

ARTICLE

BROKEN BUDGETING

Natasha Sarin*

As peacetime deficits rose over the course of the last half century, policymakers searched for tools to assess how close—or far off—new budget, tax, and spending proposals would bring them to fiscal sustainability. This search led to the birth of modern scorekeeping, a complex and highly technical exercise undertaken by neutral government analysts known as scorekeepers. Because its origins are tied to rising deficits, scorekeepers are governed by rules that focus their attention on myopic cost/benefit analysis, rather than long-term policy evaluation. Over the years, many have criticized the process and questioned the accuracy of scores in particular arenas.

This Article offers a more provocative and fulsome take. While ostensibly neutral, the primacy of scorekeeping and scorekeepers has created impediments to legislating a progressive vision of government. Progressive policymaking has at its core government interventions that give society the ability to reap benefits down the line—like investments in children, or in combatting climate change—benefits that accrue in the long-term and are difficult to quantify. Presently, scorekeepers register these types of interventions as costs to the fisc rather than profitable investments, and that hinders their adoption. This is not the fault of scorekeepers, who have limited scope to act outside the rules and parameters set out by the members of Congress they serve. But it is a critique of those rules, which through manipulation and misunderstanding create a process that is far from neutral: instead, one that skews policy outcomes against progressive reforms that invest in future generations and in redressing inequality.

In this piece we lay out the ways in which deficit-centrality has shaped the

* Professor of Law, Yale Law School, with a secondary appointment at the Yale School of Management. Previously, Professor Sarin served as Deputy Assistant Secretary for Economic Policy and later as a Counselor to Treasury Secretary Janet Yellen at the United States Treasury Department, where her work focused on narrowing the gap between the taxes owed by the American public and those collected by the Internal Revenue Service. Prior to her government service, Sarin was Professor at the University of Pennsylvania Carey Law School and the Wharton School. Her research centers on public finance and financial regulation, with work on tax policy, household finance, insurance, and macroprudential risk management. Her scholarship has been covered in *The New York Times*, *The Washington Post*, and the *Financial Times*, among other publications.

federal budgeting and scorekeeping process, synthesize the deficiencies of this approach, and offer a way forward. Our piece is a call-to-arms, as academics have an important role to play in helping policymakers arrive at a more holistic approach to policy analysis, as opposed to the narrow focus on cost estimates that guides policy discussions today.

TABLE OF CONTENTS

INTRODUCTION	209
I. HISTORY OF SCOREKEEPING AND THE CONGRESSIONAL BUDGET PROCESS	215
A. The Creation of CBO	216
B. <i>Gramm-Rudman-Hollings</i>	220
C. <i>The Budget Enforcement Act</i>	223
II. SCOREKEEPING TODAY.....	226
A. A Scorekeeping Primer.....	227
B. The Scorekeeping Process' Failure to Grapple with Uncertainty: The Case of Capital Gains	238
C. The Scorekeeping Process' Failure To Consider Benefits: The Case Of The Child Tax Credit.....	244
D. The Scorekeeping Process' Failure to Consider Long-Term Effects: The Case of Tax Compliance	250
E. The Scorekeeping Process' Failure to Fully Grapple with Macroeconomic Feedback Effects: The Case of High-Skilled Immigration	257
III. WHERE DO WE GO FROM HERE?	263
CONCLUSION.....	270

INTRODUCTION

“Budgets are not merely affairs of arithmetic, but in a thousand ways go to the root of prosperity of individuals, the relation of classes, and the strength of kingdoms.”

—Gladstone¹

It is impossible to spend time around policy debates in Washington and not realize the central role played by revenue estimates on public discourse and ultimately policymakers’ decision-making. These “scores” are generated by official government “scorekeepers” at the Congressional Budget Office (CBO), the Joint Committee on Taxation (JCT), the Office of Management and Budget (OMB), and Treasury’s Office of Tax Analysis (OTA). Legislative scores are so dispositive of policy outcomes that members of Congress often refuse to vote prior to seeing these analyses, and congressional journalists breathlessly await the verdict of scorekeepers on key policy provisions.²

The birth of modern scorekeeping—and its centrality in the policy process—dates back to the origin of peacetime deficits in the United States about half a century ago. Prior to this moment, although the Constitution called for a “regular Statement and Account of the Receipts and Expenditures of all public money,”³ the process of federal budgeting was uncoordinated and inchoate.⁴ So although

1. A.E. Buck, *The Development of the Budget Idea in the United States*, 113 ANNALS AM. ACAD. POL. SOC. SCI. 31, 31 (1924) (quoting an early recognition of the significance of budgets, significance that for many years was disregarded).

2. See, e.g., Garrett Downs & Meredith Lee Hill, *Inside the Fall Farm Bill Spring*, POLITICO (July 31, 2023, at 10:00 EDT), <https://perma.cc/RL78-RX9N> (detailing lawmakers’ frustration with delay from the Congressional Budget Office (CBO) in scoring the 2023 farm bill, with members of Congress noting that the “biggest challenge [to the legislation] is [CBO] scores” and lawmakers from “both chambers . . . complaining that CBO scores for the farm bill are taking too long and delaying the committee’s work on getting a farm bill draft together”); see also Amber Phillips, *What is the CBO, and How Could Its Score Derail Democrats’ Social Safety Net Bill?*, WASH. POST (Nov. 19, 2021), <https://www.washingtonpost.com/politics/2021/11/05/what-is-cbo-how-could-its-score-derail-democrats-spending-bills/> (noting the primacy of CBO scores and the ways in which scorekeepers’ prior verdicts have swayed both legislative outcomes and popular opinion); Philip Rocco, *Congress Is Waiting on the CBO for Its Build Back Better Report—But How Did Fiscal Scorekeepers Come to Be So Powerful in Politics?*, THE CONVERSATION (Nov. 16, 2021, at 08:19 EST), <https://perma.cc/TC4W-JQ7B> (discussing legislators’ refusal to vote on Build Back Better legislation until CBO had scored it). Indeed, Professors Abbe Gluck and Lisa Bressman have highlighted the primacy of scorekeeping and noted that congressional drafters “routinely change bill text to bring legislation within a budgetary goal,” quoting one drafter emphasizing that “[i]n tax and spending programs you live and die by the score.” Lisa Schultz Bressman & Abbe R. Gluck, *Statutory Interpretation from the Inside—An Empirical Study of Congressional Drafting, Delegation, and the Canons: Part II*, 66 STAN. L. REV. 725, 764 (2014).

3. U.S. CONST. art. I, § 9, cl. 7.

4. Kate Stith, *Rewriting the Fiscal Constitution: The Case of Gramm-Rudman-Hollings*,

some of the aforementioned scorekeeping bodies existed, their remit was narrow and their work unlikely to meaningfully shape policy outcomes.

All that changed as federal deficits began to climb, which forced policymakers to grapple with new fiscal realities. To judge how the fiscal picture was evolving—and the ways in which new proposals would add to or detract from fiscal sustainability—modern scorekeeping was born. Scorekeepers were needed to analyze the impact of budgetary choices on deficits, first to determine whether policy choices were in compliance with budget rules that were adopted, and even later, as these rules fell away, to try and provide a dynamic measure of how our fiscal position was changing for policymakers on both sides of the aisle who voiced concern with rising federal debts.⁵

76 CAL. L. REV. 593, 601 (1988) (noting that “[t]his uncoordinated budget process was adequate for the needs of the 19th century; the federal budget usually remained in surplus (except during wartime) as tariff revenues outpaced spending.”).

5. Indeed, every president in the last half century has publicly emphasized the importance of cutting deficits and balancing the federal budget. See Adam Cancryn, *Biden’s New Deficit Hawk Persona Has Some Progressives Feeling Some Bad Deja Vu*, POLITICO (Feb. 28, 2023, at 04:30 EST), <https://perma.cc/44S9-YLJA> (quoting President Biden: “My economic plan is working. . . . It’s reducing the deficit. It’s fiscally responsible.”); *Trump’s Budget Boosts Defense Spending, Cuts Billions from Domestic Programs*, PBS NEWSHOUR (Feb. 27, 2017, at 20:06 EDT), <https://perma.cc/55HQ-5CK7> (quoting President Trump: “With \$20 trillion in debt . . . [t]he government must learn to tighten its belt, something families all over the country have had to learn to do”); Barack Obama, *Remarks by the President on Fiscal Policy*, WHITE HOUSE (Apr. 13, 2011, at 13:48 EDT), <https://perma.cc/2PYL-BBQW> (“We have to reduce our deficit, and we have to get back on a path that will allow us to pay down our debt.”); *President Bush Outlines Budget; Calls for Government to be “Active, But Limited, Engaged, But Not Overbearing*, WHITE HOUSE (Feb. 27, 2001), <https://perma.cc/AFZA-KPKN> (quoting President George W. Bush: “Unrestrained government spending is a dangerous road to deficits, so we must take a different path”); William J. Clinton, *Address Before a Joint Session of Congress on Administration Goals*, AM. PRESIDENCY PROJECT (Feb. 17, 1993), <https://perma.cc/C697-TSMN> (“[This budget plan] puts in place one of the biggest deficit reductions . . . in the history of this country”); George H.W. Bush, *Address Before a Joint Session of the Congress on the State of the Union*, AM. PRESIDENCY PROJECT (Jan. 31, 1990), <https://perma.cc/9JJQ-WXG7> (“[This budget plan] brings th[e] deficit down further and balances the budget by 1993”); Ronald Reagan, *Remarks at a Rally Supporting the Proposed Constitutional Amendment for a Balanced Federal Budget*, RONALD REAGAN PRESIDENTIAL LIBR. & MUSEUM (July 19, 1982), <https://perma.cc/AV7Q-ERS6> (“Isn’t it time to mobilize the great American lobby and make government understand its job is to wipe out deficits and not let deficits wipe us out?”); Jimmy Carter, *The State of the Union Address Delivered Before a Joint Session of the Congress*, AM. PRESIDENCY PROJECT (Jan. 19, 1978), <https://perma.cc/4XV7-LHJ2> (“With careful planning, efficient management, and proper restraint on spending, we can move rapidly toward a balanced budget—and we will”); Gerald R. Ford, *President Gerald R. Ford’s Address Before a Joint Session of the Congress Reporting on the State of the Union*, GERALD R. FORD PRESIDENTIAL LIBR. & MUSEUM (Jan. 19, 1976), <https://perma.cc/A6CS-FPLA> (“We can achieve a balanced budget by 1979 if we have the courage and the wisdom to continue to reduce the growth of Federal spending.”); Richard Nixon, *Annual Message to the Congress on the State of the Union*, AM. PRESIDENCY PROJECT (Jan. 22, 1970), <https://perma.cc/KS2Z-3DEZ> (“We must balance our Federal budget so that American families will have a better chance to balance their family budgets.”); Lyndon B. Johnson, *Annual Message to the Congress on the State of the Union*, AM. PRESIDENCY PROJECT (Jan. 4, 1965), <https://perma.cc/DDW8-QMM7> (“We will continue

Scorekeepers have become central in the policy process in part because they are regarded as neutral arbiters who are able to provide technocratic revenue estimates and projections for the evolution of the deficit trajectory over time that are in the realm of what is reasonable. Unsurprisingly, this is a process that is laden with uncertainty and immensely challenging. Consider the following question: What is the cost of the Biden Administration’s universal preschool initiative? Answering this question requires assumptions about all of the following: Which states will opt into this program? What portion of parents will choose to send their children to public preschool? Should the fact that parents can more easily spend these years in the labor force while their children are in preschool be accounted for in the revenue estimation process? What of the fact that years of schooling are related to improved lifetime outcomes for the children that receive it? Amidst this uncertainty, scorekeepers are required to land on a precise estimate: in this case, they concluded that expanded federal subsidies for universal preschool would come with a price tag of \$109 billion over a decade.⁶

The point of the scorekeeping exercise is not that this \$109 billion is the ultimate truth of what the cost of such a provision would be. The assumptions are too shaky to even aspire for “right” or “true” in revenue estimation, and scorekeepers are transparent about their lack of anything close to perfect foresight. They are explicit that these estimates are generally meant to be understood as “middle-of-the-road” with respect to a range of possible outcomes, as opposed to singularly accurate.⁷

If not because they produce numbers that are right, why do scorekeepers have such centrality in the legislative process—and such deep trust from policymakers? It is because these nonpartisan revenue estimators are trusted to not put spin on the ball, so to speak.⁸ That neutrality has ensured that scorekeeping as an art (for it is really more art than science) has gained ever more primacy in policy debates.⁹

along the path toward a balanced budget in a balanced economy.”).

6. CONG. BUDGET OFF., ESTIMATED BUDGETARY EFFECTS OF TITLE II, COMMITTEE ON EDUCATION AND LABOR, H.R. 5376, THE BUILD BACK BETTER ACT, AS POSTED ON THE WEBSITE OF THE HOUSE COMMITTEE ON RULES ON NOVEMBER 3, 2021 (RULES COMMITTEE PRINT 117-18), AS AMENDED BY YARMUTH AMENDMENT 112 (2021).

7. For example, CBO notes that all of its “cost estimates are to some degree uncertain,” often “represent[ing] the midpoint of such a range so that the estimate is just as likely to be too high as too low.” *CBO Describes Its Cost-Estimating Process*, CONG. BUDGET OFF. (Apr. 2023), <https://perma.cc/M83C-C7JG>.

8. As historian Philip Joyce has noted in his history of the Congressional Budget Office, the CBO—and scorekeepers generally—have made the active choice to be an arbiter in the policy process, but not to recommend policy outcomes. PHILIP G. JOYCE, *THE CONGRESSIONAL BUDGET OFFICE: HONEST NUMBERS, POWER, AND POLICYMAKING* 8 (2011) (describing the centrality of the CBO as revelatory to “the significance of honest numbers, about faceless but neutral and dedicated civil servants, about organizational culture, and about the importance of careful, informed analysis”).

9. RUDOLPH G. PENNER, *URB. INST., ERRORS IN BUDGET FORECASTING* 10 (2001) (noting the limits of the “art of forecasting”).

But, as this Article points out, there is spin—not by the scorekeepers, but by their process: because modern scorekeeping is focused on keeping score with respect to short-term deficits, the default set of rules and norms that guide that process push against a progressive policy agenda. That is true in the narrow sense of the scores themselves being myopic: the default today is a ten-year budget window which, to take one example, pushes against investments in future generations, because children will not be in the labor force to generate gross domestic product (GDP) gains in the next ten years. But it is also the case that the “right” variables for policy analysis are broader than GDP metrics—the normative case for social safety net expansions that combat inequality is not best made by narrow focus on GDP growth. Instead, it requires a more fulsome consideration of broader policy objectives, like what the impact of a fully refundable child tax credit will be on child poverty, or what the impact of green energy subsidies will be on emissions goals.

We begin this Article in Part I by tracing the history of modern scorekeeping. We point out that the growth of peacetime deficits created a need to develop reliable measures of deficits and changes in revenues and expenditures over time. When strict rules were in place (at least theoretically),¹⁰ keeping score became an essential input into the legislative process. Even as those rules have fallen away, the primacy of scorekeepers remains, because policymakers have built up the muscle of deficit analysis.¹¹ They ask—and eagerly await—knowledge of the impact of new legislative proposals on deficits, and celebrate the (relative rare) instances when scorekeepers judge new spending packages as paid for.¹² The deficit-centrality of these historical developments helped policymakers and scorekeepers develop reliance on a singular analytical measure, such that they judge spending and tax packages based on their impact on federal revenues over a short-run horizon, even when other analytical possibilities are available to them.

We then turn in Part II to detailing the traditional remit of scorekeepers. We explain scorekeepers’ approach to revenue estimation and detail the ways in which an overwhelming focus on net fiscal costs guides their work. We then home in on a few areas where the limitations of the scorekeeping process lead to

10. This was not, however, true in practice, as the penalties for not achieving deficit reduction targets tended to be sidestepped by Congress, as we describe *infra* in Part I.

11. Although the federal government has operated without formal budget rules in the last two decades, cost estimates provided by scorekeepers remain important policy inputs. In some cases this is because legislative processes mandate their consultation; for example, reconciliation cannot include measures that raise deficits in any year after the period covered by the instructions unless those “outside-the-window” costs are offset. RICHARD KOGAN & DAVID REICH, CTR. ON BUDGET & POL’Y PRIORITIES, INTRODUCTION TO BUDGET “RECONCILIATION” 6 (2022), <https://perma.cc/DS9W-2TZM>.

12. Indeed, Speaker Johnson’s recent celebration of the \$78 billion bipartisan tax package that passed the House in January 2024 reflects this tendency, as he extolled the virtues of the “pro-growth tax reform” that is paid for by “[ending] a wasteful COVID-era program, saving taxpayers tens of billions of dollars.” Benjamin Guggenheim & Nicholas Wu, *Johnson Plans to Bring Bipartisan Tax Package to House Floor Wednesday*, POLITICO (Jan. 31, 2024, at 14:05 EST), <https://perma.cc/Q67V-LW8M>.

perverse policy outcomes and create biases against progressive policy preferences. Specifically, traditional scorekeeping underestimates the potential revenue gains from tax reform, fails to consider the benefits associated with investments in children, and largely ignores the macroeconomic feedback effects associated with spending proposals.

Here, our contribution is two-fold. From an analytical perspective, we quantify the magnitude of traditional scorekeeping's limitations. This is hard to do, because it requires building out the infrastructure to reproduce the complex work of scorekeepers while revisiting key assumptions, methodological choices, and ways of framing the results. The analysis we present here represents original empirics based on novel approaches to policy evaluation being developed at the new Budget Lab at Yale.¹³

Second, from a theoretical perspective, although past work has considered certain scorekeeping deficiencies in isolation,¹⁴ we unify and frame these deficiencies as particular examples of a general problem: the federal policy process is narrowly focused on short-term deficit impacts. Practically, this has the effect of thwarting a progressive policy agenda that hinges on the value of long-term investments in people and in developing state capacity. These are arenas where the most relevant outcome variable of interest simply is not GDP growth over a ten-year budget window. So, a process that is so deficit-centric—at the exclusion of any sort of more holistic policy analysis—can be neutral in its inputs without being neutral in its outputs. That is, even though scorekeepers do not act with any set of political biases that favor a small government and push against an expansion of the social safety net, the process itself can push to exactly that outcome, and it does, systematically, in ways that have been underappreciated by the academic literature and, more importantly, by policymakers themselves.

13. The Budget Lab at Yale is a new think tank estimating the fiscal and social impacts of federal government policies, expanding traditional revenue and macroeconomic estimates to demonstrate how policies affect incomes across the distribution and deliver returns on investments. The Budget Lab is led by Martha Gimbel, Natasha Sarin, and Danny Yagan. The empirical analysis presented in this Article provides an initial sketch of the Budget Lab's capacity based on approaches developed by the Lab's early work. Our results are illustrative for the purpose of discussion as the Lab continues to refine its models.

14. Indeed, one of the authors has written extensively on scorekeepers' tendency to understate the benefits of tax reform measures. See Natasha Sarin & Mark J. Mazur, *The Inflation Reduction Act's Impact on Tax Compliance—and Fiscal Sustainability* 21 (May 15, 2023) (unpublished manuscript) (on file with author); see generally Natasha Sarin & Lawrence H. Summers, *Understanding the Revenue Potential of Tax Compliance Investment* Nat'l Bureau of Econ. Rsch., Working Paper No. 27571, (2020); Natasha Sarin, Lawrence Summers, Owen Zidar & Eric Zwick, *Rethinking How We Score Capital Gains Tax Reform*, 36 TAX POL'Y & ECON. 1, 2 (2022) [hereinafter Sarin et al., *Rethinking*]. Many others have also considered these deficiencies and made the case for reform. See, e.g., Doug Elmendorf & Heidi L. Williams, *A Serious Case for Dynamic Scoring*, SLOW BORING (Aug. 28, 2023), <https://perma.cc/JW4N-48SE>; Scott Levy, Note, *Spending Money to Make Money: CBO Scoring of Secondary Effects*, 127 YALE L.J. 936 (2017).

So, where do we go from here? We turn to this critical question in Part III. One way to summarize the deficiencies of the scorekeeping process that we identify is that today, the process captures only part of the equation: policymakers ask scorekeepers to quantify short-term costs, but not long-term benefits, which can be pecuniary (future GDP growth) or not as clearly so (impact on child poverty, or maternal health outcomes). What is needed, then, is to empower policymakers to demand—and arm the public with—better measurement of the full range of costs and benefits from interventions. Academics have much to contribute: there are myriad examples of research helpfully probing the assumptions that scorekeepers rely on in their estimation.¹⁵ There is also a large body of work that considers a broader range of benefits from government interventions than the narrow focus on GDP that is the scorekeepers' traditional remit.¹⁶ What is missing, then, is a bridge between this research and policy analysis: a systematic approach to quantifying the distributional and so-called “non-pecuniary” gains from policy interventions.

Ultimately, it is our hope that this Article serves as a call to arms for academics to produce work that helps the scorekeeping process innovate along the lines we suggest. When scorekeepers speak today it is with significant weight, and there is little scope for them to independently produce answers to questions that legislators do not ask. One value of thoughtful outside research is that it can encourage policymakers to request more fulsome analysis.

We also suggest ways in which the scorekeeping process itself can be reformed to preserve its vision of neutrality. Given the history of scorekeeping that we retell here, is not surprising that the default scorekeeping rules prioritize

15. See, e.g., Sarin et al., *Rethinking*, *supra* note 14, at 1 (challenging the elasticities relied upon by scorekeepers in analyzing capital gains tax reform by noting the ways in which the nature of capital gains realizations has shifted relative to earlier assumptions); John E.T. Bistline, Neil R. Mehrotra & Catherine Wolfram, *Economic Implications of the Climate Provisions of the Inflation Reduction Act*, BROOKINGS PAPERS ON ECON. ACTIVITY, Spring 2023, at 79 (estimating costs of the Inflation Reduction Act higher than those estimated by official scorekeepers while providing further evidence of the Act's social benefits in reducing carbon emissions); *Return on Investment: Re-Examining Revenue Estimates for IRS Funding*, INTERNAL REVENUE SERV. (Feb. 2024), <https://perma.cc/KEC6-MDSM> (reevaluating the traditional approach to scoring investments in IRS enforcement, taking into account deterrent effects and efficiency gains emphasized by some academics).

16. Recent studies of the Child Tax Credit emphasize its role in alleviating child poverty. See, e.g., CHRISTOPHER WIMER, SOPHIE COLLYER, DAVID HARRIS & JIWAN LEE, COLUM. UNIV. CTR. ON POVERTY & SOC. POL'Y, *THE 2021 CHILD TAX CREDIT EXPANSION: CHILD POVERTY REDUCTION AND THE CHILDREN FORMERLY LEFT BEHIND 1* (2022), <https://perma.cc/LMV3-SLVR> (finding that the expansion of the Child Tax Credit in the American Rescue Plan Act reduced child poverty by forty-three percent in 2021). A long literature on the Earned Income Tax Credit studies the credit's effect on the labor force attachment of low-income workers. See, e.g., Jacob Bastian & Katherine Micheltore, *The Long-Term Impact of the Earned Income Tax Credit on Children's Education and Employment Outcomes*, 36 J. LABOR ECON. 1127, 1127 (2018) (estimating that childhood exposure to credit expansions increases the likelihood of being employed as a young adult by one percent).

deficit estimation. Defaults clearly matter, as the default rules continue to dominate the modern practice of scorekeeping. But the defaults can also be changed. By illustrating the limitations of the traditional requests scorekeepers receive and the rules under which they are expected to operate, we encourage policymakers to make better use of the scorekeeping process, with assumptions and outputs tailored to the policies being evaluated.

