Do Public Employees “Game” Performance Budgeting Systems? Evidence From the Program Assessment Rating Tool in Korea

Hosung Sohn¹ and Kwang Bin Bae²

Abstract
We examine whether performance budgeting systems such as the Program Assessment Rating Tool (PART) induce public employees to engage in “gaming” behavior. We propose an algorithm for detecting gaming behavior that makes use of the discrete nature of the PART system in Korea (KPART) and the revealed patterns of the distribution of the KPART scores. By employing the test developed by McCrary, we find suspicious patterns in the density of the KPART scores and evidence points to the fact that manipulation is prevalent in the KPART system. Our analysis suggests that public employees are sensitive to negative incentives and that great care must be taken when designing performance budgeting systems.

Keywords
performance-based budgeting system, strategic behavior

Introduction
Since the influential work of Osborne and Gaebler (1992), many government reforms have been adopted to hold government accountable for its performance. The Hoover Commission, Planning Programming Budgeting System, Management by Objectives, and Zero-Based Budgeting are the examples of government reforms that aim to promote government performance. One of the major reforms is the Government Performance and Results Act (GPRA) of 1993. The purpose of the Act is to promote effectiveness and efficiency of federal programs and spending by measuring program performance, providing publicly the reports on their progress, and holding agencies accountable for achieving program goals. The GPRA, however, failed to produce information relevant to improving program performance (Dull, 2006; Gilmour & Lewis, 2006b). The Program Assessment Rating Tool (PART), instituted by President George W. Bush and run through the Office of Management and Budget, is one of the policies introduced for the purpose of overcoming the limitation of the GPRA. As with the GPRA, one of the main goals of PART is to rigorously and systematically evaluate the effectiveness of public programs and establish a systematic link between information provided by program assessments and budget processes.

The adoption of PART is based on the argument that incorporating performance-based system in the budgeting process and linking the performance to budgeting would promote the performance of public programs (Broom & McGuire, 1995; Grizzle & Pettijohn, 2002; Joyce, 1993). For this reason, previous studies focus on whether the PART information, indeed, is used for linking performance information to budgeting activities (e.g., Gilmour & Lewis, 2006a; Lavertu & Moynihan, 2012). Note that the usefulness of the PART information hinges critically upon whether the information provided by agencies truly reflects program performance. PART may not provide adequate information that is crucial for evaluating performance of programs, however, because of the special features inherent in the public sector. Such features include, but are not limited to, multiple principals and multiple goals (Dixit, 2002), ambiguity in agency goals, and difficulty in quantifying the measures of performance (Frederickson & Frederickson, 2006; Radin, 2000). Under these features, extracting veritable performance measures from the public sector is extremely difficult, and as such, the PART information may not appropriately reflect the performance level of public programs.

Also, performance management systems such as PART may give rise to one particular form of unintended consequences: gaming. Program administrators, who are likely to be budget maximizing agents, may behave strategically under the PART system, in particular, because performance is explicitly linked to the budgeting process (Niskanen, 1975). There are many reasons that a system like PART may induce program managers to engage in gaming the system. First, the primary measure of success in the public sector may be determined by whether an agent has secured its budget or not. As a consequence, the performance management

¹Korea Institute for Health and Social Affairs, Sejong, South Korea
²North Carolina Central University, Durham, NC, USA

Corresponding Author:
Kwang Bin Bae, Department of Public Administration, North Carolina Central University, Durham, NC 27707, USA.
Email: kbae@nccu.edu
system under PART can be considered as a high-stakes game because PART enables decision makers to attach budgetary consequences. Second, PART is distinct from the GPRA in the sense that PART renders a judgment on whether programs are effective (Breul, 2007). And because program managers are sensitive to the judgment they receive, they have an incentive to game the system in an effort to receive positive judgment. Third, it is the program managers who bear the burden of proof on whether the programs are effective (Heinrich, 2012). This, in turn, implies that the program managers are aware of the evaluation factors that contribute to a positive rating. The system, therefore, provides more room for program managers to engage in undesirable behaviors. This is especially likely as principals cannot effectively determine whether the information provided by the agents are true or not because of information asymmetry (i.e., the principal–agent problem). Finally, in the context of the PART system, the raw score required for a program to receive a positive rating is oftentimes preannounced. If this is the case, the system is likely to create incentives for program managers to engage in manipulating their performance because the managers are aware of the eligibility cutoff.

If, in fact, gaming is prevalent in the PART system, it is indisputable that the information provided in PART will neither be sufficient for correctly evaluating the performance of government programs, nor appropriate for use in rewarding or punishing program administrators. In this article, therefore, we examine whether gaming behavior is prevalent in the PART system. To fulfill our goal, we use the institutional setting in South Korea. The general structure of the PART system in Korea (KPART) is simple—if the score a program receives is below a certain threshold, the program gets a penalty in the form of budget cuts. Studying the KPART system is favorable for investigating the gaming behavior because the system is likely to induce strategic behaviors given that many of the features mentioned above are inherent in the KPART system. Note that this study uses the “historical” setting to examine the gaming behavior of public officials, as KPART (not to mention PART) is no longer implemented in Korea. The purpose of our study, however, is to identify the prevalence of such behavior—a difficult task—using novel statistical methods, rather than examining the limitation of the KPART system. We therefore believe that studying KPART—although no longer implemented—will provide many lessons for policymakers in the United States as well as around the world in improving performance budgeting systems.

Detecting the prevalence of manipulation, however, is easier said than done for many reasons. It is obvious that if gaming is committed during the evaluation process, it is likely that actions among program administrators are conducted covertly. Such behavior might be disclosed if a whistle blower divulges the cheating. Even if this is the case, however, uncovering the “hard evidence” is still challenging.

In this article, we propose an algorithm for detecting gaming behavior that makes use of the discrete nature of the institutional setting in Korea and the revealed patterns of the distribution of the PART scores combined with the nonparametric statistical test developed by McCrary (2008). Although the method was initially developed for testing the manipulation of the assignment variable density function in the context of regression discontinuity design, the method is very useful in the context of this article because a density function of the revealed pattern itself is the object of interest. Using administrative records of the PART scores maintained by the Ministry of Strategy and Finance and employing the test developed by McCrary, we find suspicious patterns in the density of the PART scores. To evaluate the significance of our results, we also conduct a series of sensitivity tests as well as falsification tests, and evidence still points to the fact that gaming is quite likely. Although the strategy used in this study and the corresponding results do not “prove” that there are illicit acts, the analysis strongly suggests that performance budgeting systems such as PART may induce unintended behavioral distortions.

The rest of the article is organized as follows. We first provide a review of literature relevant for the purpose of this study. Next, we present institutional details of the KPART system. Third, we introduce the data and statistical method used in this study, followed by a discussion of statistical results. Finally, we highlight implications derived from our study and conclude.

**Literature Review**

Promoting the performance of public programs has been a major concern among public administrative reformers, and performance-based budgeting (PBB) system has been suggested as a way to achieve such goal (Joyce, 1997). Expecting that PBB system would enhance the performance of public programs, one of the major performance management reform initiatives introduced during the Bush administration is PART. At the initial stage of PART, many advocated the PART system and anticipated that it would bring performance information into the budgeting process, and consequently, leading to a “results-oriented” government (e.g., Breul, 2007; Dull, 2006).

