Management Accounting Research 44 (2019) 1–11

Contents lists available at ScienceDirect

Management Accounting Research

journal homepage: www.elsevier.com/locate/mar

Reward type and performance: An examination of organizational wellness programs☆

William G. Heninger, Steven D. Smith, David A. Wood*

Brigham Young University, Marriott School of Management, School of Accountancy, 540 TNRB, Provo, UT, 84602, USA

ARTICLE INFO

Keywords:
Wellness programs
Incentives
Gift cards
Cash incentives
Tangible rewards

ABSTRACT

We examine organizational control in the context of wellness programs—organizational initiatives designed to improve the physical and mental health of employees. In a field study setting, we examine the associations of three different types of incentives (cash, gift cards, and tangible rewards) with wellness program performance. We find that employees who successfully complete program challenges are associated with greater weight loss. We also find participants choosing gift cards are associated with the greatest program success, even though cash rewards are selected more than twice as often as gift cards. Tangible rewards are the least frequently selected reward and are associated with lower performance than gift cards but relatively similar performance to cash. These results support theories of individual choice and motivation, and suggest that employees' incentive choices are not necessarily aligned with the strongest motivational power.

1. Introduction

An emerging literature in the area of management control examines the effects of different reward types on employee performance (e.g., see Jeffrey, 2009; Shaffer and Arkes, 2009; Presslee et al., 2013; Kelly et al., 2015; Choi and Presslee, 2017). Many firms now use rewards other than cash bonuses (e.g., gift cards, vacations, or consumer goods) to provide incentives in pay-for-performance systems (Incentive Federation, 2013). While prior research shows that most employees express a preference for cash incentives, evidence on the performance effects of reward type is mixed, and some studies show that different reward types can lead to variation in employee performance that is not always congruent with stated preferences (Jeffrey, 2009; Kelly et al., 2015). In this study, we examine employees' reward type preferences and the association of reward type and employee performance in a unique setting: organizational wellness programs.[1]

Wellness programs represent efforts by organizations to promote and increase the health and well-being of their employees. These programs often include short-term wellness initiatives with individual or group-level rewards for goal achievement. A vast majority of large businesses now offer wellness programs (Jakobson, 2013), and evidence suggests that these programs can yield significant productivity gains and health cost savings (Baicker et al., 2010; Berry et al., 2010). However, many firms have not observed significant benefits, in large part due to difficulty in motivating employee participation (some estimate that average employee participation is only around 20 percent (Jakobson, 2013)). Evidence on the performance effects of different reward types may therefore be useful to firms seeking to increase wellness program engagement by employees.

Prior research on reward type finds that, while most employees express a preference for cash incentives, cash is not always the most effective motivator of behavior (Jeffrey, 2009; Kelly et al., 2015; Shaffer and Arkes, 2009). This research suggests that employees may express a preference for cash because its fungibility makes it the more economically rational choice, while more hedonic and mentally separate rewards are actually more appealing and thus more motivating (Shaffer and Arkes, 2009). We similarly rely on theories of mental accounting and hedonic processing to develop our hypothesis regarding the association between reward types and wellness challenge performance.

We analyze data from 1855 participants in the employee wellness program of a large university who participated in up to six different challenges resulting in 8686 observations. The program offers rewards to employees and their spouses for acquiring points during six-week periods referred to as wellness challenges. Points can be accumulated in two ways: (1) by achieving the set goal for the period; and (2) by achieving personal goals. To gain enough points for each six-week challenge, participants must be successful in achieving at least some of the set challenge goals

☆ We thank Henri Dekker (editor) and two anonymous reviewers, along with Willie Choi, Khim Kelly, Adam Presslee, and Tim Seidel for providing suggestions on how to improve our manuscript. We also express our gratitude to Christian Clayton and Michael Thomas for valuable research assistance.

* Corresponding author.
E-mail addresses: heninger@byu.edu (W.G. Heninger), stevesmith@byu.edu (S.D. Smith), davidwood@byu.edu (D.A. Wood).