A related question that we pose in this Article, but do not precisely answer, is whether the type of analysis that we call for is best suited for scorekeepers or some other analytical body that can arm policymakers with a more accurate reflection of the wide range of costs and benefits associated with a particular intervention. In some contexts, we speculate that scorekeepers are well-suited to this work. For one, what we call non-pecuniary benefits of policy proposals are often essential inputs into preparing a cost estimate.¹⁷ But in a world where scorekeepers are massively capacity constrained even in providing cost estimates, asking these same individuals to present more holistic policy analysis may be a bridge too far. And when official scorekeepers are limited to answering the questions posed by the members of Congress they serve, external voices might be best equipped to provide relevant information to the public.

These are hard issues of institutional design. But they are ones where the stakes are high, and as such they demand thoughtful consideration. Importantly, preserving the status quo is not a choice that is free from normative implications. As this Article shows, the current process of scorekeeping and its primacy in policymaking biases against progressive policy interventions in ways that make it harder to invest in future generations or build out the care economy than it is to cut taxes for the wealthiest and large corporations. It is no surprise then that in recent legislative battles, we have observed the (relative) ease of the latter and the difficulty of the former. Scorekeepers as players are rightly celebrated as neutral arbiters of the legislative game. But the rules of that game—the norms that govern scorekeeping—are not neutral at all. They tilt against policies that redistribute, invest in future generations, and promote equity—a tilt that it is essential to reverse in the years to come.

I. HISTORY OF SCOREKEEPING AND THE CONGRESSIONAL BUDGET PROCESS

To understand the current flaws of government budget scoring, it is necessary to understand how we got here. The centrality and biases of

17. Indeed, in assessing the impact of the Affordable Care Act, CBO did ultimately release its estimate of the impact of the legislation on the uninsured population (a non-GDP “benefit” of Obamacare), and it was well-situated to do so in part because it had to arrive at an estimate of the increase in the insured population in order to produce a cost estimate. Such cases we speculate are more common than not: to assess the cost of early childhood education programs, an essential input into the estimate is how many children will be enrolled in those programs. To assess the cost of clean energy subsidies, an essential input into the estimate is how many consumers will buy electric cars that would not have otherwise, and so forth.

scorekeeping are not accidents, nor are they reasoned results of thoughtful *ex ante* decision-making. Rather, they are the direct result of the historical context through which the federal budget process formally emerged and subsequently evolved. As the budget process and scorekeeping alongside it have developed through time, an initial focus on deficit reduction has been repeatedly reinforced, such that scorekeeping today is synonymous with revenue estimation, and revenue estimation is thus determinative of the policy decisions being made.¹⁸ Three key evolutions in the history of federal budgeting illustrate this reinforcement in action and elevated the role of scorekeepers in ways that continue to reverberate today: the creation of CBO, the enactment of the Gramm-Rudman Hollings Act, and the enforcement of pay-as-you-go (PAYGO) rules. In each of these moments, the work scorekeepers produced and the ways in which this work was absorbed by policymakers was shaped by the singular goal of deficit reduction. Mounting need to track and reduce deficits thus led to analytic tools that enable deficit estimation at the expense of other forms of policy evaluation.

A. The Creation of CBO

The Constitution itself references the importance of public budgetary accounting, stating that “a regular Statement and Account of the Receipts and Expenditures of all public Money shall be published from time to time.”¹⁹ Nevertheless, for many years, the federal budget process lacked formality and cohesion. Though the Constitution provides Congress with the “power of the purse strings” to tax, borrow, and spend as “necessary and proper” to carry out its enumerated powers and provide for the “common Defence and general Welfare,”²⁰ it is silent as to the practices Congress should undertake to exercise these powers. There was no encouragement that revenues line up or exceed spending, no mandate for balanced budgets or fiscal discipline, not even a set of rules for determining the size of the deficit and how it was to evolve over time. Professor Kate Stith, in work in the late 1980s, provided an overview of an “uncoordinated” historical process that through the late twentieth century considered spending and revenue legislation on an “ad hoc basis” as debates arose around specific laws, rather than through a formal budget process.²¹ This sort of start-and-stop approach to fiscal discipline was fit for purpose for much

18. Paul Starr describes a process of “entrenchment,” whereby constitutive features of politics and society are reinforced and become resistant to change, that serves a useful lens for understanding how initial scorekeeping conventions came to dominate modern approaches through historical path dependence. PAUL STARR, *ENTRENCHMENT: WEALTH, POWER, AND THE CONSTITUTION OF DEMOCRATIC SOCIETIES* 1-2 (2019) (defining “entrenchment”); *see also id.* at 11-13 (“Lock-in and the costs of change”).

19. U.S. CONST. art. I, § 9, cl. 7.

20. U.S. CONST. art. I, § 8.

21. Stith, *supra* note 4, at 601.

of this nation's early history because deficits were essentially zero, outside of wartime spending increases.²²

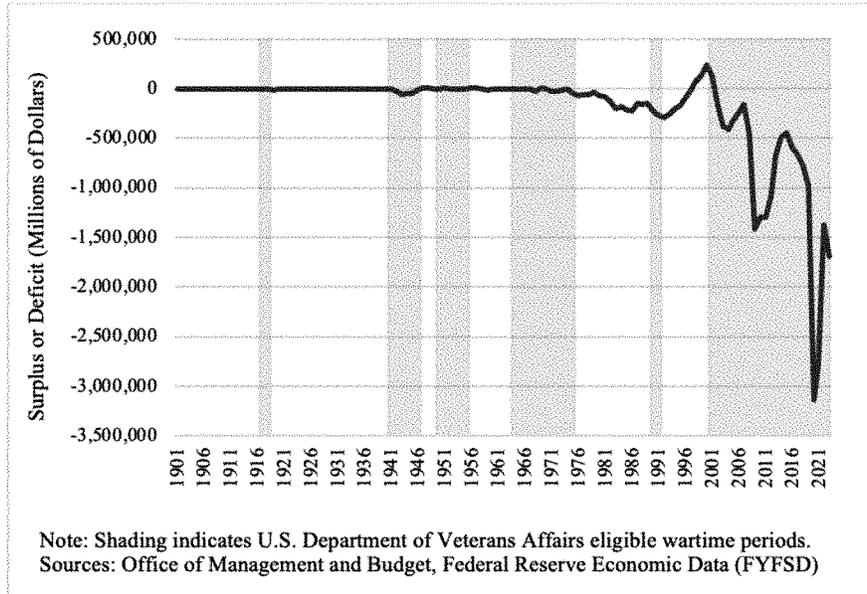


Figure 1: U.S. Annual Deficit in Millions of Dollars

22. *Id.* (“This uncoordinated budget process was adequate for the needs of the 19th century; the federal budget usually remained in surplus (except during wartime) as tariff revenues outpaced spending.”).

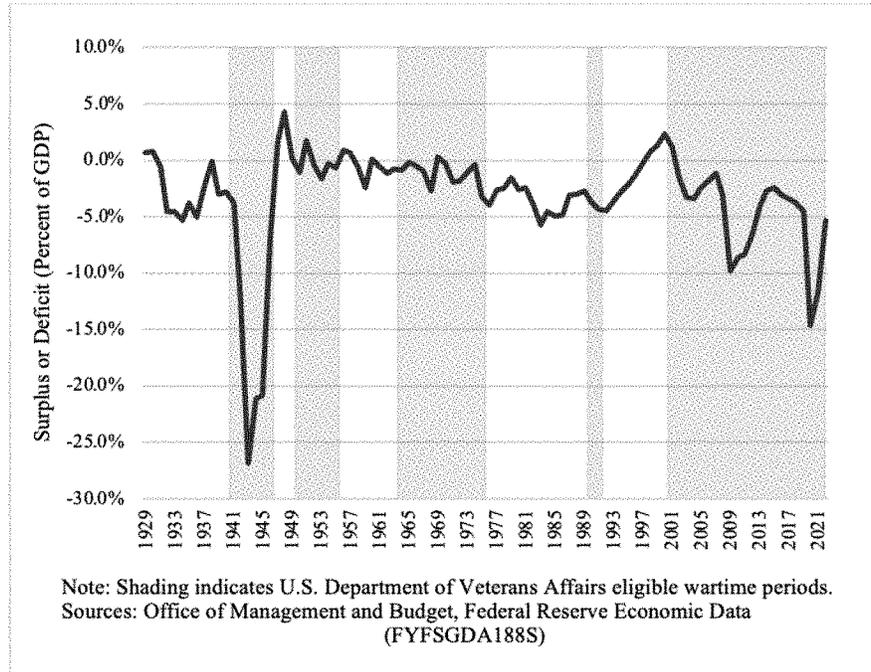


Figure 2: U.S. Annual Deficit as a Percentage of GDP

In the 1970s, federal deficits began to climb, which focused attention on the importance of a sustainable fiscal path and shaped the ways in which a new budget process became formalized.²³ In hearings surrounding budget reform in the early 1970s, discussion of deficits dominated, with legislators and other policymakers decrying “sizable deficits in the Federal budget [that] continue to plague us,”²⁴ bemoaning a “situation where budgetary deficits have become the rule,”²⁵ and blaming the executive for “sending down deficit budgets.”²⁶ Members of Congress feared that they had not only lost control over deficits, but also over the entire budgetary process in relation to the executive branch.

Congress’s response came in the Congressional Budget and Impoundment Control Act of 1974, which was focused on shifting budgetary control from the executive branch back to Congress. It created and largely outlines the

23. *Id.* at 615-16 (describing how budget reform was viewed as a potential check on growing appropriations through “backdoor spending mechanisms”).

24. *Improving Congressional Budget Control: Hearings Before the Joint Study Comm. on Budget Control*, 93d Cong. 22 (1973) (statement of Arthur Burns, Chairman, Bd. of the Fed. Rsrv. Sys.).

25. *Improving Congressional Budget Control: Hearings Before the Joint Study Comm. on Budget Control*, 93d Cong. 44 (1973) (statement of Sen. William Brock).

26. *Improving Congressional Budget Control: Hearings Before the Joint Study Comm. on Budget Control*, 93d Cong. 87 (1973) (quoting Sen. Humphrey).

congressional budget process in place today.²⁷ The basic idea was a top-down approach to budgeting, with a structure in place for Congress to provide meaningful topline guidance on aggregate fiscal totals as opposed to authorizing spending in a piecemeal manner that failed to take into account how debt and economic conditions had evolved, and how outlays and revenues compared in any given year.²⁸

Perhaps more importantly, though, the Act created the infrastructure to support this budget process, as its drafters realized that shifting control back to Congress required not just a prescription to do so, but an expansion of technical expertise and capacity to enable the requisite budget analysis. That included the creation of House and Senate committees that would oversee budgeting and a Congressional Budget Office to serve as Congress's version of the Executive's Office of Management and Budget (OMB), which provided economic information that fed into executive budget processes.²⁹

Under the 1974 Act, CBO is required to produce a cost estimate to the extent practicable for all bills reported out of committee.³⁰ As a result, any bill with revenue effects is accompanied by a CBO report providing a concise estimation of what those effects would be.³¹ This enabled Congress to exert more control over the budget by giving it the tools to understand how individual pieces of legislation affected the overall deficit outlook.

But, importantly for our purposes, it also focused attention to the revenue impacts of any given bill, even those that were not proposed in the context of a deficit-reducing budgetary scheme, rather than any other form of analysis that Congress could have used to expand its technical expertise and empower itself relative to the executive branch. Whereas in earlier years, the revenue impacts of an individual piece of legislation would have flown under the radar, the creation

27. Congressional Budget and Impoundment Control Act, 2 U.S.C. § 601 (1974). The Act was also specifically responsive to the problem of impoundments—Nixon's refusal to obligate funding that had been appropriated by law in order to preserve his own spending priorities. See Christopher Wlezien, *The Politics of Impoundments*, 47 POL. RSCH. Q. 59, 60 (1994) (describing Nixon's impoundments as significant and substantive expressions of policy disagreement, eliminating spending priorities rather than merely reducing their appropriations). The Act thus limited presidential authority to withhold appropriations to specific cases of rescissions and deferrals, subject to congressional approval.

28. Although prior to the Act, each appropriations committee had operated in a decentralized manner, focusing on particular budget accounts in its purview, the new process required a budget resolution reported out of the House and Senate budget committees according to these committees' determinations for overall spending guidelines. This annual non-statutory congressional action, termed a "congressional budget resolution" would set targets for new budgetary authority, annual revenues, annual outlays, the deficit, and the gross public debt, with the objective that the appropriations would be consistent with these macro targets.

29. Congressional Budget and Impoundment Control Act of 1974, Pub. L. No. 93-344, § 201, 88 Stat. 297, 302, (1974) (codified at 2 U.S.C. § 601).

30. *Id.* at § 403.

31. *Id.* at § 308.

of CBO ensured that this would never again be the case. The 1974 Act also introduced the concept of a budget window in requiring CBO to provide cost estimates for five fiscal years beginning with the bill's effective date.³² As we will see, the importance of this window was magnified with further budgetary process developments and is of crucial relevance today.

At the same time that the 1974 Budget Act created a new federal budget process, it also replicated existing infrastructure that was previously housed at the executive's OMB. As a result, the practice of scorekeeping was duplicated in ways that doubly reinforced analysis an understanding of proposed legislation's impact on the deficit. Thus, while the federal government's analytic capacity was significantly bolstered through the creation of CBO, the scope of its analytic lens homed in on budgetary effects alone. Eventually, CBO's need to estimate deficits spread to other federal scorekeeping bodies, for example, demanding similar analytic focus of JCT to produce revenue estimates of tax legislation to feed into CBO's deficit estimates.³³

B. *Gramm-Rudman-Hollings*

Although it laid the foundation for a robust scorekeeping process that today is at the center of legislative debates, as a means of reigning in deficits, the Budget Control Act of 1974 achieved far less than its advocates had hoped. Both the House and the Senate routinely skirted the Act's procedural requirements.³⁴ And though the creation of CBO enabled lawmakers to very transparently see how their decisions affected revenue and spending totals, beyond that, it provided no incentive for those decisions to run in a particular fiscal direction.³⁵

Deficits had begun to climb in the late 1970s, and they grew even more

32. *Id.* at § 403.

33. Our focus here is on CBO, as its developments are central and in some ways determinative of developments in other scorekeeping bodies. JCT's history is illustrative. JCT was originally conceived of as an investigative entity to combat perceived inefficiencies and fraud in the Bureau of the Internal Revenue, the predecessor of the IRS. Later legislative developments added investigation of tax administration more generally to JCT's remit. Since then, JCT's statutory mandate has not changed much. *History*, JOINT COMM. ON TAX'N, <https://perma.cc/2UJF-MD6Q> (archived on Aug. 10, 2025). However, in practice, JCT's work has changed significantly: in particular, JCT prepares revenue estimates of tax legislation at the request of members of Congress. *Revenue Estimating Process January 2023*, JOINT COMM. ON TAX'N (Jan. 27, 2023), <https://perma.cc/77SH-M6WN>. Similarly, OMB works closely with Treasury's Office of Tax Analysis so that the deficit effects of both spending and revenue proposals can be accounted for.

34. See Stith, *supra* note 4, at 618 ("In part, the failures of the budget process after 1974 can be traced to Congress' own failure to abide by the 1974 legislation. Both Houses routinely violated the Act's procedural requirements.").

35. CBO historian Philip Joyce makes this point in analyzing how the creation of CBO, alongside other developments in the budget process, elevated the importance of scorekeeping in the legislative process. Philip G. Joyce, *Congressional Budget Reform: The Unanticipated Implications for Federal Policy Making*, 56 PUB. ADMIN. REV. 317, 320 (1996).

substantially in the years following the 1974 Act, which is attributable to both the tax cuts passed in the Economic Recovery Tax Act of 1981 and the build-up in defense spending in the Reagan Administration.³⁶ Congress took two actions in response, both of which magnified the role of deficit-centric scoring in the budget process.³⁷ First, in the early 1980s, Congress informally adopted a new procedure, termed reconciliation, that required a greater portion of the budget to fall under annual congressional supervision.³⁸

Second, Congress faced the larger question of how best to overhaul federal budgeting to rein in federal spending. During a debate concerning an increase in the debt limit, Senators Phil Gramm, Ernest Hollings, and Warren Rudman took a more forward-looking view towards fiscal issues.³⁹ They introduced a bill that would mandate phased-in deficit reduction and balance the budget. And in late 1985, this next major evolution in federal budgeting, known as Gramm-Rudman-Hollings (GRH), was passed, with a core commitment to eliminating the federal

36. See Stith, *supra* note 4, at 623 n.187 (“A good argument can be made that the huge deficit growth in the 1980s was due to increased defense expenditures and tax cuts—two of President Reagan’s three fiscal goals (the third being a balanced budget)—in combination with an economic recession.”).

37. There is also empirical support for the proposition that deficits themselves cemented CBO’s authority as an official scorekeeper. See Philip Rocco, *Keeping Score: The Congressional Budget Office and the Politics of Institutional Durability*, 53 *POLITY* 521, 707 (“Consistent with the argument that rising deficits helped generate demand for analyses from the CBO, the results of all three models show that the size of the federal budget deficit is positively and significantly related to the number of *New York Times* articles mentioning the CBO in a given year.”).

38. Stith, *supra* note 4, at 620. It is worth noting that in the intervening four decades, reconciliation has become an increasingly important legislative tool because it provides a way to authorize new congressional spending in a land of slim majorities in the Senate. See TORI GORMAN, CONG. RSCH. SERV., R48444, *THE RECONCILIATION PROCESS: FREQUENTLY ASKED QUESTIONS 1* (Mar. 6, 2025) (“[D]ebate on a reconciliation bill is limited to 20 hours . . . which means cloture is not required to reach a final vote”). Because reconciliation bills are not subject to filibuster and are subject to only limited amendments, this process has advantages for facilitating quick adoption, which has happened twenty-three times since the budget reconciliation process was introduced by the 1974 Act. See MEGAN S. LYNCH & JAMES V. SATURNO, CONG. RSCH. SERV., R44058, *THE BUDGET RECONCILIATION PROCESS: STAGES OF CONSIDERATION 9-11* (2021) (describing the 20-hour Senate limit on debates, motions to proceed as not debatable; limits on amendments and germaneness, and no historical need to invoke cloture); GORMAN, CONG. RSCH. SERV., *supra*, at 4 & tbl.A-1 (“Since its first use in 1980, Congress has considered 28 reconciliation bills—23 have been enacted into law, four vetoed, one did not pass the Senate”). As Professor Kate Stith has commented, somewhat ironically, enhanced, annual scrutiny of the budget process made it harder for Congress to regain control of spending, because such regular supervision mandated that Congress return and renegotiate, a process that requires both time and significant political capital. Stith, *supra* note 4, at 620-21.

39. Darrell M. West, *Gramm-Rudman-Hollings and the Politics of Deficit Reduction*, 499 *ANNALS AM. ACAD. POL. & SOC. SCI.* 90, 95 (1988) (discussing the contentious atmosphere in Congress when Gramm-Rudman-Hollings (GRH) was introduced, no agreement having yet been made); Leonard E. Burman, *Congress’s Debt Limit Problem Is Toddler Fiscal Policy*, TAX POL’Y CTR., URB. INST. & BROOKINGS INST. (Sep. 22, 2021), <https://perma.cc/56PR-R7UX> (characterizing the enactment of GRH as leveraged by the debt limit).

deficit by 1991.⁴⁰

In its introduction of automatic triggers for deficit reduction, GRH made the careful estimation of budget deficits even more central, both elevating the role of scorekeeping and entrenching scorekeeping's focus on deficits. As Professors Jim Poterba, David Stockman, and Charles Schultze explained, GRH was novel in its introduction of a mechanism for ensuring that deficits did not exceed targets. Under GRH's staged approach, the President submitted a budget with a forecasted deficit within the GRH target, OMB and CBO projected the actual deficit using enacted tax and budget legislation, and an average of OMB and CBO's deficit projections higher than the target triggered a requirement that the President issue a sequester order.⁴¹

GRH thus magnified the 1974 Act's focus on deficits in the budget by attaching real consequences to deficit-increasing scores CBO produced. It thus made the scoring process the essential determinant of whether automatic deficit reduction would take place or not: the threat of sequestration required both CBO and OMB to provide snapshot estimates of spending baselines and deficit excesses that would form the policy baseline and help guide determinations about the need for sequestration. Sequestration would hang in the balance depending on the netting of spending and revenues, vesting immense power in scorekeepers' revenue and outlay determinations.

In this way, the revenue estimation process provided (and continues to provide) a lot of leeway to scorekeepers in estimating overall expenditures and revenues collected.⁴² Further, the binding nature of the budget rules if anything upped the stakes for these determinations, increasing the primacy of scores in the budget process.⁴³ That primacy pushed against the acknowledgement of the wide

40. Balanced Budget and Emergency Deficit Control Act of 1985, Pub. L. No. 99-177, 99 Stat. 1038. GRH introduced the process of "sequestration," providing automatic spending cuts across all covered programs to cut the deficit if Congress and the President did not find a fiscally sustainable budgetary outcome within a prescribed "maximum deficit amount." *Id.* § 253(a)-(b)(1), 99 Stat. 1078-79 (providing that, if the estimated deficit exceeds the "maximum deficit amount," the President must order a sequestration to eliminate the excess). The overarching objective of GRH was perhaps a naïve one: because the budget process had become so political in ways that entrenched existing programs and required backroom negotiations and pork exchanged for congressional support, imposing and somewhat automating legal and institutional constraints against deficit growth was an attempt to try and depoliticize an inherently political process. The idea was that rather than trying to shame Congress and the President into supporting a budget imposing meaningful fiscal discipline, GRH would automate that discipline by creating a regime to guarantee deficit reduction.

41. This staged approach is described in more detail in James M. Poterba, David Stockman & Charles Schultze, *Budget Policy*, in *AMERICAN ECONOMIC POLICY IN THE 1980S*, at 235, 263-64 (Martin Feldstein ed., 1994).

42. For examples of scorekeepers exercising that discretion in this period, see *infra* note 46 and accompanying text.

43. In providing scorekeepers with leeway, scorekeeping under the GRH provided opportunities to "game" scores to produce tailored outcomes. For example, to the extent budget savings were achieved in the early years following the amended GRH's enactment, those "savings" were achieved, at least in part, through manipulated scoring like accounting

uncertainty associated with these estimates and focused attention narrowly on the estimates, as opposed to on broader policy analysis of the particular costs or benefits of spending and revenue proposals.

