Ending Wasteful Year-End Spending: On Optimal Budget Rules in Organizations[*]

Christoph Siemroth[†]
University of Essex

February 7, 2022

Abstract

What can organizations do to minimize wasteful year-end spending before the annual budget expires? I introduce a two-period model to derive the optimal budget roll-over and audit rules. A principal tasks an agent with using their budget to fulfill the organization's spending needs, which are private information of the agent. The agent can misuse funds for private benefit at the principal's expense. The principal decides upfront which share of unused funds the agent can roll over to next year, and which spending amounts to audit in order to punish fund misuse. The optimal rules are to allow the agent to roll-over a share of the unused funds, but not necessarily the full share, in most cases to audit only sufficiently large spending, and to exert maximum punishment if fund misuse is detected. An extension with endogenous budget levels shows that strategically underfunding the agent can be optimal.

Keywords: Auditing, Budget Carry-Forward, Budget Roll-Over, Fund Misuse, Moral Hazard, Year-End Spending
JEL Classification: D82, G31, H50, H83

[*] I thank Yair Antler and Albin Erlanson for thoughtful discussions on this topic, as well as Stuart Baumann, Daniel Garrett, Jayant Ganguli, Michael Gibbs, Piero Gottardi, Hans Peter Grüner, Wim van der Stede, and seminar audiences at Essex and Mannheim for useful comments. This work was supported by the Economic and Social Research Council [grant number ES/T006048/1].

[†] University of Essex, Department of Economics, Wivenhoe Park, Colchester, CO4 3SQ, UK. E-mail: christoph.siemroth@essex.ac.uk.

Non-technical summary and policy recommendations

This study investigates how to minimize wasteful year-end spending in organizations. It applies to any situation where a principal grants an annual budget and delegates spending decisions to an agent to fulfill the principal's mission. Examples are an organization giving funds to the IT department to keep computers running, or a CEO giving funds to the marketing department to increase sales or brand recognition.

If budgets expire at year-end, anecdotal and empirical evidence shows that agents spend sizable parts of their budgets at the end of the fiscal year on unneeded or low priority items, thus wasting funds. This "use-it-or-lose-it" behavior is exacerbated if agents expect future budgets to be cut in case of unused funds. Principals should therefore publicly commit to not cut budgets if funds are returned or saved.

Aiming at changing the agent's incentives, this study investigates three instruments to curb wasteful year-end spending: (1) Allowing unused funds to be rolled-over to next year, (2) auditing the agent's spending and punishing in case of fund misuse, and (3) deliberately underfunding the agent, i.e., setting a low budget.

First, allowing fund roll-over can be effective in preventing wasteful spending if there is a non-trivial chance that next year's budget is not enough to cover all spending needs otherwise. Unlike auditing, allowing fund roll-over is a relatively cheap instrument. And it has few drawbacks: It can induce agents to save unneeded funds rather than to misuse them, so in the worst case these funds are misused next year, which is no worse than misuse this year. Hence, the recommendation is that fund roll-over should be allowed. It can be optimal to allow only a partial roll-over of funds, so that for example only 75% of unused funds are rolled over to next year, and the remaining 25% are returned to the principal at year-end. However, for most parameter values allowing a full fund roll-over is optimal.

Second, if the agent has a sufficiently generous budget and audit costs are not too high, then auditing high spending amounts after the fiscal year, and harshly punishing fund misuse if uncovered by the audit, is optimal. Low spending amounts, which are far below budget and suggest an efficient use of funds, should not be audited to save audit costs. If agents are underfunded, then there is little room for waste and no auditing is needed. Auditing large spending is also unnecessary if fund roll-over is effective. The audit policy should be announced in advance, since the goal is to deter fund misuse rather than punish it.

Third, setting a low budget to prevent wasteful spending might do more harm than good. It implies that there is a chance the agent is unable to afford essential items to fulfill his mission. Hence, it might reduce wasteful spending at the cost of also preventing useful spending, so should be used with caution. Underfunding the agent can be optimal if funds in the organization are scarce or if the agent's mission is of a low priority.

Finally, a first step in combating wasteful spending is making agents aware of what sort of spending is wasteful, otherwise they cannot change their behavior in the right direction.

1 Introduction

Many agencies or organizational divisions are granted annual budgets by their principals—which expire at the end of the fiscal year—to fulfill their mission. A public sector example is US Congress (principal), which grants funds and tasks the Department of Defense (agent) with defending the nation. In the private sector, CEOs give their heads of marketing a budget to promote the firm brand. In universities, deans or grant funders give funds to academics for research, staff, and conferences.

Funds that are not spent by fiscal year-end are lost to the agent. Consequently, a "use it or lose it" mentality can be observed among agents, with a rush to spend the remaining funds in the last month or week of the year. Such year-end spending sprees are well documented in the public sector (e.g., Liebman and Mahoney, 2017; Baumann, 2019 and references therein), and anecdotes suggest the phenomenon is also common in the private sector (e.g., Digiday, 2017 for marketing).

Such expiring budgets can therefore have undesirable consequences. First, the agent might hastily spend the remaining funds on low value items (e.g., Liebman and Mahoney, 2017), because there is no useful purpose left at year-end, or there is no time to execute the spending well (e.g., because the best contractor is unavailable on short notice). Second, the agent might spend the funds on currently unneeded but durable "assets," in the hopes the expiring budget can still yield a benefit in the future. For example, Hurley et al. (2014) report a case where a military officer was ordered to buy a train-wagon-load of toilet paper at year-end. While such tricks might not be a complete waste of funds, toilet paper as a currency is less fungible and storable than money, and hence loses value. Third, the agent might misuse the funds for personal gain and little value to the principal. An example might be a $9000 chair bought by the Pentagon at year-end (Military Times, 2019), where one would think a $5000 chair would have done as well. Another example is academic conferences at beach or ski resorts, which have an unusually light session schedule. Or new gadgets bought to satisfy the curiosity of a tech nerd working in an organization's IT department without benefit to the organization. The model in this paper is motivated by this third interpretation, but is also consistent with the first and second.

These examples show there are cases where funds spent at year-end could have been put to better use by the principal (if funds had been returned) or by the agent in the future (if given the flexibility). That is, the use of funds by the agent is not efficient under annual expiring budgets. In a survey among US Department of Defense staff responsible for spending, 95% said there was a problem with year-end spending (McPherson, 2007). On average, interviewees estimated that 32% of year-end spending was on low priority items or at least partially wasted.

But what other rules could be adopted to incentivize agents to use funds more efficiently? Should agents be allowed to keep unused funding, and how much of it? Can fund roll-over replace costly auditing to reduce inefficient fund use? Under which conditions should agent spending be audited? To study these questions, this paper introduces a new model to determine the optimal budget roll-over and audit rules that principals can adopt to improve spending efficiency. This is an issue of great importance for most organizations where spending has to be delegated. If inefficient fund use could be reduced, then organizations could achieve the same at a lower cost, or could achieve more at the same cost.

In practice and in the model, the source of the principal-agent conflict is that the principal does not know the agent's exact spending needs, so grants the agent some discretion in spending. The agent does not know of better uses for unneeded funds outside of his agency, nor does he take into account the cost of these funds like the principal does. Hence, the agent spends everything even if it is not needed to fulfill the principal's task.[1]

This paper models the principal-agent interaction as a game of asymmetric information. The model has two years, and the agent receives an exogenous budget in each of them, which expires at year-end. The agency's spending need θ_y in year y is a random draw from a continuous distribution. This captures that future spending needs are uncertain, for example it is unclear how many computers will break down in the agency's office. The realization θ_y is privately observed by the agent in year y, who then decides on a spending amount for that year, subject to the budget constraint. Any spending up to the spending need θ_y fulfills those spending needs and generates a high value for the principal, but any spending above θ_y is a misuse of funds, as it yields no value to the principal. The agent receives a high marginal value from fulfilling spending needs, and a lower but positive marginal value from spending more. Hence, the agent wants to fulfill his mission first and foremost, but also values the $9000 office chair or the "conference" at the beach, whereas the principal does not.

Without further additions, this model generates a year-end spending surge as observed in practice: As budgets expire at year-end, the agent rationally spends everything and misuses funds, even if the spending need that year is considerably below the budget. This is an undesirable outcome for the principal, as any spending above need generates no value, but the funds are costly to her (e.g., due to credit costs or opportunity costs).

To investigate measures to mitigate this waste of funds, the principal commits to the following rules before the agent moves. First, the principal sets a roll-over rule $\Delta \in [0, 1]$, the share of unused funds that will be rolled-over to the second year, whereas share $1 - \Delta$ is returned to the principal. Hence, for any unused dollar at the end of the first year, Δ

[1] At least two more motives for excessive year-end spending are mentioned in the literature. First, the agent's fear to have his budget cut next year if not all funds are used, which is known as the ratchet effect (e.g., Freixas et al., 1985). Second, US politicians appear to see unused funds as the agent not doing his job, applying pressure to spend, and thus providing a disincentive for agents to use their funds efficiently. While this may sound silly to economists—viewing more spending as desirable irrespective of the value it generates or its opportunity costs—it may be rational from a political economy point of view. As unused federal funding reverts to the treasury only after 5 years, the only way to get a benefit for the constituency in the current 4 year term is to pressure agencies to spend their annual budgets (e.g., McPherson, 2007).

dollars are added to the agent's budget in the second. Second, the principal sets an audit rule for each year, and these rules specify for which spending amounts a *costly* audit will be triggered, and for which no audit will be triggered. An audit is the only way for the principal to observe the spending need θ_y and thus to identify fund misuse. The audit happens only once spending for the year is done by the agent and if the audit rule prescribes an audit for that year's spending amount. Third, the principal sets the costly punishment inflicted on the agent if he is audited and caught misusing funds. The combination of roll-over rule, audit rules, and punishment that jointly maximizes the principal's ex ante utility, taking into account the agent's reaction to these rules, is the optimal policy.

Based on this model, I find that the optimal roll-over rule features some, but not necessarily full roll-over. That is, a positive share of the unused funds should be available to the agent next year, but not necessarily the full share. This possibility of fund roll-over gives the agent a reason to save rather than misuse unneeded funds. Hence, as sometimes suggested by practitioners (e.g., Jones, 2005; McPherson, 2007), allowing for fund roll-over can reduce fund waste and increase principal utility in the model. But there has been no analysis of an optimal roll-over rule, nor has partial roll-over been suggested as preferable to full roll-over. The optimal roll-over share is weakly increasing in the agent's marginal value from fund misuse and in the budget. Partial roll-over can be optimal if the agent receives little value from fund misuse and if the principal's audit and fund costs are high.

The intuition why partial roll-over can be optimal is as follows. Suppose there is no auditing. If there are very few spending needs in year 1, then if all unspent funds are rolled-over ($\Delta = 1$), the agent would only save some for roll-over and misuse the rest. This is because the agent needs only so many additional funds to cover most expected spending needs next year, so at some point fund misuse is more attractive than rolling over more. But if the roll-over rule "taxes" the roll-over ($\Delta < 1$), then it forces the agent to save more in order to have the same amount of funds available next year. This additional saving crowds out fund misuse, which benefits the principal. Consequently, taxing the roll-over can sometimes, but not in all cases, be optimal. In practice, it would also address concerns that agents accumulate too many savings over time.

With one exception, the optimal audit rules are threshold rules, which audit all yearly spending above a threshold (which may differ between both years), but do not audit below the threshold. The optimal punishment is harsh enough so that no agent wants to be caught misusing funds. Not only does a large punishment make audits a perfect deterrent, it also saves the principal punishment costs. In effect, audits happen in equilibrium, but any agent who is audited has legitimately spent a lot due to a large spending needs realization, so there is no punishment in equilibrium. The principal only audits large enough spending amounts, because auditing is costly and because the expected fund misuse conditional on small observed spending is small. Consequently, there is still some scope for fund misuse for agents in years with low spending needs while staying under the audit threshold.

The intuition why threshold rules are optimal is as follows. Suppose, to the contrary, the audit rule was to audit spending above \$3m, not to audit spending between \$3m and \$2m, and then again to audit between \$2m and \$1m. This is not a threshold rule due to the gap between the audit regions. If the spending need is only $\theta_y = \$1$m, then the agent would misuse funds to spend \$3m overall, but still not get audited. In fact, the agent would get away with fund misuse for any spending need realization $\theta_y < \$3$m, and the audit region between \$2m and \$1m is completely ineffective, i.e., spending in that region never happens. Hence, any audit rule R is in outcome equivalent to a threshold rule, where the threshold is the largest spending amount that is not audited under R. Consequently, the optimal audit rule can be represented as a threshold rule.

Under the optimal threshold audit rules, the principal tends to audit more (i.e., for smaller spending amounts) the lower the cost of auditing and the larger the cost of funds. A larger cost of funds implies a larger loss for the principal from fund misuse, hence she is willing to audit more to prevent misuse. Moreover, if the annual budget is enough to cover all potential spending needs, then the optimal audit thresholds for both years are identical, otherwise the principal tends to audit more in the first than in the second year. This is because auditing in the first year not only discourages fund misuse, but also leads to more fund roll-over, which helps to satisfy more spending needs in the second year. For small budgets, the principal tends not to audit at all, because there is little scope for fund misuse by the agent, as the probability of the spending needs realizing below the budget is small. For larger budgets, the principal tends to audit as long as audit costs are not too large and the cost of funds is large enough.

There is one exception where the audit rule is not a threshold rule in the first year, if the roll-over rule is effective in inducing the agent to save funds for roll-over. In this case, the agent does not spend his entire budget in the first year—unless the spending needs of the agency require it—even if there is no auditing, because he wants to have more funds available next year. The optimal audit rule in this case is an interval rule, which audits spending amounts just above and below what a saving agent would spend absent auditing. Consequently, the optimal interval rule audits spending amounts in the interior of the budget set, and might not audit the very largest spending amounts close to the total budget, where the roll-over incentive prevents fund misuse. Hence, the roll-over rule non-trivially interacts with the optimal audit rule. An interval rule, like a threshold rule, induces the agent to spend above a threshold if and only if a large spending needs realization requires it. But the interval rule might save costs by not auditing all spending amounts above the threshold.

In an extension, I endogenize the amount of the annual budget, to determine if it can be optimal to strategically underfund the agent. And indeed, if the ratio of audit costs to cost of funds, and the cost of funds, are large enough, then it is optimal for the principal to grant a budget that is smaller than the maximum possible spending need realization. The smaller budget implies there is a positive probability the agent will not be able to meet all spending needs. Still, this is better for the principal than a large budget with a high chance of fund misuse, as auditing is too costly. In cases where either the audit costs or the cost of funds are sufficiently small, however, a large budget enough to fulfill all spending needs is optimal with or without extensive auditing, respectively.

This model does not exhibit a ratchet effect, in the sense that agents spend more because they worry their budget next year will be cut otherwise. This is because the spending need realizations in both years are independent, and the agent does not know the second year realization when spending in the first year. Nevertheless, without proper policy the model generates wasteful year-end spending as observed in practice, hence it is useful and suitable to investigate policies to improve spending efficiency.

The policy recommendations for organizations trying to reduce wasteful year-end spending in practice are as follows. First, rolling over unused funds should be allowed, as it is weakly better to do so. The worst that can happen is that funds are misused next year rather than this year, which is no worse than the status quo with use-it-or-lose-it behavior. And allowing roll-over might in fact improve spending quality, by fulfilling spending needs next year which could not have been fulfilled without the roll-over. Moreover, while auditing requires additional manpower, allowing fund roll-over is relatively cheap. Second, auditing and punishing fund misuse can help for sufficiently small audit costs. Usually only large spending amounts should be audited, but not small spending amounts far below the budget, which indicate funds were likely used in the principal's interest. If the roll-over incentive is effective in inducing some savings, then auditing the very largest spending amounts could be unnecessary. If audit costs are large or the cost of funds are very small, then no auditing is optimal. Third, deliberately setting a low budget—which implies there is a positive probability not all spending needs can be fulfilled—can be optimal if the audit costs and the cost of funds are large, or the agent's mission is of low priority.

1.1 Literature

This paper contributes to the year-end spending literature. Liebman and Mahoney (2017) is one of the most important studies in this literature, with both theoretical and empirical contributions. Empirically, they document US federal procurement year-end spending surges, and that IT spending made in the last week of the fiscal year is of lower quality than usual. They also show that allowing some roll-over of unused funds to the next year reduces year-end spending surges. Liebman and Mahoney (2017) also develop a model in which a precautionary savings motive for the agent generates more spending of less value at year's end. There is no fund misuse nor disagreement about the value of spending; rather, the principal-agent conflict consists in agents not taking into account the cost of funds. They show that allowing roll-over can increase welfare. In this paper, I contribute the first analysis of the optimal roll-over rule in this context as well as an analysis of optimal audit rules, which interact with the roll-over rule, whereas there is no auditing in their model.

Baumann (2019) also empirically documents year-end spending surges in UK governmental agency spending, and investigates whether a precautionary savings motive is causing them. He finds that an alternative explanation—that agencies procrastinate in spending their funds—also plays a role. He suggests that spending later in the fiscal year should be more costly than spending earlier in the year, to counteract the tendency to procrastinate.

Hurley et al. (2014); Brimberg and Hurley (2015) propose models with expiring budgets where investments need some preparation, so funds cannot simply be spent at year's end when all important spending needs are known with certainty. They show that a rational risk-neutral planner who aims to maximize the value of the spending will—due to the uncertainty—be conservative in spending early and often have unused funds at the end of the year. They also show that pressure to minimize unused funds at the end of the year leads to lower value spending. Their studies are critiques of expiring budgets, but their focus is not on analyzing alternatives, which is what I add in this paper.