[1] We note we are only able to study associations and not causal relationships in this study given that our data is based on a previously collected archive of data rather than random assignment to experimental conditions.

https://doi.org/10.1016/j.mar.2019.02.001
Received 22 May 2017; Received in revised form 12 February 2019; Accepted 15 February 2019
Available online 22 February 2019
1044-5005/

(i.e., personal goal attainment is not sufficient to accumulate enough points). Examples of challenge goals set by the institution include a daily number of steps taken, an amount of weight loss or non-weight gain, or a number of days on which the employee must complete some form of exercise. Participants also can set personal goals during each challenge period, and a point system that combines personal goals and challenge goals is used to evaluate performance. Importantly, at the beginning of each challenge participants are able to choose which reward type they would like to receive upon successfully completing the challenge. Specifically, participants may choose to be rewarded with a cash bonus to be paid at the end of the year, a cash bonus applied to a university card that can be used similarly to a gift card[2] anywhere on campus (e.g., at eateries, bookstores, etc.), or one of a selection of tangible rewards of equivalent monetary value (e.g., a gym bag, George Foreman Grill, hydration pack, food processor, first aid kit, or a Leatherman utility knife). The collected data include reward choices, points received, pounds lost, and whether the challenge goal was achieved for each program participant per challenge. In total, we analyze one year of data that spanned two calendar years and consisted of six independent challenges.

We first establish that program participants who complete wellness challenges lose more weight than those who do not. This result confirms the wellness benefits of challenge completion in our setting, and establishes challenge completion as an appropriate dependent measure for examining the association of reward choice with wellness program performance. We also find that 60 percent of wellness program participants choose cash rewards at the beginning of a challenge, followed by 30 percent and 10 percent selecting gift cards and consumer goods, respectively. This result confirms previous research findings of a general stated preference for cash over other reward types.

We also find that reward type is associated with wellness program performance, as measured by challenge completion. Across the sample as a whole, participants who select gift cards and cash rewards perform similarly to each other, but better than participants who choose tangible rewards as the program incentive. Interestingly, we find that *employee* participants' performance is higher when gift cards are selected than when either cash or tangible rewards are chosen. We also find that the performance differences are weaker for *spousal* participants, which is consistent with the mental accounting and motivational theories underlying our prediction (in particular, the degree to which the reward can be considered as separate from the employee's salary). Thus, among employees, cash is selected most frequently, but is not associated with the highest performance.

This paper contributes to the emerging literature on reward type in organizations in a number of ways. To our knowledge, ours is the first study to consider the association between reward types and performance in an organizational wellness program setting. As companies strive to achieve greater efficiencies with smaller workforces (Sheehan et al., 2000; Gandolfi, 2013), the importance of healthy and productive employees increases, and organizational encouragement of healthy lifestyles is likely to continue. Thus, our study represents an important and timely expansion to the literature on the role of incentives within the organizational control environment.

Using a field study setting in which we are able to observe actual behavior, we find that employee wellness program participants perform at a higher level when they have chosen to be rewarded with gift cards than with cash or tangible consumer goods, despite choosing to be rewarded with cash significantly more often. These results are consistent with theoretical arguments of economic fungibility and mental accounting (Thaler, 1985). Importantly, the choice data in our study represent actual reward choices made by participants, rather than survey or post-experimental expressions of preference, which are common in studies that assign reward types randomly. To our knowledge, this is the first study that examines performance relationships across three different reward types in a naturally occurring setting where rewards are self-selected.

Our study should be of interest to managers seeking to increase participation in organizational wellness programs. All three of the reward types offered in our field setting were selected with some frequency, suggesting a general benefit to the offering of reward type choices. However, we find evidence that employees choosing to be rewarded with gift cards reap the greatest health benefits from the program. Our results are also consistent with the argument that cash and tangible rewards may have offsetting costs and benefits (vis-à-vis their fungibility and the separate mental accounts into which they are categorized), and that gift cards represent a balance of the strengths and weaknesses of different reward types.

2. Background

The number of organizational wellness programs has increased significantly in the past decade. Data gathered by the RAND Corporation indicates that 92 percent of companies with more than 200 employees offer wellness programs to their employees (Jakobson, 2013). Brino (2015) finds that companies spend an average of $693 per employee on wellness programs, and that this number jumps up to $878 per employee for companies with more than 20,000 employees. Most recently, PWC announced that they will spend $45 million on wellness programs for their employees (Inside Public Accounting, 2018).

Advocates claim that, due to better health and reduced absenteeism, company wellness programs achieve high productivity ROIs, substantial savings in employee benefit costs, and improvements in worker productivity (Baicker et al., 2010; Berry et al., 2010; Gubler et al., 2018). Baicker et al. (2010) report that every dollar spent on wellness programs achieves an average ROI of $3.27 for medical costs and $2.73 for absenteeism. Johnson & Johnson claims that over the last 20 years its wellness program has saved the company $250 million, for a total per dollar return on investment of $2.71 (Berry et al., 2010). Other research reports worker medical costs fall by $3.27 for every $1 spent on wellness programs and that absenteeism costs are reduced by $2.73 for every dollar spent (Kocakulah and Powers, 2015). Finally, Gubler, Larkin, and Pierce (2018) report that participation in wellness programs increases worker productivity by about ten percent.