C. The Budget Enforcement Act

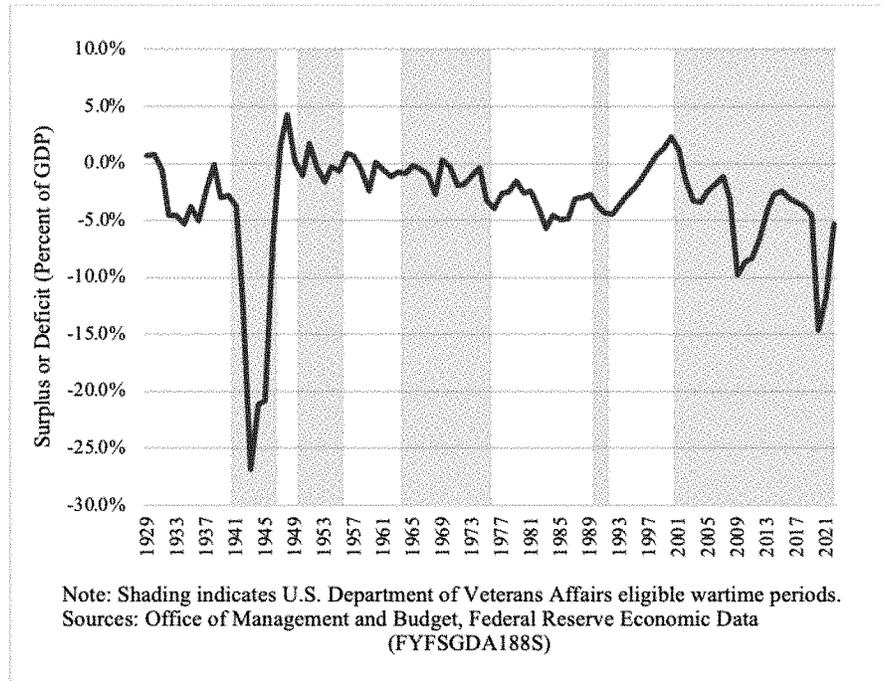


Figure 3: U.S. Deficit Reduction in the President's Budget, 1983-1991

While GRH was in effect, actual deficits came within \$10 billion of targets only once. Even so, only one significant sequester in fiscal year (FY) 1986 was actually implemented (with OMB recommending cuts of \$5.85 billion each in defense and non-defense spending).⁴⁴ In other years, sequester was either

tricks and questionable revenue raisers, including asset sales and shifting some agencies off budget. James M. Poterba, *Do Budget Rules Work?* 19 (Nat'l Bureau of Econ. Rsch., Working Paper No. 5550, 1996). According to one estimation, half of the deficit reduction following GRH came as the result of one-shot fiscal measures like asset sales. Robert D. Reischauer, *Taxes and Spending Under Gramm-Rudman-Hollings*, 43 NAT'L TAX J. 223, 227-28 (1990). Further, CBO's economic and technical adjustments in the decade following GRH increased more than four-fold, suggesting that part of the GRH legacy was "to place greater pressure on forecasters to use optimistic assumptions that would make deficit reduction easier." Poterba, *supra*, at 23.

44. CONG. BUDGET OFF., SEQUESTRATION REPORT FOR FISCAL YEAR 1986: A SUMMARY 8 (1986). For a detailed summary of year-by-year spending and sequestration decisions during the years in which GRH was in effect, see G. William Hoagland & Loren Adler, *Origins of the Sequester*, BIPARTISAN POL'Y CTR. (Feb. 27, 2013), <https://perma.cc/2H9D-9U2Q>.

avoided through accounting tricks that made deficits disappear without meaningfully changing revenues or outlays, rescinded, or legislatively overridden.⁴⁵ By FY 1991, following the recession in 1990, GRH would have called for cuts of *fifty percent*—a clearly untenable outcome, especially considering how many programs were exempt from sequestration.⁴⁶ More broadly, GRH's sequestration mechanism did nothing to force agreement between the President and Congress, and it provided no guidance as to why a deficit target had not been met and where an appropriate cut might accordingly be made.⁴⁷

Disenchanted, Congress turned to a new approach to deficit reduction. But it was one that continued to magnify the importance (and preserved the biases) of modern scorekeeping.

In 1990, the Budget Enforcement Act (BEA) adjusted the budget rules to place caps on discretionary spending and created a PAYGO (pay-as-you-go) rule for entitlement spending excluding Social Security.⁴⁸ The structural difference relative to the GRH was that the BEA did not require hitting numerical deficit targets on an annual basis; instead, it required that enacted policies not raise deficits relative to projections within a specified time frame (five years). In some sense, this goal was a more attainable one because the contraction of existing revenue sources or more broad changes in economic conditions that caused deficits to grow mechanically would not tie legislators' hands with respect to sequestration.⁴⁹ Congress had learned, in other words, that the "baseline" against which OMB and CBO would score policies ought to consider factors beyond lawmakers' control.

From a scoring perspective, though, the biases of modern scorekeeping were, if anything, augmented by this new approach. If GRH made prominent the

45. Hoagland & Adler, *supra* note 44. For example, estimators could make their assumptions about economic growth slightly more optimistic, increasing revenue projections and reducing the likelihood of sequester. CONG. BUDGET OFF., *THE ECONOMIC AND BUDGET OUTLOOK: FISCAL YEARS 1994-1998*, at 89 (1986).

46. In a hearing before the House Committee on the Budget concerning a proposed balanced budget amendment to the Constitution in 1997, Allen Schick identified the many exemptions in GRH's sequestration formula as a primary reason that GRH failed to meaningfully reduce deficits. Allen Schick, *Hearing on the Proposed Balanced Budget Amendment to the Constitution*, BROOKINGS INST. (Feb. 5, 1997), <https://perma.cc/L5TS-MLMJ>.

47. CBO emphasized this fundamental weakness of the GRH approach in defending the new approach underlying the Budget Enforcement Act (BEA). CONG. BUDGET OFF., *supra* note 45, at 88.

48. Budget Enforcement Act of 1990, Pub. L. No. 101-508, 104 Stat. 143.

49. See CONG. BUDGET OFF., *supra* note 45, at 86 ("Deficit targets, though they still exist, play no role under the BEA through 1993. Through that year, the President must adjust them for changes in economic and technical assumptions and conceptual revisions . . . As long as the targets are fully adjusted, the deficit estimate may increase substantially—because of a deteriorating economy, for example—and no action would be mandated to offset the bleaker deficit outlook.").

importance of budgetary baselines against which deficit targets would be judged, BEA introduced policy myopia by crystallizing the importance of the budget window in the scorekeeping process.

The BEA was meant to discourage the adoption of policies that would increase deficits. But what it did specifically was discourage policies that would increase deficits over the five-year budget window.⁵⁰ That created a new opportunity for budget gimmickry, because proposals that would lower revenues or raise spending outside this horizon were not similarly penalized. It also narrowed policymakers' focus, as policies that reduced deficits in the long-run (but at a short-run cost), would register as no better than those that had short-run costs today, but without any long-run benefits.

More generally, the BEA ratcheted up the need for and attention on scorekeeping, with one former CBO economist commenting, "These numbers are more important than they ever were, if only because of the new budget rules."⁵¹ But yet again, these rules failed to achieve their underlying goal of fiscal discipline. Although the BEA was officially in place for a decade after its enactment, formally expiring in 2002,⁵² it began to erode much earlier.

The fact that the federal government achieved a budget surplus dampened enthusiasm for continued fiscal discipline, and because Congress was permitted even while the BEA was still in place to simply set aside its rules on a case-by-case basis, it was possible to deviate from the precepts of budget balancing even while the rules were technically in place.⁵³ For example, during the Bush Administration in 2001, Congress advanced large tax cuts without considering revenue offsets.

But PAYGO rules were revived in 2009 in the Obama Administration, in response to increased deficits following the Bush tax cuts, making changes in the annual federal deficit critically important once more.⁵⁴ The Statutory PAYGO

50. Former CBO Director Robert Reischauer, in testimony before the House Subcommittee on Legislation and National Security, recognized this point early in the BEA regime: "Consequently, the closer we get to 1995, the easier it becomes for proponents of reduced taxes or increased mandatory spending to push the cost of their proposals beyond the PAYGO window." *Budget Enforcement for Reconciliation: Hearing Before the Legis. & Nat'l Sec. Subcomm. of the H. Comm. On Gov't Operations*, 103d Cong. 70 (1993) (statement of Robert D. Reischauer, Dir., Cong. Budget Off.).

51. Julie Kosterlitz, *Educated guesswork*, 23 NAT'L J. 2408, 2408 (1991) (quoting Paul B. Ginsburg).

52. *The "Pay-As-You-Go" Budget Rule*, CTR. ON BUDGET & POL'Y PRIORITIES (Aug. 12, 2019), <https://perma.cc/4EMV-HZAF>.

53. Former CBO Director Dan Crippen testified before the House Committee on the Budget, which was considering extending the BEA, that fiscal surpluses lend to abandonment of deficit-controlling procedures. *Extending the Budget Enforcement Act: Testimony Before the H. Comm. on the Budget*, 107th Cong. 3 (2001) (statement of Dan L. Crippen, Dir., Cong. Budget Off.).

54. Barack H. Obama, *Message to the Congress on Proposed Pay-As-You-Go Legislation*, WHITE HOUSE (June 9, 2009), <https://perma.cc/R8P7-7GAK> (announcing to Congress the submission of President Obama's PAYGO proposal). PAYGO principles

Act of 2010 required any revenue reductions to be accompanied by increases in revenue or decreases in mandatory spending, that is, spending outside of the appropriations process, including Social Security and Medicare outlays.⁵⁵ The rules were intended to bring back the pay-as-you-go norm of the BEA while curbing earlier gimmicks, prohibiting legislation from moving costs outside the budget window.⁵⁶

Even so, the Statutory PAYGO Act, like its predecessors, has done little to actually curtail deficits. Its procedural mechanisms are easily sidestepped by Congress, which can vote to waive the PAYGO rules.⁵⁷ Nonetheless, the legacy of PAYGO and earlier historical developments has left us with a scorekeeping apparatus that is very good at doing one narrow thing: estimating the effects of proposed policies on deficits within a short timeframe. In the next section, we describe the scorekeeping process today and turn to its limitations.

II. SCOREKEEPING TODAY

In practice, over the course of recent years, Congress has operated without formal spending limits. That is not to say that the budget process functions devoid of the kind of focused attention from policymakers on revenue impacts that was a hallmark of the budget rules era of the 1970s-1990s. Although formalized budget rules are no longer in place, the scorekeeping process remains hugely impactful because policymakers are attuned to the revenue impact of policy proposals, and public pressure pushes in favor of deficit reduction (and certainly not deficit expansion) even when not congressionally mandated.

We believe this focus is errant, and that it has significant ramifications for policy outcomes. It is important to note that our critique of the increasing reliance on scorekeeping is not a critique of *scorekeepers*, but rather a critique of the

functioned as the core of the BEA, demanding matching of revenues and expenditures within short-term time frames.

55. *The Statutory Pay-As-You-Go Act of 2010: A Description*, OFF. MGMT. & BUDGET, <https://perma.cc/Y7X8-Y8PV> (archived Aug. 11, 2025). This update to PAYGO addressed some of the shortcomings of the previous iteration—prohibiting, for example, legislation subject to PAYGO from moving costs outside of the budget window. See *What Is PAYGO?*, TAX POL’Y CTR. (Jan. 2024), <https://perma.cc/FNP9-CH8V>. The same biases pushing against deficit-increasing policies within the budget window, however, persisted.

56. See *What Is PAYGO?*, *supra* note 55.

57. Notably, Congress waived the rules in 2017 to avoid having to pay for the Tax Cuts and Jobs Act (TCJA), which cost the federal government an estimated \$1.5 trillion in revenue over the 10-year budget window. JOINT COMM. ON TAX’N, JCX-67-17, ESTIMATED BUDGET EFFECTS OF THE CONFERENCE AGREEMENT FOR H.R. 1, “THE TAX CUTS AND JOBS ACT” 8 (2017). This estimate comes from the JCT, and as many have noted, is a product of numerous accounting tricks, such as sunset provisions designed to reduce costs, and it ignores dynamic effects. The Committee for a Responsible Budget estimated that the Tax Cuts and Jobs Act (TCJA) could end up costing \$2.2 trillion over ten years. *Final Tax Bill Could End Up Costing \$2.2 Trillion*, COMM. FOR A RESPONSIBLE FED. BUDGET (Dec. 18, 2017), <https://perma.cc/DQW6-HKEV>.

overwhelming focus on net fiscal costs that traditional scorekeeping emphasizes and policymakers typically demand of scorekeepers. As we have demonstrated, the federal scorekeeping apparatus acquired this focus during critical moments in history when deficits were front of mind, so it is unsurprising that the process is now narrowly focused on deficit estimation, governed by rules that bias against short-run deficit increases that may well pay off in the longer term, and positioned to elevate the importance of the underlying score.

What is counted by scorekeepers, counts. Because scorekeepers are typically narrowly tasked with producing cost estimates over a short-term budget window, their policy analysis is incomplete. It informs policymakers of the short-run costs of particular interventions, while failing to consider their longer-term benefits. That is not just wrong from an analytical perspective, but it more importantly, creates biases in a scorekeeping system that is supposed to be neutral.

In particular, while scorekeepers themselves are neutral, the task they have been assigned in focusing on short-term deficits as their primary outcome variable of interest creates clear biases against policies that invest in the future and promote equality. An alternate scorekeeping process might instead evaluate long-term costs and benefits, the way a private company analyzes potential investments on the basis of their net present value. While others have described the shortcomings of our current federal process relative to such an alternative vision, this paper is the first to examine scorekeeping's limitations in the context of the history that produced them and as an obstacle to progressive policies across the board, not just in a handful of previously identified instances. Our argument is that scorekeeping's focus on short-term deficits systematically skews policy outcomes against progressive reforms that invest in future generations and redress inequality.

For some, skewing toward a more stagnant, smaller government may be normatively preferable. But for policymakers who demand policies that seek to empower and enlarge government, there is a missing understanding and lack of emphasis on the ways in which the traditional scorekeeping process inherently tips the scales towards opposing normative choices. In order to explain these biases and how they arise, and to empower this latter group of policymakers to make full use of scorekeepers' analytic capacity, we begin by providing a comprehensive review of scorekeeping, revealing its strengths, limitations, and the ways in which its current structure bears significantly on policy outcomes.

A. A Scorekeeping Primer

When we refer to “scorekeeping,” we have in mind the analytic process by which the budgetary effects of legislation are assessed.⁵⁸ The centrality of

58. The Congressional Research Service defines scorekeeping as “the process of measuring the budgetary effects of pending and enacted legislation against the baseline.” BILL HENIFF JR., CONG. RSCH. SERV., 98-560, BASELINES AND SCOREKEEPING IN THE FEDERAL

scorekeeping in the policy process means that policy debates tend to focus overwhelmingly on the information produced by scorekeepers—generally revenue estimates—precluding other relevant discussion on policy merits. The history outlined above traces how CBO developed its mandate through a historical focus on deficits. A similar story can be told of OMB, the executive branch’s major federal scorekeeper.⁵⁹ Both CBO and OMB rely on additional scorekeepers—the Joint Committee on Taxation and Treasury Office of Tax Analysis respectively—that are specifically tasked with scoring revenue proposals. These revenue scores are taken as inputs by CBO and OMB in estimating the effects of omnibus legislation and overall deficit levels. While the specific guidelines for executive branch scorekeepers at OMB and OTA and congressional branch scorekeepers at CBO and JCT differ somewhat, their overall task is generally well-summarized by CBO, which emphasizes that its congressional mandate requires scorekeepers to “prepare cost estimates for legislation at certain points in the legislative process,” with each estimate “tell[ing] a concise story about a legislative proposal’s likely effects on federal outlays or revenues.”⁶⁰

It is hard to overstate the preeminence of official scores (and scorekeepers) on the policy process. Professors Cross and Gluck have noted that the role CBO has come to occupy has ensured that its estimates “play[] a pivotal role in a bill’s success or failure.”⁶¹ Given this, official scorekeepers are often involved in legislative discussions from the outset, working to explicitly ensure that bills are drafted in ways that meet revenue targets.⁶² Surveying congressional drafters, Professors Bressman and Gluck find that CBO scoring impacts are an oft-cited reason for statutory revisions and ambiguities, with one respondent indicating that “[i]n tax and spending programs you live and die by the score.”⁶³ Indeed, in recent years, the preeminence of the scores and the way they are treated when PAYGO rules are enforced has been blamed for the demise of a host of

BUDGET PROCESS 1 (2012).

59. While our focus is on CBO, the historical developments that shaped CBO’s role in the scorekeeping process applied in equal force to OMB. For example, GRH’s sequestration trigger depended upon the deficit estimates of both CBO and OMB.

60. CONG. BUDGET OFF., 53519, HOW CBO PREPARES COST ESTIMATES 1 (2018).

61. Jesse M. Cross & Abbe R. Gluck, *The Congressional Bureaucracy*, 168 U. PA. L. REV. 1541, 1578 (2020).

62. Bressman & Gluck, *supra* note 2, at 764 (quoting congressional drafters, “We have a number in advance and we work back and retrofit the policy to the score.”).

63. *Id.* at 763-64. Indeed, Professor Gluck further argues that CBO involvement in the legislative process is so consequential as to merit a statutory interpretive presumption to construe legislative ambiguities consistent with assumptions made in the underlying CBO score. Abbe Gluck, *The “CBO Canon” and the Debate over Tax Credits on Federally Operated Health Insurance Exchanges*, BALKINIZATION (July 10, 2012, at 20:55 ET), <https://perma.cc/BD5H-SS88>. Others have built upon this idea. *See, e.g.*, Clint Wallace, *Congressional Control of Tax Rulemaking*, 71 TAX L. REV. 179, 183-84, 225-30 (2017) (proposing a “JCT canon” that would resolve statutory ambiguities according to JCT assumptions).

progressive spending initiatives, including expanded veterans benefits⁶⁴ and robust Affordable Care Act subsidies.⁶⁵

The process scorekeepers use to form a cost estimate has steps that are easily enumerated, albeit difficult to execute. According to CBO, the process entails an attempt by scorekeepers to:⁶⁶

1. Understand legislative provisions by studying the language of the legislation.
2. Research potential effects from a variety of sources and outside experts.
3. Analyze and quantify effects.
4. Communicate results.

This process has several key weaknesses, not because scorekeepers are unable to conduct these steps or motivated by personal biases, but because the process itself contains biases that conceal hidden normative preferences inconsistent with progressive policy outcomes. First, the deficit estimates scorekeepers generate, which require presentation as precise point estimates, are laden with uncertainty; as a result, scorekeepers' estimates are viewed as objective truth when in reality they depend on value judgements in ways that are not appreciated. As former CBO director Rudolph Penner describes, scorekeeping has always been part art, part science, in that official government scores are provided by economists who develop rigorous analytical models reliant on a trove of government data.⁶⁷ Even with sophisticated technology and econometric methods, the unpredictability of outcomes years into the future make accurate cost estimation an impossible task.⁶⁸ Further, because laws

64. John B. Wells, *As it Applies to Veterans, it Is Time for Pay-Go to Go*, HILL (Jan. 24, 2019, at 11:30 ET), <https://perma.cc/XA9L-THEY>.

65. Josh Bivens, *The Bad Economics of PAYGO Swamp Any Strategic Gain From Adopting It*, ECON. POL'Y INST. (Dec. 18, 2018, at 10:43 ET), <https://perma.cc/YV9G-JSRD>.

66. CONG. BUDGET OFF., *supra* note 60, at 2. JCT follows a similar approach. *See Revenue Estimating*, JOINT COMM. ON TAX'N, <https://perma.cc/D5BH-TBGG> (archived Aug. 11, 2025). OMB and OTA play parallel roles in the executive branch. OMB describes scoring as "the process of estimating the change in Government spending and collections resulting from enacted or proposed legislation, compared to what would happen in the absence of that legislation." OFF. OF MGMT. & BUDGET, EXEC. OFF. OF THE PRESIDENT, CIRCULAR A-11: SECTION 21 – OVERVIEW OF SCORING LEGISLATION 1 (July 2024), <https://perma.cc/UAR2-PWPZ>. Meanwhile OTA provides revenue estimates for the revenue proposals in the president's budget. *Revenue Proposals*, U.S. DEP'T TREASURY (Mar. 2023), <https://perma.cc/9GD7-5NZL>. Importantly, the scorekeeping community views itself as closely related, in that at the staff level scorekeepers across agencies spend time discussing their estimates and methodological approaches with one another. Although deviations remain, they are relatively infrequent and their reasons generally understood.

67. Penner, *supra* note 9, at 16-19 (describing the pattern of errors in budget forecasting, the inevitability of such errors, and implications for Congress and the media).

68. *See* Kosterlitz, *supra* note 51, at 2408 (noting that in economic forecasting, technological sophistication is only "part of the battle," with the accuracy of estimates also

themselves tend to evolve from the moment in which their cost is estimated to the time that they are eventually enacted, “it’s practically impossible to compare past estimates with current realities.”⁶⁹ A related challenge is that statutes often require interpretation by agencies who write regulations that ultimately implement those directives, and those interpretive questions, which take time and agency expertise to answer, will factor into the ultimate cost and revenue estimates in ways that it is difficult for scorekeepers to accurately account for ex-ante.⁷⁰

Scorekeepers themselves have long discussed the uncertain nature of the scorekeeping enterprise.⁷¹ In her book *Inside OMB: Politics and Process in the President’s Budget Office*, Shelley Lynne Tomkin describes OMB staffers reporting that “spend-out rate estimates of how much an agency or program will spend over remaining parts of the fiscal year become a judgment call.”⁷² This is just one example of the numerous assumptions that must be made to estimate the impact of different policies.⁷³ Those assumptions are especially complex when one considers relatively novel policy proposals that do not have easy historical analogues from which to extrapolate from. Scorekeepers well understand this uncertainty, and they describe it in thinking about their work, noting that what

depending on the quality of data and assumptions that go into technical models).

69. *Id.* A recent example of this challenge is demonstrated in the difficulty of estimating the effects of green energy tax credits in the Inflation Reduction Act. Recent cost estimates of these provisions are significantly higher than estimates from when the IRA was enacted. *See* Bistline et al. *supra* note 15, at 79 (providing an estimate about three to four times higher than the official CBO and JCT score). This is in part because official scorekeepers prepared revenue estimates of changes in policy relative to these academic estimates that are a total cost number (inclusive of baseline policies and sensitive to the interaction between various new provisions that impact the same underlying activity). *See* STAFF OF THE JOINT COMM. ON TAX’N, FACTORS CONSIDERED WHEN ESTIMATING THE REVENUE EFFECTS OF THE ENERGY PROVISIONS OF PUBLIC LAW 117-169 AND SUBSEQUENT DEVELOPMENTS 16-20 (2023). But it is also because the uptake of some of the credits has been higher than expected due to accompanying regulatory changes governing electric vehicle production. Jim Tankersley, *Why the Cost of Biden’s Climate Law Keeps Going Up*, N.Y. TIMES (Feb. 8, 2024), <https://perma.cc/KR7U-BADS>.

70. *See supra* note 63.

71. Former CBO Director Doug Elmendorf has written about this uncertainty. *See* Douglas W. Elmendorf, “Dynamic Scoring”: *Why and How to Include Macroeconomic Effects in Budget Estimates for Legislative Proposals*, BROOKINGS PAPERS ON ECON. ACTIVITY, Fall 2015, at 95.

72. SHELLEY LYNNE TOMKIN, *INSIDE OMB: POLITICS AND PROCESS IN THE PRESIDENT’S BUDGET OFFICE* 155 (1998).

73. One assumption that is of particular consequence in the scorekeeping process is the baseline. Baselines serve as the anchor to policy changes that scorekeepers are scoring. Baselines themselves rely upon myriad assumption that are impossible to precisely know, such as the level of economic growth in the absence of a proposal. This uncertainty makes baselines ripe for manipulation in ways that artificially inflate or deflate the effect of policies measured against the baseline. For example, use of a “current policy” baseline that incorporates into the baseline policies that are due to expire under current law allows policy expirations to register as costless. *See* David Kamin & Rebecca Kysar, *Temporary Tax Laws and the Budget Baseline*, 157 TAX NOTES 125, 129 (2017).

they offer is often an “educated guess” with respect to reasonable outcomes from policy changes.⁷⁴ Or scorekeepers note that given a range of plausible assumptions, their approach is to use a reasonable middle-of-the-road assumption to help enable extrapolation.⁷⁵

Uncertainty with regards to scorekeeping assumptions mechanically leads to uncertainty in the estimates themselves. In economic modeling, when researchers try to understand the relationship between variables, they summarize the impact as a point estimate, a singular number, but acknowledge that it comes with a (potentially large) confidence interval, reflecting the fact that confounding factors render it impossible to know the precise impact of a policy change on an outcome of interest. The same is true in the budget scoring process, and the potential confounders are numerous—the baseline could be incorrectly chosen, the policy could impact more (or fewer) households than one anticipates based on program take-up, it could take time for agencies to deploy new resources that it receives, and so on. CBO acknowledges as much and notes that its process forces it to guesstimate a specific cost estimate—a singular number, rather than a range—because such precision is “necessary for the enforcement of budget rules during Congressional deliberations.”⁷⁶ The problem with this approach is that point estimates create a false sense of certainty, especially when they are amplified by lawmakers and public commentators and upheld as a final verdict on policy impact. This tendency of focusing overwhelmingly on a precise point estimate that scorekeepers view as the midpoint of plausible policy outcomes can make policymakers overconfident about parameters that are much more uncertain than they appreciate. Moreover, scorekeeping’s presentation of point estimates masks the myriad normative choices within what is viewed as a highly technical exercise.

