Session Number: 3
Session Title: Measurement of government output
Session Organizer(s): David Caplan, Tim Smeeding
Session Chair: David Caplan

Paper Prepared for the 29th General Conference of
The International Association for Research in Income and Wealth

Joensuu, Finland, August 20 – 26, 2006

Measurement of Government Output in Education and Health: Alternative Approaches

Michael Christian, Bruce Baker, Barbara Fraumeni,
Alyssa Holdren, and Matthew Williams

For additional information please contact:

Michael Christian
Bureau of Economic Analysis, 1441 L Street NW, Washington DC 20230, USA
E-mail: Michael.Christian@bea.gov
Fax: 1-202-606-5366
Telephone: 1-202-606-9643

This paper is posted on the following websites: http://www.iariw.org

Introduction

The national income accounts of the United States currently estimate both the nominal and real value of government services using the value of the goods, services, and labor consumed by governments to produce those services. The resulting measure of government output, called the *input* measure, assumes that productivity in the government services sector is constant at zero. For example, the input measure implies that schools and hospitals cannot produce more education and health care services without employing more inputs. It also implies that schools and hospitals inevitably produce more education and health care services if they do employ more inputs.

An alternative to the input measure is a *volume* measure of output, which is an index that attempts to directly measure the output of government services. A volume measure allows government productivity to increase or decrease over time. In the sections that follow, we present new volume measures for public education, following up on earlier work presented in Fraumeni *et al* (2004). The new measures suggest that public education output grew at an annual rate of between 1.1 and 1.5 percent over 1980-2001, which is substantially slower than the 2.5 percent annual growth rate of the input measure of public education output. We also discuss the possibilities for alternative price and volume measures for health care services.

Volume Indexes for Elementary and Secondary Education

The simplest volume index for the output of public elementary and secondary education is a count of students enrolled in public elementary and secondary schools. This count has grown significantly more slowly than the input index for public elementary and secondary education. Between 1980 and 2001, the number of students

enrolled in public elementary and secondary schools grew at an annual rate of 0.7 percent.[1] In contrast, state and local government consumption and sales for public elementary and secondary education grew at a rate of 2.4 percent per year.[2] More detailed growth rates for these two series are presented in Table 1.

There are several drawbacks to measuring the output of education with a simple count of students. One is the failure of such a measure to capture possible increases in the quality of educational services provided. Another is the implicit assumption that education is the same across different grades and kinds of education. Both of these problems suggest that it may be a good idea to use a more sophisticated measure that makes some adjustments for changes in education over time.

Accounting for Special Education

One of the most frequently discused changes in public elementary and secondary education over the past twenty years is the accommodation of students with special needs. The percentage of public elementary and secondary students who received special-education services increased from 9.4 percent to 12.1 percent between 1980 and 2001.[3] A review by Chambers *et al* (2004) finds that the cost of educating a special education student has been estimated as being between 1.9 and 2.3 times the cost of educating a regular education student. This suggests double-weighting special education students as an easy way to account for the rise of special education. Doing so increases the annual growth rate of a count of students from 0.7 percent to 0.9 percent.

[1] Elementary and secondary enrollment data are from the state nonfiscal surveys of the Common Core of Data and are published in various editions of the *Digest of Education Statistics*, which is published annually by the National Center for Education Statistics of the U.S. Department of Education.
[2] State and local consumption and sales for education are from unpublished data in the National Income and Product Accounts of the Bureau of Economic Analysis of the U.S. Department of Commerce.
[3] Special education data are from the annual reports to the U.S. Congress on implementation of the Individuals with Disabilities Act by the Office of Special Education Programs of U.S. Department of Education.

Adjusting Output for the Quality of School Inputs

It is also possible that the quality of education has changed within regular and special education. One way to adjust for this kind of change in quality is to adjust for the quality of school inputs. For example, the pupil-teacher ratio in public elementary and secondary schools declined from 18.7 to 15.9 between 1980 and 2001.[4] How might this have affected the quality of education? Rivkin *et al*'s (2005) study of Texas elementary school students suggested that a one-student reduction in class size that persists over fourth through seventh grade would normally increase mathematics test scores by 0.02 standard deviations.[5] Presuming a class size of 20, this suggests that a one-year, one-percent drop in class size would improve test scores by 0.001 standard deviations.[6]

Translating a standard deviation of test scores into a greater volume of education output is a challenge. One approach is to compare the economic returns to test scores and years of education. Bowles, Gintis and Osborne's (2001) literature review suggests that the economic return to a standard deviation of cognitive skill is about equal to the economic return to a year of education. We could interpret this to mean that a standard deviation of test scores is the equivalent of one year of education. It is probably most appropriate to think of this as a lower bound on the rate of substitution; if the distribution of test scores is normal, it implies that an eighth grader in the 15th percentile is slightly

[4] Teachers data are from the state nonfiscal surveys of the Common Core of Data and are published in the *Digest of Education Statistics.*

[5] The .02 estimate can be found by adding the first four coefficients in the third column of Table VII of Rivkin *et al* (2005). The sum is -.0197.

[6] If average class size is 20, then a one-student reduction in class size is a five-percent reduction in class size. Consequently, a four-year, five-percent reduction in class size increases test scores by 0.02 standard deviations. Dividing 0.02 by four to scale down to one year and again by five to scale down to a one percent class size reduction yields 0.001 standard deviations.

less than the equivalent of a sixth grader in the 85th percentile.[7] At this rate of substitution, a one-year, one-percent reduction in class size that improves test scores by 0.001 standard deviations increases each student's education by the equivalent of 0.001 years. This could be interpreted as a 0.001 × 100% = 0.1 percent improvement in quality. If a one-percent reduction in class size improves quality by 0.1 percent, the elasticity of school quality with respect to class size is implicitly 0.1% ÷ 1% = 0.1.

Alternatively, one could translate test scores into education output simply by using the normal test score gain from a year of education. Analysis of National Assessment of Educational Progress (NAEP) math test scores suggests that a standard deviation of test scores is the equivalent of 3.3 years of schooling.[8] This probably best thought of as an upper bound on the rate of substitution; it implies that an eighth grader in the 15th percentile is slightly more than the equivalent of a *first* grader in the 85th percentile.[9] At this rate of substitution, a one-year, one-percent reduction in class size increases school quality by 0.33 percent, which implies that the elasticity of school quality with respect to class size is 0.33.

Class size is not the only variable that has changed in recent years. The number of inexperienced teachers has also changed; the preponderance of teachers with less than

[7] In the normal distribution, the difference between the 15th and 85th percentiles is 2.06 standard deviations. If one year of education is the equivalent of one cross-sectional standard deviation of cognitive skill, then two years of education—say, between the sixth and eighth grades—erases nearly all of this difference.

[8] The standard deviation of math NAEP scores for 17-year-olds is about 31 points; this was approximated by observing the percentile distribution of scores in 1996 and assuming a normal distribution. The average math NAEP score improved from 231 age at 9 to 307 at age 17. Dividing the difference between these two scores by 8 yields an annual NAEP gain of 9.5 points, which is about 1/3.3 the cross-sectional standard deviation of 31.