Despite evidence of positive outcomes, however, many wellness programs are considered ineffective. The RAND data indicate that, on average, only about 20 percent of employees participate in their company's wellness programs (Jakobson, 2013). Moreover, the per employee cost of wellness programs has increased over time (FMR, 2015). Some have argued that wellness programs are ineffective because they are not supported by their company cultures, and/or that employees do not truly want to change unhealthy habits (Butler, 2014). Others argue that the effectiveness of wellness programs is more a function of the provision of proper information and appropriate incentives for employees to participate (Johnson, 2014).

We examine the latter possibility, specifically focusing on how incentive type is associated with performance in wellness programs. The study of incentive effects in control systems has produced a rich literature in management accounting (Shields, 1997; Sisaye, 2005; Verbeeten, 2008; Maas and van Rinsum, 2013; Otley, 2016; Chen et al., 2017). Our research extends this literature into a new and important determinant of organizational performance (i.e., employee wellness).

3. Research setting

Our research setting is the wellness program at a large, private university in the western United States. The program is offered to all

[2] Prior research has defined gift cards as tangible rewards (e.g., see Presslee et al., 2013). We separate these two reward types because our theoretical predictions (discussed in the next section) suggest they may not function identically. In our setting, employees and spouses both have university cards (spouses can use cards for benefits such as various discounts on and off campus, use of physical facilities, free local public transportation, etc.).

university faculty, full-time staff, and spouses. Participation is not mandatory, but is encouraged.

The wellness program comprises six separate "challenges" over the calendar year. Each challenge is an independent event with its own sign-up, tracking, completion, and reward. Participants can complete as many or as few challenges as they choose throughout the year, but challenges cannot be made up or performed in a different order than the program dictates.[3] Each challenge lasts about two months.

To successfully complete a challenge, a participant must accumulate a predetermined number of points. Points can be earned in different ways, depending on the challenge. The majority of challenges require the completion of a certain amount of exercise (usually measured in steps or minutes of exercise per day) and achievement of a personal goal. Examples of personal goals include weight training, weight loss, getting sufficient sleep, eating fruits and vegetables, self-selected goals, etc.[4] At the beginning of each challenge, each participant selects a personal goal, the measurement of exercise (steps or minutes), and the reward. Points are earned by completing a sufficient amount of exercise and personal goals. If a participant gains enough points during the period, he or she receives the selected reward.[5]

For each challenge, there are three different rewards from which a participant chooses (simultaneously with challenge enrollment and the setting of the personal goal): $30 cash, which is paid in the first paycheck of the next calendar year; an immediate $30 credit to the participant's university ID card which can be used to make purchases anywhere on campus[6] (we refer to this reward as a gift card); or the consumer good that is offered for that challenge.[7] During the time we measured performance, the consumer goods included: a CalPak Riviera Gym Bag, a George Foreman Grill, a High Sierra Hydration Pack, a Brentwood Food Processor, a Family First Brand First Aid Kit, and a Leatherman Wingman utility knife.[8] Participants know what tangible reward is offered when they make their reward selection. The reward types offered are similar to those offered in other wellness programs (Hall, 2008; Finkelsteing and Kosa, 2003, p. 57; Towers, 2012).

We collected program data for 1855 participants pursuing six challenges during 2014. These participants participated in a total of 8686 challenges.

[3] The program does offer a special fitness evaluation, planning, and performance option to take the place of two challenges. This evaluation is meant to be a more thorough training and health program as it involves body fat tests, meeting with student coaches, and additional efforts. This is separate from the challenges we examine (i.e., participants cannot enter this and the challenges) and thus, we do not include any participation in this event in our data analysis. Our discussions with program administrators revealed that relatively few employees or spouses enroll in this program.

[4] Personal goals represent a minor component of each challenge; that is, the majority of points are earned by completing the primary challenge task, which is constant across all participants.

[5] There is no explicit monitoring of either personal or challenge goals. Thus participants can create easy goals and/or lie about their performance. Although there is no direct monitoring, we do note that the university has an honor code that all employees are required to follow and regularly review, which should help to mitigate lying. Also, we are not aware of any reason why participants would lie differentially across conditions, suggesting our results are unlikely to be fully explained by participants in one condition being less honest than those in other conditions.