There are studies in mechanism design which show how to deal with principal-agent conflicts, although these more abstract models typically deviate somewhat from the budget setting. Bird and Frug (2019) investigate how an agent can be incentivized to provide effort to implement investments, which are beneficial to the principal but costly to the agent. Rewards are costly to the principal but beneficial to the agent. They assume that investment and reward opportunities arrive stochastically over time, and are observable for the principal only once undertaken by the agent. Otherwise, only the agent observes these opportunities. They show the optimal mechanism is to allow the agent to take reward opportunities for a limited time after implementing investments for the principal. While their model is not explicitly about budgets or year-end spending, it might give wasteful year-end spending a new interpretation as rewards for previous agent performance.

My model and those in the year-end spending literature take it as given that the agent receives a budget and some discretion in spending it, and investigate improvements within that structure. This should make the recommendations derived from this model easier to implement in practice. Malenko (2019) takes a step back and asks whether agents should receive budgets in the first place, or whether spending should get micromanaged more and funded by the principal directly. In his environment, there is a principal-agent conflict because the agent has preferences for overspending. The optimal mechanism separates small and large investments, so that small investments are funded at the agent's discretion from his budget, which is replenished over time. Large investments are funded by the principal directly, but only after an audit confirms it is worth to be implemented. If the large project is not worth it, the agent is punished for recommending it. Interestingly, this optimal mechanism does not have annual budgets; in fact, funds never expire. Hence, it could be viewed as an annual budget with complete and indefinite roll-over.

The auditing part of the model is related to the theoretical costly state verification literature. The context of the state verification differs, with the original Townsend (1979) about contingent contracting between two parties, about single object allocation among many agents without monetary transfers (Ben-Porath et al., 2014; Li, 2020), or about the problem of extracting privately known wealth from an agent (Border and Sobel, 1987). The stylized setting of the latter paper finds that agents who report larger wealth are audited less often. Similarly, in my budget setting, agents who spend less (i.e., have more funds remaining) tend to get audited less in the optimal policy, although the relationship here is not monotonic but determined by a threshold. A main difference is that the principal in Border and Sobel (1987) seeks to identify the private information of the agent, whereas the optimal policy in my budget model is about preventing moral hazard. A new insight in the budget context is that allowing fund roll-over can make auditing unnecessary, or reduce the need for auditing. Moreover, the optimal audit rule changes depending on the effectiveness of the roll-over incentives.

2 The model

2.1 Set-up

There are two risk neutral expected utility maximizing players, the principal (she) and the agent (he). This is a very general setting, which applies to most organizations where an agent is given a budget to fulfill a mission, with some leeway in spending it.

Time. Time is discrete. There are two periods in which the agent moves, which I call years, $y = 1, 2$. A minimum of two years is needed to investigate whether allowing to roll-over funds to the next year can mitigate inefficient year-end spending.[2] The year is the time frame for which a budget is granted to be spent by the agent, and in practice many organizations grant budgets for one year. The principal moves in $y = 0$.

State of the world: Spending needs. In every year, a random variable θ_y realizes, which is identically and independently distributed according to a continuous uniform distribution on $[0, u]$, with $u > 0$. Let the density function be g and the cumulative distribution function be G.[3] The agent observes the realization θ_y in year y, hence does not know realization θ_2 in year 1 (i.e., the agent cannot foresee the future). Moreover, θ_y is private information of the agent, so the principal does not observe it. θ_y is the *spending need* arising in year y in the agent's division or agency. For example, the more computer hardware breaks down and needs to be replaced, or the more temporary hires need to be

[2] An earlier version of this model additionally divided the year into two subperiods (the last to be interpreted as year-end), each with an independent spending need realization. Adding these subperiods yielded the same spending and fund misuse decisions as this model, because it was a weakly dominant strategy for the agent to postpone fund misuse to year's end (exactly as observed empirically). To ease the exposition, I dropped this extra division into subperiods. Still, fund misuse in this model can be interpreted as year-end spending.

[3] Because the distributions of the state are known by the principal and the realizations are independent, there is no ratchet effect in this model. That is, the spending amount or realization θ_1 in year 1 does not tell the principal anything new about the realization θ_2 in year 2.

made due to absences, the larger the realization of θ_y. The realization θ_y is what the principal would like the agent to spend (to be made precise in the utility function below), and in one interpretation represents the spending needed for the agent to fulfill his mission.

Agent strategy space. The agent decides how much to spend in every year y. The spending strategy is a mapping from the available budget $b_y \in \mathbb{R}_0^+$ (defined in next paragraph) and the realized spending needs into a spending amount, $s_y : [0, b_y] \times \Theta_y \to [0, b_y]$ for $y = 1, 2$. The agent's spending amount s_y is observable to the principal. The agent cannot spend more than his budget b_y for the year, but is able to spend more than is needed (θ_y) for personal benefit (see the agent utility function below). Hence, I call $s_y > \theta_y$ a *misuse of funds*. Consequently, the single action of setting spending amount s_y simultaneously determines to what degree the principal's mission is fulfilled (by fulfilling spending needs) and to what degree the agent acts against the principal's interests.

Budget and fund roll-over. The agent is exogenously granted an annual budget $b \in (0, u]$. This budget is available to the agent every year. In $y = 2$, the available budget b_2 might additionally include unused and rolled over funds from the previous year, whereas in $y = 1$ only b is available. Thus,

$$\begin{aligned} b_1 &= b, \\ b_2 &= b + \Delta(b_1 - s_1), \end{aligned}$$

where $\Delta \in [0, 1]$ is the fraction of unused funds from $y = 1$ allowed to be rolled over to the next year. The roll-over rule, represented by Δ, is part of the principal's strategy, and is set by the principal at the beginning of the game.

Auditing. At the end of the year (i.e., once s_y is set by the agent), the principal can audit at a cost of $c_A > 0$ to determine if there was a misuse of funds. The audit technology is perfect and reveals θ_y, and thus also the amount of fund misuse $s_y - \theta_y$. The principal commits to a deterministic audit rule in $y = 0$, one for each year, which maps every possible total spending s_y that year into an audit decision, $A_y : [0, b_y] \to \{0, 1\}$. Hence, the agent is fully aware of the audit rules before moving.

Punishment. In case any fund misuse was detected via audit, a punishment $p > 0$ is inflicted on the agent. Punishment is costly for the principal. In practice, for example, suspending an agent will require hiring a temporary replacement, and other punishments like firing will generate legal costs. I assume that reducing the agent's utility by p costs the principal $p \cdot c_p$, with $c_p > 0$. Thus, larger punishments are more costly.[4] The principal commits to p at the beginning of the game.

Principal strategy space. This paragraph summarizes the above actions by the principal. The principal acts in $y = 0$, and commits to audit rules A_1, A_2, a punishment if the audit found misuse of funds p, and a fund roll-over rule $\Delta \in [0, 1]$. The collection

[4] As will become clear later, whether costs scale in punishment or not does not matter for the results.

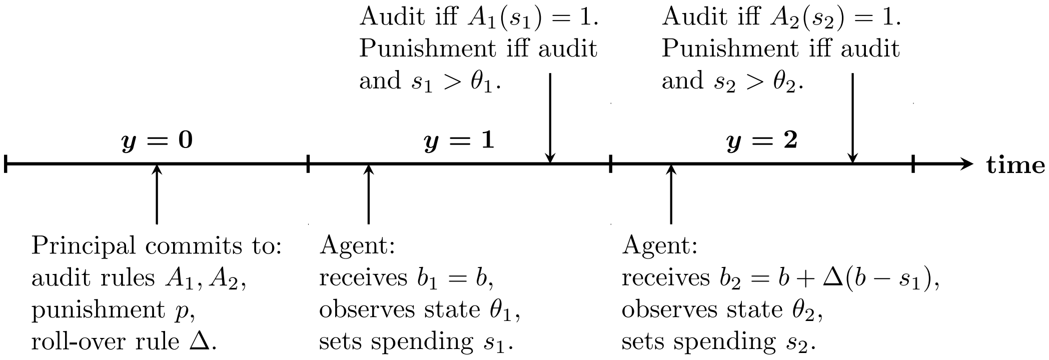

Figure 1: Timing of principal and agent moves, as well as audits and punishment.

$\{A_1, A_2, p, \Delta\}$ will also be called "policy," and the principal's equilibrium strategy is the optimal policy, i.e., the policy that is optimal given the agent strategy (best response). Figure 1 plots the timeline of the model. Commitment is important here, because otherwise there is a time-inconsistency problem with costly auditing.

Agent utility function. The agent cares about fulfilling the spending needs of the principal, i.e., about doing his job, but also cares about additional spending with personal benefits. As described in the introduction, year-end spending is often on items that the principal might not value very highly, but the agent more so. Examples include more luxurious office chairs, new technical gadgets that are fancy but not needed to do the job, or "business trips" to and "conferences" at beach resorts.[5] The agent's utility function is

$$\begin{aligned} U_{\text{agent}} &= \sum_y \left[\min\{s_y, \theta_y\} + \alpha \cdot \mathbf{1}_{\{s_y > \theta_y\}} \cdot (s_y - \theta_y)\right] - \sum_y \left[\mathbf{1}_{\{A(s_y)=1, s_y > \theta_y\}} \cdot p\right] \\ &= \sum_y \left[s_y \cdot \mathbf{1}_{\{s_y \leq \theta_y\}} + \mathbf{1}_{\{s_y > \theta_y\}}[\alpha(s_y - \theta_y) + \theta_y]\right] - \sum_y \left[\mathbf{1}_{\{A(s_y)=1, s_y > \theta_y\}} \cdot p\right], \end{aligned}$$

with $0 < \alpha < 1$, where $\mathbf{1}_{\{.\}}$ is the indicator function. That is, in the first sum, every dollar spent on fulfilling spending needs (i.e., as long as $s_y \leq \theta_y$) yields a marginal utility of 1, whereas every dollar spend above θ_y is fund misuse and yields a marginal utility of $0 < \alpha < 1$. Thus, the agent benefits from additional spending, but not as much as from fulfilling the principal's spending needs.[6] This implies the agent will attempt to misuse funds only once all spending needs are fulfilled. This is consistent with the "use it or lose it" phenomenon of agents observed at the end of the fiscal year, who scramble to spend the remaining funds on something of value to them before they expire, but not at the beginning of the year when fresh funding is available and might crowd out legitimate spending. The

[5] Sometimes this personal spending is not an immediately obvious misuse of funds, which motivates the need for an audit before it can be determined whether that spending really is in the interest of the principal.

[6] While in practice there are undoubtedly agents who value the personal spending more than fulfilling the spending needs of their principal, this is not the focus of this paper. Moreover, it should be possible for a principal to verify whether the agent actually keeps their agency running, and punish if not, whereas it is harder to determine whether too much was spent to keep the agency running, which is the focus of this study.

agent's utility function, moreover, includes the punishment in case of fund misuse and audit (second sum).

Principal utility function. Following Liebman and Mahoney (2017), I model the principal as having a cost of funds $0 < \lambda < 1$. In the framing of a government, this means there is no hard budget constraint, but rather additional funds could be raised via higher taxes or more borrowing, which however comes at marginal cost λ. Similarly, a corporation could change strategy or borrow more at a cost.[7] As a consequence of these costs, the principal only wants spending done with value exceeding λ. Hence, the principal's utility is

$$\begin{aligned} U_{\text{principal}} &= \sum_y \min\{s_y, \theta_y\} - \sum_y \left[\lambda s_y + \mathbf{1}_{\{A(s_y)=1\}}(c_A + \mathbf{1}_{\{s_y>\theta_y\}} \cdot pc_p)\right]. \\ &= \sum_y s_y \mathbf{1}_{\{s_y\leq\theta_y\}} + \theta_y \mathbf{1}_{\{s_y>\theta_y\}} - \sum_y \left[\lambda s_y + \mathbf{1}_{\{A(s_y)=1\}}(c_A + \mathbf{1}_{\{s_y>\theta_y\}} \cdot pc_p)\right]. \end{aligned}$$

Thus, like the agent, the principal receives a marginal utility of 1 from fulfilling spending needs, but unlike the agent, receives no utility from the misuse of funds.[8] Any returned funds are valued at a marginal rate of $0 < \lambda < 1$, or equivalently, the used funds cost λ. The principal-agent conflict is thus about the additional spending above θ_y, but both agree on the spending up to θ_y. Besides the cost of funds, the second sum also includes the costs for auditing and punishment.

Equilibrium concept. I use the standard Perfect Bayesian Equilibrium.

3 Agent reaction and principal utility given agent reaction

3.1 Agent reaction function

Take the principal strategy A_1, A_2, p, Δ as given. Assume the audit rule has a threshold spending value $\underline{a}_1, \underline{a}_2$ for each year, such that there is always auditing above but not weakly below that audit threshold (to be confirmed later), and that $\underline{a}_y \leq b$.

3.1.1 Year 2

For the decision whether to misuse funds in $y = 1$, i.e., whether $s_1 > \theta_1$, first calculate the benefit of rolling over funds to the next year, for which we need the utility and optimal spending strategy in year 2.

[7] The cost of funding can also be interpreted as the opportunity cost of using available funds, e.g., funds could be used by another agent yielding a utility of λ to the principal. However, in this paper I do not explicitly introduce multiple agents.

[8] Results are similar if the marginal utility of the additional spending to the principal is positive but below λ, so that the principal does not prefer the additional spending. Hence, setting the marginal utility to zero simplifies the exposition, but is not crucial for the results.

If the principal sets punishment $p > 2\alpha b$ (to be confirmed later), then the agent will never misuse funds above the audit threshold. In this case, the agent optimally spends

$$s_2(\theta_2) = \begin{cases} \min\{\theta_2, b_2\} & \text{if } \theta_2 \geq \underline{a}_2, \\ \underline{a}_2 & \text{if } \theta_2 < \underline{a}_2. \end{cases}$$

The agent expected utility in year 2, as a function of the principal audit rule $\underline{a}_2$, and given the agent best response, is therefore

$$\int_0^{\underline{a}_2} (\theta_2 + (\underline{a}_2 - \theta_2)\alpha)\mathrm{d}G(\theta_2) + \int_{\underline{a}_2}^{b_2} \theta_2 \mathrm{d}G(\theta_2) + \int_{b_2}^{u} b_2 \mathrm{d}G(\theta_2).$$

3.1.2 Year 1

Now that year 2 spending s_2 is defined as a function of b_2, we can consider year 1 spending s_1. The marginal utility from spending above θ_1, but below the auditing threshold, is α. The marginal benefit from saving funds in the amount of x, rolling over Δx, and thus setting $b_2 = b + \Delta x$ is

$$\begin{aligned} &\frac{\partial}{\partial b_2}\left[\int_0^{\underline{a}_2} (\theta_2 + (\underline{a}_2 - \theta_2)\alpha)\mathrm{d}G(\theta_2) + \int_{\underline{a}_2}^{b_2} \theta_2 \mathrm{d}G(\theta_2) + \int_{b_2}^{u} b_2 \mathrm{d}G(\theta_2)\right] \cdot \frac{\partial b_2}{\partial x} \\ &= \left[b_2 g(b_2) - b_2 g(b_2) + \int_{b_2}^{u} g(\theta_2) d\theta_2\right] \cdot \Delta = (1 - G(b_2))\Delta. \end{aligned}$$

This marginal benefit is clearly non-negative, less than one, and strictly decreasing until $G(b_2) = 1$, in which case it is zero. Determining the amount of rolled over funds x, so that the marginal expected utility from the roll-over equals the marginal utility of misusing funds in year 1:

$$\alpha = \Delta(1 - G(b + \Delta x)) \iff \frac{G^{-1}(1 - \frac{\alpha}{\Delta}) - b}{\Delta} = \hat{x}. \tag{1}$$

Let $\bar{x} = \min\{\max\{\hat{x}, 0\}, b\}$ to ensure there is no negative saving (max) and that not more than the budget is saved (min). Moreover, define $G^{-1}(c) = -\infty$ if $c < 0$, which implies that $\bar{x} = 0$ for any $\Delta < \alpha$. The optimal spending decision by the agent in year 1 is thus

$$s_1(\theta_1) = \begin{cases} \min\{\theta_1, b\} & \text{if } \theta_1 > \underline{a}_1, \\ \underline{a}_1 & \text{if } \theta_1 \leq \underline{a}_1, \underline{a}_1 \leq b - \bar{x}, \\ \max\{\theta_1, b - \bar{x}\} & \text{if } \theta_1 \leq \underline{a}_1, \underline{a}_1 > b - \bar{x}. \end{cases}$$

In the first line, the spending needs are above the audit threshold, so any additional misuse of funds would be detected and punished, and hence does not happen. In the second line, the spending needs are below the audit threshold, and there is enough budget to both misuse funds up to the audit threshold and to roll over enough funds to get all marginal benefits above α in year 2. Hence, there is misuse of funds (until spending equals the audit threshold) and some fund roll-over, but the possibility of fund roll-over does not actually reduce the misuse of funds, only auditing does. In the third line, the spending needs are below the audit threshold, but there is not enough budget to both misuse funds until spending equals the audit threshold and to roll over enough to get a marginal utility of α. Hence, there is less fund misuse than there would otherwise be due to the possibility of rolling over funds (i.e., due to $\Delta > 0$). Thus, the agent deliberately misuses fewer funds to have more funding next year and to fulfill spending needs then.

3.2 Principal: Optimal policy

To write out the principal expected utility function, given the agent reaction function, it is useful to first prove three results that narrow down the optimal policy of the principal. The first two characterize the structure of the optimal auditing rule, whereas the last establishes the optimal punishment. All proofs are in the appendix.

Lemma 1. *If $\bar{x} = 0$, then in both years, the optimal audit rule A_y can take the form of a threshold rule, so that $A_y(s) = 1$ for any $s > \underline{a}_y$ and $A_y(s') = 0$ for any $s' \leq \underline{a}_y$.*

Intuitively, for the agent only the largest spending amount $\underline{a}_y$ that does not get audited matters when deciding how much to misuse. If there are audits for smaller spending amounts $s < \underline{a}$, then these spending amounts are simply avoided by the agent by misusing even more funds so that $s = \underline{a}$. In the remainder of the paper, I restrict attention to threshold rules that do not audit in case of equality, i.e., $A_y(s_y) = 0$ if $s_y = \underline{a}_y$.