The value of scorekeepers in the process of revenue estimation is not, then, that they are the ultimate arbiter of what the true revenue consequences of different investments or tax changes are going to be—any hope of arriving at truth is wholly unrealistic in a world that is so uncertain. But their benefit in the policy apparatus is that they are regarded, at least in recent history, as neutral arbiters who are able to provide estimates that are in the realm of what is reasonable without putting a thumb on the scale too aggressively for a preferred policy outcome the way an advocate may be tempted to do. The staying power

74. Kosterlitz, *supra* note 51, at 2411 (quoting one policy analyst: “You look at the social science literature to get some ideas, but if [the policy’s impact on behavior] has not been tested, then you have to make an educated guess.”).

75. See Philip Joyce, *The Congressional Budget Office at Middle Age 1* (Hutchins Ctr. on Fiscal & Monetary Pol’y at Brookings Inst., Working Paper No. 9, 2015), <https://perma.cc/FMN5-J6TV> (“CBO frequently stresses that its point estimates are the midpoint in a range, and that there is a great deal of uncertainty inherent in their projections.”). One JCT director, Tom Barthold, described JCT’s analysis as grounded in “middle-of-the-road economic work.” Brian Faller, *The Man Who Could Dash the GOP’s Tax Dreams*, POLITICO (June 1, 2016, at 17:26 EDT), <https://perma.cc/F74E-SWW3>.

76. CONG. BUDGET OFF., *supra* note 60, at 8.

of scorekeeping is very dependent on this perception of neutrality. Indeed, the emergence of congressional scorekeepers, which fundamentally changed the way in which the budget process evolved in this country, reshuffled power in large part because of a belief that the executive had become too political in its budgetary scorekeeping function in ways that diminished its perception as an honest broker in the system.⁷⁷ But regardless of whether scorekeepers today are truly neutral, the process they abide by is far from neutral, demanding scorekeepers to make value judgments without making clear to policymakers and outside observers the extent to which such judgments are not only made but also outcome determinative—a situation that creates strong status quo bias once initial judgments are made.

Turning an assumption-laden process like uncertain revenue estimation into one that produces a singular point estimate has normative implications. This process requires scorekeepers to make a choice about where to land on uncertain assumptions, and in doing so, it puts scorekeepers in a position to make powerful choices where there are inherently contestable value judgments at stake. Selecting among plausible assumptions is perceived to be a technical and neutral task, and it is this technicality that conceals biases of the revenue estimation process that have not historically been probed. This Article fills that void, exploring how different sets of assumptions result in drastically different revenue implications with regards to capital gains tax reform.

Another source of scorekeeping bias is the policy outcomes measured (and not measured) given a historical focus on deficits. Scorekeepers are most often instructed to measure variables with dollar signs—revenues, outlays, GDP. Scorekeepers at JCT also provide distributional estimates of certain tax proposals. But lacking contrary direction from policymakers, scorekeepers typically neglect other variables, including those that plausibly contribute to a policy's merits, or indeed motivate its adoption in the first place. Because scorekeepers have a reputation for producing revenue estimates, policymakers fail to request information about impacts on, say, child poverty, adult illiteracy, or carbon emissions. This does not merely produce an informational gap, but also biases traditional scorekeeping against policy outcomes other than deficit

77. At the time of CBO's creation, Nixon's Office of Management and Budget (OMB) was viewed as controlled by Nixon, who had politicized the budget process, making Congress suspect the accuracy of OMB cost estimates. Olivia B. Waxman, *This Is Why the Congressional Budget Office Was Created*, TIME (May 24, 2017, at 11:30 EDT), <https://perma.cc/3JWF-9TJE> ("In the 1970s, the Watergate scandal damaged the credibility of the executive branch, and there was a feeling on Capitol Hill that 'information coming out of [the Office of Management and Budget] couldn't be trusted because it was being controlled by Nixon,' as [historian] Joyce puts it."). A survey of OMB professionals at the time found that the increasingly prominent role of political officials in the Nixon OMB "downgraded the contribution made by career officials in OMB." Larry Berman, *OMB and the Hazards of Presidential Staff Work*, 38 PUB. ADMIN. REV. 520, 520-21 (1978). And Congress was well aware of OMB's newly politicized reputation, with congressional leaders charging the OMB with "go[ing] behelven to nobody but the President." *Id.* at 522 (quoting Congressman John Melcher of Montana).

reduction, including policies that would encourage a stronger safety net, ensure that fewer children are impoverished, and promote enhanced equity. That is, it biases in favor of low-cost policies and against high-cost ones, even if a social planner would prefer the latter because of the broader benefits that would result. Scorekeeping is essentially cost-benefit analysis without the benefit analysis. Normatively, then, the incompleteness of the scorekeeping exercise pushes in favor of smaller government as new government spending registers as cost without the commensurate benefit, a fact that this Article makes clear in examining analysis of the Child Tax Credit.

A third scorekeeping bias is the budget window. The production of a revenue score requires a timeframe within which revenues will be collected and boundaries delineating the effects that will be counted. This first encourages policymaking myopia, with attention paid to short-term effects within the budget window and blatant disregard of long-term effects. The budget window references the period of time that should bind spending and revenue decisions made in a budget resolution.⁷⁸ The window determines which effects are included in a score and which are not, depending on when effects occur. A ten-year budget window implies that policies with costs in year one and benefits in year eleven are viewed equivalently by the scorekeeping process as policies with costs in year one and no benefits whatsoever. As a result, policies that require monetary investments today that are expected to deliver long-run returns—investments, for example, in children, education, and the climate—emerge as key losers from the scorekeeping perspective. A private company considering an investment with long-run returns would not ignore those returns, but would estimate their value in present-day dollars—it would not shortsightedly rule out investing in a new product line simply because profits are not immediately anticipated. But in the federal government, spending decisions are front-loaded in ways that illogically fail to account for longer-term consequences; for example, investments in children are scored as increasing the deficit because of the expenditure required at the outset, but longer-term gains from those investments, in the form of incentivizing greater labor force participation or economic production for and eventually, by, those children, are not reflected. Similarly, in considering investments in tax reform, we will see how long-run benefits are shut out of consideration by a short-term budget window.

Furthermore, effects may be manipulated to fall within or outside of a budget window, allowing scorekeeping rules to dictate whether or not they will be

78. The length of the budget window has varied over time. Today, legislators work with a ten-year window, but that length has inched up over time. In the 1970s, when the modern system of federal budgeting was born, resolutions covered just one year, growing to three years in the 1980s. Beginning in the 1990s, windows frequently covered five years, but some were longer. For a summary overview of the federal budget window, see *What is the Budget Window?*, PETER G. PETERSON FOUND., <https://perma.cc/C2QC-GFZ2> (archived Aug. 11, 2025). The Senate rules have required assessment of budgetary effects in the first and second decade after the ten-year budget window. See S. Con. Res. 11, 114th Cong. § 3109 (2015) (enacted).

seriously considered. Even when conducted neutrally by strictly objective analysts, scorekeeping provides significant leeway with respect to the range of estimates that may be derived, and the timing of effects in the context of budget windows presents a prime example. This opportunity for manipulation was present when the budget rules were in full force given the interest in deficit reduction; it remains available today given the prominence of reconciliation and the scorekeeping rules that dictate its outcomes.⁷⁹ The budget window creates very clear incentives for legislators, and derivatively for scorekeepers, when thinking through policy design. The goal for those looking to game the impact on deficits is to create policies that generate revenue in the window and push expenditures outside the window.⁸⁰

A fourth and final scorekeeping bias that we will consider is that the ways in which scorekeepers are currently instructed to account for feedback effects of policies are incomplete, undercounting the benefits of policies that promote economic growth. In estimating the effects of policy changes, scorekeepers most intuitively account for the mechanical effects. An increase in the corporate tax rate, for example, mechanically increases the amount of revenue that is raised from corporate income. Scorekeepers also pay close attention to behavioral effects, or the ways in which policy changes may induce microeconomic responses in economic activity. So while an increase in the corporate rate would mechanically raise revenue, accounting for behavioral effects, it might decrease revenue if corporations respond to the policy change by reorganizing as pass-through entities that face a lower tax burden. This makes sense because policy clearly induces behavioral responses, and in some cases, it induces broader macroeconomic effects as well. An increase in the corporate tax rate, for instance, may reduce the overall amount of investment in the economy. But this

79. Reconciliation is a way to enact tax or spending legislation with a majority, avoiding the possibility of a legislative filibuster. Senate rules mandate that reconciliation cannot increase deficits beyond the period prescribed in the legislation (e.g., the ten-year budget window), which both places import on the deficit impact that scorekeepers measure and encourages congressional gamesmanship to avoid out-of-window deficit effects, for example through sunseting legislative provisions. David Wessel, *What is Reconciliation in Congress?*, BROOKINGS (Jan. 13, 2025), <https://perma.cc/7DKV-WJ4V>.

80. As Professor Allen Schick pointed out in congressional testimony, this is an extreme example of a more broad-based phenomenon:

[I]n some years, scoring has swallowed up the budget process. It would not be an overstatement to say that congressional budgeting has turned mostly into a scoring exercise. . . It has created a cottage industry on Capitol Hill and in the K street corridor on how you structure legislation to get the score that you want. Provisions are phased in or then sunseted, back-loaded, or front-loaded so that CBO produces the score that you want. This practice is highly damaging. I believe, to the legitimacy of congressional budget process. I don't have a solution for it, but basically we have empowered the scorers of the process, rather than the makers of policy.

Budgeting in the Congress: Reflections on How the Budget Process Functions: Hearing before the H. Comm. on the Budget, 109th Cong. 14 (2005) (statement of Allen Schick, Professor, Univ. of Md.).

last set of macroeconomic effects are relatively neglected by scorekeepers when generated by government spending.

Advocates of tax cuts have recognized the bias inherent in ignoring macroeconomic effects, and accordingly scorekeepers now account for these effects when analyzing corporate tax changes and other pieces of tax legislation. But advocates of government spending have failed to raise these same arguments, and scorekeepers generally do not incorporate the full suite of macroeconomic effects of spending legislation. This is the case even where spending policies very obviously influence macroeconomic variables, such as when high-skilled immigration policies increase the U.S. population and stimulate labor productivity. Consequently, the bias imposed by the scorekeeping process's failure to estimate macroeconomic feedback effects of all policies is also a normative bias, privileging policies that enhance economic growth through tax cuts over policies that enhance economic growth through a progressive government spending.

Related to these four biases is a practical limitation of scorekeeping as it is currently conducted. Official scorekeepers do not have the flexibility to estimate whatever policies under whichever assumptions they wish, unlike, say, think tanks or academics. Rather, official scorekeepers are beholden to the policymakers they serve—in the case of OMB and OTA, the president, and in the case of CBO and JCT, members of Congress. This means, that for the most part, official scorekeepers can only analyze the policies put in front of them by policymakers, and can only analyze them in ways requested by policymakers. This is the case even when policymakers lack the analytic expertise to understand how a given policy might most usefully be analyzed. In such cases, policymakers might reasonably be tempted to rely upon default analytic requests, namely revenue estimates, rather than tailor the set of outputs scorekeepers generate according to the policy being analyzed.

More generally, scorekeepers operate under rules devised by policymakers. Crucially, members of Congress specify the rules governing the scorekeeping process in budget resolutions. These rules include the length of the budget window and the extent to which macroeconomic and other feedback effects should be incorporated into a revenue estimate. Only in understanding these rules and the consequences of how the default rules currently operate can policymakers update the rules in ways that facilitate relevant information-gathering regarding for policies of interest. As we will see, some policymakers have begun to fashion new scorekeeping rules according to their normative preferences, but progressive policymakers seeking to expand government investment in long-term outcomes lag behind.

Collectively, these observations have many important implications for the practice of scorekeeping, but there are two conclusions that should *not* be drawn from scorekeeping's biases. First, to the extent there are deficiencies in the process, they belong to the process itself; not to the scorekeepers. While it is plausible that of the hundreds of scorekeepers involved in revenue estimation in

Washington, there are some (even many!) with particular normative biases that push in favor of some assumptions relative to others that tend to paint their preferred policies in a more favorable light, scorekeepers as a whole are regarded as providing objective analysis and praised for their nonpartisan estimation.⁸¹ We agree that scorekeepers are faithful to their mandate over any personal biases, but argue that the scorekeeping mandate itself is biased in ways that have not been appreciated, forcing scorekeepers to ascribe an unwarranted sense of certainty to their point estimates, concealing the underlying judgments made, and to restrict their outputs in ways that undercount the benefits of long-run investments and therefore push against progressive priorities.

Second, though the limitations of scorekeeping are considerable, they do not render scorekeeping a useless or impossible task.⁸² Importantly, though the scorekeeping's rules are tainted with historically derived biases, members of Congress can change the rules at their will, and they can even more easily request bespoke analyses from scorekeepers. Our call in this Article is for a more holistic approach to policy analysis, not for the abandonment of scorekeeping altogether.

It is worth noting that the challenges of scorekeeping are not contained within the policy context, but are inherent difficulties of measurement more generally. A long literature across a range of disciplines, including natural

81. A senior fellow at the conservative American Enterprise Institute has noted that “CBO’s mission is critical. It produces independent and nonpartisan analyses of budget and economic policies and estimates the fiscal impact of legislation considered by Congress.” Alex Brill, *Give the CBO Long-Range Tools*, AM. ENTER. INST. (Apr. 16, 2014), <https://perma.cc/6BRL-XDQC>. A report from the left-leaning Center for American Progress similarly commented: “The CBO’s reputation for providing nonpartisan research and technical assistance to legislators is unparalleled, and many experts credit the CBO with helping protect checks and balances and legislative independence.” SAM BERGER, SETH HANLON & GALEN HENDRICKS, CTR. FOR AM. PROGRESS, REFLECTIONS ON THE CONGRESSIONAL BUDGET ACT: LEARNING FROM PREVIOUS BUDGET REFORM EFFORTS 4 (2018), <https://perma.cc/ND44-WWBU>.

82. The argument that it is fundamentally impossible to succinctly estimate how a given policy will affect the economy and that such estimation should therefore not be attempted is well-articulated in the literature critiquing the government’s use of cost-benefit analysis, which notes that while costs tend to be easily ascertainable, benefits are often intangible, making them difficult to quantify and monetize and causing cost-benefit analysis to inherently understate net benefits or favor the rich by quantifying benefits through “willingness-to-pay”. See, e.g., Zachary D. Liscow, *Is Efficiency Biased?*, 85 U. CHI. L. REV. 1649, 1650-51 (2018) (noting inequitable biases of investment decisions made through efficiency maximization measured according to willingness-to-pay); Kwangseon Hwang, *Cost-Benefit Analysis: Its Usage and Critiques*, 16 PUB. AFFS. 75, 75 (2016) (critiquing cost-benefit analysis on the grounds that “numbers don’t tell us everything” and many things are unquantifiable); Frank Ackerman & Lisa Heinzerling, *Pricing the Priceless: Cost-Benefit Analysis of Environmental Protection*, 150 U. PA. L. REV. 1553, 1553-56 (2002); (noting the difficulty of using cost-benefit analysis to quantify the benefits of environmental protection); Nathanan Aldred, *Incommensurability and Monetary Valuation*, 82 LAND ECON. 141, 142-45 (2006) (analyzing critiques of cost-benefit analysis that describe environmental benefits as monetarily incommensurable); Robert H. Frank, *Why Is Cost-Benefit Analysis So Controversial?*, 29 J. LEGAL STUD. 913, 913-15 (listing the “incommensurability problem” among the reasons cost-benefit analysis is so controversial).

science,⁸³ sociology,⁸⁴ and statistics⁸⁵ articulate these challenges, discussing the ways in which what is able to be measured, what is actually measured, and how measurement is conducted biases the decisions made and results inferred from such measurement. In short, measurement can be difficult, and distillation of a measurement into a single number obscures the challenges, assumptions, and uncertainty that pervade that number. But while other disciplines have probed their measurement biases, scorekeeping has not, and largely continues to be viewed as an entirely objective endeavor, free from normative leanings.

We are not the first to push back on scorekeeping's perceived objectivity, as others have observed how the scorekeeping process is deficient in analyzing particular policies. Many, including one of us, have previously written about the

83. Historical examination of scientific disciplines exposes the inherent biases of seemingly "objective" empiricism. An early observation in this area is offered by Professor Kuhn, who argued that there is always a discrepancy between theoretical predictions and empirical observations, necessitating scientists to seek "reasonable agreement" in comparing theoretical and experimental outcomes despite a lack of consensus in how to distinguish a failed theory from mere measurement error. Thomas S. Kuhn, *The Function of Measurement in Modern Physical Science*, 52 *ISIS* 161, 165-66 (1961). Professor Schaffer focuses on the way in which context influences the production and meaning of measurement in analyzing the biases introduced by 19th century astronomers seeking to correct human error with measurement devices calibrated according to certain standardizations; he concludes that "[m]easurement is given its meaning when situated in specific contexts of styles of work and institutions. Disciplines give meanings to values, and often resist attempts by others to redefine these meanings or to gain authority over measurement. Quantification is not a self-evident nor inevitable process in science's history, but possesses a remarkable cultural history of its own." Simon Schaffer, *Astronomers Mark Time: Discipline and the Personal Equation*, 2 *SCI. IN CONTEXT* 115, 115-18 (2008). Professor Porter similarly notes that the bureaucratic need for standardization imposes a certain arbitrariness in scientific measurement, pointing for example to the use of one and five percent significance levels. Theodore M. Porter, *Quantification and the Accounting Ideal in Science*, 22 *SOC. STUD. OF SCI.* 633, 645 (1992).

84. Sociological analyses of measurement also raise caution against overreliance on measurement and quantification. See, e.g., Steve Woolgar, *Beyond the Citation Debate: Towards a Sociology of Measurement Technologies and Their Use in Science Policy*, 18 *SCI. & PUB. POL'Y* 319, 321-23 (1991) (exposing a key weakness of the use of citation counts to measure the impact of scholarly work—that the quality of scholarly work is subjective and impossible to be wholly captured through citation analysis) and Linda Derksen, *Towards a Sociology of Measurement: The Meaning of Measurement Error in the Case of DNA Profiling*, 30 *SOC. STUD. OF SCI.* 803, 806 (arguing that the inevitable existence of measurement error contributes to the process whereby subjective representations, such as those produced by DNA typing, come to be viewed as objective: "The standard error of measurement is a number that stands for variations arising from diverse (often unknown cultural, practical and environmental sources, and yet the number itself gives precise, numerical form to an error term that incorporates subjective judgments and supports the stability and credibility of 'objective' forensic results.").

85. The statistics field has long been familiar with the concept of measurement bias. See, e.g., Roger E. Millsap & Howard T. Everson, *Methodology Review: Statistical Approaches for Assessing Measurement Bias*, 17 *APPLIED PSYCH. MEASUREMENT* 297, 297 (1993) (reviewing the statistical methods for detecting measurement bias, defined as "systematic inaccuracy of measurement" with regards to the functioning of a test as applied to groups that have been matched on the relevant attribute being tested).

limitations of scorekeeping including in some of the policy areas we outline below.⁸⁶ Our contribution here is a broader frame for understanding these scorekeeping biases, arising not on a case-by-case basis out of unique policy formulations and circumstances, but as a universal byproduct of a normatively biased analytic lens. The result of these biases is that policy outcomes are systematically skewed in favor of a more stagnant, more myopic, and ultimately smaller federal government. Scorekeeping has advanced these biases in ways academics and policymakers—even those who would normatively prefer a larger and more proactive government—have failed to grapple with meaningfully and therefore perpetuate. This Article seeks to rectify that failure. We next explore specific examples that highlight these key biases of the scorekeeping process in practice by considering four critical policy areas: capital gains taxation, tax administration, the Child Tax Credit, and high-skilled immigration. In doing so, we suggest ways that policymakers can extract more useful information from the scorekeeping process, as well as how external entities can supply this information to the public when members of Congress fail to secure it.

B. The Scorekeeping Process' Failure to Grapple with Uncertainty: The Case of Capital Gains

As we have seen, scorekeeping estimates are typically accompanied with large implicit confidence intervals, but are nonetheless required to be presented as singular point estimates that conceal the many value judgments that must be made to arrive at that number. But scorekeeping is not viewed as an exercise in making value judgements, it is viewed as a neutral and technical enterprise. The results scorekeepers produce are not viewed as reliant upon uncertain assumptions, but as scientific and objective truth. One way to appreciate this is to consider scorekeepers' approach to estimating the revenue impact of capital gains tax reform.

The appropriate approach to capital gains taxation is of perpetual interest in tax policy. Current tax law contains a host of preferences that advantage capital relative to wage income, and capital income is extremely concentrated at the top: wages and salaries taxed at ordinary rates make up less than 10 percent of the income of the top 0.001 percent, but capital gains make up more than 70 percent of income of the top 0.001 percent annually.⁸⁷ Our approach to capital taxation is currently structured with a significant tax avoidance opportunity: when individuals pass down capital assets to their heirs, the basis "steps up" such that tax liability owed on appreciated gains during one's lifetime is fully wiped out

86. See Sarin & Mazur, *supra* note 14, at 13-14 (challenging the scorekeeping process's failure to take into account many of the indirect revenue effects associated with increased tax enforcement).

87. Lily L. Batchelder & David Kamin, Taxing the Rich: Issues and Options 4 (Sep. 11, 2019) (unpublished manuscript) (on file at <https://perma.cc/8LJM-5DZX>).

upon the death of the donor.⁸⁸ To take a precise example, if Mark Zuckerberg saw his stake of Meta rise from \$1 per share to \$300 per share and chose to sell a share, the share would be subject to capital gains taxes on \$299 of gains. If he instead holds that share and passes it to his child upon his death, the share's basis will readjust, so that the gains that accumulated during Zuckerberg's lifetime will never be taxed.

This disparate tax treatment is of course inefficient in addition to being inequitable, since it distorts the ways in which capital owners deploy assets during their lifetime.⁸⁹ For example, entrepreneurs can choose to bequeath ownership stakes to family members, not because these family members will be the most productive future owners, but to avoid tax liability on their shareholdings.

Raising revenue by reforming the treatment of capital income has been a focus of policymakers and tax experts alike.⁹⁰ That is unsurprising, given how significant a potential pot of revenue untaxed capital gains represent: forty percent of the wealth of the top one percent takes the form of unrealized capital gains.⁹¹

There are myriad approaches to reforming the taxation of capital gains, which range from raising capital gains tax rates to exploring options for constructive realization of unrealized gains. These approaches are often compared based on scorekeepers' assessments of their revenue potential. A key focus for scorekeepers and outside experts has been identifying the "revenue-maximizing" capital gains rate.⁹²

The intuition is that setting a capital gains rate above the revenue maximizing rate would move policymakers on the wrong side of the so-called Laffer curve, such that this higher rate (relative to the revenue maximizing rate)

88. I.R.C. § 1014(e)(B).

89. See KIMBERLY A. CLAUSING & NATASHA SARIN, HAMILTON PROJECT, *THE COMING FISCAL CLIFF: A BLUEPRINT FOR TAX REFORM IN 2025*, at 14 (2023) (describing the "lock-in" problem that results from capital owners conscientiously avoiding a realization event).