[9] If 3.3 years of education is the equivalent of one cross-sectional standard deviation of cognitive skill, then seven years of education increases cognitive skill 7 ÷ 3.3 = 2.12 standard deviations, which is slightly more than the difference of 2.06 standard deviations between the 15th and 85th percentiles.

two years of experience rose from 5.3 percent in 1980 to 8.8 percent in 2000.[10] Suppose we assumed that having a teacher with fewer than two years of experience reduced test score gains by 0.10 standard deviations.[11] If this is the case, then the semi-elasticity of school quality with respect to the proportion of teachers with fewer than two years of experience is $0.10 \times 1 = 0.1$ under the lower-bound assumption that a standard deviation of test scores is the equivalent of one year of schooling. Under the upper-bound assumption that a standard deviation is the equivalent of 3.3 years of schooling, the semi-elasticity is $0.10 \times 3.3 = 0.33$.

The above discussion suggests two volume measures that adjust for quality of school inputs. Let Q equal the output of public elementary and secondary education, RE equal regular education enrollment, SE equal special education enrollment, PT equal the pupil-teacher ratio, and XP equal the proportion of teachers with fewer than two years of experience. Under the lower-bound assumption that a standard deviation in test scores is the equivalent of one year of schooling, the output of public elementary and secondary education is

$$Q = PT^{-0.1} e^{-0.1XP} (RE + 2SE)$$

since, under the lower-bound assumption, the elasticity of school quality with respect to class size and the semi-elasticity of school quality with respect to the proportion of inexperienced teachers are both 0.1. Under the upper-bound assumption that a standard deviation in test scores is the equivalent of 3.3 years of schooling, output is

[10] Teacher experience data is from various editions of the *Status of the American Public School Teacher*, which is published quinquenially by the National Education Association. Other years are linearly interpolated.

[11] Rivkin *et al* (2005) found that, compared to teachers with more than five years of experience, test scores were 0.13 standard deviations lower when teachers had no experience, 0.06 standard deviations lower when teachers had one to two years of experience, and 0.03 standard deviations lower when teachers had three to five years of experience.

$$Q = PT^{-0.33} e^{-0.33XP} (RE + 2SE)$$

Details on the growth rates of both measures are presented in Table 1.

These adjustments modestly increase the measured growth rate of the volume measure of elementary and secondary education. The upper-bound adjustment, in particular, increases the annual rate of growth over 1980-2001 to 1.1 percent. This is still much closer to the 0.7 percent growth rate of the unadjusted count of students than it is to the 2.4 percent growth rate of the input measure. It is important to note, however, that any attempt to adjust for quality using the quality of school inputs will necessarily be incomplete, as not all school inputs are measurable. Adjusting for school inputs is an additive process, from which one starts from zero; consequently, simple adjustments like those above are likely to have small impacts.

Adjusting for the Quality of Student Outcomes

Another way to adjust for changes in quality within regular and special education is to use changes in student outcomes. Test scores are probably the most natural outcome to use. Analytically, this is a simpler adjustment than school inputs. Previously, we used school inputs to adjust for quality of education, and determined the size of the adjustment by looking at the various inputs' effects on test scores. Here, we skip the intermediate step and just use the test scores themselves to adjust for quality.

The best test score for quality adjustment is probably twelfth-grade NAEP scores, which ostensibly measure the end result of elementary and secondary education: cognitive skill at around the time of completion. The average math NAEP score improved considerably over the period of time studied: from 298 in 1982 to 308 in 1999, or by nearly a third of a standard deviation. Changes in this score can be a result of

improvement in any one of the twelve grades, so we divide the changes evenly among grades and assume that a one-standard deviation change in twelfth-grade test scores reflects a one-twelfth of a standard deviation change in test score gains in each year of education. We also assume temporarily that all change over time in test scores is caused by changes in the quality of education.

Let *TS* equal the average twelfth-grade test score, normalized to cross-sectional standard deviations. If we use the lower-bound assumption that a standard deviation of test scores is the equivalent of one year of education, a volume index adjusted for test score improvements is

$$Q = TS^{1/12}(RE + 2SE)$$

Under the upper-bound assumption that a standard deviation of test scores is the equivalent of 3.3 years of education, the adjusted volume index is

$$Q = TS^{3.3/12}(RE + 2SE)$$

Growth rates of these indexes using NAEP math test scores are presented in Table 1.[12]

The test scores adjustment increases the growth rate of the volume index by substantially more than the school inputs adjustment. After adjusting for improvements in test scores using the upper-bound adjustment, measured output of elementary and secondary education increases at a rate of 1.2 percent per year. Note, however, that this and the special education adjustment combined close less than one-third of the gap between the growth rate of an unadjusted count of students (0.7 percent) and the growth rate of the input measure (2.4 percent).

So far, we have assumed that all test score gains are a result of schooling. This is unlikely to be the case; family, peers, and environment all play an important role as well.

[12] Linear interpolations are used for years in which the NAEP was not conducted.

The ideal measure of student outcome would strip away as many non-school influences as possible and identify the particular gains from schooling.

Parents' education is an especially important non-school variable, especially when using the NAEP; the parental background of the NAEP sample has tended toward more education over time. To account for these changes, we took the separate NAEP time series for children of parents of five education categories—less than high school, graduated high school, some education after high school, graduated college, and unknown—and averaged them using as weights the distribution of NAEP children's parents by education in 1996. The effect of using this parent-adjusted NAEP time series rather than the unadjusted NAEP time series is substantial. Using the adjusted NAEP reduces the growth rate of the volume index from 1.2 percent to 1.0 percent under the upper-bound rate of substitution between test scores and years of education.

The big picture on volume indexes for elementary and secondary education in the United States is illustrated in Figure 1, which plots three volume measures—an unadjusted count of students, a count adjusted for school inputs using the upper-bound adjustment, and a count adjusted for raw test scores using the upper-bound adjustment—alongside the currently employed input measure. The three volume indexes have more in common with each other than with the input measure. Even with adjustments for quality, growth in the output of public elementary and secondary education when measured by volume is much slower than growth as it is currently measured under the input approach.

Volume Indexes for Higher Education

Measuring the output of public higher education by volume is a different challenge from measuring the output of public elementary and secondary education by

volume. The most substantial difference is that instruction is only one of many functions of higher education. State and local colleges and universities exist to teach students, but they also exist to conduct research and act in the public service.

In computing the volume index of output, we assume that the proportion of the nominal public higher education output that is dedicated to instruction of students is equal to current expenditure by public institutions for instruction and student services divided by current expenditure by public institutions for instruction, student services, research, and public service. This proportion, devised by To (1987), was used by Winston and Yen (1995) to identify the component of operating and capital costs that is dedicated to instruction at individual institutions. Across all institutions, this proportion dropped from 0.75 in 1980 to 0.70 in 2000, which may indicate a decline in the relative importance of instruction in public higher education.[13] We also split the input measure of the real output of public colleges and universities between instruction and non-instruction using this proportion.