Next, in year 1 and if agents want to roll-over funds absent auditing (i.e., if $\bar{x} > 0$), then the optimal audit rule is an interval rule and takes a slightly different shape.

Lemma 2. *If $\bar{x} > 0$, then in year 1 the optimal audit rule A can take the form of an interval rule, so that $A_1(s_1) = 1$ for any $s_1 \in (\underline{a}_1, \overline{a}_1)$, and $A_1(s'_1) = 0$ for any $s'_1 \notin (\underline{a}_1, \overline{a}_1)$, with $(b - \bar{x}) \in [\underline{a}_1, \overline{a}_1]$ and, for any $\underline{a}_1 \leq b - \bar{x}$,*

$$\overline{a}_1(\underline{a}_1) = \frac{2b\Delta(1 + \Delta) + 2u(\alpha - \Delta) - \underline{a}_1\Delta^2}{\Delta^2}. \tag{2}$$

An interval rule audits all spending amounts $s_1 \in (\underline{a}_1, \overline{a}_1)$. An interval rule with $\overline{a}_1 > b$ can induce the same agent spending decisions as a threshold rule with the same threshold $\underline{a}_1$. So the family of interval rules is broader, and can achieve outcomes that the threshold rule cannot achieve if $\bar{x} > 0$. As will be shown later, an interval rule that audits only interior spending amounts but not the largest spending amounts can be optimal, and this is not achievable by a threshold rule.

The important thing here is that auditing must be concentrated around the spending amount $b - \bar{x}$. This is the spending amount that no agent with $\theta_1 \leq b - \bar{x}$ wants to exceed absent auditing, because the agent wants to save and roll-over rather than misuse more than that amount (see the construction of $\bar{x}$ above). To decrease fund misuse further, an audit rule must audit the spending amounts below $b - \bar{x}$ to push agent spending down for small spending need realizations. But at the same time, it must also audit amounts above $b - \bar{x}$ to make spending $s_1 = \underline{a}_1$ incentive compatible. It is this constraint that determines $\overline{a}_1$ as a function of $\underline{a}_1$ in the Lemma. Hence, only interval rules, which audit on both sides of $s_1 = b - \bar{x}$, are effective at reducing fund misuse further, and hence the optimal audit rule is from the family of interval rules if agents save absent auditing ($\bar{x} > 0$).

Lemma 3. *The optimal policy uses a large punishment $p > 2\alpha b$, so that punishment never occurs in equilibrium.*

This result is quite straightforward: In the spirit of Becker (1968), larger punishment effectively discourages fund misuse, and at the same time saves punishment costs, hence is strictly better than lower punishment that does not always discourage fund misuse. This is also a feature of optimal policy in Malenko (2019)'s audit model.

To focus the analysis of the optimal policy, I will restrict attention to the cases $\underline{a}_y \leq b$. Allowing $\underline{a}_y > b$ adds a lot of additional case distinctions to the analysis with comparatively little additional insight. This constraint is without loss of generality for $\underline{a}_1$, since $\underline{a}_1 = b$ implies no auditing, and $\underline{a}_1 = 0$ implies always auditing, and all cases in between are covered. However, if $b_2 > b$ and $u > b$, then this constraint might bind for $\underline{a}_2$.

Assumption. *The exogenous upper bound of the audit threshold is b, i.e., $\underline{a}_y \leq b$.*

Given Lemma 1, 2 and 3, the optimal policy has a large punishment and the audit rules A_y are threshold or interval rules, where the optimal thresholds $\underline{a}_y$ are to be determined. Moreover, the roll-over rule $\Delta \in [0, 1]$ is still to be determined.

3.3 Principal expected utility given agent reaction

First, define the second year expected utility (EU) as a function of the second year budget as

$$V(b_2) := \int_0^{\underline{a}_2} (\theta_2 + \lambda(b_2 - \underline{a}_2)) \mathrm{d}G(\theta_2) + \int_{\underline{a}_2}^{b_2} (\theta_2 + \lambda(b_2 - \theta_2) - c_A) \mathrm{d}G(\theta_2) + \int_{b_2}^{u} (b_2 - c_A) \mathrm{d}G(\theta_2).$$

Now we can write out the principal EU given the agent reaction function over both years. If $\bar{x} = 0$, the principal uses a threshold audit rule in both years:

$$\begin{aligned} EU = & \int_0^{\underline{a}_1} [\theta_1 + V(b + \Delta(b - \underline{a}_1)) + \lambda(1 - \Delta)(b - \underline{a}_1)] \mathrm{d}G(\theta_1) \\ & + \int_{\underline{a}_1}^{b} [\theta_1 - c_A + V(b + \Delta(b - \theta_1)) + \lambda(1 - \Delta)(b - \theta_1)] \mathrm{d}G(\theta_1) \\ & + \int_b^u [b - c_A + V(b)] \mathrm{d}G(\theta_1) - 2\lambda b. \end{aligned} \tag{3}$$

In the first line, the year 1 spending needs realize below the audit threshold, $\theta_1 \leq \underline{a}_1$, so the agent fulfills all spending needs and misuses funds until spending reaches the audit threshold, but no higher, to avoid punishment. The principal does not incur audit costs in year 1 as the audit threshold $\underline{a}_1$ is not exceeded. Since the agent does not spend everything, amount $\Delta(b - \underline{a}_1)$ is rolled over to year 2, and amount $(1 - \Delta)(b - \underline{a}_1)$ is returned to the principal at marginal utility λ.

In the second line, spending needs realize between the audit threshold and the available budget b. Hence, the agent fulfills the spending needs but does not misuse funds, since fulfilling spending needs already puts the spending above the audit threshold, and additional misuse would trigger punishment. The principal incurs audit costs. Since the agent does not spend everything, amount $\Delta(b - \theta_1)$ is rolled over to year 2.

In the third line, the spending needs exceed the budget b. Hence, the agent cannot fulfill all spending needs due to the budget constraint, but fulfills as many needs as possible by spending everything. The principal incurs audit costs (unless $\underline{a}_1 = b$). No funds are rolled over. The cost of funds over both years is $2\lambda b$.

Next, if $\bar{x} > 0$, then the principal uses an interval audit rule, as defined in Lemma 2, with $\underline{a}_1 \leq b - \bar{x} \leq \overline{a}_1$, in year 1, so the principal EU is

$$\begin{aligned} EU = &\int_0^{\underline{a}_1} [\theta_1 + V(b + \Delta(b - \underline{a}_1)) + \lambda(1 - \Delta)(b - \underline{a}_1)]\mathrm{d}G(\theta_1) \\ &+ \int_{\underline{a}_1}^{\overline{a}_1} [\theta_1 - c_A + V(b + \Delta(b - \theta_1)) + \lambda(1 - \Delta)(b - \theta_1)]\mathrm{d}G(\theta_1) \\ &+ \int_{\overline{a}_1}^{b} [\theta_1 + V(b + \Delta(b - \theta_1)) + \lambda(1 - \Delta)(b - \theta_1)]\mathrm{d}G(\theta_1) \\ &+ \int_b^u [b + V(b)]\mathrm{d}G(\theta_1) - 2\lambda b. \end{aligned} \tag{4}$$

No audit cost is paid for large spending amounts exceeding $\overline{a}_1$, because the principal can be sure this spending is justified, since the agent wants to save these funds for roll-over unless a large θ_1-realization requires such high spending. Hence, misuse of funds does not have to be discouraged via audit if the savings and roll-over incentives already achieve this, which saves the principal audit costs.

4 Optimal policy

To ease the exposition, I split the presentation of the optimal policy (roll-over rule Δ^*, optimal year 2 audit threshold $\underline{a}_2^*$, optimal year 1 audit threshold $\underline{a}_1^t$ with threshold rule, optimal year 1 audit threshold $\underline{a}_1^i$ with interval rule) into separate propositions. In a first step, I maximize principal EU for each instrument separately, taking the others as given, and then put these results together to get to the joint optimum.

4.1 The optimal roll-over rule

Proposition 1 (**Optimal roll-over rule**).

i. If $\bar{x} > 0$ and $\overline{a}_1(\underline{a}_1) < b$ at $\Delta = \Delta_{\max \bar{x}} < 1$, and $\bar{x} < b$ at $\Delta = 1$, where

$$\Delta_{\max \bar{x}} := \min\left\{\frac{2\alpha u}{u-b}, 1\right\} > \alpha, \tag{5}$$

then the optimal budget roll-over rule is a $\Delta^* \in (\Delta_{\max \bar{x}}, 1]$. If, in addition, λ is sufficiently large, then a $\Delta^* \in (\Delta_{\max \bar{x}}, 1)$ is optimal.

ii. If $\bar{x} = 0$ at $\Delta = \Delta_{\max \bar{x}} < 1$, or $\Delta_{\max \bar{x}} = 1$, or $\bar{x} = b$ at $\Delta = 1$, or $\overline{a}_1(\underline{a}_1) > b$ at $\Delta = \Delta_{\max \bar{x}}$, then $\Delta^* = 1$ is optimal.

iii. If $b = u$, or $\underline{a}_1 = b$ and $\bar{x} = 0$ at $\Delta = \Delta_{\max \bar{x}}$, then any $\Delta^* \in [0, 1]$ is optimal.

Proposition 1 shows that either a potentially interior $\Delta^* \in (\Delta_{\max \bar{x}}, 1]$ or the maximum $\Delta = 1$ is optimal. In some cases, the choice of Δ does not matter, so any $\Delta \in [0, 1]$ is optimal. Consequently, allowing at least some roll-over of unused funds is weakly better.

For the agent to want to save and roll-over any funds, the share of unused funds that is rolled-over must exceed α, which is the marginal value the agent receives from misusing funds. Hence, in this model with linear utility functions, if agents at the margin get 50% of utility from misusing funds compared to fulfilling spending needs, then *more than* 50% of unused funds should be allowed to be rolled over for use next year to incentivize agents not to misuse funds this year.

4.1.1 Why partial roll-over of unused funds may be optimal

To understand why sometimes $\Delta = 1$ is optimal and sometimes $\alpha < \Delta < 1$, we have to distinguish two purposes of the roll-over rule. First, whatever funds are saved in year 1—either due to the threat of auditing or because the agent wants to roll-over funds—given that these funds are saved, it is weakly better for agent and principal alike to roll-over as much as possible. This is because more roll-over increases the chances to fulfill additional spending needs in year 2, which is better than fund misuse for both. It is only weakly better, because if the budget is already very large, then the expected benefit of rolling over more funds is small or even zero once all potential spending needs can be fulfilled ($b_2 = u$).

Second, the roll-over rule also determines $\bar{x}$, how much the agent wants to save rather than misuse in year 1, in order to roll-over funds and to be able to fulfill more spending needs next year. Hence, the possibility of rolling over funds induces the agent to misuse fewer funds, which can allow the principal to save audit costs without increasing fund misuse. The amount saved by the agent $\bar{x}$—if realization θ_1 is small enough—is not in general increasing in $\Delta \in [0, 1]$. It can in fact be locally decreasing, so that there is a unique $\Delta_{\max \bar{x}} < 1$ which

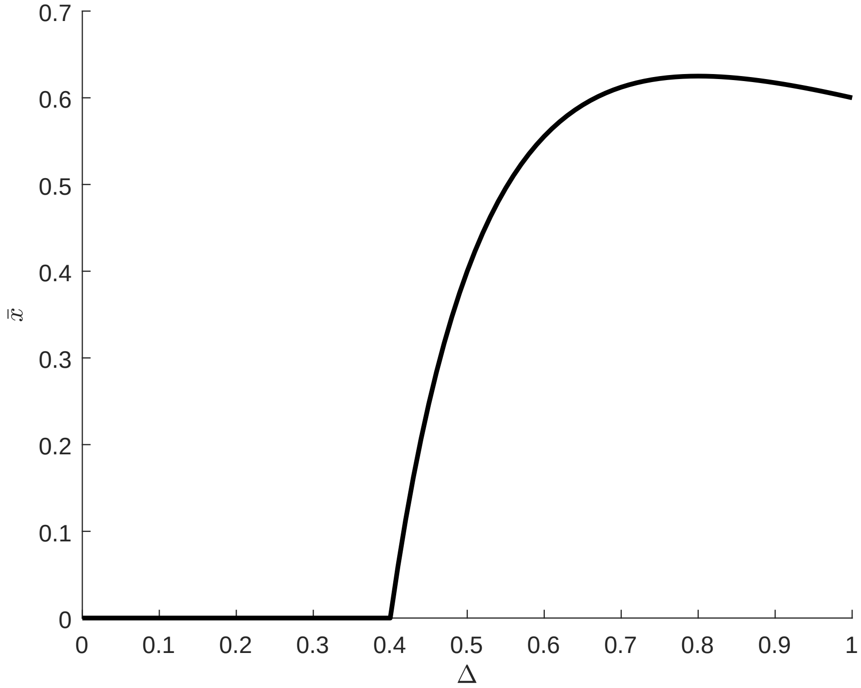

Figure 2: Plot of agent saving $\bar{x}$ depending on Δ. Parameter values: $u = 2, b = 1, \alpha = 0.2$. The maximum is at $\Delta_{\max\ \bar{x}} = 0.8$.

maximizes the saved amount $\bar{x}$. Figure 2 plots an example. Consequently, in these cases, the principal has to trade off the maximum $\Delta = 1$ (first purpose) with $\Delta = \Delta_{\max\ \bar{x}} < 1$ (second purpose).

The intuition why $\Delta_{\max\ \bar{x}}$ can be less than 1 is as follows. The roll-over rule Δ essentially determines how efficient the budget transfer is from one year to the next. If budget b is small, then the agent very much wants to roll-over unneeded funds to be able to meet more spending needs next year, even if the roll-over means forgoing fund misuse for private gain. But making the transfer too efficient in these cases would leave the agent more room for fund misuse while still having enough funds next year, so the principal optimally "taxes" the roll-over to induce more saving.

Why must the optimal Δ^* exceed α? Rolled-over funds can be used to fulfill additional spending needs at a marginal value of 1 next year. Moreover, every dollar saved leads to Δ dollars additionally available next year, so the expected agent marginal value of rolling over funds is bounded above by Δ.[9] The agent compares this expected marginal value from rolling over to misusing funds at a marginal rate of α, so it is clear that Δ must exceed α for the agent to want to roll-over any funds at all and stop misusing funds in year 1.

[9] The expected marginal value from rolling-over is in fact strictly lower than Δ, because the probability of needing the rolled-over funds to satisfy spending needs in year 2 is less than unity. See (1) for the exact expression.

4.1.2 Optimal roll-over rule

Proposition 1 part i. are the cases where $\Delta < 1$ can be optimal. These cases require that $\Delta_{\max \bar{x}} < 1$, otherwise there is no trade-off between inducing more saving and rolling over more funds. $\Delta_{\max \bar{x}} < 1$ requires a sufficiently small agent utility from fund misuse, α, because that is the opportunity cost of rolling over, and a sufficiently small budget b. The smaller budget is needed, so there is a larger chance the agent cannot meet his spending needs next year, hence he is more willing to roll-over even for lower values of Δ. This is not a precautionary savings motive, as the agent is risk neutral, but an expected value calculation that rolling over is more attractive if the budget is small.

The trade-off between $\Delta < 1$ and $\Delta = 1$, as discussed above, consists of preventing more fund misuse (at a utility of λ) and saving audit costs, or of fulfilling more spending needs (at a net utility of $1 - \lambda$). Consequently, if λ is sufficiently large, then $\Delta < 1$ is optimal, since fulfilling additional spending needs next year is not as important as preventing fund misuse and saving audit costs this year.

$\Delta < 1$ can be optimal only if b and α are not too large, so that the agent wants to save funds at all, i.e., only if $\bar{x} > 0$ for $\Delta_{\max \bar{x}} < 1$. But b and α also cannot be too small, since otherwise the agent wants to save all unused funds even if $\Delta = 1$, which would then be optimal. Hence, $\Delta^* < 1$ also requires $\bar{x} < b$ at $\Delta = 1$.

In case ii., which captures all cases where $\Delta_{\max \bar{x}} = 1$ maximizes the amount saved by the agent, the maximum $\Delta^* = 1$ is optimal. This is because there is no conflict between inducing the maximum saving by the agent and rolling-over everything that is saved. Moreover, if $\Delta_{\max \bar{x}} < 1$ but no saving by the agent can be induced for any Δ—for example if b is large—then $\Delta^* = 1$ is also optimal.

Finally, there are some cases where Δ does not affect the principal's payoff. The simplest of these cases is the maximum budget $b = u$. In this case, all spending needs can always be met, so rolling over has no benefit for the principal, and for the same reason the agent does not save funds. Hence, any Δ can be set.

4.1.3 Comparison to budget rules in practice

As the literature section shows, most public sector organizations in the US do not allow agents to roll-over unused funds, which conflicts with the finding here where at least some roll-over is weakly better. But there are a few documented cases where fund roll-over was allowed in practice.

The UK governmental roll-over rule, introduced in 2010, allows for a full roll-over of unused funds ($\Delta = 1$), but for only for up to 0.75% (large agencies) or up to 4% (small agencies) of the budget (Baumann, 2019). In the cases where my model finds $\Delta^* < 1$ to be optimal, the UK implementation might be imperfect in two respects: First, $\Delta = 1$ does not discourage as much fund misuse as $\Delta^* < 1$ as just explained, and second, the 4% upper bound means the UK rule can at most prevent fund misuse in the amount of 4% of the budget. A larger percentage could potentially prevent more fund misuse.[10]

There is at least one case where a (local) governmental body allowed its agencies to roll-over some, but not all, of their unused funds to the next year. The State of Washington's Saving Incentive Program from 1997 allowed agencies to retain 50% of unused funds for next year. Interestingly, the 50% were set not because that number is optimal in reducing fund misuse, as in this model. Instead, the other 50% were promised to the education sector, which made this reform politically feasible (e.g., Jones, 2005, p.152; Miller et al., 2007).