[6] Examples of places that the gift card may be used include numerous campus restaurants, the campus store (which sells a range of products such as books, clothing, food, electronics, gifts, and novelty items), university events, etc.

[7] The different timing of the delivery of the various reward types is a limitation of the data. As reported in section 6.1, we conduct supplemental analyses to alleviate the concern that our results can be explained by such timing differences.

[8] The retail value of each of these goods was about $30 at the time of the challenges. While this amount was not explicitly listed in the program for participants, they could easily search for the items online to find their approximate values. As discussed in the next section, the hedonic value of the rewards is an important part of our theory. The consumer goods offered in the wellness program are likely to have varying degrees of hedonic value.

4. Theory and hypothesis

To develop our prediction, we consider the mental categorization and fungibility of each reward type. Theories of mental accounting argue that people categorize economic outcomes into different mental "accounts" (Thaler, 1985, 1999), and that once outcomes are categorized into different accounts, people will consider and treat them differently even though the funds are economically identical. For example, a person finding a $20 bill on the road might mentally categorize that money as "fun" or "extra" money and spend it on something for which they would not spend $20 they earned from work, which they would categorize as salary or "work" money. Fungibility refers to the ability to use a reward in different ways or for different purposes; that is, it is the degree to which a reward can be used in mutual substitution with other goods and services. We specifically consider how cash, gift cards, and tangible rewards differ in their fungibility and mental accounting effects, and how those differences are likely to be associated with participants' wellness challenge performance.

At the beginning of each wellness challenge, participants are likely to choose the reward they believe will provide the most motivation and the highest likelihood of achieving the challenge goals and receiving the selected reward. Thus, in our setting, the null hypothesis that reward type will have no significant association with performance is particularly plausible. That is, since each participant chooses the reward they think will best motivate them to achieve their challenge goals, it is reasonable to predict that performance will be equal across all reward types (note that participant choice therefore becomes an effective control for ex ante motivation level in our analyses). However, theories of motivation from both psychology and economics suggest that *during* the wellness challenge (i.e., when motivation to pursue challenge goals is needed), different reward types will be associated with differences in motivation and performance.

Mental accounting theory suggests that different types of rewards (specifically cash and tangible rewards, which could include both gift cards and consumer goods) are likely to be categorized into different mental accounts based on how separately they can be considered from salary. Choi and Presslee (2017) provide experimental evidence that individuals are more likely to categorize tangible rewards as different from salary, relative to cash rewards. Furthermore, they find that performance on an effort-sensitive task is greater when rewards are categorized separately from salary. These results suggest that the perception of separateness from salary increases the affective appeal, and thus the motivational force, of the reward. Applied to our setting, these results suggest that the more separately rewards are mentally categorized from salary, the more likely they are to improve wellness challenge performance.

Of the three wellness program reward types we consider, cash is likely to be perceived as being most similar to salary. Alonzo (1996) and Odell (2005) both report that cash rewards from work are viewed as "more salary." Cash rewards are usually denominated in the same currency as salary, and in our setting they are paid as part of the employee's normal paycheck.

Gift card and consumer good rewards are less likely to be categorized similarly to salary than cash rewards. White (2006) suggests that gift cards are perceived more as "spendable money" relative to cash because they are restricted in application and are considered wasted if not spent. Gift card expenditures also differ from cash expenditures in that they cause less spending guilt (Valentin and Allred, 2012). In our setting, gift cards are not paid out with employee paychecks but are kept on a separate card with a separate balance. Tangible rewards have similar separateness characteristics, and are likely to be perceived even less like salary than gift cards, since they are physical objects that are not denominated in dollars and which must be picked up on campus rather than received automatically. Based on the criteria of *separateness from salary*, then, gift cards and tangible rewards should both be expected to motivate greater performance than cash rewards.[9]

[9] A related mental accounting argument focuses on the psychological costs of quitting, or failing to complete a challenge. In the case of cash, failure to complete a challenge still leaves a mental account with salary in it, mitigating

However, separateness from salary should not be considered as the only factor that determines the motivational force of a reward. We argue that motivation will also be affected by the participant's consideration of the range of possibilities each reward offers; that is, the *fungibility* of the reward. Neoclassical economic theory predicts that cash will be preferred to noncash incentives of equal market value, and will provide greater motivation, because of its high option value (List and Shogren, 1998; Walfogel, 1993, 1996). That is, the recipient of a cash reward can buy whatever tangible reward may be offered, as well as any number of other items or experiences that provide more utility. The gift cards in our setting have some fungibility given their relative breadth of usability, while the tangible rewards effectively have none—the reward cannot be exchanged in any way.

