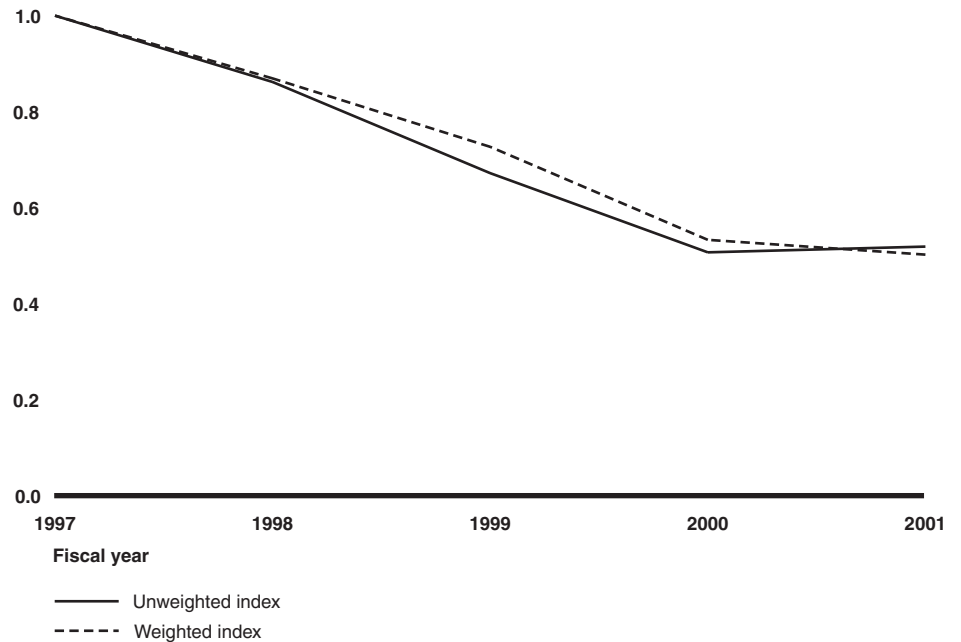
Figure 1: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for All Individual Returns



Source: GAO illustration base on analysis of IRS data (Tax Compliance Activities Report, 2002).

Figure 2 provides additional evidence to support the conclusion that the shift to more EIC exams is the reason for the difference in productivity shown in figure 1. Between fiscal years 1997 and 2001, the weighted and unweighted indexes track each other very closely when the EIC exams are removed. Both show a decline in productivity of about 50 percent over this period. The available data were not sufficient to control for other factors that may have influenced exam productivity. For example, RRA98 imposed additional requirements on IRS's auditors, such as certifications that they had verified that past taxes were due.

Figure 2: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for Individual Returns (without EIC)