Basic Measures for Instruction

Like elementary and secondary education, the simplest volume measure of the instructional function of public higher education is an unweighted count of students. The annual growth rate of this count was 1.2 percent between 1980 and 2001, which is quite a bit slower than the 2.3 percent annual growth rate of the input measure for instruction.[14] Note, however, that the 1.1 percent gap between a simple headcount and the input measure for higher education is considerably smaller than the analogous 1.7 percent gap

[13] Data on school finances is from the Finance surveys of the Higher Education General Information Survey (HEGIS) and its successor, the Integrated Postsecondary Education Data System (IPEDS). The data are published in various editions of the *Digest of Education Statistics.*

[14] Data on enrollments is from the Fall Enrollments surveys of HEGIS and IPEDS. Men's and women's enrollments increased at annual rates of 0.8 percent and 1.6 percent over 1980-2001.

for elementary and secondary education. Double-weighting graduate enrollments and converting to full-time equivalents (FTEs) by counting part-time enrollments as one-third of a full-time enrollment has virtually no impact on the growth rate of public higher education instruction; the annual growth rate remains 1.2 percent. The composition of enrollment across full-time, part-time, undergraduate, and graduate enrollment is remarkably static over time. More details on these series are presented in Table 2.

Using degrees instead of enrollments to measure public higher education instructional output changes matters slightly. A simple count of degrees earned grows at an annual rate of 1.4 percent per year.[15] Weighting the count has little impact, as the composition of total degrees earned across associate's, bachelor's, master's, first-professional, and doctoral degrees has also remained very static over time. A series that weights degrees by the number of years typically required to completion, with graduate degrees counting double, still grows at an annual rate of 1.4 percent.[16]

Creating a Hybrid Index of Enrollments and Degrees

A hybrid index of enrollments and degrees can be constructed if a satisfactory approach to weighting enrollments and degrees is found. One criterion for assessing a weighting approach is how well it reflects the rate at which people would willingly substitute years of education for earned degrees. This rate of substitution could be measured by comparing economic rates of return from years of education and earned degrees. Such estimates exist in the literature on sheepskin effects, which is critically reviewed by Flores and Light (2004).

[15] Data on degrees from the Earned Degrees surveys of HEGIS and IPEDS, which are published in various editions of the *Digest of Education Statistics*.

[16] In this index, associate degrees recieve a weight of 2, bachelor's degrees a weight of 4, master's degrees a weight of 4, first-professional degrees a weight of 6, and doctoral degrees a weight of 8.

One of the best papers on sheepskin effects is by Jaeger and Page (1996), who used a matched sample from the March 1991 and 1992 demographic supplements to the Current Population Survey (CPS). Over this sample, Jaeger and Page regressed log hourly wages against a set of variables that included dummies for the number of years of education completed and dummies for degrees and diplomas earned. This regression estimated separate rates of return to individual years of undergraduate and graduate education; to associate's, bachelor's, and graduate degrees; and, surprisingly, to the mere act of having attending college in the first place. The last one is estimable because there are people in the data set who reported having attended college but having only completed twelve years of education, as well as people who reported having completed more than twelve years of education but who do not report having attended college.

Among white men, Jaeger and Page found that the total return to four years of undergraduate college is 17.8 percent and the additional return to two or more years of graduate school is 4.6 percent.[17] This could be interpreted as a 17.8 ÷ 4 = 4.45 percent return to a year of undergraduate schooling and a 4.6 ÷ 2 = 2.3 percent return to a year of graduate schooling. Additionally, Jaeger and Page found that white men with occupational associate's degrees earned 0.7 percent less than those with some college but no degree, while those with academic associate's degrees earned 10.8 percent more and those with bachelor's degrees earned 16.2 percent more.[18] If about half of associate's degrees are occupational and half are academic, this implies an average return to associate's degrees of (-0.7 + 10.8) ÷ 2 = 5.05 percent and a return to a bachelor's degree

[17] See the fourth column of Table 2 of Jaeger and Page (1996). The 4.6 percent return to two or more years of graduate school is calculated by subtracting the 0.178 coefficient on 16 years of schooling from the 0.224 coefficient on 18+ years of schooling.

[18] These are also derived from the fourth column of Table 2 of Jaeger and Page (1996), by subtracting the .083 coefficient on "some college, no degree" from the coefficients on undergraduate degrees earned.

of 16.2 percent.[19] Finally, Jaeger and Page find returns of 5.0 percent to master's degrees, 28.6 percent to first-professional degrees, and 6.7 percent to doctoral degrees.

The economic returns above, if estimated correctly, give us an idea of rates of substitution. For example, the economic return to a year of graduate education is about half the economic return to a master's degree. This suggests that people will value a year of graduate school at about one-half the value of a master's degree, which means in turn suggests that years of graduate school should be weighted about half as much as master's degrees in an aggregated index of years of education and earned degrees. If we use the economic returns described above as weights, the aggregated index would weight undergraduate enrollments by 4.45, graduate enrollments by 2.3, associate's degrees by 5.05, bachelor's degrees by 16.2, master's degrees by 5.0, professional degrees by 28.6, and doctoral degrees by 6.7.

Unsurprisingly, the growth rate of the resulting hybrid index, 1.3 percent, is between the growth rates for the enrollments-only and degrees-only indexes. More details on this index are presented in Table 2.

Comparing Volume Indexes and the Input Index for Public Higher Education

In Figure 2, three volume indexes for public higher education instruction are plotted: the weighted enrollment series, the weighted degrees series, and the degrees-enrollment hybrid series. The input index is also plotted. The plot as a whole is similar to that for elementary and secondary education, but not identical; it is still the case that the volume series are all more similar to each other than they are to the input series, but the difference is not as dramatic.

[19] This is approximately the case in Jaeger and Page's data; see Table 1 of their paper.

The difference between the volume and input series for higher education instruction might have been even smaller were the volume series adjusted for quality. Despite rising inputs per student in higher education instruction, the volume series all implicitly assume that the quality of public higher education is constant over time. It is difficult to adjust for quality because there are few systematic studies of the performance of college students over time; this is in part because the college curriculum is not nearly as uniform across students as the elementary and high school curriculum, and so exactly what is supposed to be tested is not very clear. If the quality of college instruction is rising over time, the difference in growth rates between the currently used input index and a properly adjusted volume index for higher education may be quite small.

Quantifying the non-instructional component of public higher education output for a volume measure is considerably harder than quantifying the instructional component. Adams and Clemmons (2006) used research papers and citations to measure the productivity of research faculty at a sample of 102 universities in the United States, which they found had risen substantially in public universities over 1981-1995. Rather than attempt to quantify the non-instructional component of public higher education, we use the input measure instead, which grew at a brisk 3.7 percent annual rate over 1980-2001.

We measure the total output of public higher education using a Fisher index of instructional and non-instructional public higher education output. When the enrollment, degrees, or hybrid volume measure is used to measure the instructional component and the input measure is used to measure the non-instructional component, the output of public higher education rises at an annual rate of between 1.9 and 2.0 percent. When the

input measure is used for both the instructional and non-instructional components, the output of public higher education rises at an annual rate of 2.7 percent. The difference in annual growth between a simple (and partial) volume measure and the currently used input measure is a small 0.7 to 0.8 percent. Since there have been no quality adjustments to the volume index for the instructional component and since there is some evidence that research productivity has been rising, a more sophisticatedly measured gap might be even smaller.