Because of its intended purpose (i.e., linking performance information to budgeting activities and improving program performance), some of the previous studies on PART focus on whether the PART scores are used properly for affecting budget choices (e.g., Gilmour & Lewis, 2006a, 2006b; Greitens & Joaquin, 2010; Lavertu & Moynihan, 2012). While the conclusions as to the question on whether PART promoted the program performance are not unanimous among these studies, every study concludes that the PART information is not extensively used in the budgeting process (Moynihan, 2013).

As a matter of course, research on the effectiveness of PART is predicated on the belief that the PART score rightly reflects program performance and that evidence presented by program managers is of high quality. Assessing the quality of PART information, therefore, should precede the evaluation of PART usage (Joyce, 1997). Deriving quantifiable and
With an intent to improve gaming as manipulating performance measurement systems, we complement Bevan and Hood’s definition and define ties. In the context of performance management systems, for manipulating evaluation scores (there exist opportunities and there are rooms system attaches budgetary consequences that public managers to engage in undesirable behaviors, namely, gaming the system. This concern is shared by many researchers. For example, Radnor (2008) conceptually shows that the performance measurement gives rise to organizational gaming. Behn and Kant (1999) present 10 pitfalls of performance-based system in the context of contracting out in the public sector. One of the pitfalls is that some measures used in performance contracting may distort behavior of agents.

Bevan and Hood (2006) offer an excellent definition of gaming in the context of governance by targets. They define gaming as “reactive subversion such as hitting the target and missing the point” (Bevan & Hood, 2006, p. 521). They further note that gaming can arise under two types of agents. First, there are “reactive gamers” who agree with the goals set by the central administrators, but engage in gaming the target system if they have opportunities and reasons to do so. The second type is “rational maniacs.” These agents do not share the goals set by the central administrators and try to game data to conceal their activities. Those gaming the KPART system are likely to consist of the former type: the system attaches budgetary consequences that public managers are sensitive to (there exist reasons) and there are rooms for manipulating evaluation scores (there exist opportunities). In the context of performance management systems, we complement Bevan and Hood’s definition and define gaming as manipulating performance measurement systems with an intent to improve reported performance without promoting actual performance.

There are some research that investigate the prevalence of gaming in the public sector. Using the district-level data in Texas public schools, Bohte and Meier (2000) examine whether school districts manipulate the passing rates of students in standardized exams. They find that the districts do cheat the passing rate by preventing some students from taking the exams. By examining the annual system of publishing performance measures for public health care organizations in England, Bevan and Hood (2006) conclude that, although there were improvements in the reported performance, these improvements are due to the gaming behavior. Based on the empirical analysis of the performance bonus system adopted in the Workforce Investment Act (WIA) program, Heinrich (2007) also finds public employees being manipulating applicant pools by limiting some people from accessing the WIA program.

The three studies above highlight that high-powered incentive systems are likely to induce public employees to game the system. It is, however, somewhat hard to identify, from the three studies, specific factors that may give rise to gaming behavior. Identifying such factors is desirable, in particular, given the ongoing efforts to promote incentive systems in the public sector through the use of performance measures.

Although the public administration and policy literature has paid little attention to investigating specific factors that may lead to gaming, some research in education examined such factors (e.g., Jacob, 2005; Jacob & Levitt, 2003). Most of these research show that agents exploit distinctive features inherent in the system that are oftentimes flawed. This is especially likely given the fact that designing the “manipulation-proof” system is an extremely difficult task. Courty and Marschke (2004), for example, analyze the federal job training program, initiated by the Job Training Partnership Act (JTPA), and find that training agencies engage in gaming by manipulating the timing of the reporting of trainees’ outcomes. Manipulation of timing was possible because timing requirements were not specified in the JTPA.

Given the extant literature, our study contributes to the public administration and policy literature in three ways. First, we empirically examine, for the first time, the prevalence of gaming behavior in PBB systems. We argue that the results from our study highlight to what extent public employees are sensitive to the budgeting process. Second, we investigate factors that induce program managers to game the system by making use of the institutional setting in Korea. Namely, we show that informational advantages are the major driver for undesirable consequences. We believe lessons derived from our study would be useful for designing the manipulation-proof PBB system. Third, we propose a way to test, statistically, for the gaming behavior in the public sector by employing the cutting-edge nonparametric statistical method that we believe is notably favorable for the purpose at hand.

Institutional Background

Korea set up the KPART system in 2005, and the system was introduced for two main purposes: promoting the performance of public programs, and increasing the linkage between performance information and the budgeting process. Note that KPART has been designed by benchmarking the U.S. PART system. As such, KPART is extremely similar to PART with respect to its operation and shares features that are comparable to that of the United States.

In Table 1, we provide a comparison of the PART system between the United States and Korea. First, Office of Management and Budget (OMB) is in charge of PART under GPRA, whereas in Korea, Ministry of Strategy and Finance (MOSF) manages KPART under the National Fiscal Law. Second, program evaluation is conducted by agencies in
charge of the programs, and the supervisor reviews the evaluation (for both countries). Third, PART evaluation is based on 25 questionnaires on four indices (purpose and design, strategic planning, management, and results/accountability); KPART evaluation is based on 13 questionnaires on three indices (planning, execution, and results). Fourth, both countries rate in five categories. Last, there is no strong linkage between the PART scores and the budgeting process. KPART, on the other hand, imposes budgetary consequences for programs with scores below 60.

Based on Table 1, it is evident that the PART and KPART systems are operationally similar. The distinction, however, comes from the degree of the linkage between the PART scores and the budgeting process. Under the U.S. PART system, receiving low PART scores does not necessarily imply budget cuts. Contrarily, KPART scores enter the equation that determines the budget for the following fiscal year; budgets for poor-performing programs are reduced by 10%—at the extreme, programs may be abolished. Thus, a strong linkage between the KPART scores and budgetary consequences are likely to induce public employees to go great lengths to avoid being designated as poor-performing programs.

Figure 1 shows whether there is indeed a strong linkage between the KPART scores and the budgeting process by Table 1.
plotting the average share of programs that have faced budget cuts by the baseline KPART scores. For example, the share is about 0.65 for the PART score from 30 to 31. This implies that about 65% of the programs—that received a score between 30 and 31—faced budget cuts. As expected, there is a strong negative relationship between the share and KPART scores. That is, the share of programs whose budgets were reduced in the following fiscal year is considerably higher for those whose scores are low. Note that some high-scoring programs (programs with scores above 60) faced budget cuts and some low scoring programs (programs with scores below 60) did not. This is because the KPART score is not the only factor that determines the budget for the coming fiscal year. Nevertheless, the negative relationship displayed in Figure 1 clearly indicates that there is a close connection between KPART and the budgeting process.