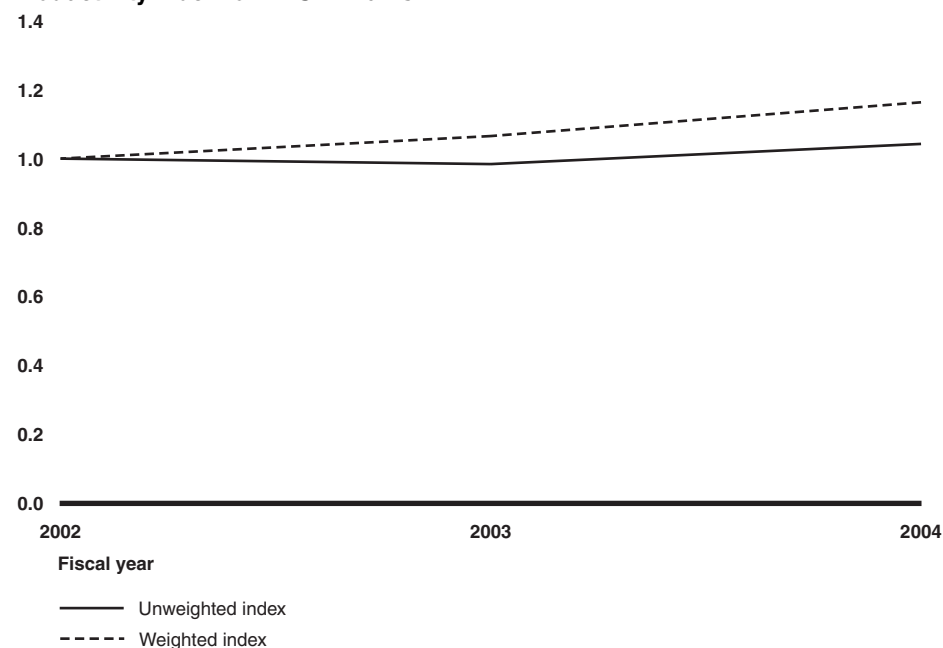


Source: GAO illustration base on analysis of IRS data (Tax Compliance Activities Report, 2002).

Figure 3 compares unweighted and weighted productivity indexes for exams done in LMSB division. As figure 3 shows, between fiscal years 2002 and 2004, the unweighted productivity index increased by 4 percent, while the weighted index increased by 16 percent. This difference appears largely due to the individual exams and small corporate exams done in LMSB. Over the period, total exams were declining but the mix of exams was changing. LMSB was shifting away from less labor-intensive individual returns and small corporation returns to more complex business industry

and coordinated industry return exams.¹⁷ This shift caused the unweighted index to give an incomplete picture of productivity. Here, the shift masked the larger productivity increase as shown by the weighted index.

Figure 3: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for LMSB Exams

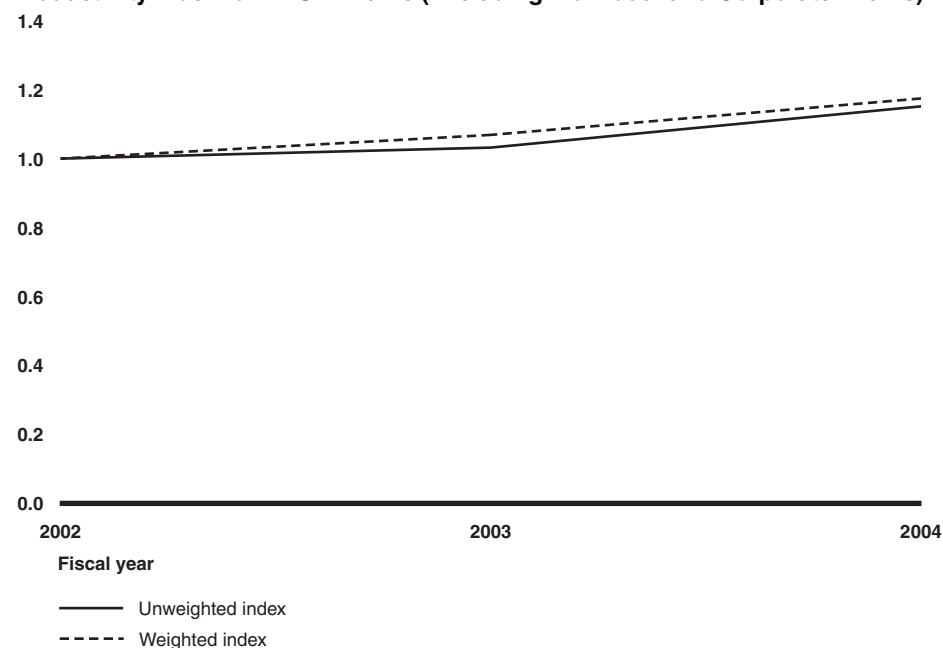


Source: GAO illustration based on analysis of IRS data (LMSB Critical Measures).