90. See, e.g., U.S. DEP'T TREASURY, *GENERAL EXPLANATION OF THE ADMINISTRATION'S FISCAL YEAR 2010 REVENUE PROPOSALS 77* (2009), <https://perma.cc/MD35-M9T9> (detailing the Obama Administration's proposals to increase the top tax rate on capital gains and dividends by five percentage points); Natasha Sarin, Lawrence H. Summers & Joe Kupferberg, *Tax Reform for Progressivity: A Pragmatic Approach*, in *TACKLING THE TAX CODE: EFFICIENT AND EQUITABLE WAYS TO RAISE REVENUE* 317, 334-39 (Jay Shambaugh & Ryan Nunn eds., 2020) [hereinafter Sarin et al., *Tax Reform for Progressivity*] (proposing an "overhaul" in the taxation of capital gains by raising tax rates and eliminating step-up in basis, among other policy changes); David Kamin, *How to Tax the Rich*, 146 *TAX NOTES* 119, 124 (2015) (proposing to treat bequest or gift as a realization event); Clausing & Sarin, *supra* note 89, at 14-16 (proposing carryover basis for capital gains, increasing tax rates on capital gains, and eliminating the carried interest loophole).

91. Jesse Bricker et al., *Wealth and Income Concentration in the SCF: 1989-2019*, FED. RES. V.: FEDS NOTES (Sep. 28, 2020), <https://perma.cc/V8JE-K7M9>.

92. External voices include nonprofits that proffer revenue estimates like the Tax Policy Center, the Tax Foundation, and the Penn Wharton Budget Model.

results in less tax revenue because of how taxpayers choose to adjust their capital gains realizations when faced with a higher tax rate.⁹³ Given the opportunity to wait until death and defer capital gains taxes entirely, taxpayers will be more incentivized to wait and hold in a world with higher tax rates.

This “elasticity” of capital gains realizations to the capital gains rate is an extremely important parameter in capital gains tax reform scoring, because this is the parameter that dictates how estimators believe that taxpayers will adjust realizations in the face of tax rate changes. The current “realization elasticity” used by JCT is approximately -0.7.⁹⁴

The implications of this elasticity assumption are succinctly summarized in prior work as follows: “If tax rates increased by 100%, a crude application of this elasticity implies that realizations would fall by 70%. In concrete terms, roughly \$1.25 trillion of annual realizations would shrink to around \$375 billion due to an increase in capital gains tax rates from 20% to 40%.”⁹⁵

Clearly this elasticity is going to bear heavily on the revenue estimate that scorekeepers will adopt for a proposed increase in capital gains tax rates: the fact that higher capital gains rates discourage realizations means that higher rates will decrease the capital gains tax base and, if the rate is set above the revenue-maximizing rate, these behavioral effects could fully overwhelm any gains from higher tax rates.

Importantly, an elasticity of -0.7 is quite large, as it suggests that the tax base has the capacity to shrink by seventy percent in the face of a 100% increase in capital gains rates. Equally importantly, this elasticity assumption is an estimate, one that is vigorously debated among economists. JCT economists themselves do not believe their elasticity is an incontestable and unchanging fact, and it is not one on which there are uniform views among the scorekeeping community. Indeed, scorekeepers themselves are regularly engaged in ongoing research regarding the elasticity of capital gains to the tax rate. For example, Tim Dowd, with Robert McClelland and Athiphat Muthitacharoen, estimates an elasticity of -0.72 using a panel of tax returns from 1999 to 2008.⁹⁶ But an update to this

93. The Laffer curve graphically illustrates the relationship between tax rates and revenue that the government collects. Rates so high that they are on the “wrong side” of the Laffer curve lose revenue rather than raise it, for example because extreme tax rates on labor income will push individuals to choose untaxed leisure rather than labor. *See generally* ARTHUR B. LAFFER, HERITAGE FOUND., *THE LAFFER CURVE: PAST, PRESENT, AND FUTURE* (2004), <https://perma.cc/4MAR-TFC5> (describing the economic theory underlying the Laffer curve).

94. JOINT COMM. TAX’N, JCX-42-21, ESTIMATED BUDGETARY EFFECTS OF AN AMENDMENT IN THE NATURE OF A SUBSTITUTE TO THE REVENUE PROVISIONS OF SUBTITLES F, G, H, I, AND J OF THE BUDGET RECONCILIATION LEGISLATIVE RECOMMENDATIONS RELATING TO INFRASTRUCTURE FINANCE AND COMMUNITY DEVELOPMENT, GREEN ENERGY, SOCIAL SAFETY NET, RESPONSIBLY FUNDING OUR PRIORITIES, AND DRUG PRICING, SCHEDULED FOR MARKUP BY THE COMMITTEE ON WAYS AND MEANS ON SEPTEMBER 14, 2021, at 1-12 (2021).

95. Sarin et al., *Rethinking*, *supra* note 14, at 2.

96. Tim Dowd, Robert McClelland & Athiphat Muthitacharoen, *New Evidence on the Tax Elasticity of Capital Gains*, 68 NAT’L TAX J. 511 (2015).

methodology by Tim Dowd and Robert McClelland extending the time frame over which taxpayers may respond to a tax change results in a preferred elasticity of -0.6.⁹⁷ In another paper, Tim Dowd and Robert McClelland study capital gains realizations in 2012, estimating the effect of an anticipated increase in the long-term capital gains rate to produce a short-term elasticity of -0.47 and long-term elasticity of -0.79.⁹⁸

Meanwhile, past work that we quote above by economists Natasha Sarin, Larry Summers, Owen Zidar, and Eric Zwick has demonstrated that many existing estimates of this elasticity are likely to be overstated because they fail to reflect the changing landscape of capital gains: for example, in recent years, more than half of capital gains have accrued through pass-through distributions that cannot be timed by taxpayers.⁹⁹ Further, past studies of the sensitivity of capital gains realizations to tax rates on which the JCT estimate relies focus on the short-run impact on realizations, but fail to recognize that realizations that do not happen in year one when rates rise are likely to occur in the medium-term rather than be deferred indefinitely.¹⁰⁰

These conflicting elasticities illustrate the status of research in this area—contested and incomplete, with more unknowns than knowns and much work left to be done. They also demonstrate how the elasticity assumption is itself dependent on conditions embedded in the underlying context, conditions that are subject to change and human inconsistency.¹⁰¹ This Article does not seek to settle truth in the search for the true elasticity, but to highlight the range of estimates that exist and the ultimate impossibility of knowing for certain, in any given moment in time or context, which one is right. Even so, revenue estimators must choose an elasticity, and this choice produces monumental effects in the resulting revenue estimate.

To see the wide range of uncertainty associated with a proposal to raise capital gains tax rates, it is helpful to consider the implications of different realization elasticities. To do so, we follow economist Jane Gravelle, who synthesizes the elasticities from the literature based on studies conducted after

97. TIM DOWD & ROBERT MCCLELLAND, URB. INST. & BROOKINGS INST.: TAX POL'Y CTR., THE SENSITIVITY OF THE TAX ELASTICITY OF CAPITAL GAINS TO LAGGED TAX RATES AND MIGRATION 7-8 (2024), <https://perma.cc/9ZDH-5953>.

98. Tim Dowd & Robert McClelland, *The Bunching of Capital Gains Realizations*, 72 NAT'L TAX J. 323, 346-47 (2019).

99. Sarin et al., *Rethinking*, *supra* note 14, at 17.

100. *Id.* at 1 (“[M]any prior studies focus primarily on short-run taxpayer responses, and so miss revenue from gains that are deferred when rates change.”). *But see* Tim Dowd & Zach Richards, Contextualizing Elasticities for Policymaking: Capital Gains and Revenue-Maximizing Tax Rates, at 4-5, 15-18 (Jan. 15, 2021) (unpublished manuscript) (on file at <https://perma.cc/RK36-XXQ5>) (arguing that elasticity estimates derived in particular settings may lack external validity in generating an overall revenue maximizing rate).

101. Again, JCT itself acknowledges this fact. Its use of the elasticity parameter varies depending upon asset type, time horizon of tax reform proposal, and other context-specific facts.

the 1980s.¹⁰²

We apply this wide range of elasticities to a policy reform recently scored by JCT: raising the top tax rate on long-term capital gains and qualified dividends to 25%, which scored as generating \$123.4 billion of new tax revenue over the course of the decade.¹⁰³ Only altering the elasticity parameter, with no other changes to assumptions that are relevant to the ultimate score (e.g., the trajectory of capital gains as a share of gross domestic product) shows that the revenue associated with this capital gains increase could range from a *gain* of \$557.5 billion to a *loss* of \$1.1 billion.

Elasticity	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	10-year budget window
0	39.7	46.6	50.2	53.3	55.5	57.6	59.9	62.3	64.9	67.6	557.5
0.22	28.8	34.5	36.8	39.1	40.6	42.2	43.9	45.7	47.7	49.6	409.1
0.25	27.4	32.9	35.1	37.2	38.7	40.2	41.8	43.6	45.4	47.3	389.5
0.31	24.6	29.7	31.6	33.5	34.8	36.2	37.7	39.2	40.9	42.6	350.8
0.39	20.8	25.5	27.1	28.7	29.7	31.0	32.2	33.6	35.0	36.5	300.1
0.47	17.2	21.5	22.6	24.0	24.8	25.8	26.9	28.0	29.2	30.4	250.3
0.5	15.8	19.9	21.0	22.2	22.9	23.9	24.9	25.9	27.1	28.2	231.9
0.55	13.6	17.5	18.2	19.3	19.9	20.7	21.6	22.6	23.5	24.5	201.4
0.64	9.6	13.0	13.4	14.2	14.5	15.2	15.8	16.5	17.3	18.0	147.7
0.68	7.9	11.1	11.3	11.9	12.2	12.8	13.3	13.9	14.5	15.2	124.2
0.72	6.2	9.2	9.2	9.7	9.9	10.3	10.8	11.3	11.8	12.4	100.9
0.75	4.9	7.8	7.7	8.1	8.2	8.6	9.0	9.4	9.8	10.3	83.6
0.9	-1.3	0.8	0.1	0.0	-0.3	-0.2	-0.1	-0.1	0.0	0.1	-1.1

Table 1: Annual Revenue Gains from Raising the Long-Term Capital Gains Rate to 25% Under Different Elasticity Assumptions (Billions of Dollars)

102. JANE G. GRAVELLE, CONG. RSCH. SERV., R41364, CAPITAL GAINS TAX OPTIONS: BEHAVIORAL RESPONSES AND REVENUES (2021). The estimates in Gravelle's review cover the timespan from 1991 to 2020, though it is worth noting that more recent estimates may be given more weight given as more reflective of the current capital gains landscape. See Sarin et al., *Rethinking*, *supra* note 14, at 1 (describing the current landscape and ways in which it has shifted relative to earlier assumptions).

103. JOINT COMM. ON TAX'N, *supra* note 94, at 7.

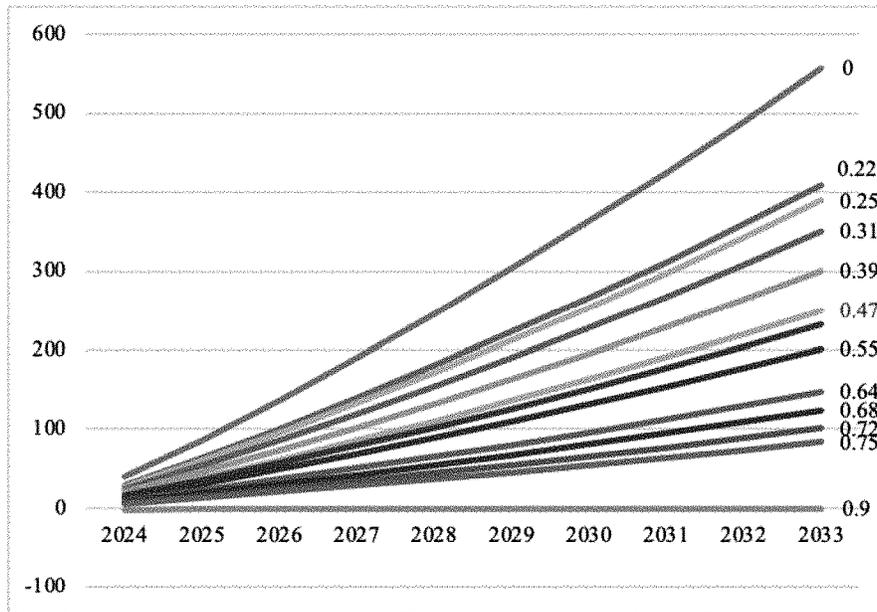


Figure 4: Cumulative Revenue Gains from Raising the Long-Term Capital Gains Rate to 25% Under Different Elasticity Assumptions (Billions of Dollars)

Source: Revenue estimates are calculated using the Budget Lab’s tax microsimulation model. Estimating a revenue change from a change in the tax rate does not merely require applying the new rate to current capital gains realizations, as the level of realizations will change in response to the new rate according to the elasticity of realizations to the capital gains tax rate. To simulate filer-level change in realizations, we first calculate effective marginal tax rates (EMTRs) on long-term capital gains under current law (τ_b) and the counterfactual policy scenario (τ_p). We then calculate the percent change in realizations by applying the assumed elasticity β to the change in EMTR under the following functional form assumption: $1 - e^{\frac{\beta}{0.238}(\tau_p - \tau_b)}$.

Is such wide uncertainty edifying or ultimately destructive to the apparatus of scorekeeping, in that it makes tangible such a wide range of outcomes? Certainly, scorekeepers have made clear that this type of sunlight is not necessarily a helpful disinfectant—as they worry that arming policymakers with a broad range allows legislators to “pick-your-own-adventure” with respect to the elasticities that will most support policy priors.¹⁰⁴ On the other hand,

104. Former CBO director Doug Elmendorf, for example, has observed:

“The main reason that CBO usually does not provide ranges of estimates is that we are still developing ways to help legislators make effective use of our quantification of uncertainty. Part of the challenge is that providing ranges for estimates sometimes muddies, rather than enhances, general understanding of our analysis. For example, when we report ranges, people

acknowledgement of uncertainty can help policymakers draft proposals in such a way that they are most likely to achieve certain outcomes. For example, the incentives to lock in capital gains are heightened in a world where waiting to pass them on to a decedent allows one to escape tax liability; in a world where tax bases carry over, those behavioral concerns are mitigated.

Furthermore, providing a range of estimates in lieu of a point estimate helps policymakers and the public understand how sensitive estimates are to their underlying assumptions. Ideally, this enables scorekeepers to provide relevant information to the policymaking process without redirecting all policy debate to the score. When people understand that a revenue estimate is just an estimate, they might become more attuned to other, potentially more relevant analytic inputs. We next turn to some of these alternative inputs.

C. The Scorekeeping Process' Failure To Consider Benefits: The Case Of The Child Tax Credit

The Child Tax Credit was legislated through the Taxpayer Relief Act of 1997 as a means of helping families cover the cost of raising children.¹⁰⁵ There is a wealth of academic evidence that points to the CTC's positive impact on children, both in early childhood and later life—and also on parents' outcomes. And yet, scorekeepers register the CTC as a net cost to the government, not a benefit, because the benefits of the CTC appear outside the budget window in variables scorekeepers don't measure.

Since its enactment, the CTC was expanded many times with bipartisan support.¹⁰⁶ One recent Child Tax Credit expansion was enacted in the 2017 Tax Cuts and Jobs Act (TCJA), which temporarily doubled the maximum Child Tax Credit per child from \$1,000 to \$2,000, (this increase is set to expire in 2025).¹⁰⁷ For the lowest-income families, the credit exceeds the total amount of taxes owed: under the TCJA, the CTC is “partially refundable,” so taxpayers can receive up to \$1,400 of the credit as a refund.¹⁰⁸ JCT scored this expansion as

who would prefer that our estimate be smaller tend to cite the bottom of the range, and people who would prefer that our estimate be larger tend to cite the top of the range. That can make the public discussion of our analysis quite confusing, and we have a limited ability to clear up that uncertainty.”

Doug Elmendorf, *Communicating the Uncertainty of CBO's Estimates*, CONG. BUDGET OFF. (Dec. 15, 2014), <https://perma.cc/E5KV-KQDW>. A related concern is that admitting to a large range of estimates undermines confidence in the underlying policy. In the words of then-Minority Counsel of the House Ways and Means Committee Charles N. Kahn III, “uncertainty gives opponents the upper hand.” Kosterlitz, *supra* note 51, at 2413.

105. See MARGOT L. CRANDALL-HOLLIICK, CONG. RSCH. SERV., R45124, *THE CHILD TAX CREDIT: LEGISLATIVE HISTORY 2* (2018) (connecting the motivation behind the Child Tax Credit to the rising costs of raising children, in part due to increased tax burdens).

106. For a comprehensive history of the Child Tax Credit (CTC), see *id.*

107. Tax Cuts and Jobs Act of 2017, Pub. L. No. 115-97, 131 Stat. 2073.

108. For an explanation of how partial refundability works, see, *How did the TCJA*

costing \$570 billion in revenue.¹⁰⁹

The American Rescue Plan (ARP) built on the TCJA to further expand the CTC. First, it increased the maximum credit amount to \$3,600 per child for children under age six and to \$3,000 per child for children under age six. Second, it expanded the universe of children covered, including seventeen-year-olds for the first time. And third, it made the credit fully refundable, so that even families whose tax liability was below the credit amount that they were eligible for would receive the full amount of the credit.¹¹⁰ The credit was also newly advanced, with monthly payments disbursed throughout the year as opposed to a one-time tax refund.¹¹¹ The ARP's changes to the CTC's structure were sunset after a year: they were in place only for 2021, with the structure reverting back to the TCJA's in 2022.¹¹² The cost of this expansion as estimated by JCT was \$110 billion.¹¹³

The ARP's expansion had an immediate and historic impact on child poverty: the Census Bureau estimated that following the CTC expansion, child poverty declined by forty-six percent to a record low of five percent in 2021.¹¹⁴ Ninety percent of this reduction was attributed to the CTC expansion legislated in the ARP.¹¹⁵ Full refundability was especially important because it enabled the expanded credit to reach the lowest-earning families who previously had received only a portion of the credit. The Center on Poverty and Social Policy at Columbia University noted that the ARP's expansion reduced child poverty, food insufficiency, and financial hardship, with Black and Latino children at the bottom of the income distribution especially likely to have benefitted from the policy.¹¹⁶

Given its success, President Biden included in his Build Back Better framework a proposal to perpetuate permanently the ARP's changes to the

change taxes of families with children?, TAX POL'Y CTR. (Jan. 2024), <https://perma.cc/4KLH-PFUH>.

109. JOINT COMM. ON TAX'N, JCX-67-17, ESTIMATED BUDGET EFFECTS OF THE CONFERENCE AGREEMENT FOR H.R. 1, THE "TAX CUTS AND JOBS ACT" 1 (2017).

110. American Rescue Plan Act of 2021, Pub. L. No. 117-2, 135 Stat. 4.

111. *Id.* at 146.

112. *Id.* at 144.

113. JOINT COMM. ON TAX'N, JCX-67-17, ESTIMATED REVENUE EFFECTS OF THE CONFERENCE AGREEMENT FOR H.R. 1319, THE "AMERICAN RESCUE PLAN ACT OF 2021," AS AMENDED BY THE SENATE, SCHEDULED FOR CONSIDERATION BY THE HOUSE OF REPRESENTATIVES I (2021).

114. Kalee Burns, Liana Fox & Danielle Wilson, *Expansions to Child Tax Credit Contributed to 46% Decline in Child Poverty Since 2020*, U.S. CENSUS BUREAU (Sept. 13, 2022), <https://perma.cc/HL4Z-VAG9>.

115. *Id.*

116. Megan A. Curran, *Research Roundup of the Expanded Child Tax Credit: One Year On*, 6 POVERTY & SOC. POL'Y REP., at 5, 22, 24, 27, 45-46 (2022), <https://perma.cc/Z82F-WGTV> ("Children of color stood to see even greater gains: child poverty among Black children could be cut by more than fifty percent in a pre-pandemic context; among Latino children by more than 45 percent; and among Native American children by more than 60 percent.").

credit.¹¹⁷ Critics of this proposal turned to scorekeepers to pass judgement on its economic effects: then-ranking member of the House Ways and Means Committee Kevin Brady commissioned a revenue analysis from the Joint Committee on Taxation to consider the effects of a permanent extension of the ARPA CTC.¹¹⁸

The verdict from the scorekeepers was that such a policy change would carry a significant price tag: JCT produced a “conventional” revenue estimate (assuming no impact of the policy change on macroeconomic aggregates) and assigned it a \$1.25 trillion ten-year cost.¹¹⁹

JCT also offered a version of this estimate that counted macroeconomic effects, but only some, concluding that the policy would decrease the average annual growth rate of GDP by 0.04 percentage points, costing an additional \$113 billion over the ten-year budget window.¹²⁰ JCT explains its conclusion as follows: “Aggregate labor supply falls primarily due to a positive income effect from the increase in the value of the credit, and a negative substitution effect from the reduction in the after-tax marginal return to labor.”¹²¹ Said another way, when households have access to additional dollars, textbook economic models suggest that they will choose leisure and consumption rather than working additional hours.

But missing from this analysis is any consideration of broader economic effects that do not come from this one margin (the tradeoff between leisure and work). Specifically, if parents are more attached to the labor force because they have additional resources to expend towards paying for childcare, or if there are any economic benefits from a reduction in child poverty, those are entirely missing from JCT’s narrow analysis. JCT is explicit about this omission,¹²² but noting this clear limitation of their estimates did not prevent critics of the credit’s expansion from claiming that it would be both “costly in fiscal terms and harmful to the overall economy.”¹²³

Economists who have built models to consider the cost of child poverty

117. *The Build Back Better Framework: President Biden’s Plan to Rebuild the Middle Class*, WHITE HOUSE, <https://perma.cc/4JX3-28DE> (archived Aug. 11, 2025).

118. JOINT COMM. TAX’N, *MACROECONOMIC ANALYSIS OF A PERMANENT CHILD TAX CREDIT EXPANSION* (2022).

119. *Id.* at 7.

120. *Id.* at 1, 7.

121. *Id.* at 5.

122. *Id.* at 1 (“Because the macroeconomic models used have been developed with an emphasis on tax policy, some outcomes of public policy decisions are not modeled. For example, this ignores potential human capital losses from parents leaving the workforce and any potential long-run benefits from a reduction in child poverty.”). *But see* Rachel Moore & Brandon Becoraro, *A Tale of Two Bases: Progressive Income Taxation of Capital and Labor Income*, 49 PUB. FIN. REV. 335, 377 (2021) (describing JCT’s Overlapping Generations Model, which does incorporate childcare costs).

123. Alex Brill, *Macroeconomics of the Child Tax Credit*, AM. ENTER. INST. (Nov. 10, 2022), <https://perma.cc/234S-TA4T>.

supply much information in place of JCT. In important work, economists Harry J. Holzer, Diane Whitmore Schanzenbach, Greg J. Duncan, and Jens Ludwig concluded that child poverty “costs” the U.S. economy nearly four percent of GDP annually, with a reduction in productivity, the costs of crime, and the costs of added health expenditures as roughly equal contributors.¹²⁴ Their estimates seek to establish a link between poverty among children and adult behavior that generates economic costs. Importantly, then, these estimates are both under-inclusive, in that they do not capture any short-run impact on parents’ labor force attachment when the government provides more generous child-rearing support, and over-inclusive relative to traditional budget scorekeeping norms, in that they speak to long-run impacts on later-life outcomes when children are lifted out of poverty.