To further gauge the effect of the KPART score on the probability of facing budget cuts, we run a regression of a binary variable, denoting whether budgets were reduced, on dummy variables indicating the rating of programs. Table 2 presents the results. We find the estimated effect of 0.155 which implies that programs rated as “poor or below” are 15.5% more likely to face budget cuts, on average, than programs that received the “average” grade, with the statistical significance well below the 1% level. We do not find any statistically and practically significant effects for the programs rated as “excellent or above.” We also run the regression using the amount of budget cuts as an outcome variable on the same explanatory variables, and find that the amount of budget is reduced by about 15 billion won (or 15 million dollars), on average, for the programs rated as “poor or below.” Compared with the “average-rated” programs, the estimated effect for being rated as “excellent or above” is statistically insignificant. Last, we reject the joint-test of the hypothesis that the program performance indicators jointly had no effect, implying that the program performance indicators do affect the outcome variables.

All in all, Figure 1 and the results in Table 2 clearly show that there is a strong linkage between the KPART score and the budgeting process. The results in Figure 1 and Table 2 are quite different from the findings by Gilmour and Lewis (2006a). They find a statistically significant impact on budget decisions, but the associated budget decision is not driven by the “results” component of PART scores. Rather, the decision is influenced by the “program purpose” component. The reason for the difference in the magnitude of the effect of PART scores between the United States and Korea, we believe, is that the PART system in Korea is created to explicitly affect budget decisions.

Because of the strong impact of the KPART scores on budgetary process, therefore, it is plausible that the KPART system is more likely to induce public employees to engage in undesirable behaviors. As such, the institutional setting in Korea is favorable for examining whether program managers are sensitive to the negative incentive inherent in the KPART system. Analyzing the KPART system, in particular, would allow us to investigate whether public employees are sensitive to budgetary consequences because of the negative incentive, that is, budget cuts, and may provide policy implications for effectively designing PBB systems.

### Data and Empirical Strategy

#### Data

We use administrative records of the KPART score maintained by the MOSF. The database maintained by the MOSF contains scores from 2005 to 2011. While the numbers vary by year, it includes results for approximately 3,000 programs that were evaluated during the time frame. Some of the key variables in the database are as follows: program names, total scores given to each program, ratings granted (in five categories), subscores (in three categories) that contribute to total scores, initial budget amounts, and finalized budget

---

**Table 2. The Effect of the PART Score on Budget Cuts.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Programs faced budget cuts (1 = yes)</th>
<th>Amount of budget cuts (in billion won; LCU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programs received the poor grade</td>
<td>0.155***</td>
<td>−14.963***</td>
</tr>
<tr>
<td>(1 = yes)</td>
<td>(0.018)</td>
<td>(3.909)</td>
</tr>
<tr>
<td>Programs received the excellent</td>
<td>−0.042</td>
<td>−6.610</td>
</tr>
<tr>
<td>grade (1 = yes)</td>
<td>(0.027)</td>
<td>(4.469)</td>
</tr>
<tr>
<td>Year fixed effects</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>F-test</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Sample size</td>
<td>3,257</td>
<td>3,154</td>
</tr>
</tbody>
</table>

**Note.** “LCU” denotes local currency unit. Standard errors in parentheses. For all regressions, a dummy variable that indicates whether programs received an average grade is the reference category. The numbers presented in the F-test row are p values retrieved from the joint-test of the hypothesis that the program performance indicators jointly had no effect. $1 = 1,000$ won. 

***indicates statistical significance at the 1% level.
amounts that reflect the ratings. We provide descriptive statistics of these variables in Appendix A.

In Table 3, we show the average share of programs, with the corresponding standard deviations, by year and rating. The average share of programs over the 7-year period that received “very poor,” “poor,” “average,” “excellent,” and “very excellent,” is 0.08, 0.28, 0.52, 0.08, and 0.02, respectively. As the shares indicate, half of the programs were rated “average.” There were 3,260 programs that were rated between 2005 and 2011.

Note that, in principle, every program is rated every 3 years. So if a program started, for example, in 2005, then that program is evaluated, again, in 2008. So there is a chance that some programs are evaluated multiple times. But because programs in Korea, in general, do not last for more than 2 years, there are very few programs that are evaluated more than once. In fact, we find that less than 1% of the programs in each year have been evaluated multiple times.

One interesting point to note from the table is that the difference in the shares between the “poor” and “average” ratings widen over the years. This is an atypical trend given the difference among other ratings are relatively stable across years. And as can be seen from the table, the difference in the shares becomes particularly salient starting from 2008. This insinuates that program managers who received a “poor” or “average” rating may be gaming the KPART system, which calls for a more thorough examination of the gap.

**Hypothesis 1**: Public managers, even if budget maximizing, do not game the performance budgeting system.

Testing the hypothesis above is an ambitious task. Manipulation of the PART score, if done in a sophisticated manner, will not likely leave tell-tale clues that provide convincing evidence of cheating. For this reason, many governments operate some form of whistle-blowing policy to capture illicit activities. Oftentimes, however, these policies are not strong enough to catch such activities for many reasons (e.g., the problem of the prisoner’s dilemma). Acknowledging the difficulty involved in detecting the gaming behavior, some studies, especially in the education-related literature, offer insightful ways of isolating such behaviors (e.g., Angoff, 1974; Wollack, 1997). Jacob and Levitt (2003), for example, examine answer sheets given by

<table>
<thead>
<tr>
<th>Year</th>
<th>Very poor</th>
<th>Poor</th>
<th>Average</th>
<th>Excellent</th>
<th>Outstanding</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>0.153 (0.361)</td>
<td>0.379 (0.486)</td>
<td>0.372 (0.484)</td>
<td>0.076 (0.265)</td>
<td>0.020 (0.140)</td>
<td>554</td>
</tr>
<tr>
<td>2006</td>
<td>0.111 (0.315)</td>
<td>0.443 (0.497)</td>
<td>0.340 (0.474)</td>
<td>0.083 (0.277)</td>
<td>0.023 (0.149)</td>
<td>576</td>
</tr>
<tr>
<td>2007</td>
<td>0.054 (0.225)</td>
<td>0.283 (0.451)</td>
<td>0.487 (0.500)</td>
<td>0.116 (0.320)</td>
<td>0.060 (0.239)</td>
<td>579</td>
</tr>
<tr>
<td>2008</td>
<td>0.034 (0.181)</td>
<td>0.230 (0.421)</td>
<td>0.595 (0.491)</td>
<td>0.115 (0.319)</td>
<td>0.026 (0.160)</td>
<td>383</td>
</tr>
<tr>
<td>2009</td>
<td>0.041 (0.198)</td>
<td>0.164 (0.371)</td>
<td>0.737 (0.441)</td>
<td>0.041 (0.198)</td>
<td>0.018 (0.131)</td>
<td>342</td>
</tr>
<tr>
<td>2010</td>
<td>0.063 (0.244)</td>
<td>0.179 (0.384)</td>
<td>0.709 (0.455)</td>
<td>0.046 (0.211)</td>
<td>0.002 (0.046)</td>
<td>474</td>
</tr>
<tr>
<td>2011</td>
<td>0.082 (0.275)</td>
<td>0.219 (0.414)</td>
<td>0.625 (0.485)</td>
<td>0.071 (0.257)</td>
<td>0.003 (0.053)</td>
<td>352</td>
</tr>
<tr>
<td>M</td>
<td>0.092 (0.274)</td>
<td>0.287 (0.452)</td>
<td>0.528 (0.499)</td>
<td>0.080 (0.272)</td>
<td>0.024 (0.152)</td>
<td>3,260</td>
</tr>
</tbody>
</table>

**Note.** Standard deviations in parentheses.
students in the Chicago public schools to empirically identify the prevalence of teacher cheating. Their approach exploits unexpected fluctuations in student test scores as well as unusual patterns of answers given by students.