Volume Indexes for the Entire Public Education Sector

Measuring the output of the entire public education sector involves combining three components: elementary and secondary education, higher education, and "other" education, which includes public libraries. Combining the three components is a straightforward application of the Fisher index. We use the input measure for "other" education, which ranges between 3.7 and 4.3 percent of nominal education output over the period studied. When "other" education is combined with the non-instructional component of higher education—the other part of education output for which we do not create a volume index—the resulting sum ranges between 10.1 percent and 11.9 percent of nominal education output. Put roughly, the "volume" indexes we present for the entire public education sector are more approximately 90/10 volume/input indexes.

In Table 3, we present growth rates for two combined volume indexes for public education. One combines the slowest-growing volume indexes: the unadjusted count of elementary and secondary students and the weighted FTE count of enrolled college and graduate students. The other combines the fastest-growing volume indexes: the count of elementary and secondary students adjusted for raw NAEP math scores, and the weighted

count of earned college degrees. The slower-growing volume measure grows at a rate of 1.1 percent, while the fast volume measure grows at a rate of 1.5 percent. By comparison, the growth rate of the input index for public education is 2.5 percent.

The three general indexes are plotted in Figure 3. Unsurprisingly, the overall picture is not much different from the separate pictures for elementary and secondary and for higher education. The two volume measures resemble each other more closely than than they resemble the input measure, and grow at a considerably slower pace. Overall, the results suggest strongly that volume measures of public education output grow substantially slower than the currently employed input measures.

Does the growth gap between the input measure and our volume measures for education suggest that there is a problem with either from a measurement perspective? We do not necessarily think so. It is not the goal of a fully quality-adjusted output volume measure to replicate the input measure; indeed, there would be no point to estimating a volume measure were it not for the possibility that it might be different from the input measure. The availability of two different measures for education from two different approaches to measurement offers many chances for insight in the public education sector.

Measuring Health in the United States

At this point, we would like to change topics and briefly discuss alternative price and output indexes for health care in the United States. Unlike the case of many European countries, only a small percentage of health care produced in the United States is produced by governments. In 2004, combined personal and government consumption expenditures for health care was $1.75 trillion, of which only $234 billion, or 13 percent,

was government consumption and sales.[20] A larger percentage of health care is paid for by governments; the combination of government consumption and social benefits for health was $692 billion, or 39 percent of total consumption of health care.[21] The remaining 61 percent of health care consumption is paid for privately, often with insurance that is acquired through one's employer or on one's own. Of the $1.75 trillion in total health care expenditure, $1.48 trillion is on health care services.

Because the majority of health care in the United States is privately produced and purchased, most research on health measurement in the United States has focused on properly measuring prices rather than volumes. Triplett (2001) notes that the prices of similar goods and services tend to be more strongly correlated with each other than the volumes of similar goods and services. This, in turn, means that average price movements in an incomplete and not necessarily representative sample of services are more likely to reflect movements in actual average prices than similarly mismeasured volume movements. With this in mind, it may be more fruitful to concentrate on measuring medical prices, which can be used to calculate medical output by deflating medical expenditure.

Much of the recent literature on health care pricing in the United States has focused on pricing the complete treatment of an individual disease or condition, such as cataracts (Shapiro *et al*, 2001), heart attacks (Cutler *et al*, 1998), depression (Berndt *et al*, 2002), or schizophrenia (Frank *et al*, 2004). This is opposed to pricing individual

[20] National Income and Product Accounts, Tables 2.4.5, 3.10.5, and 3.17, and unpublished data. Personal consumption expenditure for health care includes opthalmic products and orthopedic appliances, drug preparations and sundries, and medical care services. Government consumption expenditures are nondefense only, and government sales are state and local health and hospital charges only. Unpublished data was only needed to account for $1.2 billion in federal government health care sales.
[21] National Income and Product Accounts, Table 3.17.

procedures, such as particular surgeries or diagnostic procedures or drug prescriptions. The advantage of pricing a full treatment is that it takes into account technological changes that allow less expensive and more effective procedures to substitute for more expensive and less effective procedures. As a result, measures that price complete treatments for diseases or conditions estimate rates of price growth that are often lower and sometimes negative.

Implementing Disease-Based Measures of Medical Prices

Although there have been many studies such as those mentioned above that attempt to price complete treatments for individual diseases or conditions, there have not been many attempts to create price and volume indexes across diseases for the entire health care sector. In this section, we describe an attempt to measure an alternative disease-based price index for health care services that are produced by government hospitals.

The alternative price index is a relatively straightforward Fisher index of mean costs of hospital stays by ailment, weighted by the volume of hospital discharges by ailment. The advantage of the alternative price index is its ability to account for technological changes that allow for fewer treatments and less expensive treatments to be used in treating diseases; as a result, price growth could be expected to be lower under the alternative index than under a more traditional index. However, the alternative index unexpectedly grew faster than the currently used price index for government hospital services over 1997-2003. This faster growth is possibly a result of improvements in the quality of care or shifts in the composition of patients that are not accounted for in either index but which may cause greater distortions in the alternative index. The unexpected

result suggests that simple disease-based price indexes run a substantial risk of mismeasuring price change, and that a successful disease-based price index is likely to require substantive adjustments for changes in the quality of care and for changes in the composition of patients over time.

The data used to calculate this index are originally from the Nationwide Inpatient Sample (NIS), a database on hospitals created by the Healthcare Cost and Utilization Project (HCUP). The HCUP is sponsored by the Agency for Healthcare Research and Quality (AHRQ) of the U.S. Department of Health and Human Services. It combines data on health care from both private and government sources into a unified set of databases on health care. The NIS database includes data on 5 million to 8 million inpatient stays at 800 to 1,000 community hospitals in the United States for each year since 1988. It also includes weights for each hospital that allow the NIS to be used as a stratified sample of all community hospitals in the United States, with the strata defined by ownership/control, bed size, teaching status, urban/rural location, and region. In 2003, the NIS included data on hospitals from 37 states; in the past, the NIS sample covered fewer states.

The AHRQ uses the NIS to estimate the total number of discharges and mean charges per hospital stay at community hospitals for each year since 1997. More interesting from the perspective of health care services measurement, AHRQ also estimates the number of discharges and mean charges per hospital stay by hospital ownership (public, for-profit, non-profit) and by Diagnosis Related Group (DRG).[22] DRGs are codes for diagnoses; for example, a patient whose DRG is 21 has viral meningitis, while a patient whose DRG is 103 is receiving a heart transplant.

[22] Estimates from the NIS can be accessed at AHRQ's HCUPnet website at http://hcup.ahrq.gov/.

We use these estimates by AHRQ to create price indexes for government hospital services. We treat hospital stays for each DRG as a distinct service, with price equal to the mean charge for a stay at a government hospital for that DRG. The prices of hospital stays by DRG are aggregated into a price index for all hospital stays using a Fisher index, with the number of discharges from government hospitals by DRG used as volume weights.