In forming a prediction about the relative performance effects of the three reward types, then, we consider how they compare across both dimensions (separateness from salary and fungibility). Cash (tangible) rewards are the most (least) favorable from a fungibility perspective, but are the least (most) favorable from a separateness from salary perspective. In contrast, gift cards appear to be relatively high on both dimensions in our setting. Though less fungible than cash, they can be used in a variety of ways; and while less separate from salary than tangible rewards, they are received and used in a way that is distinct and which can generate strong affective appeal.[10] This discussion leads us to predict that gift cards will be associated with the highest wellness program performance. Given the offsetting costs and benefits of cash rewards and tangible rewards, however, we do not make a formal prediction about how those two reward types will compare.

Hypothesis. *Wellness program performance will be greater for gift card participants than for either cash or tangible reward participants.*

5. Method and design

5.1. Participants

We analyze the reward selections and wellness challenge performance of 1855 university faculty, staff, and spouses who participated in up to six different challenges resulting in 8686 observations. Table 1 provides demographic information about the participants. Nearly 70 percent of participants were current university employees. The average age of participants was 47, and 45.5 percent of participants were male. The average weight at the beginning of the challenges was approximately 177 pounds. Also, nearly 30 percent of the time participants selected the gift card as the reward type, 9 percent of the time they selected the tangible reward, and 61 percent of the time they selected cash. The success rate across all challenges in our sample is 85.8 percent.

We were not able to acquire population statistics for the faculty, staff, and spouses of the university personnel, but we have no reason to believe they differ widely from this diverse sample. Sample demographics are presented in Table 1.

Data for our analyses includes two samples. The larger sample includes all variables other than the weight variables. Far fewer participants provided weight information, thus we analyze the smaller sample that includes weight values to establish that performance is related to an important health outcome and analyze the larger sample, omitting controls for weight, to test the hypothesis. The correlation matrix in Table 2 shows that, other than a high correlation between starting weight and gender, the correlations are quite modest between variables, suggesting that not controlling for weight in the tests of our hypothesis should not be problematic for our inferences.

Table 1
Sample Descriptive Statistics.

Variable	N	Mean	Std Dev	Minimum	Maximum
Employee	8,686	69.5%	46.0%	0	1
Age	8,686	47.2	11.0	26	70
Male	8,686	45.5%	49.8%	0	1
SamePrize	8,686	61.6%	48.6%	0	1
PrevChallenges	8,686	2.20	1.68	0	5
WeightStart	3,125	176.9	36.4	104	295
WeightLoss	3,125	−1.02	2.95	−14	6
Complete	8,686	85.8%	34.9%	0	1
GiftCard	8,686	29.9%	45.8%	0	1
TangibleReward	8,686	9.2%	28.9%	0	1
Cash	8,686	60.9%	48.8%	0	1

Responses are trimmed at the 1% and 99% levels for continuous variables age, weight, and weight loss because of outliers. All missing values are excluded from analysis. The 8686 observations relates to 1855 unique individuals. The 3125 observations relate to 1477 unique individuals. Variables are defined as following:
Employee = A dummy variable equal to one if the participant is a current employee and zero otherwise.
Age = The age, in years, of each participant.
Male = A dummy variable equal to one if the participant is male and zero if they are female.
SamePrize = A dummy variable that is equal to one if the participant selected the same reward for all challenges measure and zero if they did not.
PrevChallenges = The number of challenges completed before the current challenge (e.g., for challenge 4, if they completed challenges 1, and 3 the variable would have a value of 2).
WeightStart = The weight, in pounds, of each participant at the beginning of a challenge.
WeightLoss = The weight, in pounds, lost by each participant during the challenge period. A positive (negative) value indicates a participant gained (lost) weight.
Complete = A dummy variable equal to one if the participant completed the challenge and zero otherwise.
TangibleReward = A dummy variable equal to one if the participant selected to receive the tangible reward and zero otherwise.
GiftCard = A dummy variable equal to one if the participant selected to receive the gift card reward and zero otherwise.
Cash = A dummy variable equal to one if the participant selected to receive the cash reward and zero otherwise.