Our approach to identifying the prevalence of gaming behavior in the KPART system is similar to the one described above in the sense that we use the patterns of the KPART scores. Note, however, that we do not have access to the actual evaluation sheets and information submitted by program managers. Accordingly, we rely on the revealed density (i.e., data points) of the distribution of the KPART scores and examine whether there exist unusual patterns in the density.

To fulfill our goal, we exploit simple discontinuous rules (i.e., the budget cut is determined by a single cutoff point) adopted in the KPART system and apply the specification test developed by McCrary (2008). Suppose the probability of facing budget cuts \( D_i \) changes from zero to one as the assignment-determining variable (i.e., KPART scores) crosses some threshold \( c \) (i.e., in our study, \( c = 60 \)). The intuition behind the test is that if program managers know when they will face budget cuts and can manipulate their value of the KPART scores, then they may be able to place themselves just above (or below) the threshold \( c \). In that case, the programs just above the threshold will disproportionately consist of those gaming the rule. In the KPART system, programs rated as “poor” will experience budget cuts only when their scores fall below 60. Accordingly, managers of programs with scores just below 60 face an incentive to manipulate their scores to fall right above 60 so that they do not face budget cuts.

To determine the statistical significance of the unusual patterns at the cutoff, McCrary (2008) suggests a specification test examining the density of the assignment variable as it crosses the cutoff point. If program managers are manipulating their values of the assignment variable (i.e., the KPART scores) to fall just above the eligibility cutoff, then it is very likely that we observe a “discontinuity” in the distribution of the KPART scores as it crosses the cutoff point. If the distribution of the KPART scores, on the other hand, is smooth as it crosses the cutoff point, it is less likely that program managers are gaming the assignment variable. The main idea behind this test is that if program managers have “imprecise” control over the KPART scores, it is unlikely that we will see irregular patterns in the assignment variable.

While the density test proposed by McCrary (2008) is developed for the purpose of complementing the existing specification checks in the regression discontinuity design, the test is valuable in an application where the discontinuity in the data points is itself the purpose of interest. Saez (2010), for example, examines tax avoidance by making use of the discontinuity in the reported income. Furthermore, the test is useful in situations in which predetermined covariates that are available are not suitable for use in the substantive topic under study. In the context of this study, estimating the discontinuity is itself the object of interest, and predetermined variables that we have are not relevant for examining the issue at hand. The test, therefore, suits our purpose.2

As it is the case in the regression discontinuity-type analysis, the validity of the test developed by McCrary (2008) hinges critically upon a graphical analysis; the graph that fails to exhibit a visually perceptible break in the data points at the cutoff point is basically not credible, regardless of the formal regression results that test for the discontinuity at the cutoff. Hence, we first generate a graph that plots the frequency of programs by the KPART scores to see whether there is a visual break at the eligibility cutoff.

Next, as suggested by McCrary (2008), we conduct a local polynomial regression, separately for the left and right side of the cutoff point, using the histogram bin midpoints (i.e., \( x \)-axis) as an explanatory variable and the relative frequency (i.e., \( y \)-axis) as an outcome variable. Formally, let \( x \) be the KPART scores. At every point \( x < c \), we run the following kernel regression:

\[
\hat{F}_{L}(x) = \alpha_0(x) + \beta_1(x)(X_i - x) + \ldots + \beta_p(x)(X_i - x)^p K\left(\frac{X_i - x}{h}\right),
\]

where \( \hat{F}_{L}(x) = \frac{N}{Nh} \) is the frequency observations within \( b \) being the binwidth of the \( x \)-axis, \( h \) and \( K(\bullet) \) is the bandwidth and kernel function, respectively. \( X_i \) denotes the histogram bin midpoint with \( k \) being the number of bins to the left of the cutoff point. \( p \) is the order of the polynomial. The density estimate at point \( x \) is then \( \hat{f}(x) = \alpha_0(x) \). Now, for every point \( x \geq c \), we repeat the same analysis using the regression of the form,

\[
\hat{F}_{R}(x) = \alpha_0(x) + \beta_1(x)(X_i - x) + \ldots + \beta_p(x)(X_i - x)^p K\left(\frac{X_i - x}{h}\right),
\]

and recover the density estimate \( \hat{f}(x) = \alpha_0(x) \). The test statistic used for testing the discontinuity at the cutoff point will be \( \hat{\theta} = \ln \hat{f}(x) - \ln \hat{f}(x)^\dagger \), and we normalize the test statistic by the standard error, \( \sigma_\theta = \sqrt{4.8 \times ((1/\hat{f}(x)^\dagger) + (1/\hat{f}(x)^\dagger)) / (Nh)} \).

We then conduct statistical inference on \( \hat{\theta} / \sigma_\theta \) based on a standard \( t \)-distribution.

As can be seen from Equation (1) and (2), we need to make choices on four parameters when estimating the discontinuity at the cutoff: binwidth \( b \), bandwidth \( h \), kernel function \( K \), and the degree of polynomial \( p \). Simulations conducted by McCrary (2008) indicate that the choice of \( b \) is not too important, but the choice of \( h \) can be extremely important. While McCrary offers a plug-in bandwidth
estimator, the literature on regression discontinuity-type designs suggests that researchers test for the sensitivity of the discontinuity estimates using different bandwidth choices. McCrary suggests using the triangle kernel and local linear regression (i.e., $p=1$). Again, we estimate the discontinuities experimenting with different kernels and degree of polynomials.

Finally, we conduct falsification tests by examining whether there are other significant discontinuities at other cutoffs. If the discontinuity estimated for the true cutoff point (e.g., 60) is indeed driven by manipulation, we should not see discernible discontinuities at other cutoff points. Or at least, the estimated discontinuity at the true cutoff should be significantly larger than those estimated for other cutoff points.

**Results**

The strength of the regression discontinuity design-type analysis revolves around the fact that we see discontinuous changes in the density if there is any effect as the assignment variable crosses some threshold $c$. Consequently, a graphical analysis should be a focal point in investigating the discontinuities. Hence, we conduct a graphical analysis before providing an in-depth statistical analysis of the discontinuities.

**Graphical Analysis**

As previously mentioned, programs face consequences (i.e., budget cuts) if they are rated as “poor or below” based on the KPART scores. To avoid these ratings, programs are required to receive a score of 60 or higher. As a consequence, there exist significant incentives for program managers to exhibit gaming behavior around this cutoff. To examine such behavior, we present, in Figure 2, the density of the KPART scores around 60. We also show the predicted values from a local polynomial regression estimated separately for the left and right of the cutoff. We use different bandwidth choices for the left and right of the cutoff because densities become sparse for scores below 50; for the right of the threshold, we limit the range to 80 because 80 is also used as the eligibility cutoff to reward outperforming programs.