This approach seeks to account for cost-saving technological improvements in hospital treatment. For example, our approach treats all hospital stays for DRG 88, chronic obstructive pulmonary disease, as the same service, even if the procedures performed in some hospital stays for chronic obstructive pulmonary disease are different from the procedures performed in others. Suppose technological changes make it possible to successfully treat pulmonary patients with a smaller and less expensive regimen of procedures. The reduction in average charges for pulmonary patients that results will appear in our data as a reduction in the price of hospital services. A price reduction is the appropriate understanding of this technological change if patients do not care about which procedures are performed so much as the actual treatment of the ailment. Had we instead treated each individual procedure performed during hospital stays as a distinct service, the shift to fewer and less expensive procedures would have appeared as a reduction in the volume of hospital services rather than as a reduction in price.

The price index currently used by BEA to deflate government hospital expenditure and calculate the volume of government hospital services is the Producer Price Index (PPI) for hospital services. This index, which is measured by the Bureau of

Labor Statistics (BLS) of the U.S. Department of Labor, is a bit less fluid in defining individual hospital services than our index. The PPI is calculated from a survey of hospitals, each of which supplies price data for a subset of DRGs. At the time a hospital enters the survey, BLS selects a discharge bill from the hospital's records—usually the last one from the previous month—for each DRG in the subset. The items on this bill are re-priced every month thereafter by this hospital to measure changes in the price of a hospital stay for this DRG. Consequently, short-term changes in the PPI for hospital services measure changes in charges for fixed bundles of hospital procedures corresponding to DRGs.[23] Unlike our index, it does not account for changes over time in the procedures used to treat hospital patients within DRGs. In particular, it will not recognize cost-saving shifts from more expensive to less expensive procedures within DRGs as price decreases.

Given that our index should reflect cost-saving shifts in procedures that the PPI does not, one would expect price growth measured by our index to be slower than price growth measured by the PPI. However, this is not the case. Figure 4 plots both the PPI for general medical and surgical hospitals and our more fluid hospital price index over 1997-2003. The PPI grows at an annual rate of 3.0 percent, while the alternative index grows more than twice as quickly at a much faster annual rate of 6.7 percent.

Why the surprising result? One possibility is that the alternative index may incorrectly recognize quality improvements as price increases. Suppose, for example, that medical technology improves by making new procedures available to treat particular ailments, and that the new procedures are more expensive than the old ones. If this is the

[23] In the long term, re-sampling will account for certain shifts from more expensive to less expensive procedures. See Fixler and Ginsburg (2001) for more details on the hospital services PPI.

case, then patients in the DRGs affected by the technology will receive more procedures and more expensive procedures than before. These changes will lead to higher charges per hospital stay, and will appear in our index as a price increase. However, this may not truly be a price increase, as the new procedures may result in higher-quality treatment. In quality-adjusted terms, the new procedures may be a price *decrease* rather a price increase. If the new procedures do bring about an effective price decrease, the PPI is more accurate than the alternative index; the PPI prices a fixed bundle of procedures, so the switch to newer, more expensive procedures is ignored and has no effect, positive or negative, on measured prices.

Another possible reason for the surprising result is the possibility that less severely ill patients are increasingly being treated in physician's offices or as hospital outpatients rather than as hospital inpatients. This increases the proportion of patients staying at hospitals who are severely ill. Because these patients are more costly to treat, the average charges for patients by DRG will be rising, which in turn will appear in our index as a price increase. Note, however, that the shifting of patients from expensive hospital inpatient services to less expensive hospital outpatient and physician's office services may reflect a reduction in price over time, particularly if there are only marginal differences in the quality of hospital inpatient care, hospital outpatient care, and physician's office care. Changes in the composition of patients will not affect the PPI for hospitals, which prices a fixed bundle of procedures and does not take into account changes in that bundle to accomodate a more severely ill composition of patients.

A third possibility is that the NIS data that we used simply do not measure cost as precisely as the surveys that BLS uses to produce the PPI for hospitals. Not only does

our index grow more quickly, but the variance of yearly changes is greater as well. A topic for further work in using NIS data to measure hospital prices is reducing the amount of noise in the estimates. One likely area of investigation is the growth of the NIS sample, which covered 22 states in 1997 and 37 states in 2003; this growth may have contributed to the volatility of our results.

The results for our alternative price index suggest that a successful disease-based price index may need to be adjusted for changes in the quality of treatments and for changes in the degree of sickness of patients in the health care sector being measured. Consequently, implementing a disease-based price index is likely to require a close look at the characteristics and outcomes of patients being treated.

Direct Volume Indexes for Health Care Services in the United States

Even in the presence of prices, a volume measure for the United States may be preferable to a measure from price deflation because the market for health care in the United States is not a traditional competitive market. Insurance creates a moral hazard problem, as neither insured patients nor the doctors who treat them have much incentive to take costs into account when pursuing additional treatments. Even if patients did take costs into account, they are unlikely to know enough about medicine, even with their doctor's help, to decide what treatments they would most prefer. There is also the "technological imperative" to use treatments that are on the cutting edge, even if more out-of-date but cheaper treatments may better meet cost-benefit criteria.[24]

All of these problems suggest that prices may not reflect marginal valuation by the consumers of health care. This makes price deflation a problematic strategy for measuring health care. Dividing expenditure on a particular good by a price index for

[24] All three of these problems and their ramifications are mentioned in Pauly (1999).

that good does not yield the actual physical volume of that good. Instead, it yields a volume index for that good. This is not a problem if one plans to aggregate volume indexes of several goods using the accompanying price indexes as weights, as in a Fisher or Tornqvist index, because whatever rescalings of volume exist in the volume indexes will be cancelled out by equal and opposite rescalings in the price indexes. Weighting volume indexes with price indexes is numerically equivalent in this case to weighting actual volumes with actual prices.

However, prices should only be used as weights for aggregating volumes if the prices of the goods volumed reflect the actual value of the goods to the consumers. For the reasons stated above, this may not be the case in the peculiar market for health care goods and services. As a result, we may prefer a volume index that weights the volumes of individual health care services by a measure of value other than prices. To do so requires knowledge of the actual volumes of the services being aggregated rather than the scaled volume indexes that come from price deflation. We need to know exactly how many cancer treatments, depression treatments, cataract treatments, etc. were performed, so that they can be properly weighted.

For this reason, a measure of health care output that aggregates individual health care goods and services using something other than price as a weight ought to be a direct volume index. Perhaps the most appropriate alternative weight is the amount of quality-adjusted life-years (QALYs) created by the individual health care good or service. A measure similar to a volume index that aggregates using QALYs as weights is discussed

in Pauly (1999).[25] This index would require direct measurement of health care goods and services, a rather tall order.