5.2. Challenge completion as a performance measure

Given that the program is designed around wellness challenges, we focus on challenge completion as the primary binary independent variable for our analysis. In order to establish challenge completion as a legitimate performance measure, we test for a positive relationship with participants' health outcomes. We use weight loss as the primary dependent measure, given its continuous nature and the broad acceptance of weight loss as a generally beneficial health outcome (Higgins et al., 1993).

To establish this relationship, we analyze the following regression model (clustering standard errors by participant):

$$\text{WeightLoss} = B0 + B1*Complete + B2*Employee + B3*Age + B4*Male + B5*SamePrize + B6*PrevChallenges + B7*WeightStart + Challenge\ Control\ Variables + e$$

In this model, the dependent variable is a measure of how much weight each participant lost during the challenge. The *WeightLoss* variable is positive if participants gained weight and negative if they

(*footnote continued*)
the pain of the loss of the reward. In the case of gift cards and tangible rewards, the associated mental accounts are subsequently more "empty," resulting in a greater psychological cost of failure and thus motivating higher levels of effort. We thank a reviewer for recommending this additional line of reasoning.

[10] Another factor that is likely to limit the motivational influence of tangible rewards in our setting is that the items do not convey a significant level of status, nor are they items of high hedonic value, as tangible rewards are in other contexts (i.e., vacations, expensive electronics, etc.). Both of these factors have been shown to influence the motivational power of rewards (Kosfeld and Neckermann, 2011; Bradler et al., 2016).

Table 2
Correlation Matrix.

Variable	Complete	Tangible Reward	Gift Card	Cash	Employee	Age	Male	SamePrize	PrevChallenges	WeightStart
TangibleReward	**−0.04**									
GiftCard	0.01	**−0.21**								
Cash	0.01	**−0.40**	**−0.81**							
Employee	−0.02	**0.03**	0.02	**−0.03**						
Age	**0.07**	**0.05**	**0.03**	**−0.06**	**0.12**					
Male	**0.04**	**−0.05**	0.01	0.02	**0.39**	0.01				
SamePrize	**−0.20**	**0.20**	**0.15**	**−0.26**	**0.02**	0.01	**−0.02**			
PrevChallenges	**0.11**	0.01	**−0.05**	**0.04**	0.00	**0.03**	0.01	**−0.14**		
WeightStart	−0.02	0.01	**0.04**	**−0.04**	**0.24**	**0.06**	**0.48**	0.00	**−0.04**	
WeightLoss	**−0.05**	0.00	−0.03	0.03	0.00	0.02	0.00	−0.02	**0.17**	**−0.15**

All variables included in models are presented in a Pearson correlation matrix. Bold indicates p-value < 0.05. Variables are defined in Table 1. We include all observations despite the violation of independence assumptions. If we only examine one response per person, the correlations are similar in magnitude to those reported. More specifically, we create three different samples of unique observations, keeping the first response by each participant, the last response by each participant, or a random response by each participant. We then compare these three samples to those reported (a total of 165 comparisons) and find that in only five cases is the magnitude difference greater than 0.1, and in only two cases is the magnitude difference greater than 0.16 (magnitude differences of 0.385 and 0.334). In both of those cases, the correlation is still below 0.5 between PreviousComplete and Complete and PreviousComplete and PreviousChallenges. Thus we believe it is appropriate to present the correlations above as a reasonable correlation between variables that violate the independence assumption.

lost weight. The key independent variable for testing the expected relationship is *Complete*, a dummy variable taking the value of one if the participant completed the challenge and zero otherwise. We expect a statistically significant positive coefficient on *Complete*.

We also include several control variables in this regression based on data availability. The full definition of each variable is included in Table 1. We included several demographic variables in the regression in an attempt to rule out that the possibility that results were caused by individual differences between participants. The available data include whether the participant is an employee or not, age, gender, and starting weight. We also include variables to control for whether participants selected the same prize for each challenge in the sample (*SamePrize*), as well as how many previous challenges each participant had completed (*PrevChallenges*). These were included to control for selection effects (i.e., to control for participants who may mechanically choose the same reward out of routine or habit, rather than an effort to maximized the possibility of success) and motivational effects from previous success or failure.