Figure 2 clearly shows evidence of the discontinuity at 60. It is visually unambiguous that there is a large and perceptible break at the cutoff. Moreover, the overall shape of the density near the threshold seemingly points toward the prevalence of manipulation. Observe that the local averages are fairly smooth between 50 and 57. The average tends to decline at 59, followed by a huge jump in the density of the assignment variable at 60. Compared to other scores, the bin-to-bin jumps at 60 is considerably large enough to warrant the jump as unusual. Note, further, that the local average increases onward and drops significantly after 63. We argue that the reason we observe larger jumps between 61 and 63 compared with 60 is that program managers who received 59 or 60 may have inflated their scores by two to three points, fearing that being placed exactly at 60 would induce suspicious eyes from the authority. All in all, the pattern bespeaks that programs that received scores 60 to 63...
disproportionately consisted of those gaming the rule because there are no particular reasons to believe that we should observe such irregular patterns around these scores.5

One important point to note is that the fact that we observe a huge discontinuity at the eligibility cutoff does not imply that all of the program managers that received scores just above the cutoff point are those who engaged in gaming. There may be managers who in fact succeeded in raising “actual” performance. The statistical test we employed does not allow for distinguishing those who raised “actual” performance from those who raised “reported” performance without promoting actual performance. But if manipulating the score is difficult, observing such a huge discontinuity at the cutoff point is less likely. As such, we argue that there are program managers that engaged in gaming the system especially among those who scored between 60 to 63.6

Note, furthermore, that the fact that we observe discontinuity in the density at the 60-point cutoff does not imply that all the programs that received scores just above 60 consisted of manipulators who would have received scores just below 60. It is possible that programs that would have received scores well below 60 (e.g., 50) may have engaged in manipulative behavior. The density test we employed cannot prove whether the program managers who received scores well below 60 gamed the system, as the identification of discontinuity is based on the density points around the cutoff. We argue, however, that such incidence is less likely. We think that public managers who received a score that is far away from the 60-point cutoff (e.g., those who received 50) face few incentives to manipulate their scores, because manipulating the score of 10 points—to be placed above 60 to avoid budget cuts—is much harder than manipulating the score of 2 or 3 points (e.g., from 58 points to 60 points).

In Figure 3, we examine densities of the PART scores around the other cutoff, that is, 80 (“excellent or above”). While the programs rated as “excellent or above” are not necessarily rewarded, the provisions of the KPART system do state that these programs may receive awards.7 Therefore, the cutoff at 80 also calls for an investigation into whether program managers game the system.

The benefit of analyzing the discontinuity at the 80 cutoff is that it may allow us to explore—by comparing the magnitude of the discontinuity with that of the cutoff at 60—whether public employees are more sensitive to negative incentives than to positive incentives in the budgetary process. Knowing which of the incentives public employees are sensitive to would provide policy implications for effectively designing performance budgeting systems.

Contrary to the striking gap observed at the 60 threshold, we do not observe any significant hikes around the 80 cutoff. Rather, the estimated local polynomial curve is continuous at the eligibility cutoff. While the relative frequency does leap between 79 and 80, the size of the leap hardly constitutes a surprise given the gaps observed for other cutoffs let alone for the 60 cutoff in Figure 2. This may suggest that program managers are more sensitive to negative incentives. With respect to performance

**Figure 3.** Density of the assignment variable at the 80 cutoff.

*Note.* The density of the assignment variable is plotted with a binwidth of one point. Local polynomial fit is conducted using the bandwidth of 10 (for the left of the cutoff), 20 (for the right of the cutoff), triangle kernel function, and second-order polynomial.
budgeting systems, therefore, it may be desirable for the authority to use negative—rather than positive—incentives to motivate program managers. Note, however, that programs rated as “excellent or above” are not unconditionally rewarded. In contrast, programs receiving the poor rating or below do incur a penalty—in the form of budget cuts in subsequent years. If awards come with no strings attached for the highly rated programs, we might have observed a more discernible discontinuity at the threshold. Therefore, we caution against deriving strong conclusions from the comparison between Figure 2 and Figure 3.

As an attempt to supplement evidence of gaming behavior, we examine the association between subscores and total scores in Figures 4 and 5. Programs receive scores in three distinct categories: planning, performance, and execution. These three scores are added up to constitute total scores. If manipulation seems far-fetched, we should see a reasonable amount of positive correlation between each of the subscores and total scores. Subscores on “planning” and “performance” are plotted against total scores in the left and right panels of Figure 4. The revealed patterns in Figure 4 show a smooth relationship between each subscore and total scores with most of the local averages located around the 45-degree line—if data points are all located on the 45-degree line, then it implies a perfect collinearity. Another point to note from both panels is that we do not see any perceivable discontinuities around the thresholds.

In Figure 5, we plot scores on the execution category by the assignment variable. Contrary to the two figures in Figure 4, Figure 5 shows a tremendous amount of noisiness in the local averages of the execution score. There is no particular reason to believe, a priori, that we should observe such fluctuations in the execution scores given the extremely stable and smooth patterns displayed for the planning and performance scores in Figure 4. Moreover, while Figure 4 shows that there are no discontinuities in the planning and performance scores around the two cutoffs, Figure 5 displays an overwhelmingly large discontinuity at the eligibility cutoff. Hence, these three figures strongly suggest that program managers engage in manipulation of the KPART system, and the channel through which gaming is executed is via the execution score.

**Estimation Results**

It may have been the case that the anomalous piling up of the densities around the cutoff at 60 may have been generated due to sampling error (i.e., chance variation). To gauge the extent to which the sudden increase in the densities is driven by such error, we conduct a statistical test of the discontinuity using the method proposed by McCrary (2008). To be more specific, we estimate discontinuity estimates in the density function of the assignment variable using various specifications on kernel functions, degree of polynomials, and bandwidth choices.

In Table 4, we provide discontinuity estimates derived from running the local polynomial regression using the bandwidth equal to 10. We estimate several discontinuity estimates to assess the sensitivity of the estimates to the choices on kernel functions and degree of polynomials. Discontinuity estimates at the 60 cutoff are presented in Panel A. According to the estimates in Panel A, while the effect estimates vary, to some extent, across polynomial specifications, we see highly stable and statistically significant estimates across kernel function specifications and within the same degree of polynomial. When estimated using the local linear regression (i.e., the degree of polynomial equal to one), the average discontinuity estimate is 0.75 with a standard error around 0.09. Based on a quadratic polynomial specification, we again find a highly significant discontinuity estimate with an average estimate equal to 1.12 with a standard error approximately equal to 0.10.8

We present discontinuity estimates for the other cutoff (i.e., 80) in Panel B of Table 4. Contrary to the highly stable findings observed at the 60 threshold, the estimated
discontinuities at 80 are immoderately sensitive to regression specifications both within and across kernel functions and degree of polynomials. Furthermore, more than half of the effect estimates cannot be rejected, statistically, even at the 10% level. The findings in Panel B are consistent with the graphical results shown in Figure 3; because there is no visual break, it is unlikely that the statistical analysis will find anything—while some of the effect estimates are statistically significant, we cannot argue from these results that there is a discontinuity in the density of the assignment variable at 80 given the highly sensitive results. Besides, note that the magnitude of the discontinuity estimates in Panel B, in general, is much smaller than those in Panel A—most of the discontinuity estimates in Panel A are twice as large as the corresponding estimates in Panel B. Hence, we argue that the manipulation is much more prevalent at the eligibility cutoff equal to 60, implying that public employees are more sensitive to negative incentives—at least as far as the budgetary consequence is concerned.