Conclusions

The previous sections presented and discussed volume measures for public education output and discussed the possibilities for price and volume measures for health care output in the United States. Volume measures of the output of the education function of government appear to grow at a slower rate than the currently employed input measure; over 1980-2001, the difference was between one and one and a half percent a year. Measuring health care by disease at the national level is a very difficult issue; we found that simply measuring disease-based prices for health care services provided by government hospitals produced results that are inconsistent with most health care price measures. BEA expects to continue investigation into volume measures and other alternative measures of government output for the United States, with the ultimate goal of providing a suite of alternative output measures for researchers and other users of the National Income and Product Accounts.

References

Adams, James, and J. Roger Clemmons (2006), "The Growing Allocative Inefficiency of the U.S. Higher Education Sector," Rensselaer Polytechnic Institute, mimeo.

Berndt, Ernst, Anupa Bir, Susan Busch, Richard Frank, and Sharon-Lise Normand (2002), "The Medical Treatment of Depression, 1991-1996: Productive Inefficiency, Expected Outcome Variations, and Price Indexes," *Journal of Health Economics* 21(3), pp. 373-396.

[25] Pauly's measure is actually more avant-garde; it would affect both nominal and real output measures of health care services.

Bowles, Samuel, Herbert Gintis, and Melissa Osborne (2001), "The Determinants of Earnings: A Behavioral Approach," *Journal of Economic Literature* 39(4), pp. 1137-1176.

Chambers, Jay, Thomas Parrish and Jenifer Harr (2004), "What Are We Spending on Special Education Services in the United States, 1999-2000?", Special Education Expenditure Project.

Cutler, David, Mark McClellan, Joseph Newhouse, and Dahlia Remler (1998), "Are Medical Prices Declining? Evidence from Heart Attack Treatments," *Quarterly Journal of Economics* 113(4), pp. 991-1024.

Fixler, Dennis, and Mitchell Ginsburg (2001), "Health Care Output and Price in the Producer Price Index," in David Cutler and Ernst Berndt, eds., *Medical Care Output and Productivity* (Chicago, Ill.: University of Chicago Press).

Flores-Lagunes, Alfonso, and Audrey Light (2004), "Identifying Sheepskin Effects in the Return to Education," mimeo., April 2004.

Frank, Richard, Ernst Berndt, Alisa Busch, and Anthony Lehman (2004), "Quality-Constant 'Prices' for the Ongoing Treatment of Schizophrenia: An Exploratory Study," *Quarterly Review of Economics and Finance* 44(3), pp. 390-409.

Fraumeni, Barbara, Marshall Reinsdorf, Brooks Robinson, and Matthew Williams (2004), "Price and Real Output Measures for the Education Function of Government: Exploratory Estimates for Primary and Secondary Education," Conference on Research in Income and Wealth, Conference on Price Index Concepts and Measurement, Vancouver, B.C., June 28-29.

Jaeger, David, and Marianne Page (1996), "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education," *Review of Economics and Statistics* 78(4), pp. 733-740.

Pauly, Mark (1999), "Medical Care Costs, Benefits, and Effects: Conceptual Issues for Measuring Price Changes," in Jack Triplett, ed., *Measuring the Prices of Medical Treatments* (Washington, D.C.: Brookings Institution Press), pp. 196-219.

Rivkin, Steven, Eric Hanushek, and John Kain (2005), "Teachers, Schools, and Academic Achievement," *Econometrica* 73(2), pp. 417-458.

Shapiro, Irving, Matthew Shapiro, and David Wilcox (2001), "Measuring the Value of Cataract Surgery," in David Cutler and Ernst Berndt, eds., *Medical Care Output and Productivity* (Chicago: University of Chicago Press)

To, Duc-Le (1987), *Estimating the Cost Of a Bachelor's Degree: An Institutional Cost Analysis* (Washington, D.C.: Office of Educational Research and Improvement, U.S. Dept. of Education).

Triplett, Jack (2001), "Measuring Health Output: The Draft Eurostat Handbook on Price and Volume Measures in National Accounts," presented at the Eurostat-CBS seminar, Voorburg, Netherlands.

Winston, Gordon, and Ivan C. Yen (1995), "Costs, Prices, Subsidies, and Aid in U.S. Higher Education," The Williams Project on the Economics of Higher Education, Discussion Paper No. 32.

Table 1: Alternative Measures of Output Volume and Price Growth in Public Elementary and Secondary Education

	Annual output growth			Annual price growth		
	1980--1990	1990--2001	1980--2001	1980--1990	1990--2001	1980--2001
Input measure:						
State and local consumption and sales for elem./sec. education	2.15%	2.65%	2.41%	5.40%	2.87%	4.07%
Volume measures:						
Unweighted count of students	0.08%	1.33%	0.73%	7.58%	4.21%	5.80%
Weighted count, 1 special ed = 2 regular ed	0.18%	1.47%	0.85%	7.47%	4.07%	5.67%
Weighted counts with adjustments for school inputs:						
Lower-bound adjustment	0.25%	1.52%	0.92%	7.39%	4.01%	5.61%
Upper-bound adjustment	0.42%	1.64%	1.06%	7.22%	3.89%	5.46%
Weighted counts with adjustments for test scores:						
Lower-bound adjustment for raw scores	0.35%	1.53%	0.97%	7.30%	4.00%	5.56%
Upper-bound adjustment for raw scores	0.72%	1.68%	1.22%	6.90%	3.85%	5.29%
Lower-bound adjustment for scores with parents' ed controlled	0.24%	1.50%	0.90%	7.41%	4.03%	5.63%
Upper-bound adjustment for scores with parents' ed controlled	0.37%	1.58%	1.00%	7.28%	3.95%	5.52%

Notes:

All measures except the input measure and the unweighted count of students count special-education students as the equivalent of two regular-education students.

All adjusted measures adjust for quality by multiplying the count of students weighted for special education by a measure of school quality normalized to 1 in 1996.

The lower-bound adjustment for school inputs weights a 10 percent decline in the pupil/teacher ratio or a 10 percentage point decline in the percentage of teachers with fewer than two years of experience as the equivalent of a 1 percent increase in the quality of education.

The upper-bound adjustment for school inputs weights a 10 percent decline in the pupil/teacher ratio or a 10 percentage point decrease in the percentage of teachers with fewer than two years of experience as the equivalent of a 3.3 percent increase in the quality of education.

The lower-bound adjustment for test scores weights a 1 standard deviation (31-point) increase in NAEP math scores for 17-year-olds as reflecting an increase in the quality of education by a factor of one-twelfth.

The upper-bound adjustment for test scores weights a 1 standard deviation (31-point) increase in NAEP math scores for 17-year-olds as reflecting a 27.5 percent ($3.3 \div 12 \times 100\%$) increase in the quality of education.

Scores with parents' ed controlled sets NAEP test takers to their 1996 distribution across five parents' education categories: less than high school, high school degree, some college, college degree, and unknown.