The authors point out that they are not able to disentangle the causal effects of low parental incomes from the broader “range of environmental factors associated with poverty in the United States,” which include “all of the personal characteristics imparted by parents, schools, and neighborhoods to children who grow up with or in them.”¹²⁵ That mingling means that the policy levers that are appropriate to address child poverty are invariably broader than an expanded Child Tax Credit that provides income support to more families. It does not, however, mean that this tool is an impotent one, and because the problem of child poverty has such large economic consequences, steps like an expanded CTC are likely to be even more impactful. In fact, these authors conclude that their huge economic estimate is, if anything “almost certainly [an] underestimate [of] the true costs associated with poverty.”¹²⁶

JCT, unlike CBO, does provide broad distributional analysis of the impact of tax policy changes, although it did not do so in its reporting on the CTC’s macroeconomic effects. If it had, it would have found it hard to identify a more progressive social safety net enhancement: in Table 2, we illustrate the progressivity of the CTC’s benefit over a five-year time frame.

124. Harry J. Holzer, Diane Whitmore Schanzenbach, Greg J. Duncan & Jens Ludwig, *The Economic Costs of Childhood Poverty in the United States*, 14 J. CHILD. POVERTY 41, 51-52 (2008).

125. *Id.* at 44.

126. *Id.* at 45. This study is just one of many to point out the significant deleterious direct and indirect economic effects of child poverty. *See also* Jeanne Brooks-Gunn & Greg J. Duncan, *The Effects of Poverty on Children*, 7 FUTURE CHILD. 55, 55, 58-9 (1997); Greg J. Duncan, Kathleen M. Ziol-Guest & Ariel Kalil, *Early-Childhood Poverty and Adult Attainment, Behavior, and Health*, 81 CHILD DEV. 306, 306 (2010) (finding childhood poverty to be causally related to reduced adult earnings work hours); Greg J. Duncan, W. Jean Yeung, Jeanne Brooks-Gunn & Judith R. Smith, *How Much Does Childhood Poverty Affect the Life Chances of Children?*, 63 AM. SOCIO. REV. 406, 406 (1998) (concluding that “family economic conditions in early childhood have the greatest impact on achievement. . .”) Gary W. Evans, *Childhood Poverty and Adult Psychological Well-Being*, 113 PROCS. NAT’L ACAD. SCI. U.S. 14949, 14949 (2016) (linking childhood poverty to negative psychological outcomes in adulthood); Gary W. Evans & Pilyoung Kim, *Childhood Poverty and Health*, 18 PSYCH. SCI. 953, 953 (2007) (linking childhood poverty to chronic stress).

The Census's conclusion that this expansion drastically decreased child poverty, then, is no surprise. Although it did not happen in the context of official scorekeeping, both government and academic experts quantified the benefits of the expanded CTC in real time. To name two examples, economist Paul Shafer and coauthors performed a cross-sectional study of Census survey data to find that household food insufficiency fell by twenty-six percent,¹²⁷ and economists Natasha Pilkauskas, Katherine Micheltmore, Nicole Kovski, and H. Luke Shaefer estimated a seventeen percent decline in material hardships experienced by low-income households, primarily driven by a similar decrease in food insecurity as documented by Shafer et al.¹²⁸ These authors also studied the impact of the more generous credit on employment, finding at least in the short run that there were no significant declines in employment as predicted by the JCT's model of macroeconomic effects.¹²⁹

<i>By income percentile</i>					
	1 year	2 years	3 years	4 years	5 years
0	10	14	19	25	31
0.2	20	23	28	33	37
0.4	26	29	35	41	45
0.6	26	28	35	41	45
0.8	32	35	43	49	54
0.9	12	25	46	53	57
0.99	1	6	20	29	37

<i>By age group (in years)</i>					
	1 year	2 years	3 years	4 years	5 years
24 and under	15	18	23	29	36
25 - 29	30	36	44	50	55
30 - 39	41	47	56	62	67

127. Paul R. Shafer et al., *Association of the Implementation of Child Tax Credit Advance Payments with Food Insufficiency in US Households*, 5 JAMA NETWORK OPEN, at 4 (2022).

128. Natasha Pilkauskas, Katherine Micheltmore, Nicole Kovski & H. Luke Shaefer, *The Effects of Income on the Economic Wellbeing of Families with Low Incomes: Evidence from the 2021 Expanded Child Tax Credit 3* (Nat'l Bureau of Econ. Rsch., Working Paper No. 30533, 2022).

129. *Id.* Although this estimate is of a temporary rather than permanent expansion of the CTC, which are generally hard to compare, it is worth noting that other estimates of a permanent expansion of the CTC do find an employment impact, but estimate a fairly modest decline. See, e.g., YALE BUDGET LAB, CHILD TAX CREDIT: OPTIONS FOR REFORM 22 (2024), <https://perma.cc/N3H2-PMB3>; Jessica Pac & Lawrence M. Berger, *Quasi-Experimental Evidence on the Employment Effects of the 2021 Fully Refundable Monthly Child Tax Credit*, 43 J. POL'Y ANALYSIS & MGMT. 192, 192-94 (2024).

40 - 49	40	46	58	65	70
50 - 64	16	20	31	39	45
65+	2	4	8	11	14
<i>By parental status</i>					
	1 year	2 years	3 years	4 years	5 years
Non-parents	0	3	12	19	25
Parents	82	89	97	98	98

Table 2: Percentage Share of Adults Today Benefitting from CTC Expansion at Any Point in Over the Next 5 Years

Note: Adults benefitting from the CTC expansion see their taxes go down as a result.

Source: To calculate the effects of the CTC expansion over time, we first determine the current share of the population that would benefit from expansion by receiving a credit not received under prior law or by receiving a larger credit. Second, we project how this benefit might change over time given how we expect eligibility to change over the course of individual lifetimes. Longitudinal distribution estimates are based on a model of annual transitions that generates a synthetic panel dimension in the Budget Lab tax microsimulation model. We use the longitudinal component of the Current Population Survey Annual Social and Economic Supplement to estimate Markov transition parameters—conditional on economic and demographic characteristics—for marital status, number of children, and family income. We also account for mortality risk using Social Security Administration actuarial tables. Then, with these transition matrices and a simulated cross-section of tax microdata at time t , we simulate time $t+1$ outcomes for all records. Finally, for each record at time t , we locate the simulated tax record at time $t+1$ most closely matching these simulated attributes and create a link. This process is repeated recursively and generates synthetic longitudinal dimension for records present at time t , allowing for the calculation of multi-period distributional metrics.

The Biden Administration’s push for an expanded CTC ultimately failed, with lawmakers referencing the policy’s high price tag and the policy’s negative impact on employment as reasons to vote against it.¹³⁰

130. It was widely reported that Senator Manchin, for example, refused to support extending the ARPA CTC without amendments to trim its cost and add a work requirement. See, e.g., Tami Luhby, *Manchin Thinks the Child Tax Credit Is Too Generous – But Republicans Are Responsible for That*, CNN (Jan. 5, 2022, at 12:05 EST), <https://perma.cc/V7NS-D6BN>; Hans Nichols, *Scoop: Manchin’s Red Lines*, AXIOS (Oct. 17, 2021), <https://perma.cc/EB7H-C2WS>.

This policy debate illustrates the important reality in scorekeeping that what is counted counts: there was a host of evidence concerning the positive benefits accruing to families and to children from the reduction in child poverty that an expanded CTC would deliver. But, at least in part because of the nature of JCT's modeling—which largely ignored benefits that do not show up in revenue estimates or within the short timeframe of the ten-year budget window—policymakers failed to grapple with these benefits meaningfully, as they were not automatically equipped with the information arguably most relevant to the policy decision at hand. As a result, five million more American children fell into poverty in 2022, with the child poverty rate more than doubling.¹³¹

The benefits of a reduction in child poverty are not impossible to quantify; indeed, academic economists have worked to develop models that do exactly that. But with a scorekeeper whose traditional remit is short-term cost estimates and whose primary toolkit is tax models that omit the effects of lifting children out of poverty, the books are cooked against investing in these sorts of policies.

JCT can—and indeed sometimes does—produce estimates of other outcomes. Its analysis of the Affordable Care Act, for example, focused on the number of uninsured individuals.¹³² But as neutral players, it is not scorekeepers' job to determine the most appropriate form of analysis—that is the role of policymakers. When policymakers fail, as they did here, to bring key pieces of information into the legislative debate, they make decisions on the basis of incomplete information. Academics can and should fill in the gaps, but they have limited ability as external estimators to provide inputs into actual policy decisions. And the language and form of academic analysis is often not well-suited to be brought to bear in policy debates that move quickly and are crafted by generalists who are not steeped in the particular models or empirical methods that we rely on. That means that there is an important missing translation between academic expertise and policymakers' decision-making. Providing that translation is not the job of scorekeepers, who answer questions that policymakers pose in an unbiased manner. But that begs the question: whose role is it, and how can policy analysis be best improved? We consider this in Part III below.

D. The Scorekeeping Process' Failure to Consider Long-Term Effects: The Case of Tax Compliance

The gross revenue the Internal Revenue Service collects is approximated at

131. Catherine Rampell, *We Let Child Poverty Soar Last Year. We Could Choose Differently.*, WASH. POST (Sept. 12, 2023), <https://www.washingtonpost.com/opinions/2023/09/12/biden-child-tax-credit-poverty-doubled/>.

132. CONG. BUDGET OFF., CBO'S AND JCT'S ESTIMATES OF THE EFFECTS OF THE AFFORDABLE CARE ACT ON THE NUMBER OF PEOPLE OBTAINING EMPLOYMENT-BASED HEALTH INSURANCE 20-25 (2012), <https://perma.cc/47GD-9GMU>.

\$4.9 trillion annually, with the agency's collections accounting for essentially all of the revenue collected by the United States.¹³³ The agency touches each and every American household and business each year, and it is tasked not just with providing taxpayer services and collecting taxes, but also with implementing important non-tax policies: the agency is among the largest distributors of federal benefits in the U.S. government, [footnote only supports claims after this point] and it was recently tasked with implementing a host of new credits in the clean energy space amidst the Biden Administration's push to encourage clean energy manufacturing and production.¹³⁴

And yet, despite its central importance, the agency today functions largely as it did in the 1960s: For example, paper returns are literally transcribed by hand,¹³⁵ and the IRS runs on the oldest technology in the U.S. government.¹³⁶ Transforming tax administration is not the type of policy that pays off overnight, and as recent pushes to fund the IRS illustrate, the failure of scorekeepers to quantify long-term benefits understates the gains to taxpayers from modernizing the IRS and creates systematic biases against tax enforcement and other progressive investments.

The IRS's approach to tax processing is antiquated because it has for much of its history found itself drastically under-resourced. Over the course of the last decade—before the Inflation Reduction Act's influx of funding for compliance efforts—the agency saw its budget reduced by over twenty percent in real terms.¹³⁷ Its number of employees had fallen to staffing levels of the 1970s.¹³⁸

This decrease did not result from increases in technological efficiency. This became obvious when the lack of automation in tax return processing created a significant backlog of unprocessed tax returns and delayed refunds when the tax system essentially shut down for months during the COVID-19 pandemic.¹³⁹ Even now, as businesses have resumed normal operations, the IRS lags significantly behind, with fewer than fifteen percent of calls to the IRS answered

133. INTERNAL REVENUE SERV., INTERNAL REVENUE SERVICE DATA BOOK, 2022, at vi, 1 (2023), <https://perma.cc/G265-KAMW>.

134. *See Inflation Reduction Act of 2022*, INTERNAL REVENUE SERV. (May 29, 2025), <https://perma.cc/88E4-BEXN>.

135. *NTA Blog: Getting Rid of the Kryptonite: The IRS Should Quickly Implement Scanning Technology to Process Paper Tax Returns*, TAXPAYER ADVOC. SERV. (Sep. 10, 2024), <https://perma.cc/958N-RBZT>.

136. U.S. GOV'T ACCOUNTABILITY OFF., GAO-23-106821, INFORMATION TECHNOLOGY: AGENCIES NEED TO CONTINUE ADDRESSING CRITICAL LEGACY SYSTEMS 12 tbl.1 (2023) (listing an age of fifty-one years for the Treasury system containing taxpayer information).

137. CONG. BUDGET OFF., 56422, TRENDS IN THE INTERNAL REVENUE SERVICE'S FUNDING AND ENFORCEMENT 1 (2020).

138. Wally Adeyemo, *Taxpayers Will See Improved Service This Filing Season Thanks to Inflation Reduction Act*, U.S. DEP'T TREASURY (Jan. 23, 2023), <https://perma.cc/T54E-5K22>.

139. Kelly Phillips Erb, *IRS Quietly Closing Offices During COVID-19 Crisis, But Union Says It's Not Enough*, FORBES (Mar. 25, 2020), <https://perma.cc/8UCN-BP45> (detailing closures of IRS call centers and return processing centers).

during the 2022 tax filing season.¹⁴⁰

The IRS's lack of resources has not only damaged its ability to provide taxpayer services, but has also crippled its capacity to meaningfully enforce the tax code. IRS audit rates fell by over fifty percent from 2010 to 2019, with the sharpest declines among individual filers with over \$5 million in income.¹⁴¹ Millionaire audits have fallen by more than eighty percent.¹⁴² The inability to police tax evasion has real fiscal costs: the IRS projected a \$688 billion tax gap in 2021, nearly \$200 billion higher than its estimates for tax years 2014-2016.¹⁴³ Over the course of the next decade, the tax gap, left unaddressed, will total \$7.5 trillion dollars, or over two percent of GDP on an annualized basis.¹⁴⁴ Some experts believe the tax gap to be even larger than official estimates suggest: This is in part because high-end evaders are able to hide their income even during an audit, which are the basis for official estimates. Indeed, former IRS Commissioner Charles Rettig has guessed that closer to \$1 trillion in owed taxes are not collected.¹⁴⁵

The need for significant investment in the IRS had been well-documented.¹⁴⁶ Given the current scale of IRS disinvestment, mounting revenue needs facing the United States, and deep inequities in tax administration (with compliance rates near 100% for wage earners but only 50% for those earning proprietorship income¹⁴⁷), there has been recent interest in providing the IRS the tools that it needs to make meaningful progress in tax enforcement. Indeed, the decline in the

140. NAT'L TAXPAYER ADVOC., ANNUAL REPORT TO CONGRESS 2022, at 74 (2023), <https://perma.cc/EJ6R-7PFK>.

141. Sarin & Mazur, *supra* note 14, at 12 tbl.2.

142. *Id.*

143. *IRS Updates Tax Gap Projections for 2020, 2021; Projected Annual Gap Rises to \$688 Billion*, INTERNAL REVENUE SERV. (Oct. 12, 2023), <https://perma.cc/286R-WJB7>.

144. *Closing the Tax Gap: Lost Revenue From Noncompliance and the Role of Offshore Tax Evasion: Hearing Before the Subcomm. on Tax'n and IRS Oversight of the S. Comm. on Fin.*, 117th Cong. 7 (2021) (statement of Charles O. Rossotti, former Comm'r, Internal Revenue Serv.).

145. Aaron Lorenzo, *IRS chief says some \$1T in taxes going uncollected annually*, POLITICO (Apr. 13, 2021), <https://perma.cc/2L8K-M96D>.

146. *See, e.g.*, Sarin et al., *Tax Reform for Progressivity*, *supra* note 90; Charles O. Rossotti, Natasha Sarin & Lawrence H. Summers, *Shrinking the Tax Gap: A Comprehensive Approach*, TAX NOTES (Nov. 30, 2020), <https://perma.cc/DA75-YCT5>; Natasha Sarin & Lawrence H. Summers, *Shrinking the Tax Gap: Approaches and Revenue Potential*, TAX NOTES (Nov. 18, 2019), <https://perma.cc/LE5M-GP9A>; Sarin & Summers, *supra* note 14; CHUCK MARR ET AL., CTR. ON BUDGET & POL'Y PRIORITIES, CONGRESS NEEDS TO TAKE TWO STEPS TO FUND THE IRS FOR THE SHORT AND LONG TERM, (2022), <https://perma.cc/9FY4-NX8J>; Jean Ross & Seth Hanlon, *The Build Back Better Act's Investments in the IRS Will Substantially Reduce the Tax Gap*, CTR. FOR AM. PROGRESS (Nov. 17, 2021), <https://perma.cc/5LWQ-SU64>; William C. Boning et al., *A Welfare Analysis of Tax Audits Across the Income Distribution*, 140 Q.J. ECON. 63; Brian Deese & David Kamin, *Principles for the 2025 Tax Debate*, TAX NOTES (Oct. 16, 2023), <https://perma.cc/73T5-JHNU>.

147. BARRY W. JOHNSON & PETER J. ROSE, INTERNAL REVENUE SERV., PUBLICATION 1415 (REV. 9-2019), FEDERAL TAX COMPLIANCE RESEARCH: TAX GAP ESTIMATES FOR TAX YEARS 2011-2013, at 20 tbl.5 (2019).

IRS's capacity has been so significant that the Biden Administration focused on the importance of funding the agency during its first two years. IRS funding provisions were first placed in the Bipartisan Infrastructure Deal¹⁴⁸ and eventually adopted as part of the Inflation Reduction Act.¹⁴⁹

One of the impediments in enacting these provisions and in generally providing the IRS with the resources it needs to retool was and continues to be the treatment of those investments by the budget scorekeeping process. The current approach has a host of deficiencies stemming from current budget rules and conventions as well as scorekeepers' hesitancy to incorporate effects of enforcement investments for which the directional effect on revenues is clear, but the magnitude of the effect is uncertain. In recent work, one of us worked with former Treasury Assistant Secretary Mark Mazur to highlight limitations of the scorekeepers' approach, noting that official estimates do not meaningfully adjust for deterrence (how taxpayer behavior responds to increased enforcement activity) and ignore the significant revenue potential of investments in service and information technology.¹⁵⁰

Another aspect of scorekeepers' underestimation has to do with their determination that the diminishing returns to additional IRS enforcement efforts are high: although initial dollars have significant returns, future investments in the agency are likely to generate much less. But it is hard to argue that there are significant diminishing returns to new IRS investments in a world in which the agency has fewer revenue agents than it did at any time since World War II¹⁵¹ and a partnership audit rate of nearly zero.¹⁵² These deficiencies exist alongside scorekeeping guidelines that actively prevent scorekeepers from incorporating efficiency gains in their revenue estimates of government enforcement initiatives—guidelines that have been critiqued before, but Congress has thus far failed to update.¹⁵³

148. Kristina Peterson & Sarah Chaney Cambon, *Senate Infrastructure Bill Drops IRS Funding, Raising Pressure for New Revenue*, WALL ST. J. (July 18, 2021, at 18:01 ET), <https://perma.cc/8VZ6-4U55>.

149. Inflation Reduction Act of 2022, Pub. L. No. 117-169 136 Stat. 1831.

150. Sarin & Mazur, *supra* note 14, at 21. *See also* Boning et al., *supra* note 146, at 96-98 (estimating that specific deterrence produces more than three times as much revenue than an initial IRS audit); Jeffrey L. Hoopes, Devan Mescall & Jeffery A. Pittman, *Do IRS Audits Deter Corporate Tax Avoidance?*, 87 ACCT. REV. 1603, 1603 (2012) (suggesting that public corporations take less aggressive tax positions when the probability of an IRS audit is higher).

151. Adeyemo, *supra* note 138.

152. U.S. GOV'T ACCOUNTABILITY OFF., GAO-23-106020, TAX ENFORCEMENT: IRS AUDIT PROCESSES CAN BE STRENGTHENED TO ADDRESS A GROWING NUMBER OF LARGE, COMPLEX PARTNERSHIPS 18 (2023) (reporting a large partnership audit rate of 0.3 percent in 2019). Indeed, the IRS is so beleaguered in partnership enforcement efforts that it deployed a single revenue agent to unpack former President Trump's 500 complex partnerships, who unsurprisingly concluded that he did not have the resources to investigate potential areas of noncompliance. Alan Rappeport, *Trump Audit Shows Depths of I.R.S. Funding Woes*, N.Y. TIMES (Dec. 22, 2022), <https://perma.cc/8LDQ-HZ5Z>.

153. In particular, the scorekeeping guidelines enshrined in Guidelines 3 and 14 provide

As a result of these shortcomings, while CBO concludes that IRS investments can deliver around \$200 billion in additional tax revenue over the next ten years, outside experts have estimates that range from three to five times that total over this horizon.¹⁵⁴ Scorekeepers themselves concur that a conservative estimate is higher than that of their official estimates.¹⁵⁵

Furthermore, estimates of enforcement investments are woefully understated in a ten-year budget window.¹⁵⁶ This illustrates how the scorekeeping process's use of a budget window biases against progressive investments with long-run returns.

One can observe the understatement of tax enforcement benefits in the standard ten-year budget window by comparing scorekeepers' and external estimators' estimated returns to IRS investments in the first decade to expected returns in the second decade and thereafter.¹⁵⁷ As Sarin and Mazur show, the revenue effects of additional IRS investments will grow over time as fixed investment costs pay out in full.¹⁵⁸ As the IRS works to modernize not just its processing infrastructure, but also its service and enforcement capacity, it will require a host of experts and auditors in highly complex areas, like partnership and global high net-worth individuals, and it will need to train them for years before they reach their full potential.¹⁵⁹ Furthermore, as the economy grows, the

that funding increases to agencies aimed at improving enforcement or decreasing fraud cannot be officially scored as deficit reducing, even though, for example, enhancing the IRS's capacity to pursue high-end tax evasion is clearly deficit reducing. Instead, paradoxically, the investment must be registered as deficit increasing, because the only dollars that are officially counted in the scorekeeping process are those expended on enforcement initiatives. *See Levy, supra* note 14, at 954-55 (describing how Guidelines 3 and 14 prohibit scoring of "secondary effects").

154. Sarin and Mazur, for example, estimate that the Inflation Reduction Act's provision of \$80 billion in IRS funding will generate at least \$560 billion and up to nearly \$1 trillion in revenue. Sarin & Mazur, *supra* note 14, at 21.

155. Phil Swagel, *The Effects of Increased Funding for the IRS*, CONG. BUDGET OFF. (Sep. 2, 2021), <https://perma.cc/M689-T3P4> (noting that the peak return on investment on IRS enforcement activities most recently relied upon by CBO was 7.1, relative to the 2.5 return on investment implied by CBO's revenue score of \$200 billion of an increased \$80 billion in IRS funding). Treasury estimates were slightly lower but directionally similar, concluding that an \$80 billion investment in the agency would generate \$400 billion in additional revenue. Natasha Sarin, *The Substantial Revenue-Raising Potential of Tax Compliance Efforts*, U.S. DEP'T TREASURY (Oct. 29, 2021), <https://perma.cc/Y3KU-QRZE>.

156. *See* Sarin & Mazur, *supra* note 14, at 24 fig.1 (presenting alternative revenue estimates of increased IRS enforcement over a twenty-year budget window).

157. It is important to note that these later decade estimates assume that enough resources exist at the IRS in the future to support the "care and feeding" of new employees that are hired during this period and to maintain new IT that the agency develops.

158. Sarin & Mazur, *supra* note 14, at 25 fig.1 (illustrating cumulative revenue estimates).

159. One estimate finds that in general, revenue agents take up to five years to become fully productive. Laura Sanders, *What \$80 Billion More for the IRS Means for Your Taxes*, WALL ST. J. (Aug. 19, 2022, at 05:30 ET), <https://perma.cc/FN95-BJRC>. Training for highly complex enforcement activities may take longer.

tax gap will grow with it. The fact that the scale of evasion is growing mechanically means that the same set of investments to address evasion will deliver more bang-for-the-buck in subsequent decades.

Following Sarin and Mazur, we illustrate over twenty years how including efficiency gains and deterrence effects dramatically magnifies the revenue potential of an \$80 billion investment in the IRS in each year and cumulatively over time. Importantly, the magnitude of these effects must be viewed in a longer time frame to be fully appreciated. In Year 10, the cumulative revenue generated from the investment is more than eight times what CBO estimates when assuming that deterrence effects are three times direct effects. But by Year 20, including such deterrence effects produces a cumulative revenue estimate that is more than twenty-five times the CBO estimate.