We also run the same specifications and test the sensitivity of our results by halving the bandwidth.9 We find several facts. First, the estimated discontinuities for the cutoff at 60 under the bandwidth of 5 are appreciably similar compared with those obtained under the bandwidth of 10. This implies that the discontinuity estimate at 60 is not sensitive to the choice of a bandwidth, and it reasonably points toward evidence of manipulation. Second, effect estimates based on the bandwidth of five are shown to be sensitive to the specifications, which is also the case in Panel B of Table 4. Some of the effect estimates are close to zero with a weak statistical significance; discontinuity estimates under flexible specifications, however, display discontinuities that are statistically significant. While one could argue in favor of the prevalence of gaming behavior at the cutoff equal to 80, we caution against drawing strong conclusions on this cutoff because of the sensitiveness of the effect estimates to the regression specifications.

Sensitivity and Falsification Tests

As it is normally the case in the regression discontinuity-design type analysis, the most significant parameter that affects the effect estimates is the bandwidth. Choosing the bandwidth is more an art than a science, and given the discreteness involved in choosing the bandwidth, there has been an effort to develop a data-driven procedure for choosing the bandwidth (e.g., Imbens & Kalyanaraman, 2012). McCrary (2008) also offers a plug-in bandwidth formula. Most research, however, recommends choosing it via visual inspection and experimenting with different bandwidths (Imbens & Lemieux, 2008; Lee & Lemieux, 2010). Hence, we test for the sensitivity of the discontinuity estimates under varying bandwidth choices. Furthermore, we complement our robustness check by conducting falsification tests. To conduct falsification tests, we create pseudo eligibility cutoffs at other scores, and then conduct the density test proposed by McCrary (2008) to derive a discontinuity estimate for each pseudo eligibility cutoff. The test is based on the idea that if program managers are indeed manipulating the KPART score especially at the 60 cutoff, then we should

![Figure 5. Average scores on “execution,” by the assignment variable.](image-url)
observe the largest discontinuity at this cutoff; for other cutoffs, we should at least observe small discontinuities.

The results of the sensitivity test are presented in Appendix B. As can be seen from the results (i.e., Panel A of Figure B1), the discontinuity estimates at the 60 cutoff is insensitive to the choice of bandwidth and all of the estimated discontinuities are statistically significant at the 5% level. Contrarily, the discontinuity estimates at the 80 cutoff (i.e., Panel B of Figure B) is quite sensitive to the choice of bandwidth and the estimates are statistically insignificant starting from the bandwidth choice of six. Therefore, we warn against drawing a strong conclusion that program managers engage in gaming behavior at the 80 cutoff.

In Appendix Table B1, we present discontinuity estimates for pseudo eligibility cutoffs. We conduct a series of falsification tests using different bandwidth choices to fully leverage our falsification tests. In the tables, we also provide a rank calculated within the model—the lower the rank, the higher the magnitude of the discontinuity estimate within the model. The falsification tests reveal two points. First, the estimated discontinuity for the cutoff at 60 is consistently the largest—both statistically and practically—across models; the case is not true for the 80 cutoff. Second, the estimates derived for other cutoffs are extremely sensitive to the choice of a bandwidth. For some discontinuity estimates, the sign changes from negative to positive, and vice versa. Furthermore, statistical significance varies to a great extent. All of these clearly evince that the estimated discontinuity for the 60 cutoff is likely to be driven by manipulation.

**Discussion**

The graphical as well as statistical analyses presented above strongly suggest that there exist irregular patterns that are statistically and practically significant. While such discontinuity in the density may have been driven by program managers who were successful in promoting true performance, we argue that disproportionate number of program managers are likely to have engaged in manipulating the system for many of the reasons mentioned above (e.g., Figure 5).

We draw two implications from our study. First, it is evident that public employees are sensitive to incentives related to the budgeting process. While the KPART system is somewhat limited in analyzing whether program managers are sensitive to positive incentives, it is clear from our analysis that they are quite contriving with respect to negative incentives. Our analyses, therefore, suggest that when implementing performance management systems using the carrot and stick approach, care must be taken, in particular, on the stick approach.

The second lesson we draw from our study is that “design” matters. We argue that the most plausible factor

---

**Table 4. Discontinuity Estimates in the Density Function of the Assignment Variable (h = 10).**

<table>
<thead>
<tr>
<th>Kernel function</th>
<th>One</th>
<th>Two</th>
<th>Three</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Cutoff = 60</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>0.655*** (0.092)</td>
<td>1.123*** (0.100)</td>
<td>1.720*** (0.127)</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.687*** (0.092)</td>
<td>1.142*** (0.101)</td>
<td>1.701*** (0.127)</td>
</tr>
<tr>
<td>Rectangle</td>
<td>0.802*** (0.090)</td>
<td>1.058*** (0.100)</td>
<td>1.512*** (0.130)</td>
</tr>
<tr>
<td>Triangle</td>
<td>0.890*** (0.093)</td>
<td>1.190*** (0.109)</td>
<td>1.423*** (0.133)</td>
</tr>
<tr>
<td>Mean of discontinuity estimates</td>
<td>0.759</td>
<td>1.128</td>
<td>1.589</td>
</tr>
<tr>
<td><strong>Panel B: Cutoff = 80</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>0.687*** (0.187)</td>
<td>−0.237 (0.151)</td>
<td>0.412*** (0.151)</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.528*** (0.179)</td>
<td>−0.188 (0.151)</td>
<td>0.463*** (0.152)</td>
</tr>
<tr>
<td>Rectangle</td>
<td>0.165 (0.160)</td>
<td>−0.060 (0.158)</td>
<td>1.900*** (0.265)</td>
</tr>
<tr>
<td>Triangle</td>
<td>0.078 (0.159)</td>
<td>0.444*** (0.169)</td>
<td>1.615*** (0.221)</td>
</tr>
<tr>
<td>Mean of discontinuity estimates</td>
<td>0.365</td>
<td>−0.010</td>
<td>1.098</td>
</tr>
</tbody>
</table>

*Note.* Discontinuity estimates are derived from running the local polynomial regression using the corresponding kernel function and order of polynomial. All estimates are obtained from using the bandwidth (h) equal to 10. Standard errors are in parentheses.

***indicates statistical significance at the 1% level.
that drives gaming behavior stems from the fact that performance management systems often embody flaws. In the KPART system, there are three major holes that program managers can take advantage of. First, program managers are able to manipulate their scores to prevent a “poor” rating because the eligibility cutoff for this category is disclosed before programs are evaluated. An aftereffect is a heap of programs placed just above the eligibility cutoff. Second, program managers are well aware of how their answers to questions—those that are used for evaluating program performance—are converted to scores for subcategories which, in turn, constitute total scores. A corollary to this awareness is a seeming manipulation we observed in Figure 5. Third, even if program managers are aware of how their answers to questions enter the formula, one may argue that gaming behavior might not arise if answers to questions are inspected by the authority in charge. Because the burden of proof was imposed on the program managers, however, it is likely that managers will do their best to justify their performance. And as a matter of course, corroborating their justifications would be challenging for the authority because of information asymmetry. All in all, we argue that if program managers do not possess these informational advantages, there might have been less room for manipulation.