¹⁷In figures 3 and 4 the exams are distinguished by size and complexity of the business and whether they are individual or corporate exams. More specifically, the types for the weighted BLS index are combinations of the following return categories under LMSB: coordinated industry (large and more complex businesses); low income (under \$10 million) corporate exams; low (under \$100,000) and high (above \$100,000) income individual exams; and business industry exams (smaller or less complex business). The output types are meant to reflect differences in degree of audit difficulty. Altogether there are five output types in this illustration. While LMSB generally serves corporations, subchapter S corporations, and partnerships with assets greater than \$10 million, it also examines all the individual officers associated with corporations as well as any individual returns that cannot be done by the other divisions or that need the particular expertise of LMSB. LMSB will also examine small corporations that are associated with larger corporations, including those related to tax shelters.

Figure 4 provides additional evidence to support the conclusion that the shift away from individual and small corporate exams is the reason for the difference in productivity shown in figure 3. Between fiscal years 2002 and 2004, when individual and corporate exams are excluded, the two indexes track more closely, with the unweighted index increasing by 15 percent and the weighted index by 17 percent.

Figure 4: Base Year Labor-Weighted (Adjusted for Type of Exam) and Unweighted Productivity Index for LMSB Exams (Excluding Individual and Corporate Exams)



Source: GAO illustration based on analysis of IRS data (LMSB Critical Measures).

There is evidence that adjusting for quality would show that LMSB's productivity increased more than is apparent in figures 3 and 4 for the years 2002 to 2004. Average quality scores available for selected types of LMSB exams show quality increasing over the 2-year period.¹⁸ Adjusting for this

¹⁸Our use of these IRS exam quality scores is to illustrate how a quality adjustment can be made and does not mean that we endorse them as adequate measures of quality. We have indicated that the methodology for computing these scores could be improved by better adjusting for the new higher level of quality implied by the new standards imposed by RRA98. See [GAO-04-287](#).

increase in quality, in addition to adjusting for complexity, would show a productivity increase for these types of exams of 28 percent over the period.¹⁹

While labor-weighted and other more sophisticated productivity indexes can provide a more complete picture of productivity changes, they do not identify the causes of the changes. These productivity indexes would be the starting point for any analysis to determine the causes of productivity changes.

Another example of the advantages of weighted productivity indexes is provided by IRS. As noted earlier, IRS has developed a weighted submission processing productivity measure. The measure adjusts for differences in the complexity of processing various types of tax returns. In an internal analysis, IRS showed how productivity comparisons over time and across the 10 processing centers depended on whether or not the measure was adjusted for complexity. For example, the ranking of the processing centers in terms of productivity changed when the measure was adjusted for the complexity of the returns being processed.

The more sophisticated methods for measuring productivity can provide IRS and Congress with better information about IRS's performance. By controlling for complexity and quality, IRS managers would have more complete information about the true productivity of activities, such as exams, that can differ in these dimensions. In addition, the weighted measures can be used to measure productivity for the organization, where many different activities are being performed. More complete information about the productivity of IRS resources should be useful to both IRS managers and Congress. More complete productivity measures would provide better information about the effectiveness of IRS resources, IRS's budget needs, and IRS's efforts to improve efficiency.

Although there are examples, such as the submission processing productivity measures, of IRS using weighted measures of productivity, IRS officials said they generally use single ratios as measures of productivity. That is consistent with our 2004 report on IRS's enforcement improvement

¹⁹We included quality adjustments for the coordinated industry exam and business industry exam and therefore the productivity measure is for those exams. No quality measures were available for the corporate and individual exams.

projects, where we reported on SB/SE's lack of productivity measures that adjust for complexity and quality.

While there would be start-up costs associated with any new methodology, the long-term costs to IRS for developing more sophisticated measures of productivity may be modest. The examples so far in this section demonstrate the feasibility of developing weighted productivity indexes using existing data. Relying on existing data avoids the cost of having to collect new data. The fact that IRS already has some experience implementing weighted productivity measures could reduce the cost of introducing more such measures.