4.2 The optimal year 2 audit rule

Proposition 2 derives the optimal audit threshold in year 2, which can condition on b_2, and is therefore conditionally independent of the year 1 audit rule and the roll-over rule.

Proposition 2 (Optimal year 2 audit rule). *As shown in Lemma 1, the optimal audit rule can be represented as a threshold rule, with $A_2(s_2) = 1$ if and only if $s_2 > \underline{a}_2$, and $A_2(s_2) = 0$ otherwise.*

i. The optimal audit threshold for year 2 is $\underline{a}_2^* = \min\left\{\frac{c_A}{\lambda}, b\right\}$ if

- *$b = u$, or if*
- *$b < b_2$, or if*
- *$b = b_2 < u$, $c_A/\lambda \leq b$ and $\frac{b_2^2\lambda}{2} + \frac{c_A^2}{2\lambda} \geq uc_A$.*

ii. The optimal threshold is $\underline{a}_2^* = b$ if $b = b_2 < u$, $c_A/\lambda \leq b$ and $\frac{b_2^2\lambda}{2} + \frac{c_A^2}{2\lambda} < uc_A$. That is, there is no auditing even if all budget is spent in this case.

Proposition 2 shows that the optimal audit rule, a threshold rule, is simple in that either the optimal threshold equals the interior solution $\underline{a}_2^* = c_A/\lambda$ or the corner solution $\underline{a}_2^* = b$.

When setting the audit threshold, a larger threshold implies less auditing, because an audit is triggered only for spending amounts exceeding the threshold. Agents with a spending need realization exceeding the threshold ($\theta_2 > \underline{a}_2$) do not misuse funds to avoid punishment. Agents with low spending needs below the threshold ($\theta_2 < \underline{a}_2$) misuse funds to stay just below the threshold. Hence, a larger threshold implies more fund misuse by the agent, who can get away with spending more without being audited. But the larger audit threshold also saves audit costs. This is the key trade-off when setting the audit threshold. The interior solution $\underline{a}_2^* = c_A/\lambda$ optimally trades off the additional fund misuse from increasing

[10] A problem with setting the 4% upper bound may have been that prior to allowing roll-over, there were typically not more than 4% of unused funds. But, as this model shows, used funds do not equal spending needs, because some funds are wasted or misused by the agent. Hence, the only way to find out if fund misuse can account for more than 4% of the budget is to allow more than 4% of the budget to be rolled-over. If spending drops in response to such a change, it would indicate that agents switched from misusing excess funds to rolling them over instead.

the threshold and the saved audit cost, and the two exactly cancel out. Consequently, a larger audit cost c_A or a smaller disutility from additional fund misuse (equal to the cost of funds λ) increases the audit threshold, i.e., leads to less auditing in the interior solution.

If $b = u$—where the yearly budget is enough to fulfill all spending needs for certain—or if $b < b_2$—where some funds were rolled-over from the previous year—then the interior solution is optimal unless $c_A/\lambda \leq b$ is binding, in which case the corner solution is optimal. These are the technically well-behaved cases, because there are no discontinuities in the principal expected utility function for $\underline{a}_2 \in [0, b]$.

However, if $b = b_2 < u$, then there can be a discontinuity at $\underline{a}_2 = b$. To understand why there is a discontinuity, note that $b_2 < u$ implies that the budget constraint for the agent is binding for all realizations $\theta_2 \in (b_2, u]$. Consequently, a probability mass of at least $1 - G(b_2)$ is spending $s_2 = b_2$. So when comparing the principal utility at $\underline{a}_2$ just below $b = b_2$ (which audits $s_2 = b$) and at $\underline{a}_2 = b = b_2$ (which does not audit $s_2 = b$), this probability mass introduces a discontinuous jump in saved audit costs. Case ii. in Proposition 2 shows that the corner solution might be optimal even if the interior solution is in fact in the interior (i.e., if $c_A/\lambda < b$). This happens in particular if b_2 is sufficiently small, whereas the interior is favored if λ is sufficiently large.

The budget b can also have an effect on optimal auditing. While the audit threshold of the interior solution does not depend on the budget, it does in part determine whether a corner solution with little auditing or the interior solution with more auditing is optimal. As $\underline{a}_2^* = \min\{c_A/\lambda, b\}$ in part i. suggests, a larger budget tends to favor the interior solution and increase audit activity. For small budgets, the corner solution prescribes less auditing, since there is not much room for fund misuse. For large budgets, the probability of the budget exceeding the spending needs increases, so fund misuse is more of a problem and auditing is more useful. Interestingly, once the interior solution is optimal, increasing the budget further does not increase expected year 2 audit costs, since in equilibrium all realizations $\theta_2 \in (c_A/\lambda, u]$ are audited, which is independent of b. Essentially, increasing b while holding the interior audit threshold fixed makes spending amounts weakly larger, but there are not more cases that need to be audited.

4.3 The optimal year 1 audit rule

Proposition 3 (Optimal year 1 audit rule if $\bar{x} = 0$). *If $\bar{x} = 0$, then as shown in Lemma 1, the optimal audit rule can be represented as a threshold rule, with $A_1(s_1) = 1$ if and only if $s_1 > \underline{a}_1$, and $A_1(s_1) = 0$ otherwise.*

i. The optimal audit threshold for year 1 is

$$\underline{a}_1^t = \min\left\{\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1\beta_3}}{2\beta_1}, \frac{c_A}{\lambda}, b\right\} \tag{6}$$

if

- $b = u$, *or if*
- $b < u$ *and* (17),

with

$$\begin{aligned}
\beta_1 &= -\frac{\Delta^2(1-\lambda)}{u} < 0,\\
\beta_2 &= -\lambda(1-\Delta) - \left(1-(1-\lambda)\frac{(1+\Delta)b}{u}\right)\Delta < 0,\\
\beta_3 &= c_A > 0,\\
a &= \min\left\{\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1\beta_3}}{2\beta_1}, \frac{c_A}{\lambda}\right\}.
\end{aligned}$$

ii. *The optimal audit threshold is* $\underline{a}_1^t = b$ *if* $b < u$, $a > b$, *and* $\neg(17)$. *That is, there is no auditing even if all budget is spent in this case.*

iii. *The optimal year 1 audit threshold is weakly lower than the year 2 threshold. That is, auditing tends to be more aggressive in year 1.*

iv. *If* $b + \Delta(b - c_A/\lambda) \geq u$, *and in particular if* $b = u$, *then the interior audit thresholds in both years are identical.*

Proposition 3 shows that the optimal audit rule in year 1 if $\bar{x} = 0$ is slightly more complex than the one in year 2. Since there is a follow-up year after year 1, the principal has to take the effects on year 2 into account when setting audit threshold $\underline{a}_1$. Setting a lower audit threshold—i.e., auditing more by auditing lower spending amounts—means there is less fund misuse by the agent, and hence more unused funds. If fund roll-over is allowed, and it is under the optimal policy (Proposition 1), then more auditing implies a larger budget in year 2, which can benefit both agent and principal by fulfilling additional spending needs. Consequently, the principal in year 1 tends to audit more (lower spending amounts) than in year 2, in order to induce more fund-roll over.

But increased auditing is optimal only if the fund roll-over which the year 1 audit threshold induces (for realizations $\theta_1 \in [0, \underline{a}_1)$) is not enough to cover all potential spending needs next year, assuming the year 1 audit threshold is set to the year 2 interior threshold ($\underline{a}_1 = c_A/\lambda$). Then, setting the same threshold as in year 2 means not all potential spending needs can be fulfilled, so the principal has a reason to audit slightly more via a lower threshold.

Otherwise, the findings for the optimal audit threshold in year 1 are very much in line with those of year 2: Either the interior solution—which tends to be smaller than in year 2—or the corner solution $\underline{a}_1 = b$ are optimal. Larger audit costs c_A tend to increase the audit threshold, whereas larger costs of funds λ tend to decrease the audit threshold. And, as in year 2, a larger b tends to favor the interior solution and hence leads to more auditing, because there is more room for fund misuse with a larger budget. Unlike in year 2, the roll-over policy Δ matters for the optimal audit threshold in year 1, because it determines whether and how much the increased year 1 auditing leads to more fulfilled spending needs in year 2.

As the next proposition shows, auditing changes qualitatively in year 1 if $\bar{x} > 0$, i.e., if agents want to save and roll-over some unneeded funds absent auditing. This proposition makes several comparisons between the optimal interval rule and the optimal threshold rule which would have been used if $\bar{x} = 0$, see Proposition 3.

Proposition 4 (**Optimal year 1 audit rule if** $\bar{x} > 0$). *If $\bar{x} > 0$, then as shown in Lemma 2, the optimal audit rule can be represented as an interval rule, with $A_1(s_1) = 1$ if and only if $s_1 \in (\underline{a}_1, \overline{a}_1)$, and $A_1(s_1) = 0$ otherwise.*

i. If $b - \bar{x} > a^t := \min\left\{\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1\beta_3}}{2\beta_1}, \frac{c_A}{\lambda}\right\}$, then the optimal lower border of the audit interval rule $\underline{a}_1^i$ is weakly larger than a^t, the interior solution of the threshold rule in Proposition 3. The principal is weakly better off compared to using a threshold rule due to the reduced audit costs.

ii. If $b - \bar{x} \leq a^t$, then the optimal lower border of the audit interval rule $\underline{a}_1^i$ is weakly smaller than a^t, and the interval rule prevents weakly more fund misuse. The principal is strictly better off compared to using a threshold rule due to the reduced audit costs.

iii. The optimal borders for the interval audit rule in year 1 are

$$\underline{a}_1^i = a^i,\ \overline{a}_1^i = \frac{2b\Delta(1+\Delta) + 2u(\alpha - \Delta) - a^i\Delta^2}{\Delta^2} = 2(b - \bar{x}) - a^i \tag{7}$$

if $a^d < a^t < a^i$, or if $a^t < a^d < a^i$ and $EU(\underline{a}_1 = a^t) \leq EU(\underline{a}_1 = a^i)$, or if $a^d < a^i < a^t$, with $EU(\underline{a}_1)$ defined in (4) *and with*

$$\begin{aligned}
a^i &= \min\left\{\frac{-\gamma_2 - \sqrt{\gamma_2^2 - 4\gamma_1\gamma_3}}{2\gamma_1}, \frac{2c_A}{\lambda}, b - \bar{x}\right\}, \\
a^d &= \underline{a}_1 \text{ such that } \overline{a}_1(\underline{a}_1) = b \text{ in (2)}, \\
\gamma_1 &= -\frac{\Delta^2(1-\lambda)}{u} < 0, \\
\gamma_2 &= -\lambda(1-\Delta) - \left(1 - (1-\lambda)\frac{(1+\Delta)b}{u}\right)\Delta < 0, \\
\gamma_3 &= 2c_A > 0.
\end{aligned}$$

iv. *The optimal borders for the interval audit rule in year 1 are*

$$\underline{a}_1^i = a^t,\ \overline{a}_1^i > b$$

if $a^t < a^i < a^d$ *and* $EU(\underline{a}_1 = a^t) \geq EU(\underline{a}_1 = a^d)$*, or if* $a^t < a^d < a^i$ *and* $EU(\underline{a}_1 = a^t) \geq EU(\underline{a}_1 = a^i)$*. That is, the optimal audit rule is equivalent to a threshold rule.*

v. *The optimal borders for the interval audit rule in year 1 are*

$$\underline{a}_1^i = a^d,\ \overline{a}_1^i = b$$

if $a^t < a^i < a^d$ *and* $EU(\underline{a}_1 = a^t) \leq EU(\underline{a}_1 = a^d)$*, or if* $a^i < a^t < a^d$ *and* $EU(\underline{a}_1 = a^t) \leq EU(\underline{a}_1 = a^d)$*. That is, the optimal audit rule is an interval rule, where only spending* $s_1 = b$ *is not audited above the audit interval.*

If $\bar{x} > 0$ and if there is no auditing, then agents with low realizations of spending needs ($\theta_1 \leq b - \bar{x}$) spend $s_1 = b - \bar{x}$, because $\bar{x}$ is exactly the amount of funds these agents would like to save and roll-over rather than misuse. Hence, even absent auditing, the roll-over rule Δ costlessly prevents some fund misuse. Lemma 2 and Proposition 4 describe how optimal auditing and this roll-over incentive interact.

If $\bar{x} > 0$, the optimal audit rule is an interval rule, which audits the spending interval $s_1 \in (\underline{a}_1, \overline{a}_1)$, but no spending outside that interval. The special case of not auditing is $\underline{a}_1 = \overline{a}_1 = b - \bar{x}$. Suppose the principal wants to prevent more fund misuse than by not auditing. Then, first, $\underline{a}_1$ has to be reduced, so that agents with small spending needs spend less than $b - \bar{x}$. However, reducing $\underline{a}_1$ alone is not enough, because then spending $s_1 = b - \bar{x}$ is more attractive than $s_1 = \underline{a}_1$ for the agent. Hence, second, to get the agent to spend $s_1 = \underline{a}_1 < b - \bar{x}$ instead—i.e., to make $\underline{a}_1$ incentive compatible—the upper border of the interval $\overline{a}_1$ has to be increased as well.

Lemma 2 shows that the incentive compatibility constraint (8) uniquely determines $\overline{a}_1$ as a function of $\underline{a}_1$ so that agents indeed spend $s_1 = \underline{a}_1$ if $\theta_1 \leq \underline{a}_1$. In fact, (2) shows that the incentive compatible $\overline{a}_1(\underline{a}_1)$ is a linear function with a slope of -1. This implies that both borders of the audit interval are equidistant to $b - \bar{x}$ and $\overline{a}_1$ can simply be determined as $\overline{a}_1(\underline{a}_1) = 2(b - \bar{x}) - \underline{a}_1$.

Consequently, an interval rule with $\overline{a}_1 \leq b$ is qualitatively different from a threshold rule used in year 2 or in year 1 if $\bar{x} = 0$, because it does not audit the largest spending amounts. Instead, it audits spending amounts around $b - \bar{x}$. The largest spending amounts in the neighborhood of $s_1 = b$ are not audited, because the principal can be sure that such spending is only done if spending needs are large and justify such large spending (e.g., if $\theta_1 \geq b$). An agent with small spending needs, on the other hand, does not misuse all remaining funds to spend that much, because such an agent would rather save and roll-over some of the remaining funds. The borders of the optimal interval rule are chosen specifically to induce agents to behave this way. The optimal interval audit rule is the audit rule that audits as little as possible to get agents with low spending needs to spend $\underline{a}_1$, making use of the agent's savings motive.

Part i. in Proposition 4 shows that the lower border of the audit interval $\underline{a}_1$ is weakly larger than the interior solution to the optimal threshold rule (Proposition 3).[11] Hence, if both rules audit at least some spending amounts, then there is weakly less auditing under the interval rule, and there is weakly more fund misuse. Nevertheless, the principal is better off with the interval rule under $\bar{x} > 0$, holding everything else constant. This can be easily seen because the principal can mimic the outcomes of any threshold rule, so any different audit rule choice by the principal implies it is better.

The reason why the interval rule audits weakly less compared to the threshold rule interior solution is as follows. Under a threshold rule, an increase of the audit threshold $\underline{a}_1$ decreases the expected audit cost proportional to that increase. But under the interval rule, the increase of $\underline{a}_1$ also decreases $\overline{a}_1$, which is determined by an incentive compatibility constraint, since a larger lower audit border is incentive compatible with a smaller upper audit border. Consequently, there is a larger decrease in expected audit costs under the interval rule when increasing $\underline{a}_1$. Yet the same $\underline{a}_1$-increase leads to the same increase in fund misuse under both rules. Thus, more auditing (smaller $\underline{a}_1$) is less attractive with an interval rule, and in the interior solution there is less auditing than in the threshold rule interior solution.

Part ii. shows that if $b - \bar{x}$ is smaller than the interior threshold rule solution, then the lower audit threshold $\underline{a}_1$ under the interval rule is smaller than the one of the threshold rule. This is immediately obvious, as the lower border of the optimal interval rule can never exceed $b - \bar{x}$, since $\underline{a}_1 = b - \bar{x}$ prevents some fund misuse but incurs no audit costs. Parts i. and ii. illustrate that the roll-over rule interacts in non-trivial ways with the optimal audit rule. The optimal audit rule strategically audits interior spending amounts while relying on the roll-over incentive to prevent fund misuse at large spending amounts for free.

Parts iii., iv., and v. derive explicit expressions for the lower and upper audit threshold of the optimal interval rule. There are three different cases because of a discontinuity in the principal expected utility function. When decreasing $\underline{a}_1$, $\overline{a}_1(\underline{a}_1)$ automatically increases according to the incentive compatibility constraint. At some point, $\overline{a}_1 = b$, so that agents with large spending needs $\theta_1 \geq b$ are not audited. Any further decrease in $\underline{a}_1$ implies $\overline{a}_1(\underline{a}_1) > b$, so that all $\theta_1 \geq b$ are audited, as they are now in the interval. And this additional probability mass of auditing causes a discontinuous jump in expected audit costs, so that interior solutions may not be optimal.