Our study, however, is not free of limitations. We acknowledge that detecting public employees’ gaming behavior is intrinsically challenging. Undeniably, we do not argue that our analysis proves the prevalence of gaming behavior in the KPART system. For instance, among those just right of the cutoff, there may be a program that received a one or two additional point from legitimate bargaining, not gaming. Our data and statistical analyses do not allow for distinguishing such legitimate bargaining from illegitimate bargaining. By carefully examining the incentives faced by public managers, as well as the density of the KPART scores, however, we argue that the evidence provided above is convincing enough to claim that manipulation exists in the KPART system. For future research, it may be informative to analyze whether removing the undesirable attributes embedded in the KPART system would prevent program managers from engaging in gaming the system.

Second, our analysis is limited in the sense that we cannot get inside the “black-box.” For example, our study does not tell whether manipulation is accomplished solely by program managers. The possibility that program managers and authorities forged a deal together cannot be ruled out. While uncovering such behavior is a difficult task, identifying such information would be useful for designing flawless performance management systems.

Last, we did not address the question of whether the manipulative behavior prevalent in the KPART system renders social welfare loss. Courty and Marschke (2004), for example, examine the question of whether strategic behaviors committed by public employees lead to welfare loss. By defining a misallocation of resources as a welfare loss, they show that earnings impacts are significantly lower for those who received training from agencies that gamed the system. Given the various tasks inherent in public programs and the difficulty and disagreement involved in quantifying the benefits (or costs) of such programs, however, conducting the cost-benefit analysis in the context of this study is extremely challenging. We, therefore, postpone such analysis for future research.

Conclusion

In this article, we study whether performance management systems such as PART induce program managers to engage in gaming. We make use of the discontinuous eligibility cutoffs adopted in the KPART system and carefully analyze the distribution of the KPART scores around these cutoffs to examine whether such systems bring about undesirable consequences. Applying the density test developed by McCrary (2008), we find an unusual and sizable discontinuity around the eligibility threshold used for penalizing underperforming programs. On the contrary, the magnitude and statistical significance of the estimated discontinuity for the other threshold—intended for use in rewarding outperforming programs—are indistinguishable. While a strong conclusion regarding public employees’ sensitivity toward positive incentives cannot be drawn from our study, the analyses strongly suggest that program managers are more likely to act on avoiding negative incentives.
Appendix A

Description of the Data

Table A1. Descriptive Statistics, by Year.

<table>
<thead>
<tr>
<th>Variable</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of departments</td>
<td>35</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>33</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>Number of programs</td>
<td>555</td>
<td>577</td>
<td>585</td>
<td>384</td>
<td>346</td>
<td>473</td>
<td>389</td>
</tr>
<tr>
<td>Average score</td>
<td>60</td>
<td>60</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>62</td>
<td>62</td>
</tr>
<tr>
<td>(13)</td>
<td>(13)</td>
<td>(13)</td>
<td>(11)</td>
<td>(10)</td>
<td>(8)</td>
<td>(10)</td>
<td></td>
</tr>
<tr>
<td>Average budget (at year t)</td>
<td>633</td>
<td>608</td>
<td>741</td>
<td>1,028</td>
<td>747</td>
<td>1,012</td>
<td>1,325</td>
</tr>
<tr>
<td>(1,767)</td>
<td>(3,365)</td>
<td>(4,006)</td>
<td>(4,010)</td>
<td>(4,744)</td>
<td>(4,998)</td>
<td>(6,588)</td>
<td></td>
</tr>
<tr>
<td>Average budget proposed by the government (at year t + 1)</td>
<td>639</td>
<td>594</td>
<td>801</td>
<td>1,119</td>
<td>825</td>
<td>1,064</td>
<td>1,310</td>
</tr>
<tr>
<td>(1,849)</td>
<td>(3,456)</td>
<td>(4,538)</td>
<td>(3,953)</td>
<td>(5,571)</td>
<td>(5,487)</td>
<td>(7,129)</td>
<td></td>
</tr>
<tr>
<td>Average finalized budget</td>
<td>656</td>
<td>596</td>
<td>800</td>
<td>1,175</td>
<td>830</td>
<td>1,073</td>
<td>1,325</td>
</tr>
<tr>
<td>(1,882)</td>
<td>(3,458)</td>
<td>(4,537)</td>
<td>(3,610)</td>
<td>(5,540)</td>
<td>(5,329)</td>
<td>(7,084)</td>
<td></td>
</tr>
</tbody>
</table>

Note. Standard deviations in parentheses. The budget in local currency unit (in one hundred million; 1,000 Won $1).

Appendix B

Sensitivity and Falsification Tests

The results of our analysis do not have applicability to the performance budgeting system that does not attach any budgetary consequences, such as in the United States. As such, our study does not necessarily imply that systems such as PART in the United States should be abandoned. Rather, we believe that our results give insights to future policymakers that considering the incentives of public managers are very important for designing effective performance budgeting system. Our analysis highlights that the system—if not designed properly—may result in undesirable consequences that invalidates the use of performance measures to evaluate public programs. For example, we find that manipulation is more likely if program managers undertake the burden for evaluation. Worrisome in this regard is the Obama administration’s initiatives to revamp the performance management system, expecting to put more responsibility of the evaluation on agencies’ shoulders and away from the Office of Management and Budget (Newcomer, 2010).

Last, we would like to emphasize that designing the performance budgeting system that is manipulation free is a hard task, as there will always be bureaucratic politics and organizational practices that make the system susceptible to gaming. Our study, we hope, may help the policymakers in detecting reasonably the indications of gaming using data and is valuable for policymakers around the world in designing more effective performance management systems.
Figure B1. Tests of sensitivity in the discontinuity estimates to bandwidth choices.

Note. The optimal bandwidth corresponds to the bandwidth choice provided by the plug-in formula suggested by McCrory (2008).
Table B1. Discontinuity Estimates at Cutoffs From 51 to 69.