As we stated previously, IRS could also improve its productivity measurement by getting better data on quality and complexity. These improved data could be integrated with the methods for calculating productivity illustrated in this report to further improve IRS's productivity measurement. However, as we acknowledged in our prior report, collecting additional data on quality and complexity may require long-term planning and an investment of additional resources. Any such investment, we noted, must take account of the costs and benefits of acquiring the data.

Conclusion

Using more sophisticated methods, such as those summarized in this report, for tracking productivity could produce a much richer picture of how IRS manages its resources. This is important not only because of the size of IRS—it will spend about \$11 billion in 2005 and employ about 100,000 FTEs—but also because we are entering an era of tight budgets. A more sophisticated understanding of the level of productivity at IRS and the reasons for productivity change would better position IRS managers to make decisions about how to effectively manage their resources. Such information would also better enable Congress and the public to assess the performance of IRS.

As we illustrate, more can be done to measure IRS's productivity using current data. However, another advantage of using more sophisticated methods to track productivity is that the methods will highlight the value of better data. Better information about the quality and complexity of IRS's activities would enable the methods illustrated in this report to provide even richer information about IRS's overall productivity.

Recommendations for Executive Action

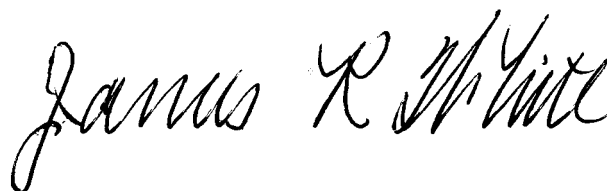
We recommend that the Commissioner of Internal Revenue put in place a plan for introducing wider use of alternative methods of measuring productivity, such as those illustrated in this report, taking account of the costs of implementing the new methods.

Agency Comments and Our Evaluation

The Commissioner of Internal Revenue provided written comments on a draft for this report in a June 23, 2005, letter. The Commissioner agreed with our recommendation to work on introducing wider use of alternative measure of productivity. Although expressing some caution, he has asked his Deputy Commissioner for Services and Enforcement to work with IRS's Research, Analysis, and Statistics office to assess the possible use of alternative methods of measuring productivity. The Commissioner recognized that a richer understanding of organizational performance is crucial for effective program delivery.

As agreed with your office, unless you publicly release its contents earlier we plan no further distribution of this report until 30 days from the date of this letter. At that time, we will send copies to interested congressional committees, the Secretary of the Treasury, the Commissioner of Internal Revenue, and other interested parties. We will also make copies available to others on request.

If you or your staff have any questions, please contact me at (202) 512-9110. I can also be reached by e-mail at whitej@gao.gov. Key contributors to this assignment were Kevin Daly, Assistant Director, and Jennifer Gravelle.



James R. White
Director, Tax Issues
Strategic Issues Team

Methods for Calculating Productivity Indexes

Productivity Indexes

Methods for calculating productivity range from computing single ratios to using statistical methods. In its simplest form, a productivity index is the change in the productivity ratio over time relative to a chosen year. However, this type of productivity index allows for only a single output and a single input. To account for more than one output, the outputs must be combined to produce a productivity index.

One method is to weight the outputs by their share of inputs used in the chosen base year. In a case where only labor input is used, following this method provides a labor-weighted output index, which, when divided by the input index, produces the labor-weighted productivity index. The use of the share of labor used in each output effectively controls for the allocation of labor across the outputs over time. For example, if productivity in producing two outputs remained fixed over time, a single productivity index may show changes in productivity if resources are reallocated to produce more of one of the outputs.¹

The Bureau of Labor Statistics (BLS) has also used labor-weighted indexes. BLS published, under the Federal Productivity Measurement Program, data on labor productivity in the federal government for more than two decades (1967-94). Due to budgetary constraints, the program is now terminated. BLS's measures used the "final outputs" of a federal program, which correspond generally to what we have called intermediate outputs in this report, as opposed to the outcomes or results of the program. BLS used labor weights because of their availability and their close link to cost weights. In particular, as with the labor weights in our illustrations, BLS used base year labor weights and updated the weights every 5 years. It relied only on labor and labor compensation, and acknowledges that the indexes did not reflect changes in the quality of labor. BLS measured productivity for a number of federal programs, ranging from social and information services to corrections. However, BLS did not produce productivity measures for IRS.