The optimal rule is either an interval rule (iii.), where the lower border is the solution to the first order condition (a^i), and the upper border is determined by the incentive compat-

[11] This does not mean that there is less auditing under the interval than the threshold rule, since the threshold rule corner solution may be optimal and does not audit at all.

ibility constraint. Or it is an interval rule that is equivalent to a threshold rule (iv.), where the audit threshold is the solution to the first order condition of the threshold rule (a^t). Or it is an interval rule (v.), where the lower border is at the discontinuity (a^d) just discussed, and the upper border is $\overline{a}_1 = b$. Case iii. applies with a $b - \bar{x}$ not too large (especially for small b) and sufficiently large c_A/λ. Since it saves audit costs, the interval rule is preferred over the de facto threshold rule for large audit costs. Case iv. occurs with a large $b-\bar{x}$ close to b and c_A/λ not too large. If $b-\bar{x}$ is close to b, then $\underline{a}_1$ cannot decrease much below $b-\bar{x}$ to increase audit activity before $\overline{a}_1(\underline{a}_1)$ exceeds b and becomes a threshold rule. And this is preferred by the principal if audit costs are sufficiently small or cost of funds sufficiently large.

4.4 Optimal policy

The previous propositions maximized principal expected utility for each instrument, holding the others constant. However, the optimal policy requires all instruments to be jointly optimal. The optimal year 2 audit threshold is independent of the other instruments, so the proposition applies directly. However, the roll-over rule and year 1 audit rule are interdependent. According to Proposition 1, the optimal roll-over rule is $\Delta^* = 1$ if $\alpha \geq 1/2$ or if b is sufficiently large, independent of $\underline{a}_1$. But if $\Delta^* \in (\Delta_{\max \bar{x}}, 1]$, then the exact optimum in that interval can depend on $\underline{a}_1$.

Overall, in the optimal policy, full fund roll-over can be consistent both with extensive or with minimal auditing. Similarly, extensive auditing can be consistent with full or only partial fund roll-over, depending on parameters. However, if the year 1 audit rule is a threshold rule, then only full fund roll-over can be optimal, since the threshold rule does not benefit from more agent saving. An interval rule in year 1, on the other hand, can be consistent with both full or partial fund roll-over. The following Corollary 1 describes the optimal policy in several distinct situations, and follows from the previous propositions.

Corollary 1. *The optimal policy sets a large punishment in case of audit and fund misuse.*

i. If $\alpha \geq 0.5$ or $b \geq u(1-2\alpha)$, then full fund roll-over is optimal ($\Delta^* = 1$) independent of the audit rules.

ii. If c_A/λ is sufficiently small and b sufficiently large, then the optimal policy is extensive auditing in both years ($\underline{a}_1^t = \underline{a}_2^* = c_A/\lambda < b$) and a full roll-over of unused funds ($\Delta^* = 1$). The audit rules are threshold rules.

iii. If c_A/λ is sufficiently small, α sufficiently small and $b > \frac{(1-\alpha)u}{2}$ small, then the optimal policy is extensive auditing in both years, with an interval rule in year 1 ($\underline{a}_1^i \leq 2c_A/\lambda$) and a threshold rule in year 2 ($\underline{a}_2^* = c_A/\lambda$). A partial roll-over of unused funds can be optimal ($\Delta^* \in (\Delta_{\max \bar{x}}, 1]$).

iv. *If c_A/λ is sufficiently large and b sufficiently small, then the optimal policy is minimal auditing in both years ($\underline{a}_1^i = b - \bar{x}, \underline{a}_2^* = b$), where agent savings to roll-over funds prevent most fund misuse in year 1, and a full roll-over of unused funds ($\Delta^* = 1$) is optimal.*

v. *If c_A/λ is sufficiently large, α sufficiently small and $b > \frac{(1-\alpha)u}{2}$ small, then the optimal policy is minimal auditing in both years ($\underline{a}_1^* = \underline{a}_2^* = b$), where agent savings to roll-over funds prevent most fund misuse in year 1, and potentially a partial roll-over of unused funds ($\Delta^* \in (\Delta_{\max\bar{x}}, 1]$) is optimal.*

5 Extensions

5.1 Endogenizing the budget

The analysis so far has shown that the other policy instruments—in particular the roll-over rule Δ and the audit thresholds $\underline{a}_1, \underline{a}_2$—depend on the annual budget b. This raises the question how the principal could strategically set the budget to maximize expected utility. A higher budget helps the agent fulfill more spending needs, which also benefits the principal, but could also lead to more fund misuse and reduces the incentives of the agent to save funds for roll-over. The annual budget b is now set in year 0, along with the other instruments.

Because of this interaction with the other instruments, maximizing principal expected utility (EU) with respect to budget b is challenging, as it shifts the optimal audit thresholds and roll-over rule. Indeed, these interactions add discontinuities to the principal EU, and the optimum will only sometimes fulfill a first order condition.

To see this, consider the plot of the principal EU depending on budget b in Figure 3. The red line represents the utility with minimal auditing ($\underline{a}_1 = \underline{a}_2 = b$), whereas the black line uses the interior solutions for the audit thresholds, unless these exceed b, in which case there is also set the thresholds equal to b. Hence, the black and red lines coincide for small b, where both do not audit, and then differ for larger b where the black line does audit. A discontinuity in the black line occurs when b increases to $b > \underline{a}_2 = c_A/\lambda$, so that all realizations $\theta_2 \in (c_A/\lambda, u]$ are audited in year 2, whereas these realizations are not audited for a smaller b if $b_2 = b \leq c_A/\lambda$. Hence, the discontinuity comes from the additional cost of auditing a mass point. This is an illustration of Proposition 2, which shows the interior solution may not be the optimal audit threshold even if it is in the interior.

The black line also shows an interior local maximum, which is not a global maximum. In the range $b < 5.8$, the optimal interior audit rule in year 1 is an interval rule which uses the fact that the agent wants to save and roll-over funds in case of small spending needs in year 1. That is, the rule audits the spending range $s_1 \in (\underline{a}_1, \overline{a}_1)$ with $\overline{a}_1 < b$, but not the largest spending $s_1 \in [\overline{a}_1, b]$, still the saving and roll-over motive of the agent prevents all realizations $\theta_1 \in (\underline{a}_1, u]$ from misusing funds. The agent spending decisions under this

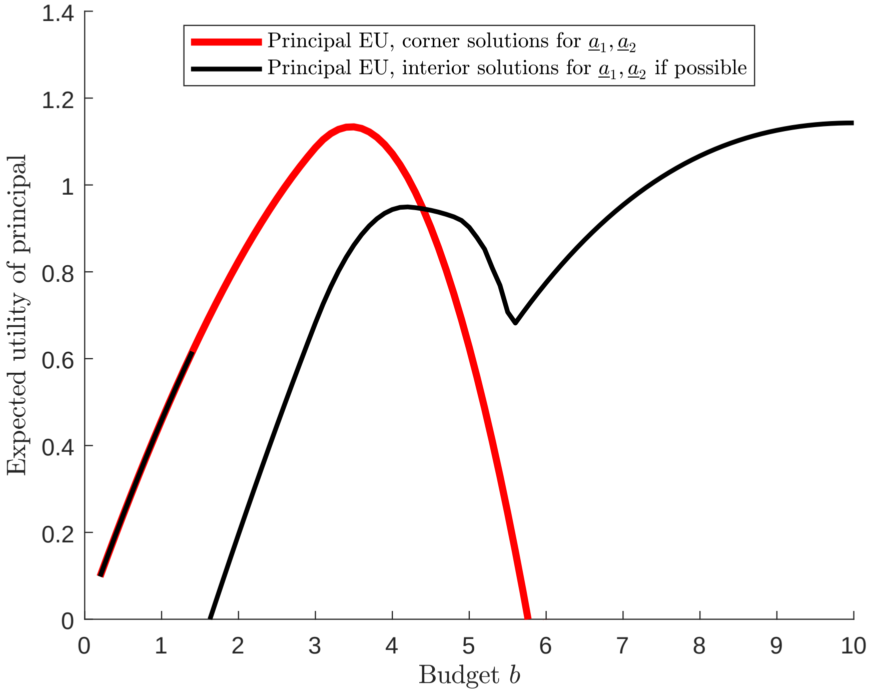

Figure 3: Principal expected utility with respect to budget b, given optimal $\Delta(b)$ and plotted separately for interior and corner solutions for audit thresholds $\underline{a}_1, \underline{a}_2$. Parameter values: $u = 10, c_A = 1, \lambda = 0.7, \alpha = 0.4$.

interval rule are the same as under a threshold rule, except the interval rule incurs a lower audit cost. The principal utility is locally decreasing as b approaches ≈ 5.8, because a larger b means the agent saving amount $\bar{x}$ decreases: More budget next year means saving and roll-over is less attractive than misusing funds in year 1. Hence, to discourage the same amount of fund misuse, the principal would have to audit more at a higher cost, or the principal would have to allow for more fund misuse, both of which is costly. For $b > 5.8$, the optimal interior audit rule is to audit all spending $s_1 \in (\underline{a}_1, b]$, even before the agent saving amount $\bar{x}$ becomes zero, so the interval rule effectively becomes a threshold rule. In this case, the principal utility is increasing in the budget b, because the audit activity, audit cost, and hence extent of fund misuse is fixed—all $\theta_1 \in (\underline{a}_1 = c_A/\lambda, u]$ are audited independent of b—but the additional budget allows more spending needs to be fulfilled.

The figure illustrates there are potentially several candidates for an optimal budget, if parameters change:

- A budget to fulfill all possible spending needs ($b = u$),
- A budget interior solution with interior audit thresholds $\underline{a}_y < b$,
- A budget interior solution with corner audit thresholds $\underline{a}_y = b$,
- A budget just below the first discontinuity where auditing starts ($b = c_A/\lambda - \varepsilon$).

Table 1: Optimal policy and agent saving and fund roll-over for all combinations of exogenous parameters $\alpha \in \{0.2, 0.8\}$ (agent marginal utility from fund misuse), $\lambda \in \{0.2, 0.9\}$ (cost of funds), $c_A \in \{0.5, 1, 1.5\}$ (audit cost), with spending needs distributed $\mathcal{U}(0, u = 10)$. A indicates auditing, NA indicates no auditing.

parameters			optimal policy					savings	remark
α	λ	c_A	b	Δ	$\underline{a}_1$	$\overline{a}_1$	$\underline{a}_2$	$\bar{x}$	
0.2	0.2	0.5	10	1	2.5		2.5	0	threshold A $y = 1, 2$
0.2	0.2	1	10	1	5		5	0	threshold A $y = 1, 2$
0.2	0.2	1.5	8	1	8		8	0	threshold NA $y = 1, 2$
0.2	0.8	0.5	10	1	6.3		6.3	0	threshold A $y = 1, 2$
0.2	0.8	1	4.7	0.8	1.2	1.2	1.3	3.5	interval NA $y = 1$, threshold A $y = 2$
0.2	0.8	1.5	3	1	0	0	3	3	interval NA $y = 1$, threshold NA $y = 2$
0.8	0.2	0.5	10	1	2.5		2.5	0	threshold A $y = 1, 2$
0.8	0.2	1	10	1	5		5	0	threshold A $y = 1, 2$
0.8	0.2	1.5	8	1	8		8	0	threshold NA $y = 1, 2$
0.8	0.8	0.5	10	1	0.6		0.6	0	threshold A $y = 1, 2$
0.8	0.8	1	2	1	2		2	0	threshold NA $y = 1, 2$
0.8	0.8	1.5	2	1	2		2	0	threshold NA $y = 1, 2$

To get a better idea of which of these cases are relevant in equilibrium, I determined the optimal budgets, roll-over rules, and audit rules numerically in Table 1 for various parameter profiles. The table covers low and high values of α and λ, as well as low, medium and high values of the audit cost c_A, with all twelve possible combinations. The largest possible spending need realization is $\theta_y = u = 10$, which is thus the maximum budget.

The first three cases have a low agent utility from misusing funds α and a low cost of funds λ. Because of a low cost of funds, a large budget is optimal for the principal, since fulfilling additional spending needs has high value $(1 - \lambda)$ and the cost of fund misuse is low (λ). In case of high audit costs, there is no auditing and the budget is not maximal to limit fund use. The agent does not want to save funds for roll-over ($\bar{x} = 0$), because even though the marginal utility from misusing funds is low, the budget in all of these cases is high enough so that rolling-over funds is not attractive, that is, the agent knows future spending needs can most likely be fulfilled even without fund roll-over.

The next three cases have a low agent utility from misusing funds but a high cost of funds. If audit costs are low, it is optimal for the principal to set a maximum budget and audit extensively to limit fund misuse. However, for larger audit costs extensive auditing is too costly. Consequently, due to the high cost of funds, the principal sets a low budget to limit fund misuse, which is very costly in this case. Because the budget is small and the agent utility from fund misuse is low, the agent wants to save and roll-over funds even absent auditing ($\bar{x} > 0$). Consequently, these are two cases where the optimal audit rule in year 1 is an interval rule. In one of these cases, with medium audit costs, the optimal roll-over rule is $\Delta = 0.8$, so that only 80% of unused funds from year 1 are rolled-over to year 2. This increases the amount saved by the agent, compared to $\Delta = 1$, and hence saves the principal audit costs or limits more fund misuse. In both of the interval rule cases, there is no auditing in year 1, because the savings motive already limits most fund misuse.

The remaining six cases have a high agent utility from misusing funds. In these cases, the agent does not want to save funds in equilibrium ($\bar{x} = 0$). Consequently, the principal sets a larger budget if the cost of funds and audit cost are sufficiently small, and audits extensively to limit fund misuse by the agent. For fund or audit costs too large, the optimal budget is small, to limit fund misuse, as auditing is too costly.

Table 1 shows that strategically underfunding the agent can be optimal in many cases, especially if the costs of funds and auditing are large. The principal then accepts that, ex ante, the agent will not always be able to cover all spending needs, but prefers this over costly fund misuse that would increase with a larger budget.

Proposition 5 generalizes some of these insights analytically.

Proposition 5 (**Optimal budget**).

i. If c_A is sufficiently small, then the optimal budget is the maximum $b^* = u$, with extensive auditing via a threshold rule ($\underline{a}_y = c_A/\lambda << b$).

ii. If c_A/λ and λ are sufficiently large, then the optimal budget is positive but below the maximum, $0 < b^* < u$, and there is no auditing.

iii. If c_A is sufficiently large and λ sufficiently small, then the optimal budget is the maximum $b^* = u$, without auditing.

Proposition 5 shows which kind of annual budgets b are optimal for the principal, depending on audit costs and cost of funds. A small cost of funds λ increases the net value of fulfilling additional spending needs and decreases the loss from fund misuse, which favors a larger budget. Small audit costs also tend to favor a larger budget, because fund misuse is a larger issue for large budgets, but audits limit fund misuse at low cost.

The first result is that a maximum budget $b = u$ with extensive auditing is optimal if audit costs are small enough, so that c_A/λ is small. As just explained, small audit costs favor the large budget. Perhaps the only surprise in this result is that a large budget can be optimal even if the cost of funds are large, because very small audit costs remove the problem of fund misuse.

Second, a smaller budget $b < u$ without auditing is optimal if the cost of funds are large and the audit costs are not too small, so that c_A/λ is large. Since unneeded funds remain less often with smaller budgets, fund misuse is not as much of a problem as with a large budget. Combined with the fact that auditing is too expensive to prevent fund misuse, a smaller budget is optimal. Hence, for these parameter values, the principal strategically withholds some funding from the agent due to moral hazard.

Third, if the audit cost is large but the cost of funds is small, then the maximum budget without auditing is optimal. The small λ implies that fulfilling additional spending needs yields a large net value to the principal, and the loss from fund misuse is small. Hence, even without auditing—due to large audit costs—the maximum budget is preferred to smaller budgets that imply less fund misuse.

5.2 Allowing probabilistic audit rules

The model requires audit rule A_y to either audit with probability 1 or with probability 0 for any given spending amount s_y. How would outcomes change if probabilistic auditing was allowed?

First, it would introduce a technical issue, as no optimal rule might exist. As long as punishments effectively deter fund misuse, setting larger punishment is cheaper than auditing more often. Hence, the principal would try to impose draconian punishments in case fund misuse is detected, and to audit only with a low probability $\varepsilon > 0$ small, keeping the expected punishment large enough to deter fund misuse. Since there is no smallest positive number, there is no optimum. However, this could be fixed by using a discrete grid of audit probabilities.

Hence, second, if such a grid allows for smaller audit probabilities, the optimal policy switches to auditing with the lowest positive probability rather than with probability 1, in combination with a large punishment to deter fund misuse. Consequently, allowing lower audit probabilities without any constraints on punishments is mathematically equivalent to a reduction in audit costs, which is already studied in the main model, and tends to lead to more spending amounts being audited. In practice, using probabilistic audits is in fact a way to reduce audit costs while still deterring some malfeasance.

5.3 Robustness: uniform distribution assumption

The main model assumes spending needs are continuously uniformly distributed according to $\theta_y \sim \mathcal{U}(0, u)$. This assumption was made since it allows for closed form solutions, for example for the optimal year two audit threshold. This section argues that the qualitative results generalize beyond this parametric assumption. For this discussion, suppose the spending need realizations are still independent and have an identical, continuous distribution, with pdf $g(\theta)$ and strictly increasing cdf $G(\theta)$ for all $0 \leq \theta \leq \overline{\theta}$, but no further parametric assumptions beyond that.

First, the results that the optimal audit rules take the form of either a threshold rule or an interval rule remains (Lemma 1 and 2), as the arguments do not depend on the distribution. However, the incentive compatibility constraint for the interval rule—which sets the upper border of the interval given any lower border—will be different, as it depends on the distribution. Second, the result that a large punishment is optimal persists (Lemma 3), as the Beckerian argument does not depend on the state distribution.

Third, the roll-over rule must fulfill $\Delta > \alpha$ independent of the distribution in order to be effective. No $\Delta \leq \alpha$ can induce the agent to save funds for roll-over, see (1), since fund misuse dominates in this case. Moreover, as before, the optimal roll-over rule will generally be in the interval $(\Delta_{\max}, 1]$ (Proposition 1), though the expression for $\Delta_{\max}$, which maximizes the saved amount $\hat{x}$ given in (1), will depend on the distribution. Hence, as before, the optimal rule is to let agents roll over some or all of the unused funds.