Year	CBO Estimate	Alternative Estimate with 1:1 Deterrence	Alternative Estimate with 3:1 Deterrence
0	0.0	3.5	3.5
1	2.9	14.8	17.2
2	7.8	36.4	50.7
3	13.5	68.4	105.7
4	19.0	93.6	150.4
5	24.4	109.3	178.6
6	29.6	120.8	198.1
7	33.7	133.7	219.9
8	37.6	148.0	244.1
9	35.3	160.7	267.9
10	0.0	166.4	277.7
11	0.0	173.2	288.6
12	0.0	180.4	300.0
13	0.0	186.9	310.9
14	0.0	193.7	322.2
15	0.0	200.7	334.0
16	0.0	208.0	346.1
17	0.0	215.6	358.7
18	0.0	223.3	371.6
19	0.0	231.3	384.8
20	0.0	239.5	398.5

Table 3: Yearly Revenue Estimates of an \$80 Billion Investment With and Without Efficiency Gains and Deterrence Effects (Billions of Dollars)

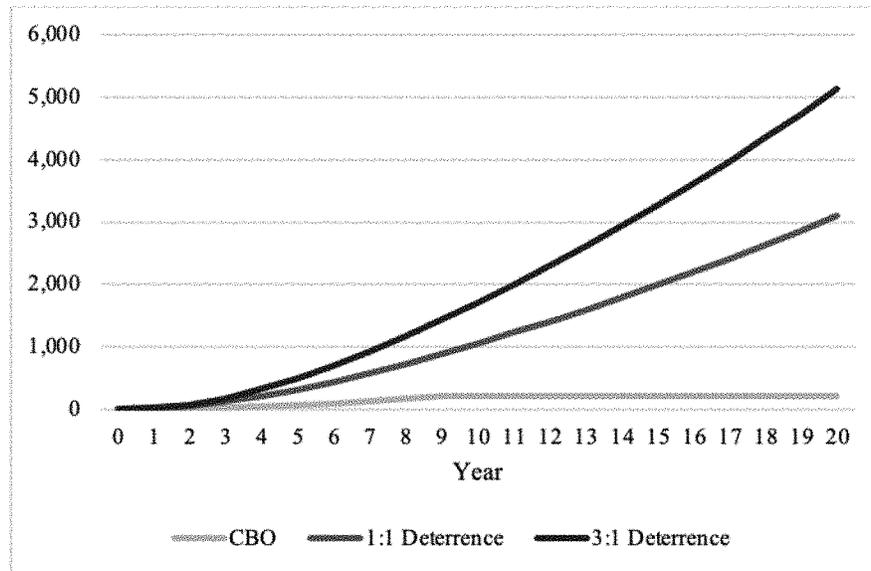


Figure 5: Cumulative Revenue Estimates of an \$80 Billion Investment With and Without Efficiency Gains and Deterrence Effects (Billions of Dollars)

Source: Budget lab calculations following Sarin and Mazur (2023). We assume \$1 billion is invested in new enforcement staff in the first year of expanded investment, \$2 billion is invested in the second year, and \$1.2 billion is invested in the third year. We then assume that each of these staffing costs grow over time such that sixty-five percent of the \$80 billion investment is spent over a decade. To calculate revenues stemming from enforcement, we first assume a return on investment from enforcement spending of 1:1 in Year 1, 3:1 in Year 2, and 5:1 in Year 3 and beyond. We then assume that twenty percent of staffing costs will be spent on support costs, including support for data, systems, and research and apply the same returns on investment. And we assume remaining investment dollars are spent on operational support with a return on investment of 2:1. Next we assume returns from improved information reporting of \$1.5 billion each year. Finally, we assume that added enforcement has a deterrence effect that begins in Year 2 and is equal to one (1:1 Deterrence) or three (3:1 Deterrence) times the revenue gain associated with enforcement and support.

Clearly, transformation of our tax processing system will not happen overnight, and reliance on a ten-year window understates the long-term benefits to be had from serious tax administration reform. The closest historical analogue to the current opportunity for reform is when Congress pushed the IRS to transition to electronic filing in the 1990s, giving the agency a ten-year statutorily

mandated target for improvement.¹⁶⁰ The agency achieved that goal, but it took time—even more than the ten years allotted.¹⁶¹ It is simply an incomplete exercise to think about the revenue associated with a project like transforming tax administration by narrowly focusing on only the first few years of additional taxes that can be collected from fixed investments that will take time to fully pay off. But this incompleteness pervades the practice of scorekeeping, biasing against tax enforcement investments, along with other progressive spending goals with anticipated future returns that are not immediate.

With each budget resolution, members of Congress select the budget window used by scorekeepers. A ten-year window budget is the current default, but it is by no means the only option available to members of Congress; their only constraint is the five-year minimum prescribed by the Congressional Budget Act.

The critique offered here does not provide clear answers as to what an appropriate budget window for evaluating tax enforcement investments and other spending initiatives would be, but suggests that experimentation is appropriate based on underlying policy goals and expectations. Indeed, over the course of debates around President Biden's Build Back Better framework, the Biden Administration did just this, presenting revenue estimates for the legislation that included second decade effects.¹⁶² The critique does make clear that failure to recognize biases in the current scorekeeping process arising out of the budget window make it harder to make improvements that will reap huge rewards in the years ahead.

E. The Scorekeeping Process' Failure to Fully Grapple with Macroeconomic Feedback Effects: The Case of High-Skilled Immigration

The examples we have discussed thus far focus on the ways (and limits to how) scorekeepers account for mechanical policy effects—the increased spending arising out of a CTC expansion or additional administrative costs necessitated by IRS infrastructural advancements. They also examine how scorekeepers, albeit imperfectly, attempt to incorporate some behavioral impacts, like the response of capital gains realizations to the capital gains tax rate. In each of these cases, scorekeepers, uncertainty and technical difficulty notwithstanding, attempt to put a number on the likely impact. When it comes to effects of spending policies that concern macroeconomic variables, however—investment, labor productivity, capital investment, economic

160. Internal Revenue Service Restructuring and Reform Act of 1998, Pub. L. No. 105-206, 112 Stat. 685.

161. The Restructuring and Reform Act implemented a goal of an eighty percent e-file rate. Fourteen years later in 2011, the e-file rate was about seventy-five percent. INTERNAL REVENUE SERV., IRS E-FILE: A HISTORY 4 (2011), <https://perma.cc/9TD4-ANF6>.

162. See *Budget of the U.S. Government: Fiscal Year 2022*, OFF. OF MGMT. & BUDGET 38 tbl.S-2 (2021) (providing in a memorandum the second decade deficit effects of the American Jobs Plan and American Families Plan).

growth—scorekeepers are more muted. The House rules currently in force instruct scorekeepers to include macroeconomic effects into their estimates for “major” legislation, which refers to a bill or joint resolution that contains a gross budgetary effect greater than or equal to 0.25 percent of projected GDP in any year.¹⁶³

By not otherwise incorporating macroeconomic effects into estimates, scorekeepers implicitly assume that macroeconomic effects do not exist for non-major pieces of legislation.¹⁶⁴ This is in part a function of the challenge of dynamic scoring, which requires significant modelling and assumptions and is inherently uncertain. But it is obviously incorrect, and importantly, it systematically biases against progressive policies that invest in future economic growth. The scorekeeping rule’s inclusion of macroeconomic effects biases against spending policies with considerable macroeconomic effects, which we can see when we consider the example of high-skilled immigration.

Economists are widely in consensus that high-skilled immigration promotes economic growth.¹⁶⁵ Rather than displacing native-born workers, high-skilled immigrant workers tend to increase overall labor market productivity. This is because immigrant and native-born workers are not perfectly substitutable: they may be employed in different jobs or undertaking different tasks, thus serving in complementary roles to each other.¹⁶⁶ This complementarity increases wages and

163. MEGAN S. LYNCH & JANE G. GRAVELLE, CONG. RSCH. SERV., R46233, DYNAMIC SCORING IN THE CONGRESSIONAL BUDGET PROCESS 13 (2023).

164. Members of Congress and Committee staffs can request additional macro-dynamic estimates in other cases, but the default is a conventional analysis without consideration of the broader macroeconomic effects. That may well be a reasonable default, given the substantial resources required to prepare macrodynamic estimates. See Elmendorf, *supra* note 71, at 110-11. But it is a default that bears on the type of proposals that are ultimately pursued by policymakers, as described *infra*.

165. See JENNIFER HUNT, HAMILTON PROJECT, RENEWING AMERICA, REVAMPING IMMIGRATION 20 (2022), <https://perma.cc/Y33P-RUF9> (proposing to increase employment-based immigration with a view towards increasing net economic benefits for native-born Americans and the United States overall).

166. Estimates of the elasticity of substitution between immigrant and native-born workers are low, implying low substitutability. See, e.g., David Card, *How Immigration Affects U.S. Cities* (Ctr. for Rsch. & Analysis of Migration Discussion Paper, Paper No. 11/07, 2007), at 15-20, <https://perma.cc/YUD4-446N>; Steven Raphael & Eugene Smolensky, *Immigration and Poverty in the United States*, 99 AM. ECON. REV. 41, 42(2009); Francesco D’Amuri, Gianmarco Ottaviano & Giovanni Peri, *The Labor Market Impact of Immigration in Western Germany in the 1990s*, 54 EUR. ECON. REV. 550, 561-63 (2010); Marco Manacorda, Alan Manning & John Wadsworth, *The Impact of Immigration on the Structure of Wages: Theory and Evidence From Britain*, 10 J. EUR. ECON. ASS’N 120, 139-46 (2012); Gianmarco I.P. Ottaviano & Giovanni Peri, *Rethinking the Effect of Immigration on Wages*, 10 J. EUR. ECON. ASS’N 152, 170-77 (2012). But see Kirk Doran, Alexander Gelber & Adam Isen, *The Effects of High-Skilled Immigration Policy on Firms: Evidence from Visa Lotteries*, 310 J. POL. ECON. 2501, 2501 (2022) (finding that winning an additional H-1B visa has at best modest effects on firm innovation and crowds out 1.5 other workers). One reason advanced to explain low substitutability is that immigrants may specialize in tasks requiring fewer communication skills using the English language. See Giovanni Peri & Chad Sparber, *Task Specialization*,

labor productivity. Evidence from an influx of highly skilled immigrants employed in science, technology, engineering, and math (STEM) fields in the United States from 2000 to 2015, for example, demonstrates that these workers were not perfectly substitutable with American-born workers. Instead, wages rose among American-born workers, in large part due to new ideas generated by immigrants.¹⁶⁷ Similarly, immigrant diversity in American cities and workplaces is associated with increased wages and productivity across the wage distribution, but especially in the highest quartile.¹⁶⁸

These productivity gains are an example of a positive labor market spillover effect arising out of increased high-skill immigration. This effect has been documented in a number of historical instances of heightened migration, including the flight of Jewish scientists from Germany during the period of Nazi rule¹⁶⁹ and the era of mass migration to the United States in the late nineteenth and early twentieth centuries.¹⁷⁰ This dynamic is also evident across regions with differing shares of immigrants in the labor force. A body of evidence finds that states and localities with higher labor market shares of immigrants exhibit greater innovation.¹⁷¹ High-skilled immigrants have furthermore been shown to drive

Immigration, and Wages, 1 AM. ECON. J. APPLIED ECON. 135, 164-64 (2009).

167. Christian Gunadi, *An Inquiry on the Impact of Highly-Skilled STEM Immigration on the U.S. Economy*, 61 LAB. ECON., at 11 (2019). See also Giovanni Peri, Kevin Shih & Chad Sparber, *STEM Workers, H1B Visas and Productivity in US Cities* (Norface Migration Discussion Paper, Working Paper No. 2013-09, 2013), at 2-3, 21-23, <https://perma.cc/S4ML-DZ3L> (providing similar evidence in U.S. cities over the period from 1990 to 2010).

168. Thomas Kemeny & Abigail Cooke, *Spillovers From Immigrant Diversity in Cities*, 18 J. ECON. GEOGRAPHY 213, 239 (2018). See also Gianmarco I.P. Ottaviano & Giovanni Peri, *Cities and Cultures*, 58 J. URB. ECON. 304, 304 (2005) (linking cultural diversity to increased productivity).

169. Petra Moser, Alessandra Voena & Fabian Waldinger, *German Jewish Émigrés and US Invention*, 104 AM. ECON. REV. 3222, 3222 (2014) (estimating that patenting by U.S. investors increased by 31 percent in migrants' fields). But see M. Daniele Paserman, *Do High-Skill Immigrants Raise Productivity? Evidence From Israeli Manufacturing Firms, 1990-1999*, 2 IZA J. MIGRATION, at 1-3 (2013) (failing to replicate this result in studying migration from the former Soviet Union to Israel in the 1990s).

170. See Dario DiDato, Andrea Morrison & Sergio Petralia, *Migration and Invention in the Age of Mass Migration*, 22 J. ECON. GEOGRAPHY 477, 477-80 (2022) (finding that native inventors benefitted from immigrants' inventive activity); Sandra Sequeira, Nathan Nunn & Nancy Qian, *Immigrants and the Making of America*, 87 REV. ECON. STUD. 382, 382-85 (2019) (linking historical migration to long-run macroeconomic benefits).

171. See, e.g., Jennifer Hunt & Marjolaine Gauthier-Loiselle, *How Much Does Immigration Boost Innovation?*, 2 AM. ECON. J. MACROECONOMICS 31, 31 (2010) (estimating that a one-percentage point increase in the share of a state's population comprised of immigrant college graduates increases the per capita patent rate by up to nearly twenty percent); Giovanni Peri, *The Effect of Immigration on Productivity: Evidence From U.S. States*, 94 REV. ECON. & STAT. 348, 348-49 (2012) (demonstrating a strong, positive association between total factor productivity and immigration); William R. Kerr & William F. Lincoln, *The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention*, 28 J. LAB. ECON. 473, 475 (2010) (finding that cities dependent upon the H-1B visa program experienced growth in total inventions when the program was expanded to admit more immigrants).

entrepreneurship in the United States,¹⁷² in part because those who are willing to internationally migrate are more likely to have entrepreneurial personalities.¹⁷³ And high-skilled immigration has been shown to have positive economic spillover effects extending even beyond the labor market, such as through capital investment and consumption.¹⁷⁴

Evidence illustrating the ways in which high-skilled immigrants promote economic growth abounds; however, this evidence is neglected by scorekeepers when estimating the budgetary effects of policies that would expand high-skilled immigration. To CBO's credit, CBO does not typically use a pure conventional score to analyze immigration proposals, and this is an arena in which scorekeepers have developed substantial infrastructure enabling them to provide a sense of the macroeconomic effects of policy changes.¹⁷⁵ A conventional score would account for mechanical effects, such as the administrative cost of issuing more visas, as well as behavioral effects, such as costs associated with additional migration induced by policy changes that make migrating easier, while leaving out changes to macroeconomic variables like population, productivity, and wages. But to ignore the population effects of a policy designed to increase the population, CBO acknowledges, is absurd.¹⁷⁶ In its recent immigration analyses, CBO relaxes its assumption that population, and along with it income and payroll tax revenues, are unchanged by the proposal being analyzed.¹⁷⁷ For example, CBO used this method to analyze the Supply Knowledge-based Immigrants and Lifting Levels of STEM Visas Act (SKILLS Visa Act), a proposal that would promote high-skilled immigration to the United States. It found that by boosting the U.S. population by almost 1 million, and with it, income and payroll tax collections, the SKILLS Visa Act would produce net revenue savings of \$110 billion over increased spending for visa processing and other government

172. For a review of the literature, see William R. Kerr, *US High-Skilled Immigration, Innovation, and Entrepreneurship*, in *THE INTERNATIONAL MOBILITY OF TALENT AND INNOVATION* 193 (Carsten Fink & Ernest Miguelez eds., 2017).

173. Peter Vandor, *Are Voluntary International Migrants Self-Selected for Entrepreneurship? An Analysis of Entrepreneurial Personality Traits*, 56 *J. WORLD BUS.*, at 2 (2021) (“The results add to the growing body of literature on the drivers of migrant entrepreneurship and provide first robust evidence that personality-based self-selection does indeed contribute to more entrepreneurship among migrants.”) (citation omitted).

174. For a review of this literature, see Max Nathan, *The Wider Economic Impacts of High-Skilled Migrants: A Survey of the Literature for Receiving Countries*, 3 *IZA J. MIGRATION* 1 (2014).

175. See, e.g., CONG. BUDGET OFF., *EFFECTS OF THE IMMIGRATION SURGE ON THE FEDERAL BUDGET AND THE ECONOMY* at 1-2, see also *id.* at 16-18 (2024).

176. CONG. BUDGET OFF., *THE ECONOMIC IMPACT OF S. 744, THE BORDER SECURITY, ECONOMIC OPPORTUNITY, AND IMMIGRATION MODERNIZATION ACT 2* (2013).

177. See also CONG. BUDGET OFF., *H.R. 2131 SUPPLYING KNOWLEDGE-BASED IMMIGRANTS AND LIFTING LEVELS OF STEM VISAS ACT (SKILLS ACT)* 4 (2014). Doug Elmendorf, Glenn Hubbard, and Heidi Williams term this approach “population-change scoring.” Douglas Elmendorf, Glenn Hubbard, & Heidi Williams, *Dynamic Scoring: A Progress Report on Why, When, and How*, *BROOKINGS PAPERS ON ECON. ACTIVITY*, Fall 2024, at 120-23.

programs.¹⁷⁸

But scores could go a step further, incorporating the effect of those population increases on productivity, employment, and other macroeconomic variables—effects that are likely to be substantial in the high-skilled immigration context.¹⁷⁹ The result would be a fully dynamic score, capturing all these macroeconomic effects and accordingly providing a lower revenue cost estimate of policies that increase high-skilled immigration. The difference between a partially and a fully dynamic score will not always be consequential, but for immigration policies, it is likely to be substantial. Dynamic scorekeeping is challenging, but CBO has invested resources into doing this important work. Doug Elmendorf, Glenn Hubbard, and Heidi Williams report that CBO has twice produced fully dynamic estimates for immigration proposals, but in both cases, these estimates were presented separately from the official estimates.¹⁸⁰

The stakes of the limitations to CBO's current approach are illustrated by the America COMPETES Act of 2022. CBO scorekeepers estimated that provisions in the COMPETES Act that would increase high-skilled immigration would cost \$3.4 billion over a ten-year budget window.¹⁸¹ These provisions would have provided an exemption from green card limits for immigrants with PhDs in STEM fields and created a new visa status for international entrepreneurs,¹⁸² thereby expanding immigration where productivity gains are likely to be largest. But these gains were ignored by CBO scorekeepers: the STEM PhD green card exemption was ultimately excluded from the bill, likely at least in part due to its high CBO score,¹⁸³ and the bill was not enacted. And as Congress turns to a fresh debate on H-1B visas, it will be increasingly imperative to accurately account for these visas' costs and benefits.¹⁸⁴

It is worth emphasizing once again that this scorekeeping flaw belongs not to the scorekeepers themselves, but to the rules they follow in constructing their estimates. Congress can and indeed has directed CBO to employ dynamic scorekeeping in some instances, but it has done so inconsistently. Under

178. CONG. BUDGET OFF., H.R. 2131 SUPPLYING KNOWLEDGE-BASED IMMIGRANTS AND LIFTING LEVELS OF STEM VISAS ACT (SKILLS VISA ACT) 1 (2014).

179. Elmendorf, Hubbard, & Williams, *supra* note 177, at 95-97.

180. *Id.* at 124.

181. CONG. BUDGET OFF., ESTIMATED BUDGETARY EFFECTS OF H.R. 4521, THE AMERICA COMPETES ACT OF 2022, AS PASSED BY THE HOUSE OF REPRESENTATIVES ON FEBRUARY 4, 2022, at 1 (2022).

182. *Analysis: Immigration Provisions in the America COMPETES Act*, NAT'L IMMIGR. F. (Mar. 29, 2022), <https://perma.cc/K7EJ-AM7N>.

183. Stuart Anderson, *Ph.D. Immigration Measure Removed, New Immigrant Visa Lawsuit Filed*, FORBES (July 14, 2022, at 12:07 EDT), <https://perma.cc/BU3Y-K6BV> (“According to a Congressional source, the House Rules Committee did not rule the amendment in order because the Congressional Budget Office (CBO) said the provision would cost \$1 billion over 10 years.”).

184. See Jack Kelly, *The Heated Debate Over H-1B Visas*, FORBES (Dec. 30, 2024, at 12:43 EST), <https://perma.cc/3PEK-CVLP>.

congressional rules currently in force, CBO is to the extent practicable required to use dynamic scoring for certain legislation with a large gross budgetary effect, but this requirement excludes appropriation acts.¹⁸⁵

This asymmetry is not an oversight, but the result of focused attention by Republican legislators to the ways in which macroeconomic effects boost the revenue potential of one of their core policy aims: tax cuts. As House Budget Committee Chairman, Paul Ryan championed dynamic scoring,¹⁸⁶ his budget resolution enabled Congress to demand dynamic scoring of any revenue (but not spending) proposal, and his proposed budget for fiscal year 2015 incorporated dynamic scoring in its estimates.¹⁸⁷ The proposal professes to balance the budget, in part through “pro-growth tax reform” that would substantially cut individual and corporate tax rates.¹⁸⁸ And when Republicans took control of the House in 2015, they immediately voted to recognize the macroeconomic effects of tax cuts in the official scorekeeping process.¹⁸⁹

Thus, in scoring the Tax Cuts and Jobs Act of 2017 (TCJA), JCT and CBO provided a conventional score that would increase the deficit by \$1.9 trillion, but a dynamic score that would increase the deficit by only \$1.4 trillion, due to the tax cut’s expected effects on economic growth.¹⁹⁰ This is in marked contrast to CBO’s approach in scoring the COMPETES Act, where growth-enhancing feedback effects could not be offset against fiscal costs. CBO is not wrong in finding macroeconomic effects associated with the TCJA: new research finds that its large corporate tax cuts did stimulate investment, though not enough to offset dynamic revenue losses from a lower corporate tax rate.¹⁹¹ But CBO’s failure to similarly regard macroeconomic effects of spending proposals creates

185. Ben Page, Unit Chief for Fiscal Pol’y Stud., Cong. Budget Off., Presentation at the NABE Foundation’s 12th Annual Economic Measurement Seminar: Dynamic Scoring at CBO 5 (July 21, 2015), <https://perma.cc/Q4G3-NSVH> (describing CBO’s dynamic scoring requirements under the 2016 budget resolution).

186. See *Chairman Camp’s Legacy Fete – Last Minute Floor Action – Internet Tax Freedom Act passes, Wyden Warns Against Internet Sales Tax*, POLITICO (Sep. 9, 2014, at 9:59 EDT), <https://perma.cc/9QWT-RC7C> (quoting Representative Ryan: “Our rules in Congress require that we don’t take into consideration behavioral changes or economic effects as a result of tax reform . . . What we want to do is change our measurement.”).

187. H.R. Con. Res. 71, 115th Cong. § 5107 (2017) (enacted); H. BUDGET COMM., THE PATH TO PROSPERITY: FISCAL YEAR 2015 BUDGET RESOLUTION 89 (2014), <https://perma.cc/27G8-MTWM>. Democrats have not brought the same fight on dynamic scoring to the spending side of the ledger that Republicans have brought to the tax side.

188. *Id.*

189. *The House’s New Rule on Dynamic Scoring*, COMM. FOR RESPONSIBLE FED. BUDGET (Jan. 8, 2015), <https://perma.cc/3CVC-MYYN>.

190. *How Did the TCJA Affect the Federal Budget Outlook?*, TAX POL’Y CTR. (Jan. 2024), <https://perma.cc/A9HL-4QZU>.