Table 2: Alternative Measures of Output Volume and Price Growth in Public Higher Education Instruction

	Annual output growth			Annual price growth		
	1980--1990	1990--2001	1980--2001	1980--1990	1990--2001	1980--2001
Input measure:						
State and local consumption and sales for higher ed. instruction	2.15%	2.48%	2.33%	5.37%	2.77%	4.00%
Volume measures:						
Unweighted count of students	1.38%	1.10%	1.23%	6.18%	4.18%	5.13%
Weighted count, part time = 1/3 full time, grad = 2 undergrad	1.20%	1.25%	1.23%	6.36%	4.02%	5.13%
Unweighted count of degrees	1.23%	1.56%	1.40%	6.34%	3.71%	4.95%
Weighted count of degrees	1.23%	1.53%	1.39%	6.34%	3.73%	4.97%
Hybrid count of students and degrees	1.23%	1.31%	1.27%	6.33%	3.96%	5.09%

Notes:

State and local consumption and sales for higher education instruction is equal to chained-dollar (1996) state and local consumption and sales for higher education times instruction's share.

Instruction's share is equal to the proportion of current expenditures for instruction, research, public service, and student services at public institutions that is dedicated to instruction and student services.

Weighted count of degrees weights associate's degrees by 2, bachelor's degrees by 4, master's degrees by 4, first-professional degrees by 6, and doctoral degrees by 8.

Hybrid count of students and degrees weights FTE undergraduate enrollment by 4.45, FTE graduate enrollment by 2.3, associate's degrees by 5.05, bachelor's degrees by 16.2, master's degrees by 5.0, doctoral degrees by 6.7, and first-professional degrees by 28.6.

Table 3: Alternative Measures of Output and Price Growth in Public Education, All Levels

	Annual output growth			Annual price growth		
	1980--1990	1990--2001	1980--2001	1980--1990	1990--2001	1980--2001
Input measure:						
State and local consumption and sales for education	2.20%	2.71%	2.47%	5.40%	2.84%	4.05%
Volume measures:						
Slowest-growing volume measure	0.56%	1.56%	1.08%	7.12%	4.01%	5.48%
Fastest-growing volume measure	1.01%	1.86%	1.45%	6.65%	3.71%	5.10%

Notes:

Slowest-growing volume measure is a Fisher index of an unweighted count of elementary and secondary students, higher education FTE enrollment (graduate years count double), and input measures of the non-instructional function of higher education and "other" education.

Fastest-growing volume measure is a Fisher index of a count of elementary and secondary students adjusted for special education and raw NAEP test scores with the upper-bound adjustment, degrees earned weighted by typical years to completion (graduate years count double), and input measures of the non-instructional function of higher education and "other" education.

Figure 1. Public Elementary and Secondary Education Output Volume Indexes (1996=100)

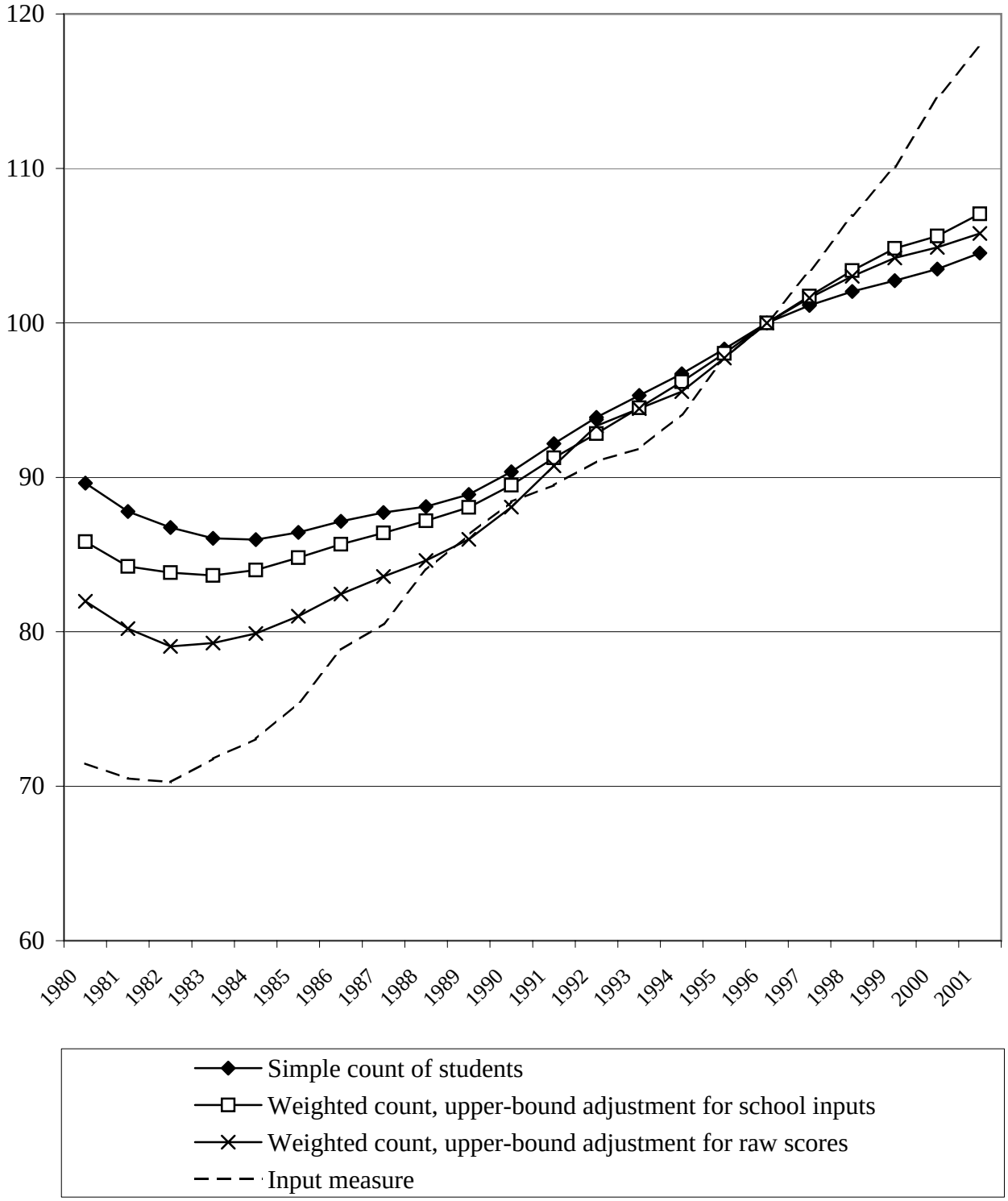

Figure 2. Public Higher Education Instruction Output Volume Indexes (1996 = 100)