¹ In a simple example of one input and two outputs over 2 years, $Qa^1 = A^1 \cdot La^1$, $Qa^2 = A^2 \cdot La^2$, $Qb^1 = B^1 \cdot Lb^1$, $Qb^2 = B^2 \cdot Lb^2$, and labor-weighted productivity change would be equal to $x \cdot A^2 / A^1 + (1-x) \cdot B^2 / B^1$, where $x = La^2 / (La^2 + Lb^2)$ then $1-x = Lb^2 / (La^2 + Lb^2)$. However, assuming additive outputs, a nonweighted productivity change would be equal to $[x \cdot A^2 + (1-x) \cdot B^2] / [y \cdot A^1 + (1-y) \cdot B^1]$, where x is defined as above and $y = La^1 / (La^1 + Lb^1)$ then $1-y = Lb^1 / (La^1 + Lb^1)$.

In addition to weighted productivity indexes, there are a number of composite productivity indexes designed to include all the inputs and outputs involved in production. This group of indexes is called Total Factor Productivity (TFP) indexes.² They are called total because they include all the inputs and outputs, as opposed to Partial Factor Productivity indexes, which relate only one input to one output. Many of the main TFP indexes, including Tornqvist, Fisher, Divisia, and Paache, require reliable estimates of input and output prices, data not available for industries in the public sector. Therefore we use the Malmquist index, which does not require that data.

Malmquist indexes are TFP indexes based on changes in the distance from the production frontier, or distance functions. These distance functions are estimated using Data Envelopment Analysis (DEA). Productivity change is represented by the ratio of two different period distance functions. The Malmquist index is the geometric average of these productivity changes (evaluated at the two different periods).³ This index can be further decomposed into efficiency and technology changes.⁴ From the decomposition of the Malmquist index, productivity change can be shown to equal the efficiency change times the technology change.

The interpretation of changes in productivity, in terms of distance functions, depends on relative distances between periods. For simplicity, assume there was no change in technology between two periods, than the productivity change equals efficiency change. In this case, when the productivity index is less than one, the distance function in the second period is smaller than the distance function in the first period. Since the distance functions are less than one, this corresponds to a distance function in the second period that is a smaller fraction than the distance function in the first period. Since movements away from one show declining productivity, a smaller fraction in the second period, with a larger

² BLS regularly produces multifactor productivity measures, another term for TFP indexes, that reflect both labor and capital inputs.

³ Mathematically, the Malmquist index is defined as: $\{[D^t(x^{t+1}, y^{t+1})/D^t(x^t, y^t)] * [D^{t+1}(x^{t+1}, y^{t+1})/D^{t+1}(x^t, y^t)]\}^{1/2}$, where x^t , x^{t+1} denote the vector of inputs at time t and $t+1$, and y^t , and y^{t+1} denote the vector of outputs in time t and $t+1$ and D^t and D^{t+1} are distance functions relative to the technology in time t and $t+1$.

⁴ Malmquist index, $M = \{[D^t(x^{t+1}, y^{t+1})/D^t(x^t, y^t)] * [D^{t+1}(x^{t+1}, y^{t+1})/D^{t+1}(x^t, y^t)]\}^{1/2} = [D^{t+1}(x^{t+1}, y^{t+1})/D^t(x^t, y^t)] * \{[D^t(x^{t+1}, y^{t+1})/D^{t+1}(x^{t+1}, y^{t+1})] * [D^t(x^t, y^t)/D^{t+1}(x^t, y^t)]\}^{1/2} = E * T$, the efficiency change, E , times the technology change, T .

fraction in the first, indicates a movement away from one over time and thus declining productivity. Thus, a productivity change less than one indicates declining productivity and therefore an efficiency change less than one also indicates declining efficiency.