Finally, we include a control variable for each challenge, as the challenges differed slightly throughout the year. This also provides control for seasonal differences since we have only one year of data and the challenges were each run for a finite period during the year.[11]

To reduce the likelihood of observed effects arising endogenously, we also construct a matched sample using propensity score matching (PSM). Specifically, we model how the included control variables predict the likelihood of completing each challenge, and then create a matched sample of observations based on the propensity score that results from this equation. The PSM matched using one-to-one matches without replacement in order to only include treatment and control conditions a single time in the final sample. Based on Austin (2011), we use a caliper distance of 0.2 times the pooled estimate of the common standard deviation of the logits of the propensity scores. In additional tests, we test the robustness of our results to these choices.

The results of this match can be seen in Table 3. The use of PSM helps alleviate endogeneity concerns caused by observable variables. Using this technique helps rule out the possibility of functional form misspecification—that is, that the results are caused by systematic individual differences in the variables we include in the PSM model (Shipman et al., 2017).[12] Reporting results using both multiple regression models and PSM helps provide corroborating evidence that our results are robust to the statistical limitations of either method.

Panel A of Table 3 shows a comparison of variables before the propensity score match by comparing standardized mean differences.[13] Differences larger than 0.25 are considered large and potentially problematic (Rubin, 2001; Stuart, 2010). So in this case, *SamePrize* and *PrevChallenges* are potentially problematic. Panel B shows the first stage model that was used to create the propensity scores. This model shows that all of the variables other than *Employee* were important predictors of whether participants completed a challenge. Panel C shows a comparison of variables for the matched sample. We assess the quality of the match by evaluating the absolute standardized mean difference and variance ratio and follow prior research by interpreting values less than absolute value of 0.25 and between 0.5 and 2 to suggest effective balance, respectively (Rubin, 2001, p. 174; Stuart, 2010, p. 11). The data indicate that the matching was successful as no standardized mean differences are greater than the absolute value of 0.07 and the variance ratios range from 0.91 to 1.18. Table 3 also provides preliminary evidence supporting the expected relationship, in that the variable *WeightLoss* is larger for the group that completed the challenge than the group that did not.

5.3. Model for hypothesis testing

To test the hypothesis, we analyze the following model (clustering standard errors by participant):

$$\text{Complete} = \text{B0} + B1{*}GiftCard + B2{*}Employee + B3{*}Age + B4{*}Male + B5{*}SamePrize + B6{*}PrevChallenges + Challenge\ Control\ Variables + e$$

In this model, the dependent variable is *Complete*, a dummy variable measuring whether participants completed the challenge or not. The key independent variable for testing our hypothesis is *GiftCard*, a dummy variable taking the value of one if participants selected the gift card incentive and zero otherwise. The hypothesis predicts a statistically significant positive coefficient on *GiftCard*. We include the same control variables in this model as in the model described earlier, except we exclude *WeightStart* because it significantly reduces our sample size.

As with our tests of the first model, we create a propensity score

[11] All possible control variables from the available data are included.

[12] An additional technique to rule out endogeneity concerns from participants' reward selection is the use of Heckman modeling. Unfortunately, this technique requires the identification of a suitable instrumental variable, which we were unable to identify given our data. Future research should consider using the Heckman technique if a suitable instrumental variable can be identified.

[13] As stated by Austin (2009), "Unlike t-tests and other statistical tests of hypothesis, the standardized difference is not influenced by sample size. Thus, the use of the standard difference can be used to compare balance in measured variables between treated and untreated subjects in the matched sample with that in the unmatched sample. Furthermore, it allows for the comparison of the relative balance of variables measured in different units" (p. 3,087).

Table 3
Propensity Score Matching: Challenge Completion.

Panel A: Descriptive Statistics Before Match.

	Complete = 1 N (Part.) = 2,803 (1,378)		Complete = 0 N (Part.) = 322 (245)		
Variable	Mean	Std Dev	Mean	Std Dev	Abs. Mean Std. Difference
Employee	72.5%	44.6%	73.0%	44.5%	0.01
Age	48.2	10.7	46.6	11.7	0.14
Male	51.9%	50.0%	43.8%	49.7%	0.16
SamePrize	59.2%	49.2%	85.1%	35.7%	0.60
PrevChallenges	2.8	2.0	1.9	1.9	0.46
WeightStart	176.7	36.2	178.6	38.2	0.05
WeightLoss	−1.07	2.99	−0.56	2.62	0.18

Panel B: First Stage Model to Create Propensity Scores (DV = Complete)

Parameter	Estimate	t-value
Intercept	2.379	4.24***
Employee	−0.197	−0.99
Age	0.015	1.98**
Male	0.543	3.15***
SamePrize	−1.228	−6.15***
PrevChallenges	0.224	3.21***
WeightStart	−0.004	−1.98**
Challenge Control Variables	Yes	
N (Part.)	3,125	(1,477)
R^2	0.106	
Area under the ROC curve	0.713	