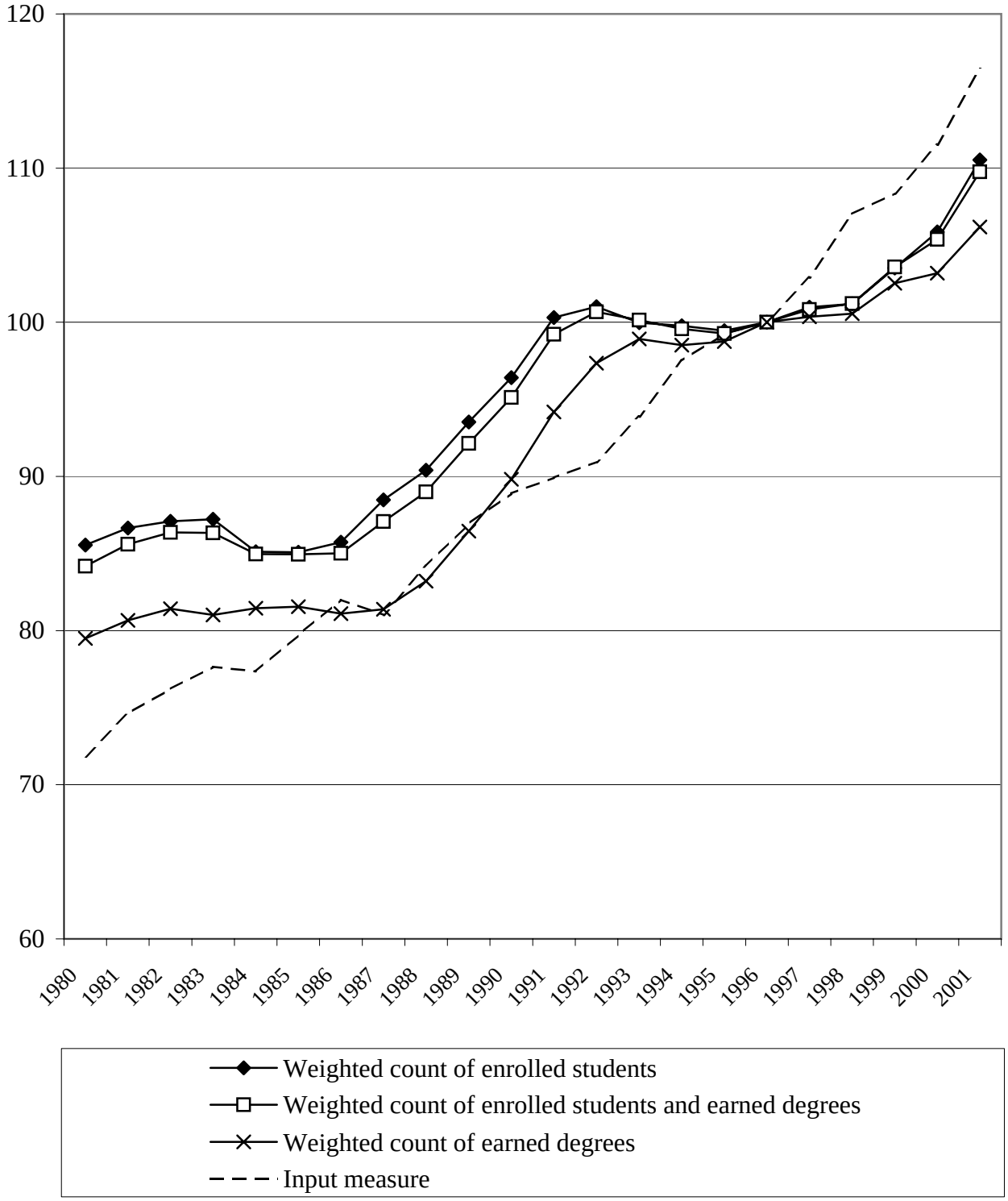

Figure 3. Alternative Total Public Education Output Volume Indexes (1996 = 100)

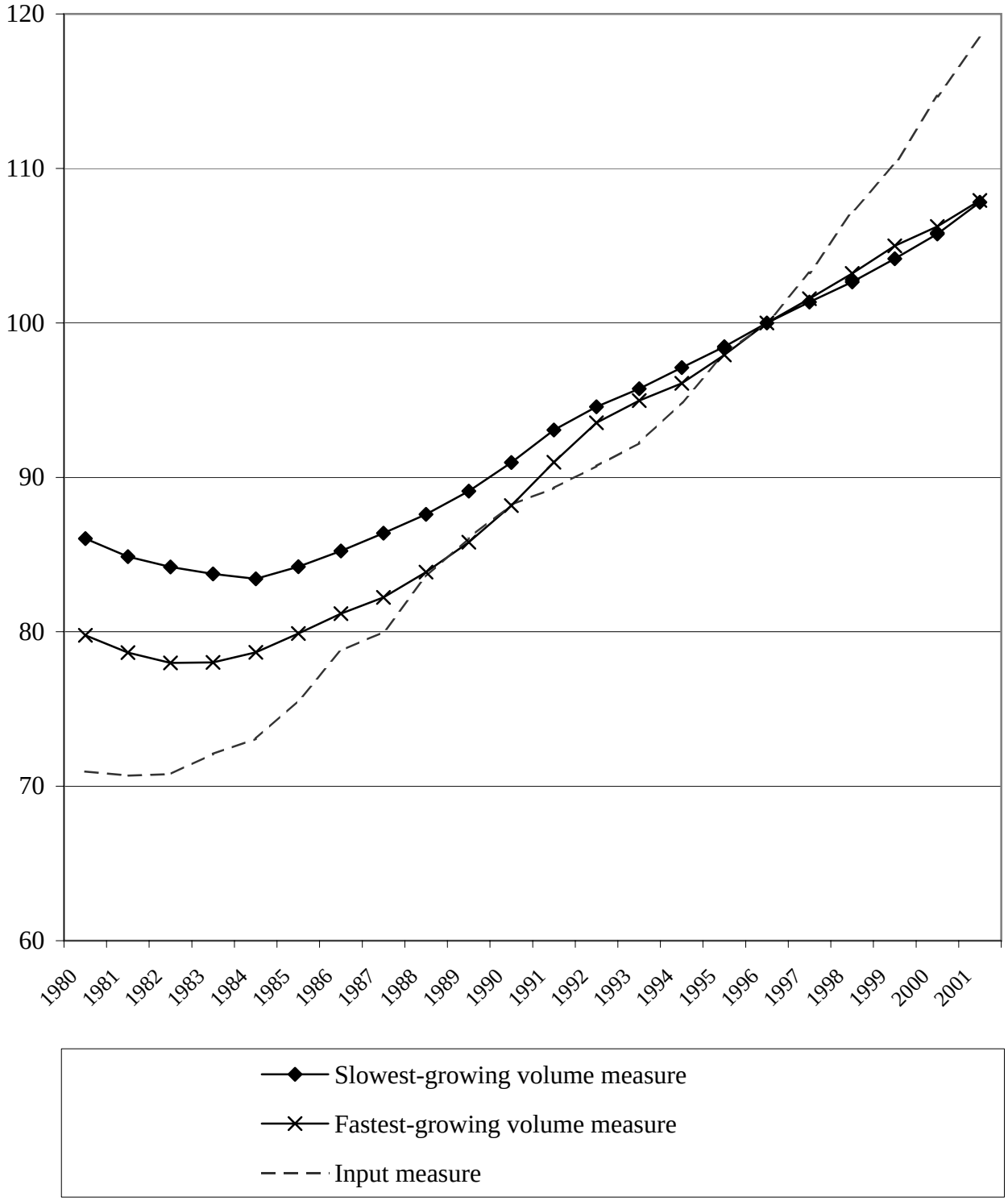

Figure 4. Comparison of Hospitals PPI and Alternative Hospitals Price Index (1997 = 100)