Alternatively, if the efficiency change was one, then the productivity change equals the technology change. Following previous analysis, a productivity change less than one indicates declining productivity. Therefore, a technology change less than one indicates an inward shift of the production frontier. If the technology change is less than one, it must be that the distance function in the first period is less than the distance function in the next period. Thus, the distance in the first period is farther away from one than is the distance in the next period, and the distance from the frontier decreased from the first period to the second period. Since the output and input bundles did not change, the frontier must shift in to produce the decrease in distance.

The Internal Revenue Service (IRS) can follow this method to generate indexes for the areas and territories and then focus on the average for an estimate of overall IRS productivity.

Estimation of Distance Functions

DEA is a nonparametric method for calculating distances from an estimated best practice production frontier. These distance functions are used to calculate malmquist indexes. Output distance functions are based on changes in output holding the amount of inputs constant.⁵ The output distance functions are estimated by a linear programming method which finds the scalar value that expands output as far as possible such that that output is still producible with the fixed level of inputs.⁶ Thus, a scalar value equal to one means that output could not be expanded any more without increasing the level of inputs. This situation indicates a firm that is efficient, producing the maximum amount of output with a given level of inputs and technology. Thus, firms with scalar values equal to one define the estimated best practice production frontier. However, a scalar value that is greater than one means that the firm could have more output than is currently produced with the same level of inputs. A firm in this situation is,

⁵ Mathematically, the distance function can be defined as: $D^t(x^t, y^t) = [\max \{ \phi \mid (x^t, \phi y^t) \in T \}]^{-1}$ and $\phi^* = (D^t(x^t, y^t))^{-1}$, with $\phi^* \geq 1$ and $D^t(x^t, y^t) \leq 1$, where ϕ denotes the value to scale output.

⁶ The linear programming problem is to $\max \phi$ subject to $\lambda x \leq x$, $\lambda y \geq \phi y$, $\lambda \geq 0$.

therefore, inefficient relative to firms with a scalar value of one. Thus, output distance functions are less than one. IRS can use this method, treating territories and areas as firms. The weights used in the linear program are designed to make each firm look its best; they represent best case scenarios.

While DEA is a nonparametric method, there is also a parametric method available called stochastic frontier analysis. Stochastic frontier analysis (regression) uses a regression model to estimate cost or production efficiency. After running the regression of performance and input data, the frontier is found by decomposing the residuals into a stochastic (statistical noise) part and a systematic portion attributed to some form of inefficiency. Stochastic frontier analysis thus requires specifying the distributional form of the errors and the functional form of the cost (or production) function. Its merits include a specific treatment of noise. While DEA's use of nonparametric methods eliminates the need to specify functional forms, one drawback is a susceptibility to outliers.

GAO's Mission

The Government Accountability Office, the audit, evaluation and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's Web site (www.gao.gov). Each weekday, GAO posts newly released reports, testimony, and correspondence on its Web site. To have GAO e-mail you a list of newly posted products every afternoon, go to www.gao.gov and select "Subscribe to Updates."

Order by Mail or Phone

The first copy of each printed report is free. Additional copies are \$2 each. A check or money order should be made out to the Superintendent of Documents. GAO also accepts VISA and Mastercard. Orders for 100 or more copies mailed to a single address are discounted 25 percent. Orders should be sent to:

U.S. Government Accountability Office
441 G Street NW, Room LM
Washington, D.C. 20548

To order by Phone: Voice: (202) 512-6000
TDD: (202) 512-2537
Fax: (202) 512-6061

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Web site: www.gao.gov/fraudnet/fraudnet.htm

E-mail: fraudnet@gao.gov

Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Gloria Jarmon, Managing Director, JarmonG@gao.gov (202) 512-4400
U.S. Government Accountability Office, 441 G Street NW, Room 7125
Washington, D.C. 20548

Public Affairs

Paul Anderson, Managing Director, AndersonP1@gao.gov (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149
Washington, D.C. 20548