Fourth, the optimal second year audit rule remains a threshold rule, but the precise threshold depends on the distribution. The appeal of the uniform distribution is that the maximization problem in the interior is strictly concave, so a unique interior solution is guaranteed, and this solution has a simple explicit expression. In the general case, the first order condition given in (12) is $-\lambda G(\underline{a}_2) + c_A g(\underline{a}_2) = 0$, which is not strictly monotone for arbitrary g, so there may be multiple interior solutions, in addition to the corner solution discussed in Proposition 2. Aside from these technical complications, the result remains that the optimal threshold is determined either by an interior solution for large enough b or the corner solution. If density g is non-increasing,[12] then based on the first order condition, the optimal audit threshold is weakly increasing in audit costs and weakly decreasing in the cost of funds, as before.

Similarly, the optimal year 1 audit rules belong to the same class, but the precise audit thresholds or intervals change with the distribution. Proposition 3 showed that there is a unique interior solution for the audit threshold, but for other distributions there may be multiple solutions. The result remains that the year 1 threshold rule tends to audit more than the year 2 rule, as long as the budget b is small enough, so that the principal wants to induce more roll-over to fulfill more spending next year. However, the degree of extra auditing will depend on the distribution.

In summary, the uniform distribution assumption lends itself to simpler expressions for interior solutions and for conditions comparing interior and corner solutions, but it is not driving the main insights in this paper.

6 Concluding remarks

Agents routinely spend sizable portions of their annual budgets in the last month or even week of the fiscal year, before they expire. There is mounting evidence that this use-it-or-lose-it spending is of low value to the principal. This paper studies various rules to prevent or at least limit wasteful year-end spending. In short, rolling over unused funds to the next

[12] For a sufficiently locally increasing density g, these comparative statics may not hold.

year should be allowed, and in some cases it might be optimal to "tax" the roll-over. Audits and punishment to deter fund misuse are a more costly option, and can be optimal for sufficiently large budgets (relative to likely spending needs), not too large audit costs, or for a sufficiently large cost of funds. The optimal audit rules interact with the roll-over rule, and if the roll-over rule is effective in inducing the agent to save unneeded funds, then less auditing tends to be optimal.

Some organizations allow their employees to spend some of their budget or expense account balance for private benefit, which is viewed as a perk and might make the organization more attractive as employer. This is consistent with the model in this paper, by including these perks in the definition of legitimate spending needs and not classifying them as fund misuse. Accordingly, this paper does not take a strong stance on what is undesirable spending, but investigates how to minimize it once it is properly defined.

One might argue that it is easier to restrict wasteful spending with rigid spending rules rather than audits and roll-over. For example, one might decree that only 10% of the budget may be spent on office chairs, a maximum of 20% on laptops, etc. But while such rigid rules might be effective in reducing unwanted spending, they undoubtedly also limit desirable spending. There is a value in spending flexibility for the agent. And if no chairs break down, then additional laptops could be ordered without more paperwork, etc. Hence, the findings of this paper help to minimize wasteful spending while retaining the benefits of some flexibility and discretion.

This paper is only a first step in the analysis of optimal budget rules. Open questions remain. What are the optimal audit rules if there is an upper bound on punishments, so that auditing is not always an effective deterrent? How does the optimal roll-over rule change in an infinite horizon setting where large savings might accumulate? How do the optimal budget rules change once a ratchet effect sets in? These are challenging questions, because the nature of the problem—budget constraints—can induce discontinuities that make first order conditions neither necessary nor sufficient for optimality. Still, the problem is an important one, and good solutions can potentially save organizations a lot of money, which can then be put to better use.

References

Baumann, S. (2019): "Putting it off for later: Procrastination and end of fiscal year spending spikes," *The Scandinavian Journal of Economics*, 121, 706–735.

Becker, G. S. (1968): "Crime and Punishment: An Economic Approach," *The Journal of Political Economy*, 76, 169–217.

Ben-Porath, E., E. Dekel, and B. L. Lipman (2014): "Optimal allocation with costly verification," *American Economic Review*, 104, 3779–3813.

Bird, D. and A. Frug (2019): “Dynamic Non-monetary Incentives,” *American Economic Journal: Microeconomics*, 11, 111–50.

Border, K. C. and J. Sobel (1987): “Samurai accountant: A theory of auditing and plunder,” *The Review of Economic Studies*, 54, 525–540.

Brimberg, J. and W. J. Hurley (2015): “Allocating operating funding in the public sector and the newsvendor problem,” *Journal of the Operational Research Society*, 66, 1035–1043.

Digiday (2017): “‘Irrational budget dumping’: End of the year means use-it-or-lose-it marketing projects,” https://digiday.com/marketing/irrational-budget-dumping-end-year-means-use-lose-marketing-projects/, accessed 10/8/2021.

Freixas, X., R. Guesnerie, and J. Tirole (1985): “Planning under incomplete information and the ratchet effect,” *The Review of Economic Studies*, 52, 173–191.

Hurley, W., J. Brimberg, and B. Fisher (2014): “Use it or lose it: On the incentives to spend annual defence operating budgets,” *Defence and Peace Economics*, 25, 401–413.

Jones, L. R. (2005): “Outyear budgetary consequences of agency cost savings: International public management network symposium,” *International Public Management Review*, 6, 139–168.

Li, Y. (2020): “Mechanism design with costly verification and limited punishments,” *Journal of Economic Theory*, 186, 105000.

Liebman, J. B. and N. Mahoney (2017): “Do expiring budgets lead to wasteful year-end spending? Evidence from federal procurement,” *American Economic Review*, 107, 3510–49.

Malenko, A. (2019): “Optimal dynamic capital budgeting,” *The Review of Economic Studies*, 86, 1747–1778.

McPherson, M. F. (2007): “An analysis of year-end spending and the feasibility of a carryover incentive for federal agencies,” Tech. rep., NAVAL POSTGRADUATE SCHOOL MONTEREY CA.

Military Times (2019): “Use-it or lose-it: DoD dropped $4.6 million on crab and lobster, and $9,000 on a chair in last-minute spending spree,” https://www.militarytimes.com/news/your-military/2019/03/12/use-it-or-lose-it-dod-dropped-46-million-on-crab-and-lobster-and-9000-on-a-chair-in accessed 10/8/2021.

Miller, G. J., D. Robbins, and J. Keum (2007): “Incentives, certification, and targets in performance budgeting,” *Public Performance & Management Review*, 30, 469–495.

Townsend, R. M. (1979): “Optimal contracts and competitive markets with costly state verification,” *Journal of Economic Theory*, 21, 265–293.

A Proofs

Proof of Lemma 1. I will show that the agent reaction to any non-threshold rule can be replicated with a threshold rule. Take any non-threshold audit rule $R(s)$, and define r as $\min r \in \mathbb{R}$ such that $R(r) = 0$ and there exists no $s > r$ with $R(s) = 0$. Then rule $R(s)$ leads to the same agent reaction as a threshold rule $A(s)$ with $\underline{a} = r$. This is because the agent best responds by spending $s_y = \underline{a} = r$ whenever $\theta_y \leq \underline{a} = r$ and $s_y = \theta_y$ whenever $\theta_y > \underline{a} = r$ under either rule.

If no minimum r as specified above exists, then use a threshold rule that also audits at the threshold, $A(s) = 1$ whenever $s = \underline{a}$, and is otherwise the same. Again take any non-threshold audit rule $R(s)$, and define $r \coloneqq \{\max r \in \mathbb{R} : R(r) = 1\}$. Then rule $R(s)$ leads to the same agent reaction as the threshold rule with $\underline{a} = r$. Hence, a threshold rule can always be part of the optimal policy. □

Proof of Lemma 2. Absent auditing, agents spend $s_1 = b - \bar{x}$ if $\theta_1 \leq b - \bar{x}$ by construction of $\bar{x}$, and $s_1 = \min\{\theta_1, b\}$ if $\theta_1 > b - \bar{x}$. In order to decrease fund misuse for $\theta_1 \leq b - \bar{x}$, a continuous interval $s_1 \in (\underline{a}_1, \overline{a}_1)$ has to be audited, with $\underline{a}_1 < b - \bar{x}$ and $\underline{a}_1 > b - \bar{x}$, as shown below. That is, only an interval rule can decrease fund misuse further. An interval rule with $\underline{a}_1 = \overline{a}_1 = b - \bar{x}$ is a special case that does not audit. Hence, the optimal audit rule is in the family of interval rules.

In the next step I narrow down what kind of interval rule is optimal. Setting $\underline{a}_1 < b - \bar{x}$ requires a specific $\overline{a}_1$ so that certain agent types do not misuse funds, i.e., so that $s_1 = \underline{a}_1$ is incentive compatible.

The difference in agent utility of spending $s_1 = \underline{a}_1$ vs spending $s_1 = \overline{a}_1$ if $\theta_1 \leq \underline{a}_1$, which needs to be non-negative for those agents to choose $s_1 = \underline{a}_1$, is:

$$\begin{aligned}\int_{b+\Delta(b-\overline{a}_1)}^{b+\Delta(b-\underline{a}_1)} \theta_2 - (b + \Delta(b - \overline{a}_1)) dG(\theta_2) + \int_{b+\Delta(b-\underline{a}_1)}^{u} \Delta(\overline{a}_1 - \underline{a}_1) dG(\theta_2) - \alpha(\overline{a}_1 - \underline{a}_1) \\ = \frac{(\overline{a}_1 - \underline{a}_1)(\overline{a}_1\Delta^2 + \underline{a}_1\Delta^2 - 2b\Delta(1+\Delta) - 2\alpha u + 2\Delta u)}{2u} = 0,\end{aligned} \tag{8}$$

a quadratic equation in $\overline{a}_1$, which is trivially fulfilled for $\overline{a}_1 = \underline{a}_1$. The other solution is

$$\overline{a}_1(\underline{a}_1) = \frac{2b\Delta(1+\Delta) + 2\alpha u - 2\Delta u - \underline{a}_1\Delta^2}{\Delta^2},$$

for which the agents with $\theta_1 \leq \underline{a}_1$ are indifferent and hence $s_1 = \underline{a}_1$ is incentive compatible. Moreover, it is easy to see that any $\theta_1 \geq b - \bar{x}$ spends $s_1 = \theta_1$ by construction of $\bar{x}$. And any $\theta_1 \in (\underline{a}_1, b - \bar{x})$ has a positive difference in expected utility between setting $s_1 = \theta_1$ and $s_1 = \overline{a}_1$ if (8) is fulfilled. This is because (8) sets $\overline{a}_1$ to exactly offset the lost marginal utility of misusing more funds (larger for $s_1 < b - \bar{x}$) with the additional expected marginal utility of rolling over more (larger for $s_1 > b - \bar{x}$). So when the range where misusing more funds is attractive is shrunk from $(\underline{a}_1, b - \bar{x})$ to $(\theta_1, b - \bar{x})$, then misusing more funds is less attractive and $s_1 = \theta_1$ is more so. That is, the agent expected utility difference between $s_1 = \theta_1$ and $s_1 = \overline{a}_1$,

$$\int_{b+\Delta(b-\overline{a}_1)}^{b+\Delta(b-\theta_1)} \theta_2 - (b + \Delta(b - \overline{a}_1))dG(\theta_2) + \int_{b+\Delta(b-\theta_1)}^{u} \Delta(\overline{a}_1 - \theta_1)dG(\theta_2) - \alpha(\overline{a}_1 - \theta_1)$$

is positive for $\theta_1 \in (\underline{a}_1, b - \bar{x})$ if (8) holds, since $\theta_1 > \underline{a}_1$. Finally, spending $s_1 > \overline{a}_1$ is clearly dominated for the agent by $s_1 = \theta_1$, and $s_1 < \underline{a}_1$ is clearly dominated by either $s_1 = \underline{a}_1$ or $s_1 = \theta_1 > \underline{a}_1$.

But there is another way of setting the interval. Suppose we determined $\underline{a}_1 < \overline{a}_1$ such that (8) holds. Keeping $\underline{a}_1$ fixed, we reduce $\overline{a}_1$ slightly to $\overline{a}_1' < \overline{a}_1$ with $b - \bar{x} < \overline{a}_1'$. As a consequence, types $\theta_1 \leq \underline{a}_1$ switch from $s_1 = \underline{a}_1$ to $s_1 = \overline{a}_1'$. But there is a marginal type $\theta_1' \in (\underline{a}_1, b - \bar{x})$ who is indifferent between $s_1 = \theta_1'$ and $s_1 = \overline{a}_1'$. Any $\theta_1 > \theta_1'$ strictly prefers $s_1 = \theta_1$.

I will now show that such interval rules, where agent types $\theta_1 \leq \underline{a}_1$ spend above the audit threshold ($s_1 = \overline{a}_1'$), are dominated for the principal by an interval rule with $\underline{a}_1, \overline{a}_1$ such that (8) holds. Under rule $(\underline{a}_1, \overline{a}_1')$ just constructed, any $\theta_1 < \theta_1'$ sets $s_1 = \overline{a}_1'$, since θ_1' is the marginal type, whereas any $\theta_1 \geq \theta_1'$ sets $s_1 = \theta_1$. The ex ante audit cost in year 1 is therefore $c_A(\overline{a}_1' - \theta_1')/u$, and the ex ante fund misuse in year 1 is

$$\int_0^{\theta_1'} \overline{a}_1' - \theta_1 dG(\theta_1) = \frac{\overline{a}_1'\theta_1' - {\theta_1'}^2/2}{u}.$$

Now consider instead the interval rule $\underline{a}_1 = \theta_1'$ and $\overline{a}_1 = \overline{a}_1'$. That is, the upper bound is the same, but the lower bound is larger, compared to the previous rule. Since θ_1' is the marginal type for whom $s_1 = \theta_1$ is weakly better than $s_1 = \overline{a}_1' = \overline{a}_1$, it follows that this rule sets (8) to zero. Consequently, all $\theta_1 \leq \underline{a}_1$ set $s_1 = \overline{a}_1$ as well, as their decision problem between $s_1 = \underline{a}_1$ and $s_1 = \overline{a}_1$ is the same as for $\theta_1 = \theta_1'$. And all $\theta_1 > \underline{a}_1$ set $s_1 = \theta_1$. Consequently, the ex ante audit cost under this rule is $c_A(\overline{a}_1 - \underline{a}_1)/u$, which is identical to the previous rule, because $\overline{a}_1 = \overline{a}_1'$ and $\underline{a}_1 = \theta_1'$. However, the ex ante fund misuse under this rule is smaller,

$$\int_0^{\underline{a}_1} \underline{a}_1 - \theta_1 dG(\theta_1) = \frac{\underline{a}_1^2 - \underline{a}_1^2/2}{u} = \frac{{\theta_1'}^2 - {\theta_1'}^2/2}{u} < \frac{\overline{a}_1'\theta_1' - {\theta_1'}^2/2}{u},$$

since $\overline{a}_1' > \theta_1'$. Therefore, any interval rule for which (8) is negative is dominated by one for which it is zero. And clearly, any rule for which (8) is positive is suboptimal as well, as it could audit less and still achieve the same spending decisions by the agent. Hence, out of all possible interval rules, we can restrict attention to those with $\underline{a}_1 \leq b - \bar{x}$ and the associated unique $\overline{a}_1(\underline{a}_1) > \underline{a}_1$ determined by (8). In other words, the interval rule that fulfills the incentive compatibility constraint with equality is best. □

Proof of Lemma 3. By contradiction, suppose punishment p is not as large, so that there exists a spending amount s such that $A_y(s) = 1$ and the agent spends s_y with $s_y = s > \theta_y$ with positive probability in equilibrium. That is, the agent misuses funds, resulting in a spending amount that triggers an audit. Clearly, setting punishment above $2\alpha b$ instead is weakly better for the principal, holding everything else constant. The benefit to the agent in misusing all funds is $\alpha(b_y - \theta_y)$, with maximum $2\alpha b$ for $b_2 = 2b$ and $\theta_2 = 0$, which is outweighed by punishment $p > 2\alpha b$. Hence, in all cases where realizations $\theta_y < s$ lead to agent spending $s_y = s$ (implying fund misuse at cost λ, audit and punishment costs c_A, p for the principal), $p > 2\alpha b$ effectively prevents fund misuse at s, and saves audit and punishment costs, which strictly increases the principal's expected utility. □

Proof of Proposition 1.

i. First, I am going to show that $\bar{x}$ is maximized for Δ set to (5). Any maximizer of $\hat{x}$ also maximizes $\bar{x} = \min\{\max\{\hat{x}, 0\}, b\}$, so from now on focus on $\hat{x}$. Plugging the inverse of the uniform CDF into the expression for $\hat{x}$ in (1) yields

$$\Delta_{\max \bar{x}} = \arg \max_{\Delta \in [0,1]} \left(\frac{1}{\Delta} - \frac{\alpha}{\Delta^2} \right) u - \frac{b}{\Delta}.$$

Only consider $\Delta \geq \alpha$, since $\alpha > \Delta$ implies $\hat{x} < 0$ and hence $\bar{x} = 0$, and so can be ruled out as maximum. Clearly, for $\Delta \geq \alpha$, $\hat{x}$ is continuous, so by the maximum theorem a maximizing Δ exists on $[\alpha, 1]$. Differentiating with respect to Δ once yields

$$-\frac{u}{\Delta^2} + \frac{2\alpha u}{\Delta^3} + \frac{b}{\Delta^2}, \tag{9}$$

with an interior solution to the necessary first order condition, and upper bound, at

$$\Delta_{\max \bar{x}} = \min \left\{ \frac{2\alpha u}{u - b}, 1 \right\},$$

which strictly exceeds α for all parameter values, and is positive due to the assumption of $b \leq u = u$. The objective $\hat{x}$ is not in general strictly concave in Δ, but it can still be shown that (5) is a maximum.