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Model A</th>
<th></th>
<th>Model B</th>
<th></th>
<th>Model C</th>
<th></th>
<th>Model D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>h = 4 Rank</td>
<td></td>
<td>h = 6 Rank</td>
<td></td>
<td>h = 8 Rank</td>
<td></td>
<td>h = 10 Rank</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>-0.781*** (0.181)</td>
<td>17</td>
<td>-0.325** (0.142)</td>
<td>13</td>
<td>-0.064 (0.124)</td>
<td>11</td>
<td>0.129 (0.112)</td>
<td>6</td>
</tr>
<tr>
<td>52</td>
<td>0.246* (0.164)</td>
<td>7</td>
<td>0.216* (0.130)</td>
<td>4</td>
<td>0.289*** (0.115)</td>
<td>3</td>
<td>0.316*** (0.106)</td>
<td>3</td>
</tr>
<tr>
<td>53</td>
<td>0.271*** (0.150)</td>
<td>6</td>
<td>0.108 (0.122)</td>
<td>9</td>
<td>0.120 (0.107)</td>
<td>6</td>
<td>0.059 (0.100)</td>
<td>8</td>
</tr>
<tr>
<td>54</td>
<td>-0.500*** (0.154)</td>
<td>14</td>
<td>-0.382*** (0.123)</td>
<td>15</td>
<td>-0.403 (0.108)</td>
<td>17</td>
<td>-0.478*** (0.101)</td>
<td>17</td>
</tr>
<tr>
<td>55</td>
<td>0.330 (0.155)</td>
<td>5</td>
<td>0.185 (0.124)</td>
<td>5</td>
<td>-0.103 (0.109)</td>
<td>12</td>
<td>-0.206*** (0.098)</td>
<td>13</td>
</tr>
<tr>
<td>56</td>
<td>-0.476*** (0.160)</td>
<td>13</td>
<td>-0.513*** (0.132)</td>
<td>17</td>
<td>-0.672*** (0.119)</td>
<td>18</td>
<td>-0.542*** (0.101)</td>
<td>18</td>
</tr>
<tr>
<td>57</td>
<td>0.626*** (0.164)</td>
<td>3</td>
<td>0.154 (0.135)</td>
<td>7</td>
<td>0.024 (0.114)</td>
<td>8</td>
<td>0.033 (0.096)</td>
<td>9</td>
</tr>
<tr>
<td>58</td>
<td>-0.581*** (0.167)</td>
<td>16</td>
<td>-0.578*** (0.143)</td>
<td>18</td>
<td>-0.242*** (0.111)</td>
<td>14</td>
<td>-0.089 (0.094)</td>
<td>10</td>
</tr>
<tr>
<td>59</td>
<td>-1.173*** (0.218)</td>
<td>19</td>
<td>-0.336*** (0.135)</td>
<td>14</td>
<td>0.022 (0.106)</td>
<td>9</td>
<td>0.158* (0.091)</td>
<td>5</td>
</tr>
<tr>
<td>60</td>
<td>1.153*** (0.182)</td>
<td>1</td>
<td>1.057*** (0.131)</td>
<td>1</td>
<td>0.983*** (0.109)</td>
<td>1</td>
<td>0.890*** (0.093)</td>
<td>1</td>
</tr>
<tr>
<td>61</td>
<td>0.353*** (0.120)</td>
<td>4</td>
<td>0.500*** (0.102)</td>
<td>2</td>
<td>0.546*** (0.090)</td>
<td>2</td>
<td>0.551*** (0.082)</td>
<td>2</td>
</tr>
<tr>
<td>62</td>
<td>0.145 (0.106)</td>
<td>10</td>
<td>0.162* (0.092)</td>
<td>6</td>
<td>0.196** (0.083)</td>
<td>4</td>
<td>0.210*** (0.077)</td>
<td>4</td>
</tr>
<tr>
<td>63</td>
<td>-0.416*** (0.115)</td>
<td>12</td>
<td>-0.429*** (0.096)</td>
<td>16</td>
<td>-0.383*** (0.085)</td>
<td>16</td>
<td>-0.341*** (0.078)</td>
<td>15</td>
</tr>
<tr>
<td>64</td>
<td>-0.931 (0.156)</td>
<td>18</td>
<td>-0.811*** (0.114)</td>
<td>19</td>
<td>-0.753*** (0.097)</td>
<td>19</td>
<td>-0.691*** (0.086)</td>
<td>19</td>
</tr>
<tr>
<td>65</td>
<td>1.066*** (0.184)</td>
<td>2</td>
<td>0.296** (0.121)</td>
<td>3</td>
<td>-0.035 (0.098)</td>
<td>10</td>
<td>-0.170* (0.087)</td>
<td>12</td>
</tr>
<tr>
<td>66</td>
<td>0.218* (0.153)</td>
<td>8</td>
<td>0.113 (0.129)</td>
<td>8</td>
<td>-0.173 (0.106)</td>
<td>13</td>
<td>-0.320*** (0.094)</td>
<td>14</td>
</tr>
<tr>
<td>67</td>
<td>-0.559*** (0.170)</td>
<td>15</td>
<td>-0.256* (0.142)</td>
<td>12</td>
<td>-0.307*** (0.120)</td>
<td>15</td>
<td>-0.461*** (0.104)</td>
<td>16</td>
</tr>
<tr>
<td>68</td>
<td>0.011 (0.185)</td>
<td>11</td>
<td>0.084 (0.150)</td>
<td>10</td>
<td>0.119 (0.131)</td>
<td>7</td>
<td>-0.116 (0.112)</td>
<td>11</td>
</tr>
<tr>
<td>69</td>
<td>0.157 (0.196)</td>
<td>9</td>
<td>0.063 (0.154)</td>
<td>11</td>
<td>0.180 (0.139)</td>
<td>5</td>
<td>0.078 (0.122)</td>
<td>7</td>
</tr>
</tbody>
</table>

Note. Discontinuity estimates are derived from running the local linear regression (i.e., first-order polynomial) using the triangle kernel function. Standard errors are reported in parentheses. The lower the rank, the higher the magnitude of the discontinuity estimate within the model.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**Notes**

1. Recognizing the importance of the credibility of performance information, Yang (2009) uses the survey data from Taiwan to analyze how bureaucratic politics and organizational practice influence performance reporting.

2. Note that it is difficult to grasp fully the method developed by McCrory (2008) without knowledge in mathematical statistics, regression discontinuity designs, and nonparametric regression methods. An excellent discussion on these related topics are presented, in an easy-to-digestible manner, in Lee and Lemieux (2010), and we refer to this article for a more complete exposition of the methodological details.

3. For the figures onward, we plot graphs using a binwidth of 1 and superimposing onto the plot the estimated values from a second-order polynomial regression with a triangle kernel. The overall look of the graphs stays the same.
regardless of one’s choices on binwidths and polynomial specifications.

4. Note, however, that the overall shape of the graph stays the same even if we use the same bandwidth choice for both sides.

5. We also tested for discontinuity by year, and the results were very similar to the pooled version.

6. We appreciate the reviewer for raising this point.

7. The provision does not state how the programs rated as “excellent or above” will be rewarded. Unfortunately, we do not have information on whether these highly rated programs received awards or not. Anecdotal evidence, however, indicates that there are very few programs that were in fact rewarded.

8. Note that we also estimated the discontinuity using four other kernel functions (i.e., Biweight, Cosine, Parzen, and the alternative Epanechnikov), and consistently find the discontinuities in the density function of the assignment variable that are highly statistically significant.

9. Estimates are not presented for the sake of manuscript length; the table is available upon requests.

10. The discontinuity estimates for the cutoffs from 70 to 89 are not presented in the Appendix for the sake of manuscript length. The authors can provide the table upon requests.

References


**Author Biographies**

**Hosung Sohn** is Associate Research fellow at the Korea Institute for Health and Social Affairs, Sejong, South Korea. He is interested in conducting program evaluations in the field of public finance using experimental and quasi-experimental research designs, especially relating to the issues of causal inference.

**Kwang Bin Bae** is an Assistant Professor in the Department of Public Administration at North Carolina Central University. His research interests are focused on human resource management, performance management, and nonprofit finance.