191. Gabriel Chodorow Reich et al., Tax Policy and Investment in a Global Economy 50 (Feb. 16, 2025) (unpublished manuscript) (on file at <https://perma.cc/W8QE-A3Y6>) (estimating a long-run increase in domestic corporate capital of about 7.4% but noting that the associated revenue effect is almost entirely offset by TCJA’s increased cost of depreciation deductions).

a bias against increased spending relative to tax cuts, even though both mechanically increase deficits.¹⁹²

A neutral and balanced approach to scorekeeping requires equal attention towards the macroeconomic effects of spending proposals. In promoting dynamic scorekeeping, Paul Ryan stated that his hope was “to improve our scorekeeping so it better reflects reality.”¹⁹³ The reality is that tax cuts may stimulate economic growth, but so too may spending proposals that grow the labor force, expand productivity, and spark innovation. A scorekeeping process that only counts macroeconomic effects of tax legislation is rigged in favor of small government.

As Doug Elmendorf notes, dynamic scoring is not without challenges.¹⁹⁴ But when spending proposals like those that increase high-skilled immigration generate significant macroeconomic feedback effects, failure to produce a dynamic score leads to a plainly incomplete revenue estimate and a clear detriment to such proposals’ success.

III. WHERE DO WE GO FROM HERE?

Government scorekeeping is and always will be complex. Inevitably, the process involves a host of assumptions that are inherently uncertain and cannot be known with any amount of precision ex-ante, and yet, the process demands precise outputs. This aspect of scorekeeping’s complexity is well appreciated by scorekeepers and by critics of scorekeeping in the policy and academic literatures. An underappreciated facet of complexity arises out of the policy choices implicit in the measurement of economic and government activity and the rules under which these policy choices are made. Although measurement is often presented and assumed to be a technical craft practiced by politically neutral experts, the process of measurement conceals a range of deep conceptual challenges and contestable normative choices. Within what is concealed is a systematic bias against policies that invest in the future and promote economic equality.

The policies highlighted here are illustrative not just of isolated, flawed features of the scorekeeping process—which many have identified in specific policy arenas—but of the fact that a scorekeeping process built around deficit

192. See also Jonathan Chait, *Why the Republican Congress’s First Act Was to Declare War on Math*, N.Y. MAG. (Jan. 7, 2015), <https://nymag.com/intelligencer/2015/01/congresss-first-act-was-to-declare-war-on-math.html> (describing House Republicans’ political motivations in adopting current dynamic scoring rules).

193. Jonathan Chait, *Paul Ryan Declares War Against Math*, N.Y. MAG (Sep. 23, 2014), <https://perma.cc/2Y2D-PWHJ>.

194. Elmendorf, Hubbard, & Williams, *supra* note 177, at 95-97, 100 (acknowledging the challenges of dynamic scoring while proposing the use of dynamic scoring for major proposals, including spending proposals, when estimating agencies have the appropriate time and resources to do so).

estimation discourages adoption of certain types of policies while encouraging others. We are the first to offer a unified perspective on scorekeeping deficiencies and show how they result in a clear, systematic bias against progressive policy outcomes. This is a convenient situation for those who are normatively in favor of lower spending on social safety nets and future generations and of smaller, less dynamic government. But those who prefer a more progressive government are unwittingly playing by a set of rules that are rigged against them, tilting policy outcomes against their preferred objectives. And not only do these progressives fail to challenge this bias, they uncomplainingly accept it, promulgating biased scorekeeping rules and lifting up biased revenue estimates in their own policy advocacy.

Our objective in this Article is to illustrate why those that care about distributional outcomes and long-term investments need to pay attention to the scorekeeping process and the ways in which they are systematically disadvantaged by its default rules and outputs. It also offers a guide for rewriting rules and scorekeeping requests to correct for these anti-progressive biases. We next turn to more structural ways to improve the scorekeeping enterprise.

This Article first functions as a call to academics, who should not disregard the scorekeeping process and must instead engage in the project of studying its estimates and helping to inform debate around their accuracy. In the first place, academics have existing research that calls into question certain scorekeeping assumptions and methodologies. As we have discussed, recent work on the elasticity of capital gains realizations is helpful for improving estimation approaches for capital gains tax reform,¹⁹⁵ just as new estimates of returns to IRS enforcement inform the associated official scores.¹⁹⁶ This research both reflects new methodologies and data as well as illuminates the extensive range of plausible estimates.

Academics can also bridge the gap between the information produced by scorekeepers and the full set of information that is useful and relevant to understanding a policy's effects. Academics are not limited by policymakers' analytic requests and are at full liberty to study a broad range of outcome variables, from child poverty to labor force attachment to carbon emissions. While official scorekeepers focused on the cost of the APRA's expanded CTC, academics illuminated the gaps in scorekeepers' analysis, studying the

195. See, e.g., Sarin et al., *Rethinking*, *supra* note 14, at 2; Ole Agernsnap & Owen Zidar, *The Tax Elasticity of Capital Gains and Revenue-Maximizing Rates*, 3 AM. ECON. REV.: INSIGHTS 399, 413 (2021) (presenting estimates of the elasticity of capital gains realizations to the capital gains tax rate that are “notably smaller than estimates used by official analysts”).

196. Boning et al., *supra* note 146, at 64-65 (providing new evidence on the returns to tax audits).

expansion's effect on poverty,¹⁹⁷ hunger,¹⁹⁸ financial stability,¹⁹⁹ and physical²⁰⁰ and mental health.²⁰¹ And finally, academics have the capacity to come up with

197. See, e.g., Jacob Bastian, *How Would a Permanent 2021 Child Tax Credit Expansion Affect Poverty and Employment?*, 77 NAT'L TAX J. 263, 264 (estimating that permanent extension of the ARPA CTC would decrease parent poverty by twenty-one percent and child poverty by twenty-eight percent); ZACHARY PAROLIN ET AL., COLUM. UNIV. CTR. ON POVERTY & SOC. POL'Y, *THE POTENTIAL POVERTY REDUCTION EFFECT OF THE AMERICAN RESCUE PLAN 1 (2021)* [hereinafter PAROLIN ET AL., *THE POTENTIAL POVERTY REDUCTION EFFECT*], <https://perma.cc/2GRJ-6LKW> (estimating that ARPA would reduce child poverty by fifty-six percent); ZACHARY PAROLIN ET AL., COLUM. UNIV. CTR. ON POVERTY & SOC. POL'Y, *MONTHLY POVERTY RATES AMONG CHILDREN AFTER THE EXPANSION OF THE CHILD TAX CREDIT 1 (2021)* [hereinafter PAROLIN ET AL., *MONTHLY POVERTY RATES*], <https://perma.cc/H568-ZJGD> (estimating that the first advance CTC payment under the ARPA reduced the monthly child poverty rate for July of 2021 by up to forty percent); ZACHARY PAROLIN, SOPHIE COLLYER & MEGAN A. CURRAN, COLUM. UNIV. CTR. ON POVERTY & SOC. POL'Y, *ABSENCE OF MONTHLY CHILD TAX CREDIT LEADS TO 3.7 MILLION MORE CHILDREN IN POVERTY IN JANUARY 2022*, at 1 (2022), <https://perma.cc/2QEX-PN7S> (estimating that the expiration of APRA increased the child poverty rate by forty-one percent); Wimer et al., *supra* note 16, at 1.

198. See, e.g., Zachary Parolin et al., *The Effects of the Monthly and Lump-Sum Child Tax Credit Payments on Food and Housing Hardship*, 113 AM. ECON. ASS'N PAPERS & PROC. 406, 409 (2023) [hereinafter Parolin et al., *The Effects of the Monthly and Lump-Sum Child Tax Credit Payments*] (estimating that the ARPA CTC reduced food insufficiency by twenty percent); Jordan M. Rook et al., *Changes in Self-Reported Adult Health and Household Food Security with the 2021 Expanded Child Tax Credit Monthly Payments*, 4 JAMA HEALTH F., at 1 (2023) (finding that survey respondents eligible for the ARPA CTC were 1.9 percentage points more likely to report being food secure); Shafer et al., *supra* note 127, at 1 (2022) (finding that the introduction of the ARPA CTC was associated with a drop in household food insufficiency of twenty-six percent); Allison Bovell-Ammon et al., *Association of the Expiration of Child Tax Credit Advance Payments with Food Insufficiency in US Households*, 5 JAMA NETWORK OPEN, at 1 (2022) (estimating that the expiration of the ARPA CTC increased food insufficiency in households with children by twenty-five percent); Leah Hamilton et al., *The Impacts of the 2021 Expanded Child Tax Credit on Family Employment, Nutrition, and Financial Well-Being 4* (Brookings Glob. Working Paper, Paper No. 173, 2022) (finding that households eligible for the ARPA CTC were 1.3 times more likely to increase fruit consumption, 1.5 times more likely to increase meat and protein consumption, and 1.4 times more likely to report an increased ability to afford balanced meals).

199. See, e.g., Parolin et al., *The Effects of the Monthly and Lump-Sum Child Tax Credit Payments*, *supra* note 198, at 410 (estimating that the ARPA CTC reduced the likelihood that households were behind on rent payments by at least ten percent); Natasha V. Pilkauskas, Katherine Micheltore & Nicole Kovski, *The 2021 Child Tax Credit, the Living Arrangements and Housing Affordability of Families with Low Incomes 19* (Nat'l Bureau of Econ. Rsch., Working Paper No. 31339, 2023) (estimating that the ARPA CTC reduced the amount of past rent and mortgage payments due by thirteen percent).

200. Lindsey Rose Bullinger & Angela Boy, *Association of Expanded Child Tax Credit Payments with Child Abuse and Neglect Emergency Department Visits*, 6 JAMA NETWORK OPEN, at 1 (finding that the ARPA CTC was associated with statistically significant reductions in emergency department visits related to child abuse or neglect among male and non-Hispanic white children); Rook et al., *supra* note 198, at 6-7 (finding that adult survey respondents eligible for the ARPA CTC were three percentage points more likely to report excellent or very good health).

201. See, e.g., Akansha Batra, Kaitlyn Jackson & Rita Hamad, *Effects of the 2021 Expanded Child Tax Credit on Adults' Mental Health: A Quasi-Experimental Study*,

new analytic frameworks that quantify things that are inherently hard to quantify using the latest methods, data, and technologies, and to draw attention to a plausible range of results, further countering the biases and capacity constraints that prevent scorekeepers from providing such ranges.²⁰² It is this capacity that the Budget Lab at Yale seeks to build out, not simply reproducing the work of scorekeepers, but taking it a step further with new approaches that account for the biases built into the scorekeeping process.

Beyond new academic work, further reform is needed, and scorekeepers, too, have a role to play in improving the neutrality of scores in the policy process. Just as it is impossible to precisely estimate policy effects, it is impossible to precisely specify an ideal scorekeeping process, as there are pros and cons associated with all potential approaches. However, we offer a few strawmen for consideration.

First, we propose that as a default matter, scorekeepers provide alternative estimates with longer-term budget windows, such as twenty years or fifty years. There is no obvious answer for what the right length is, and it is certainly true, as others have pointed out, that uncertainty goes up the farther in the future scorekeepers attempt to estimate.²⁰³ But it is equally true that certain policies have effects with clear directions over long time horizons, even if the magnitude

42 HEALTH AFFS. 74, 77 (2023) (estimating that the ARPA CTC is associated with a 1.7 percentage point decline in depressive symptoms and a 3.4 percentage point decline in anxiety symptoms among adults); Nicole Kovski et al., *Unconditional Cash Transfers and Mental Health Symptoms Among Parents with Low Incomes: Evidence from the 2021 Child Tax Credit*, 22 SSM—POPULATION HEALTH, at 7 (2023) (finding that the ARPA CTC is associated with a 0.8 percentage point decline in anxiety symptoms and a 0.6 percentage point decline in depressive symptoms among non-Hispanic Black parents); Eunho Cha, Jiwan Lee & Stacie Tao, *Impact of the Expanded Child Tax Credit and its Expiration on Adult Psychological Well-Being*, 332 SOC. SCI. & MED., at 1 (finding that the ARPA CTC significantly reduced the percentage of adults experiencing mild symptoms of psychological distress, especially among women and Hispanic adults). *But see* Benjamin Glasner et al., *No Evidence the Child Tax Credit Expansion Had an Effect on the Well-Being and Mental Health of Parents*, 41 HEALTH AFFS. 1607, 1609, 1611-13 (failing to find a short-term effect of the ARPA CTC on life satisfaction, anxiety and depression).

202. Capacity constraints appear to explain why scorekeepers do not go beyond their mandate to produce ranges of estimates. As Doug Elmendorf noted during his tenure as CBO director, because scorekeeping models are derivative of the budget rules that govern them, there are “limitations currently on our ability to quantify uncertainty and to help legislators make effective use of such quantification” since “[m]ost of the agency’s models and estimating techniques do not readily yield estimates of uncertainty.” Douglas W. Elmendorf, Dir., Cong. Budget Off., Presentation at the Hutchins Center on Fiscal and Monetary Policy at the Brookings Institution: Communicating Uncertainty in Budgetary and Economic Estimates (Dec. 15, 2014). Economist Alan Auerbach, who previously worked at JCT, has made a similar argument, pointing out that “[f]or estimates formed under constraints of time pressure and incomplete data, how is one to calculate a confidence interval? These difficulties also help to explain why estimators rarely offer a detailed description of the methodology used to construct estimates.” Alan J. Auerbach, *Dynamic Revenue Estimation*, 10 J. ECON PERSP. 141, 145 (1996).

203. *See, e.g.*, Robert McClelland, *Why We Should Stick with a 10-Year Budget Window*, TAX POL’Y CTR. (Aug. 9, 2017), <https://perma.cc/6Q9R-V374>.

of these effects is uncertain, and an approach to scorekeeping that routinely ignores such long-term effects is simply incomplete. This suggests that the right approach may be to extend the budget window only in cases where there is good reason to believe there are meaningful out-of-budget-window impacts. This would be the case for policies *designed* to have such long-run impacts, such as policies to reduce carbon emissions or to ensure that children today grow up to be healthy, productive members of society tomorrow. Currently, members of Congress can request long-term estimates, but compared to scorekeepers, they are less likely to appreciate when long-term effects are important, and more likely to want to highlight or minimize those effects for political reasons. This suggests value in updating the default approach.

We secondly suggest automatically bringing dynamic scoring to spending proposals, consistent with the approach advocated by former CBO director Doug Elmendorf.²⁰⁴ As Elmendorf has noted, dynamic scoring is quite complex.²⁰⁵ But scorekeepers are capable of producing dynamic estimates, as evidenced by their approach to tax legislation. And as Elmendorf further points out, there is no reason to think that scorekeepers' dynamic estimates are less accurate than other estimates, which may involve complicated feedback effects.²⁰⁶

Moreover, when existing dynamic estimates are compared against their respective conventional estimates, the differences are stark. Elmendorf points out that spending contained in the American Recovery and Reinvestment Act of 2009 was estimated to stimulate GDP by roughly twenty-five percent of the Act's revenue cost.²⁰⁷ Certain investments are easily characterized by their macroeconomic effects—boosting labor productivity, spurring capital investment, increasing the labor force. These obvious effects should not be ignored by the scorekeeping process, especially when they are taken into account when generated through tax legislation.

Third, we recommend that CBO more seriously engage with the benefits of proposed policies, at a minimum providing distributional estimates in the same way that JCT does for tax proposals. This would provide visibility into how

204. Elmendorf, *supra* note 71, at 93 (recommending the use of dynamic scoring to estimate “major proposals, defined as those that would have a large estimated budgetary impact excluding macroeconomic effects, and when estimates of such effects are requested by the chair or ranking member of the House or Senate Budget Committee”).

205. *Id.* at 105 (emphasizing that dynamic scorekeeping should only be undertaken when estimators have sufficient time and resources); *see also* Elmendorf, Hubbard, & Williams, *supra* note 177, at 95-97, 111.

206. Elmendorf, *supra* note 71, at 108-09 (“[CBO’s and JCT’s] methodology for conducting macroeconomic analysis reflects the consensus of informed professional thinking, and that consensus provides a useful, albeit imperfect, basis for predicting the macroeconomic effects of legislative proposals. Moreover, there is no reason to believe that their estimates of macroeconomic effects are generally less accurate than their estimates of other complex proposals, although certainly the agencies should continue to strive to improve their analyses.”).

207. *Id.* at 108.

public dollars spent are distributed across the income distribution, and for dynamic effects, where GDP gains associated with proposals are likely to provide benefits.

The limitation of distributional estimates, including existing JCT estimates, is that they remain focused on variables with dollar signs, neglecting other relevant variables, as we saw with the absence of estimates of child poverty in official estimates of the Child Tax Credit. This suggests that in certain cases, scorekeepers ought to consider other variables that are important to understanding policy effects.

It is worth remembering that scorekeepers have previously modeled new variables to accommodate the analytical needs of new policy problems. Specifically, in the early 2000s, as scorekeepers recognized the growing need to help lawmakers understand the effects of health policy, CBO developed a new microsimulation model focused on health policy.²⁰⁸ Crucially, this model produced not only cost estimates illustrating the effect of health policy on the federal budget, but also estimates of how different policies would affect the number of uninsured individuals. This was essential in understanding the effects of the Affordable Care Act, which was promoted specifically for its effects on the uninsured rate.²⁰⁹ Indeed, for a number of policies, the production of a cost estimate requires estimation of a policy's broader effects, and in such cases transparency around these essential scorekeeping inputs—for example, the number of children expected to enroll in a new preschool initiative—will go far in illustrating policy benefits.

Developing new models to study new outcomes is not an easy task, however, and we are sensitive to the fact that scorekeepers are already capacity constrained. This again suggests a role for outside voices to help inform a broader policy debate that considers variables of interest not necessary for scorekeepers to examine given their traditional mandate, such as child poverty. A number of groups in addition to the Budget Lab at Yale are already innovating in this area: Opportunity Insights seeks to expand the current cost-benefit approach to policy evaluation using a marginal value of public funds framework, resulting in a

208. See Anna Wilde Mathews, *CBO Chief Is Health-Care Referee*, WALL ST. J. (Apr. 21, 2008, at 00:01 ET), <https://perma.cc/57A7-6UAA> (quoting then-CBO director Peter Orszag in 2008 as describing health issues as “our fiscal future, and policymakers do not have as much analysis and options as they would need to make sound long-term decisions . . .”); Jessica S. Banthin, Deputy Assistant Dir., Cong. Budget Off., Methods Workshop Presentation at the 2014 Academy Health Annual Research Meeting: Microsimulation of Demand for Health Insurance: A Method Based on Elasticities (June 9, 2014) (describing the evolution and methodologies of the CBO health model).

209. See Barack Obama, *Remarks by the President on the Affordable Care Act*, WHITE HOUSE OFF. PRESS SEC'Y (Oct. 20, 2016, at 13:51 EDT), <https://perma.cc/2NU2-PQPF> (“So because of this law, because of Obamacare, another 20 million Americans now know the financial security of health insurance. So do another 3 million children, thanks in large part to the Affordable Care Act and the improvements, the enhancements that we made to the Children’s Health Insurance Program. And the net result is that never in American history has the uninsured rate been lower than it is today.”).

“unified welfare analysis” that aims to capture the benefits produced by public investments.²¹⁰

It is not enough, however, for outside groups to innovate. There must also be meaningful interaction between official scorekeepers and outside experts. When official scorekeepers put out an estimate, the stakes are very high. Thus, when external innovations are relevant to official scorekeeping analysis, there needs to be a way for such innovations to be directly included in the scorekeeping process. Analysts at CBO and JCT routinely consult with external experts,²¹¹ put out calls for research,²¹² and engage directly in academic debates.²¹³ These interactions could be expanded and formalized, perhaps with official scorekeepers raising a certain number of questions each year and publicizing changes to past approaches that are informed by current research.

Some may react to these suggestions not as revisions to the scorekeeping process, but as championing a broader approach to policy analysis writ large. After all, the narrow remit of scorekeepers is revenue estimation, and even this leaves scorekeepers capacity constrained. This is a reasonable critique, but it underscores the fact that the mandate of scorekeepers, given historical developments, is severely limited, and broader policy analysis would fill the informational gaps left open by scorekeeping’s limitations and correct the normative biases that result from its historical evolution. Perhaps what is needed, then, is a new official office that is focused on this type of broader evaluation. Our focus here is on the primacy of scorekeeping and limitations of current approaches, and we thus suggest reforms in that vein, but we acknowledge that scorekeeping is not the be all end all of approaches to improve the analytical status quo.

These are undoubtedly hard questions. But they are also ones where the stakes are incredibly high, and yet unlikely to lessen in the years to come. Those concerned with policymaking cannot focus simply on the best policies, especially when the factors informing views on what optimal policy is are not consistently considered in the process through which policies are analyzed, despite the fact that policymakers themselves control this process. More attention is needed on the path whereby policy is enacted. And as the deficit picture only becomes bleaker, scorekeeping will matter more and more, making

210. See Nathaniel Hendren & Ben Sprung-Keyser, *A Unified Welfare Analysis of Government Policies at 1-2* (Feb. 2020) (unpublished manuscript) (on file at <https://perma.cc/BV4K-ZGMJ>).

211. *10 Things to Know About CBO*, CONG. BUDGET OFF. (Jan. 7, 2025), <https://perma.cc/A6RY-WV5N> (describing the expertise of CBO staff and those that they regularly consult with); CONG. BUDGET OFF., *HOW CBO AND JCT ANALYZE MAJOR PROPOSALS THAT WOULD AFFECT HEALTH INSURANCE COVERAGE* 3, 6-7 (2018), <https://perma.cc/MK6V-YT48> (referencing CBO and JCT’s practice of consulting with outside experts and reviewing existing evidence).

212. See, e.g., John McClelland, *A Call for New Research in the Area of Taxes and Transfers*, CONG. BUDGET OFF. (July 25, 2023), <https://perma.cc/8NBU-NRUC>.

213. See, e.g., *supra* notes 97-99.

deeper thought about revenue impacts and policy analysis crucial. Though the system today is in many ways broken, we believe that there is hope for repair—though overhaul will require concerted effort by academics, policymakers, and the scorekeeping community alike.

CONCLUSION

Scorekeepers are celebrated today as providing neutral, objective analysis. While scorekeepers themselves are certainly able and unbiased analysts, the process of scorekeeping is not. In this Article, we highlight how the deficiencies of scorekeeping not only introduce noise and error into the process of government decision-making—but also systematically bias against policies that invest in the future and promote distributional equity. Lawmakers today are no longer constrained by the budget rules of earlier decades. But this central bias of scorekeeping remains, and the importance of scorekeeping in the policymaking process has only grown.

Modern scorekeeping was not designed to provide holistic or even comprehensive policy analysis. It was designed to track progress against revenue targets in the presence of high and rising peacetime deficits. Policies that “score” well today are not necessarily the best policies, they are simply the ones that do the most to reduce short-term deficits—shrinking government capacity rather than expanding it, saving money now even if costs will be reaped in the future.

We believe that this disproportionate focus on deficits is both severely harmful and plainly illogical. When the private sector analyzes potential new investments, these expenditures are viewed through a net present value framework that considers the up-front cost of the investment alongside the anticipated benefits the investment is expected to deliver in years to come. The federal government takes a completely different—and irrational—approach, ignoring long-term and hard-to-quantify benefits and instead focusing solely on costs.

Our hope is that this Article serves as a call to develop new frameworks and approaches to policy evaluation, enabling policymakers to make necessary investments for future generations. As the need for these crucial investments mounts, it becomes ever more important to get the benefit-cost calculus right and appreciate through the scorekeeping process long-term investments for what they are: sustainable policies that will ensure a more prosperous future. Policymaking that prioritizes long-term fiscal sustainability and social welfare will require us to update the scorekeeping rules by looking to the future rather than the past, overcoming the deficit preoccupations of yesterday and confronting the urgent policy challenges of tomorrow.