To show this, note both $\hat{x}$ as well as the first derivative is continuous for $\Delta > 0$. Moreover, setting derivative (9) to zero reduces it to a linear function, which implies there is only one solution to the first order condition and this solution cannot be a saddle point. Further, derivative (9) is strictly positive at $\Delta = \alpha$ for any $u \geq b$, so the solution to the first order condition is a maximum. Consequently, (5) maximizes $\bar{x}$.

Second, if $\bar{x} > 0$, then an interval audit rule with $\underline{a}_1 \leq b - \bar{x}$ is used (Lemma 2). While keeping $\underline{a}_1$ fixed, a change in Δ can change $b - \bar{x}$, and consequently $\overline{a}_1(\underline{a}_1)$ fulfilling the incentive compatibility constraint (8) changes. This Δ-change affects the expected audit cost and hence principal EU, while expected fund misuse remains the same when keeping $\underline{a}_1$ fixed.

Consider $\underline{a}_1$ such that $\overline{a}_1(\underline{a}_1) < b$. Taking the derivative of the principal EU in (4) with respect to Δ, using Leibniz' integral rule and simplifying, yields

$$\begin{aligned} &\int_0^{\underline{a}_1} \frac{\partial V(b + \Delta(b - \underline{a}_1))}{\partial b_2} \cdot \frac{\partial b_2}{\partial \Delta} - \lambda(b - \underline{a}_1)\mathrm{d}G(\theta_1) \\ &+ \int_{\underline{a}_1}^{b} \frac{\partial V(b + \Delta(b - \theta_1))}{\partial b_2} \cdot \frac{\partial b_2}{\partial \Delta} - \lambda(b - \theta_1)\mathrm{d}G(\theta_1) - \frac{\partial \overline{a}_1(\underline{a}_1)}{\partial \Delta} c_A g(\overline{a}_1). \end{aligned} \tag{10}$$

Clearly, $\frac{\partial V(b_2)}{\partial b_2} = 1 - (1 - \lambda)G(b_2) \in [\lambda, 1]$, $\frac{\partial b_2}{\partial \Delta} = b - \underline{a}_1$ or $\frac{\partial b_2}{\partial \Delta} = b - \theta_1$, respectively. Moreover, based on (8), $\frac{\partial \overline{a}_1(\underline{a}_1)}{\partial \Delta} = \overline{a}_1' > 0$ if $\frac{\partial (b - \bar{x})}{\partial \Delta} > 0$, which in turn occurs iff $\Delta > \Delta_{\max \bar{x}}$, as the first part just showed. Similarly, $\overline{a}_1' < 0$ if $\Delta < \Delta_{\max \bar{x}}$, and $\overline{a}_1' = 0$ if $\Delta = \Delta_{\max \bar{x}}$. Plugging these in, the derivative becomes

$$\begin{aligned} &\int_0^{\underline{a}_1} (b - \underline{a}_1)[1 - (1 - \lambda)G(b_2) - \lambda]\mathrm{d}G(\theta_1) \\ &+ \int_{\underline{a}_1}^{b} (b - \theta_1)[1 - (1 - \lambda)G(b_2) - \lambda]\mathrm{d}G(\theta_1) - \overline{a}_1' c_A g(\overline{a}_1), \end{aligned}$$

which is non-negative for any $\Delta < \Delta_{\max \bar{x}}$, as $1 - (1 - \lambda)G(b_2) \geq \lambda$, so all three terms are weakly positive. At $\Delta = \Delta_{\max \bar{x}}$, the first two terms are weakly positive, whereas the last is zero, so the derivative is non-negative. And for $\Delta > \Delta_{\max \bar{x}}$ sufficiently close to $\Delta_{\max \bar{x}}$, the last term is negative (since $\bar{x} > 0$ at $\Delta = \Delta_{\max \bar{x}}$ by assumption) while the first two are weakly positive. Consequently, this derivative is non-negative for any $\Delta \leq \Delta_{\max \bar{x}}$, but can be zero and negative in $\Delta > \Delta_{\max \bar{x}}$, indicating a maximum in that range.

Third, consider $\underline{a}_1$ such that $\overline{a}_1(\underline{a}_1) > b$. In this case, the change of Δ does not change $\overline{a}_1(\underline{a}_1) > b$. Taking the derivative of (4) with respect to Δ, and simplifying as in the previous part yields

$$\begin{aligned} &\int_0^{\underline{a}_1} \frac{\partial V(b + \Delta(b - \underline{a}_1))}{\partial b_2} \cdot \frac{\partial b_2}{\partial \Delta} - \lambda(b - \underline{a}_1)\mathrm{d}G(\theta_1) \\ &+ \int_{\underline{a}_1}^{b} \frac{\partial V(b + \Delta(b - \theta_1))}{\partial b_2} \cdot \frac{\partial b_2}{\partial \Delta} - \lambda(b - \theta_1)\mathrm{d}G(\theta_1) \geq 0. \end{aligned} \tag{11}$$

By assumption, $\overline{a}_1(\underline{a}_1) < b$ at $\Delta = \Delta_{\max \bar{x}}$. If, in addition, $\overline{a}_1(\underline{a}_1) > b$ at $\Delta = 1$, then an increase in $\overline{a}_1$ due to an increase in Δ can cause $\overline{a}_1 > b$, which in turn causes a discontinuous increase in expected audit costs of $c_A(u-b)/ > 0$ and a discontinuous drop in the principal EU of the same amount. This favors a $\Delta < 1$. Moreover, whether derivative (10) or (11) is valid depends on Δ. Taken together, principal EU is potentially decreasing in Δ when (10) is valid, weakly increasing when (11) is valid, and discontinuously decreasing when switching from (10) to (11). These calculations imply the optimal roll-over rule is $\Delta^* \in (\Delta_{\max \bar{x}}, 1]$.

Moreover, if $\bar{x} = b$ at $\Delta = 1$, then $\Delta^* = 1$ maximizes $\bar{x}$ and maximizes the roll-over, hence we can rule out any $\Delta < 1$ as optimal rule.

Note that $1-(1-\lambda)G(b_2)-\lambda$ becomes arbitrarily small as $\lambda \to 1$, whereas $\overline{a}_1' c_A g(b-\bar{x})$ is independent of λ (see (8) which determines $\overline{a}_1'$ and does not depend on λ). Consequently, the derivative (10) is negative at $\Delta >> \Delta_{\max \bar{x}}$, so a $\Delta^* \in (\Delta_{\max \bar{x}}, 1)$ is optimal.

ii. Since $\bar{x}$ is maximized at $\Delta_{\max \bar{x}}$, $\underline{a}_1 \leq b - \bar{x}$ with $\bar{x} = 0$ at $\Delta = \Delta_{\max \bar{x}}$ implies $\bar{x} = 0$ and $\underline{a}_1 \leq b - \bar{x} = b$ for any Δ. In this case, taking the derivative of (3) with respect to Δ yields

$$\int_0^{\underline{a}_1} (b - \underline{a}_1)[1-(1-\lambda)G(b_2)-\lambda]\mathrm{d}G(\theta_1) + \int_{\underline{a}_1}^{b} (b-\theta_1)[1-(1-\lambda)G(b_2)-\lambda]\mathrm{d}G(\theta_1) \geq 0,$$

due to $1-(1-\lambda)G(b_2) \geq \lambda$. Hence, $\Delta^* = 1$ is optimal in this case, though may not be the only optimum. If $\Delta_{\max \bar{x}} = 1$, then as shown above, principal EU cannot decrease in Δ for $\Delta < \Delta_{\max \bar{x}}$, so $\Delta^* = 1$ is optimal. If $\bar{x} = b$ at $\Delta = 1$, then $\Delta = 1$ maximizes $\bar{x}$ even if $\Delta_{\max \bar{x}} < 1$ since the min-operator in $\bar{x}$ is binding. Consequently, $\Delta^* = 1$ is optimal. Finally, if $\overline{a}_1(\underline{a}_1) > b$ at $\Delta < \Delta_{\max \bar{x}} < 1$, then $\overline{a}_1' = 0$ and hence the derivative in (10) is non-negative, hence $\Delta^* = 1$ is optimal.

iii. If $b = u$, then $\bar{x} = 0$ for any Δ. Moreover, any roll-over does not increase principal EU in year 2, as all spending needs can already be fulfilled. Hence, any $\Delta^* \in [0,1]$ is optimal.

If $\underline{a}_1 = b$ and $\bar{x} = 0$ at $\Delta = \Delta_{\max \bar{x}}$, then there is no auditing in year 1 due to $\underline{a}_1 = b$ and no fund roll-over since $\bar{x} = 0$ for any Δ. Consequently, the choice of Δ does not matter. □

Proof of Proposition 2.

i. Using Leibniz' integral rule, the marginal principal EU of changing the threshold $\underline{a}_2$ is

$$\begin{aligned}\frac{\partial EU}{\partial \underline{a}_2} = \frac{\partial V(b_2)}{\partial \underline{a}_2} &= (\underline{a}_2 + \lambda(b_2 - \underline{a}_2))g(\underline{a}_2) + \int_0^{\underline{a}_2} -\lambda \mathrm{d}G(\theta_2) - (\underline{a}_2 + \lambda(b_2 - \underline{a}_2) - c_A)g(\underline{a}_2) \\ &= -\lambda G(\underline{a}_2) + c_A g(\underline{a}_2).\end{aligned} \tag{12}$$

The maximization problem for $\underline{a}_2$ is strictly concave, as using the uniform PDF and CDF and differentiating the marginal utility in (12) yields $-\lambda/u < 0$. Hence, a unique interior solution is guaranteed, and using the uniform PDF and CDF in (12) and rearranging yields

$$\underline{a}_2^* = \frac{c_A}{\lambda}.$$

If $b = u$ or $b < b_2$, then there is no discontinuity at $\underline{a}_2 = b$, since there is no probability mass point at $s_2 = b$ given the agent reaction function. Hence, by strict concavity, the interior solution is the optimal policy, unless the constraint $c_A/\lambda \leq b$ is binding, in which case the corner solution is optimal.

However, if $b_2 = b < u$, then there is a discontinuity at $\underline{a}_2 = b$, since all $\theta_2 \in [b, u]$ agent types spend $s_2 = b$ due to the budget constraint. So even if $c_A/\lambda \leq b$ is not binding, the corner solution can be optimal due to the discontinuity. The interior solution (given $c_A/\lambda \leq b$) is optimal in this case if and only if

$$\begin{aligned}&\int_0^{\underline{a}_2^*} \theta_2 + \lambda(b_2 - \underline{a}_2^*)\mathrm{d}G(\theta_2) + \int_{\underline{a}_2^*}^{b_2} \theta_2 + \lambda(b_2 - \theta_2) - c_A \mathrm{d}G(\theta_2) + \int_{b_2}^{u} b_2 - c_A \mathrm{d}G(\theta_2) \\ &\geq \int_0^{b_2} \theta_2 \mathrm{d}G(\theta_2) + \int_{b_2}^{u} b_2 \mathrm{d}G(\theta_2) \iff \frac{b_2^2\lambda}{2} + \frac{c_A^2}{2\lambda} \geq u c_A,\end{aligned} \tag{13}$$

where the left hand side is the expected principal utility in year 2 from using the interior solution $\underline{a}_2^* = c_A/\lambda$ and the right hand side is the principal expected utility from using $\underline{a}_2 = b$.

ii. If, instead, (13) is not fulfilled, then the corner solution $\underline{a}_2 = b$ is optimal. □

Proof of Proposition 3.

i. If $b - \bar{x} \geq \underline{a}_1$, then a marginal change in $\underline{a}_1$ also changes the spending and saving of the agent, so we need to take into account the effects of a change of $\underline{a}_1$ in year 2.

$$\begin{aligned}\frac{\partial EU}{\partial \underline{a}_1} =&(\underline{a}_1 + V(b+\Delta(b-\underline{a}_1)) + \lambda(1-\Delta)(b-\underline{a}_1))g(\underline{a}_1) + \int_0^{\underline{a}_1}(V' - \lambda(1-\Delta))\mathrm{d}G(\theta_1)\\ &- (\underline{a}_1 - c_A + V(b+\Delta(b-\underline{a}_1)) + \lambda(1-\Delta)(b-\underline{a}_1))g(\underline{a}_1)\\ =&c_A g(\underline{a}_1) - G(\underline{a}_1)[\lambda(1-\Delta) + (1-(1-\lambda)G(b_2))\Delta],\end{aligned} \tag{14}$$

since $\partial b_2/\partial \underline{a}_1 = -\Delta$ for θ_1-realizations where the agent spends up to $\underline{a}_1$, and hence

$$V' = \partial V/\partial b_2 \cdot (-\Delta) = -(1-(1-\lambda)G(b_2))\Delta,$$

where b_2 for $\theta_1 \in [0, \underline{a}_1]$ is $b_2 = b + \Delta(b - \underline{a}_1)$. Substituting the PDF and CDF of the uniform distribution into the marginal utility function of $\underline{a}_1$ in (14), and setting to zero, yields the following quadratic equation:

$$c_A - \underline{a}_1\left(\lambda(1-\Delta) + \Delta\left(1-(1-\lambda)\min\left\{\frac{(1+\Delta)b - \Delta\underline{a}_1}{u}, 1\right\}\right)\right) = 0. \tag{15}$$

However, this marginal utility function is quadratic in $\underline{a}_1$ only as long as $b_2 = (1+\Delta)b - \Delta\underline{a}_1 \leq u$. If this condition is not fulfilled, then this expression continuously reduces to the same as for $\underline{a}_2$ in (12) and becomes linear in $\underline{a}_1$:

$$c_A - \underline{a}_1\lambda = 0.$$

Assuming $b_2 \leq u$, the local maximum is a solution of the quadratic equation (15), found the usual way, and is given by

$$a := \frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1\beta_3}}{2\beta_1}, \tag{16}$$

with β-terms defined in Proposition 3. This maximum is the larger of the two solutions to the quadratic equation, since $\beta_1 < 0$, so the marginal utility is positive just below the larger solution, indicating a maximum. This solution is positive, since $-\beta_2 > 0$, $-\sqrt{\beta_2^2 - 4\beta_1\beta_3} < 0$, $\beta_1 < 0$, and $|\beta_2| < \sqrt{\beta_2^2 - 4\beta_1\beta_3}$. The other solution, the local minimum, is negative, since $-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1\beta_3} > 0$. Consequently, any $\underline{a}_1$ below the minimum is negative and hence not feasible, so the local maximum (16) is in fact a global maximum for $\underline{a}_1 \in [0, b]$.

If the maximum of the quadratic equation a fulfills $(1+\Delta)b - \Delta a \leq u$, then no $\underline{a}_1 < a$ is preferable to a. This follows because the expected utility and marginal utility is continuous at $(1+\Delta)b - \Delta\underline{a}_1 = u$, the marginal utility in the quadratic range is positive for $\underline{a}_1 \in [0, a)$ as just shown, and the marginal utility is positive and decreasing for

$(1+\Delta)b - \Delta\underline{a}_1 > u$ (in the linear range) as well and hence cannot be another local maximum.

If, on the other hand, the maximum of the quadratic equation a fulfills $(1+\Delta)b - \Delta a > u$, as it might for large b, then it is not the optimal policy as the marginal utility is not quadratic in that range. However, the quadratic solution a converges to the linear solution c_A/λ from below as $G(b_2)$ increases. To show this, first note that the quadratic maximum increases in Δ since $b \geq \underline{a}_1$, and increases in b:

$$\begin{aligned} \frac{\partial(16)}{\partial b} &\propto \frac{\partial}{\partial b}\left(\beta_2 + \sqrt{\beta_2^2 - 4\beta_1\beta_3}\right) > 0 \\ \iff &\ \beta_2' + 2\beta_2\beta_2'(\beta_2^2 - 4\beta_1\beta_3)^{-1/2}/2 > 0 \\ &\iff 1 + \beta_2/\sqrt{\beta_2^2 - 4\beta_1\beta_3} > 0, \end{aligned}$$

which holds because $|\beta_2| < \sqrt{\beta_2^2 - 4\beta_1\beta_3}$ due to $\beta_1\beta_3 < 0$, hence $\beta_2/\sqrt{\beta_2^2 - 4\beta_1\beta_3} \in (-1, 0)$. The proportionality in the first line follows due to $\beta_1 < 0$, and the third line follows because $\beta_2' > 0$. Second, substituting the linear solution $\underline{a}_1 = c_A/\lambda$ into the quadratic equation (15) and solving for b yields

$$b = \frac{c_A\Delta/\lambda + u}{1-\Delta},$$

which is exactly the value of b for which $(1+\Delta)b - \Delta\underline{a}_1 = u$ at $\underline{a}_1 = c_A/\lambda$. Consequently, the quadratic maximum in (16) converges to c_A/λ from below as $G(b_2) \to 1$. Therefore, the optimal audit threshold is the minimum of the quadratic and the linear solution, as long as these do not exceed b. Adding this latter constraint, we get (6) as optimal threshold.

If $b = u$, there is no discontinuity in the principal expected utility at $\underline{a}_1 = b$, so (6) is optimal. On the other hand, if $b < u$, then the interior solution $z := \min\{a, c_A/\lambda\}$ is optimal if $z \leq b$ and

$$\begin{aligned} &\int_0^z \theta_1 + V(b + \Delta(b-z)) + \lambda(1-\Delta)(b-z)\mathrm{d}G(\theta_1) \\ &+ \int_z^b \theta_1 + V(b + \Delta(b-\theta_1)) + \lambda(1-\Delta)(b-\theta_1) - c_A\mathrm{d}G(\theta_1) + \int_b^u b + V(b) - c_A\mathrm{d}G(\theta_1) \\ &\geq \int_0^b \theta_1 + V(b)\mathrm{d}G(\theta_1) + \int_b^u b + V(b)\mathrm{d}G(\theta_1) \\ \iff &\ \frac{a}{u}[V(b+\Delta(b-z)) - V(b) + \lambda(1-\Delta)(b-z)] \\ &+ \frac{1}{u}\int_z^b V(b+\Delta(b-\theta_1)) - V(b) + \lambda(1-\Delta)(b-\theta_1)\mathrm{d}\theta_1 \geq (u-z)c_A/u. \end{aligned} \tag{17}$$

ii. If, instead, (17) is not fulfilled, then the corner solution $\underline{a}_1 = b$ is optimal.

iii. This result is straightforward, since part i. showed (16) is increasing in b and converging to c_A/λ, hence

$$\underline{a}_1^t = \min\left\{\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1\beta_3}}{2\beta_1}, \frac{c_A}{\lambda}, b\right\} \leq \min\left\{\frac{c_A}{\lambda}, b\right\} = \underline{a}_2^*.$$

iv. This is straightforward, as all conditions and expressions in this proposition simplify to their counterparts in Proposition 2 for $\underline{a}_1 = c_A/\lambda$ if $b + \Delta(b - c_A/\lambda) \geq u$. □

Proof of Proposition 4.

i. With an interval rule, by (2), $\partial\overline{a}_1/\partial\underline{a}_1 = -1$, so an increase in $\underline{a}_1$ is met with an identical decrease in $\overline{a}_1$ to fulfill incentive compatibility constraint (8). Taking the derivative of principal EU in (4) with respect to $\underline{a}_1$, while $\underline{a}_1 \leq b - \bar{x}$, and taking into account the change in $\overline{a}_1(\underline{a}_1)$ such that incentive compatibility constraint (8) remains zero, yields

$$\frac{\partial EU}{\partial \underline{a}_1} = c_A g(\underline{a}_1) + c_A g(\overline{a}_1) - G(\underline{a}_1)[\lambda(1-\Delta) + (1-(1-\lambda)G(b_2))\Delta], \tag{18}$$

as long as $\overline{a}_1(\underline{a}_1) \leq b$, otherwise the interval rule becomes a threshold rule and the derivative is as in Proposition 3. Putting in the expression for b_2 and the PDFs and CDFs yields the first order condition

$$2c_A - \underline{a}_1\left(\lambda(1-\Delta) + \Delta\left(1 - (1-\lambda)\min\left\{\frac{(1+\Delta)b - \Delta\underline{a}_1}{u}, 1\right\}\right)\right) = 0 \tag{19}$$

if $\overline{a}_1(\underline{a}_1) \leq b$, and the first order condition becomes (15) as in the optimal threshold rule computation otherwise. Condition (19) is identical to the threshold rule first order condition (15) except the benefit factor of increasing $\underline{a}_1$ is larger, as an increase of $\underline{a}_1$ simultaneously decreases $\overline{a}_1(\underline{a}_1)$ to keep (8) zero. Hence, if there is an interior solution fulfilling (19) while $\overline{a}_1(\underline{a}_1) \leq b$, then it is larger than the interior solution of the threshold rule, since (19) is positive whenever (15) is zero for the same $\underline{a}_1$.

As in Proposition 3, this first order condition is quadratic, but there is only one solution in the positive range. Consequently, the marginal utility is positive on one side of the solution and negative to the other side of the solution, hence the principal EU is larger the closer $\underline{a}_1$ to the solution of this condition. This quadratic first order condition simplifies, if $\underline{a}_1 \leq \frac{(1+\Delta)b-u}{\Delta}$, to the linear one,

$$2c_A - \lambda\underline{a}_1 = 0 \iff \underline{a}_1 = 2c_A/\lambda. \tag{20}$$

By the same arguments as in Proposition 3, since the term second term in (**??**) is equivalent to the second term in (15), the solution to the quadratic equation converges to the linear solution in (20) as $\frac{(1+\Delta)b-\Delta\underline{a}_1}{u}$ converges to 1 from below. So as in Proposition 3, the optimal audit threshold is the minimum of the quadratic and the linear solution. If none of these solutions fall into the interval $[0, b-\bar{x}]$, then the marginal utility must be increasing everywhere in this range and the corner solution $\underline{a}_1 = b - \bar{x}$ is optimal, which defines a^i in the proposition.

There is a discontinuity at $\overline{a}_1(\underline{a}_1) = b$, where a slight reduction of $\underline{a}_1$ increases $\overline{a}_1$ above b. Since $s_1 = b$ for all $\theta_1 \in [b, u]$—which is a nonempty interval because $\bar{x} > 0$ implies $b < u$—this increase in $\overline{a}_1$ increases the audit costs discontinuously by $c_A(u-b)/u$ and thus decreases principal EU by the same amount. Define $a^d := \underline{a}_1$ such that $\overline{a}_1(\underline{a}_1) = b$ in (2), which is at the discontinuity and separates the region of $\underline{a}_1$ where the first order condition is determined by (19) or (15), respectively. Recall the (interior) solution for the threshold rule is a^t, which solves (15), with $a^t < a^i$ if $b - \bar{x} > a^t$, as established above.

Therefore, if $a^d < a^t < a^i$, then a^i solving (19) is optimal in the range $\underline{a}_1 \in (a^d, b-\bar{x}]$, and is therefore better than a^t. If $a^t < a^i < a^d$, then either a^t solving (15) is optimal, whereas a^i is not as it does not solve (15) in $\underline{a}_1 \in [0, a^d)$. Or a^d is optimal due to the discontinuity. Finally, if $a^t < a^d < a^i$, then either $\underline{a}_1 = a^t$ or $\underline{a}_1 = a^i$ is optimal. Consequently, the optimal audit threshold $\bar{a}_1^i$ is weakly larger than a^t if $b - \bar{x} > a^t$.

The principal being weakly better off under $\bar{x} > 0$ follows from the fact that any threshold rule under $\bar{x} = 0$ can also be used with $\bar{x} > 0$, and whenever $\underline{a}_1^t < b$, the same outcome might be achieved with an interval rule at a lower audit cost.

ii. If $b - \bar{x} \leq a^t$, then any interval rule sets $\underline{a}_1 \leq b - \bar{x}$, and hence must fulfill $\underline{a}_1 \leq a^t$. Since $a^i \leq \min\{a^t, b\}$, the optimal interval rule has a weakly smaller $\underline{a}_1$, and hence prevents weakly more fund misuse. The principal is strictly better off, since $\underline{a}_1^t \geq a^t \geq b - \bar{x}$, according to Proposition 3. Since $\underline{a}_1 = b - \bar{x}$ incurs zero audit costs and prevents at least as much fund misuse as the threshold rule, the principal is strictly better off.

iii. As in part i., $\underline{a}_1 = a^i$ is optimal if $a^d < a^t < a^i$, or if $a^t < a^d < a^i$ and $EU(\underline{a}_1 = a^t) \leq EU(\underline{a}_1 = a^i)$. Since part i. assumed $b - \bar{x} > a^t$, which rules out $a^t > a^i$, we also have to consider $a^i < a^t$, which can occur if $a^i = b - \bar{x}$. In this case, clearly a^i is optimal, since it costlessly prevents more fund misuse than a^t does.

iv. $\underline{a}_1 = a^t$ is optimal if $a^t < a^i < a^d$ and $EU(\underline{a}_1 = a^t) \geq EU(\underline{a}_1 = a^d)$, or if $a^t < a^d < a^i$ and $EU(\underline{a}_1 = a^t) \geq EU(\underline{a}_1 = a^i)$. Since $\overline{a}_1(\underline{a}_1 = a^t) > b$ according to (8), any $\overline{a}_1 > b$ is optimal, which is outcome equivalent to a threshold rule.

v. $\underline{a}_1 = a^d$ is optimal if $a^t < a^i < a^d$ and $EU(\underline{a}_1 = a^t) \leq EU(\underline{a}_1 = a^d)$, or if $a^i < a^t < a^d$ and $EU(\underline{a}_1 = a^t) \leq EU(\underline{a}_1 = a^d)$. The optimal $\overline{a}_1$ follows from (8), and by construction of a^d equals b. □

Proof of Corollary 1. The optimal punishment is established in Lemma 3. Since the year 2 audit threshold is independent of both $\underline{a}_1$ and Δ, as it can condition on b_2, the optimal $\underline{a}_2^*$ in Proposition 2 is optimal for all $\underline{a}_1$ and Δ.

i. If $\alpha > 1/2$, then $\Delta_{\max \bar{x}} = 1$, hence $\Delta^* = 1$ according to Proposition 1. Similarly, $\Delta_{\max \bar{x}} \geq 1 \iff b \geq u(1-2\alpha)$.

ii. If c_A/λ is sufficiently small and b sufficiently large, then $\underline{a}_1^t = \underline{a}_2^* = c_A/\lambda$ is optimal for any $\Delta > \alpha$ according to Proposition 3 and 2. Moreover, for b large enough, $\bar{x} = 0$ for any $\Delta \in [0,1]$, so the year 1 audit rule is a threshold rule, and hence $\Delta^* = 1$ is optimal according to Proposition 1, ii.

iii. Small α and b imply $\bar{x} > 0$, so the year 1 audit rule is an interval rule. If $b > \frac{(1-\alpha)u}{2}$, then $\bar{x} < b$ at $\Delta = 1$. Small α also implies $\Delta_{\max \bar{x}} < 1$, hence $\Delta^* \in (\Delta_{\max \bar{x}}, 1]$ is optimal if $\bar{a}_1(\underline{a}_1^i) < b$ according to Proposition 1, i. Moreover, a sufficiently small c_A/λ and hence $2c_A/\lambda$ implies extensive auditing is optimal in both years (Proposition 2 and 4), i.e., $\underline{a}_1^i \leq 2c_A/\lambda, \underline{a}_2^* = c_A/\lambda < b$.

iv. If b is sufficiently small, then $\bar{x} > 0$, so the year 1 audit rule is an interval rule. Since c_A/λ is sufficiently large, $\underline{a}_1^i = b - \bar{x}$ and $\underline{a}_2^*$, i.e., there is no auditing (Proposition 4 and 2). Moreover, for sufficiently small b, $\bar{x} = b$ at $\Delta = 1$, hence $\Delta^* = 1$ is optimal according to Proposition 1.

v. Small α implies $\bar{x} > 0$, so the year 1 audit rule is an interval rule. As before, a large c_A/λ and small b implies $\underline{a}_1^i = b - \bar{x}, \underline{a}_2^* = b$, which also implies $\bar{a}_1(\underline{a}_1) = b - \bar{x} < b$. If $b > \frac{(1-\alpha)u}{2}$, then $\bar{x} < b$ at $\Delta = 1$. Small α also implies $\Delta_{\max \bar{x}} < 1$, hence $\Delta^* \in (\Delta_{\max \bar{x}}, 1]$ is optimal according to Proposition 1, i. □

Proof of Proposition 5.

i. Consider a small c_A, so that c_A/λ is small and the interior solutions to the audit thresholds are $\underline{a}_y = c_A/\lambda << u$ (Propositions 2 and 3). Since $b = u$, $\bar{x} = 0$, hence the relevant audit rule in year 1 is a threshold rule. The principal EU with maximum budget is

$$2\int_0^{\underline{a}} \theta - \lambda \underline{a} \mathrm{d}G(\theta) + 2\int_{\underline{a}}^{u} \theta(1-\lambda)\mathrm{d}G(\theta) - 2c_A\left(1 - \frac{\underline{a}}{u}\right)$$
$$= \frac{d^2}{u}(1-\lambda) - \frac{\underline{a}^2}{u}\lambda - 2c_A\left(1 - \frac{\underline{a}}{u}\right).$$

A smaller budget $b < u$ with $c_A/\lambda < b$ implies fund roll-over due to the low audit threshold. Instead of writing out the complex nested integral, I use an upper bound for the principal EU with budget $b < u$:

$$2\int_0^{\underline{a}} \theta - \lambda \underline{a} \mathrm{d}G(\theta) + \int_{\underline{a}}^{b} \theta_1(1-\lambda)\mathrm{d}G(\theta_1) + \int_b^u b(1-\lambda)\mathrm{d}G(\theta_1) + \int_{\underline{a}}^{u} \theta_2 \mathrm{d}G(\theta_2) - 2c_A\left(1 - \frac{\underline{a}}{u}\right)$$
$$= \frac{u^2 - b^2 + 2ub}{2u}(1-\lambda) - \frac{\underline{a}^2}{u}\lambda - 2c_A\left(1 - \frac{\underline{a}}{u}\right),$$

where in the first year not all spending needs can be fulfilled due to $b = b_1 < u$, but in the second year the budget is assumed to be $b_2 \geq u$ due to roll-over. The actual year 2 budget is lower for a positive mass of θ_1-realizations, hence this is an upper bound. After some algebra, the difference between the principal EU with budget u and budget $b < u$ is bounded below by

$$\frac{(u-b)^2}{2u}(1-\lambda),$$

which is positive and decreasing in λ. A smaller $b < u$ might have a lower audit threshold $\underline{a}_b < \underline{a} = \underline{a}_u$. In this case, the difference in principal EU is bounded below by

$$\frac{(u-b)^2}{2u}(1-\lambda) - \frac{\underline{a}_u^2 - \underline{a}_b^2}{u}\lambda + 2c_A\left(\frac{\underline{a}_u - \underline{a}_b}{u}\right),$$

where the latter two terms vanish as c_A gets small, since $\underline{a}_u = c_A/\lambda \geq \underline{a}_b$, so this difference is positive as well. Finally, the smaller budget might use an interval rule that saves audit costs yet prevents fund misuse to the same extent. In this case, too, the additional terms depend on c_A and vanish as it gets small. Hence, for sufficiently small c_A, the maximum budget with extensive auditing is optimal.

ii. For this result, I relax the exogenous constraint that $\underline{a}_y \leq b$. This relaxation allows setting $\underline{a}_2 = b_2$, which implies no auditing even if $b_2 > b$.

For sufficiently large c_A/λ, no auditing is optimal (Propositions 2 and 3). Compare budget $b < u$ and budget u, both without auditing. Since there is no auditing, the agent always misuses the remaining budget for $b = u$, so the principal EU with the maximum budget is

$$2\int_0^u \theta \mathrm{d}G(\theta) - 2\lambda u = u(1-2\lambda). \tag{21}$$

The principal EU with a lower budget $b < u$, but still with $\bar{x} = 0$, is

$$2\int_0^b \theta \mathrm{d}G(\theta) + 2\int_b^u b\mathrm{d}G(\theta) - 2\lambda b = 2b(1-\lambda) - b^2/u. \tag{22}$$

The latter exceeds the former if and only if

$$\frac{u + b^2/u - 2b}{2(u-b)} \leq \lambda,$$

which holds for a sufficiently large λ. And since

$$\frac{u + b^2/u - 2b}{2(u-b)} < \frac{u + u^2/u - 2b}{2(u-b)} = 1,$$

there exists a sufficiently large $\lambda < 1$ for any $b < u$ such that budget b is preferred to u by the principal. An even smaller $b < u$ with $\bar{x} > 0$ might be even more beneficial, but this case is not needed to prove the claim, as it does not specify whether $b < u$ with $\bar{x} = 0$ or with $\bar{x} > 0$ is optimal.

Finally, to show that a positive budget is optimal, note that $b = 0$ yields a zero utility and the derivative of (22) with respect to b is $2(1-\lambda) - 2b/u$, which is positive for small b. Indeed, this still understates the benefit of a small but positive budget if $\bar{x} > 0$, so that not all unused budget from year 1 is misused as in (22). Hence, budget b with $0 < b < u$ is optimal if λ and c_A/λ is sufficiently large.

iii. A large enough c_A and small enough λ implies no auditing is optimal. The principal EU for maximum budget u without auditing is given in (21). The principal EU with a lower budget $b < u$, so that $\bar{x} > 0$, is complex due to the nested integrals. Instead, I use the following upper bound

$$\int_0^b \theta_1 + \lambda(1-\Delta)\bar{x} \mathrm{d}G(\theta_1) + \int_b^u b \mathrm{d}G(\theta_1) + \int_0^u \theta_2 \mathrm{d}G(\theta_2) - 2\lambda b$$
$$= b(1-2\lambda) - \frac{b^2}{2u} + \frac{u^2}{2u} + \frac{b}{u}\lambda(1-\Delta)\bar{x}.$$

This upper bound exceeds the exact principal EU in two ways. First, if $\Delta < 1$, the amount returned in year 1 for any $\theta_1 \in [0, b]$ is $\lambda(1-\Delta)\bar{x}$, whereas the real expression returns this amount only for $\theta_1 \in [0, b - \bar{x}]$, and a lower amount for $\theta_1 \in (b - \bar{x}, b)$, so that the actual principal EU is lower. Second, the year 2 expected utility is larger because it assumes $b_2 = u$, whereas the actual budget due to fund roll-over is lower: $\bar{x}$ is set by the agent such that $G(b_2) < 1$, i.e., $b_2 < u$, see (1).

The difference between budget u and this upper bound for budget $b < u$ with $\bar{x} > 0$ is

$$(u-b)(1-2\lambda) - \frac{u^2 - b^2}{2u} - \frac{b}{u}\lambda(1-\Delta)\bar{x} = \frac{(u-b)^2 - 4\lambda u(u-b) - 2b\lambda(1-\Delta)\bar{x}}{2u},$$

which is positive for small enough λ. Next, the principal EU with a lower budget $b < u$ and $\bar{x} = 0$ is given in (22). The principal EU with maximum budget in (21) exceeds the EU with such lower budget if and only if

$$\frac{u + b^2/u - 2b}{2(u - b)} > \lambda,$$

that is, for a sufficiently small λ. Consequently, a maximum budget without auditing yields a higher principal EU than a lower budget with or without fund roll-over and no auditing, if λ is sufficiently small. □