Panel C: Descriptive Statistics After Match

	Complete = 1		Complete = 0			
Variable	Mean	Std Dev	Mean	Std Dev	Abs. Mean Std. Difference	Variance Ratio
Employee	73.7%	44.6%	73.0%	44.5%	0.01	1.01
Age	47.9	11.0	46.6	11.7	0.02	1.18
Male	46.9%	50.0%	43.8%	49.7%	0.06	1.02
SamePrize	79.2%	40.7%	85.1%	35.7%	0.01	0.97
PrevChallenges	3.06	1.89	2.86	1.87	0.03	1.06
WeightStart	178.3	35.2	178.6	38.2	0.04	0.91
WeightLoss	−1.55	3.15	−0.56	2.62		
N (Part.)	322 (279)		322 (245)			

***, **, * represent p-values < 0.01, 0.05, and 0.10, respectively. The propensity score matching (Panel B) matched using one-to-one matches without replacement and a caliper distance of 0.2 times the pooled estimate of the common standard deviation of the logits of the propensity scores (Austin, 2011). In Panel C, the standardized mean differences and variance ratios are presented. Values less than absolute value of 0.25 and between 0.5 and 2 suggest effective balance, respectively (Rubin, 2001, p. 174; Stuart, 2010, p. 11). Variables are defined in Table 1. We note that the number of unique participants identified in Panel A (i.e., the (Part.) number) exceeds the total number of participants because some participants are counted in each group (that is, this is the unique number of people in that group, but people can be in both groups).

matched sample for testing our hypothesis. The model is the same except the dependent variable is *GiftCard* rather than *Complete*. The model is presented in Table 4.

The results in Panel A of Table 4 show the variable *SamePrize* has particularly large difference between conditions. Panel B shows the first stage model used to create the propensity scores for matching. Finally, Panel C shows that our match was largely successful. Using the same criteria as before (Rubin, 2001, p. 174; Stuart, 2010, p. 11), the data indicate that the matching was successful as only one variable, *Age*, has a standardized mean difference greater than the absolute value of 0.25. The variance ratios suggest all matches are successful, as values range from 0.91 to 1.15. We perform multiple sensitivity analyses surrounding matching in a subsequent section.

6. Results

The results of our tests of the relationship between wellness challenge completion and weight loss are presented in Table 5. The results show a significant negative coefficient on *Complete* in both the full sample and in the PSM-matched sample. This suggests that completing the wellness challenges is associated with greater weight loss, supporting challenge completion as a legitimate measure for assessing wellness program performance. Completing a challenge is associated with a 0.877 pound reduction in weight. The challenge is held over a six-week period, which suggests that if all six challenges are completed, the participant would be expected to lose an average of 5.26 pounds in a year.[14]

The results of our hypothesis tests are presented in Table 6. Panel A shows the percentage of challenge completion based on which incentive was chosen. Participants choosing the gift card incentive have the highest completion percentage, providing initial support for our hypothesis. The difference between those who selected the gift card and the cash incentive is quite small in magnitude; however, this comparison does not control for other factors that could contribute to completion percentages.

[14] This weight loss amount is consistent with several previous studies on workplace weight loss programs. For example, other programs have observed annual weight loss amounts of 1.98 pounds (Racette et al., 2009), 2.2 pounds (Lara et al., 2008), 2.4 pounds (Niv et al., 2014), and 5.07 pounds (Barham et al., 2011).

Table 4
Propensity Score Matching: Gift Card Selection.

Panel A: Descriptive Statistics before Match

	Gift Card = 1 N (Part.) = 2,599 (855)		Gift Card = 0 N (Part.) = 6,087 (1,865)		
Variable	Mean	Std Dev	Mean	Std Dev	Abs. Mean Std. Difference
Employee	70.7%	45.5%	69.0%	46.2%	0.04
Age	47.7	10.4	47.0	11.2	0.06
Male	46.3%	49.9%	45.2%	49.8%	0.02
SamePrize	73.0%	44.4%	56.7%	49.5%	0.35
PrevChallenges	2.1	1.7	2.2	1.7	0.06

Panel B: First Stage Model to Create Propensity Scores (DV = Gift Card)

Parameter	Estimate	t-value
Intercept	−2.311	−6.79***
Employee	0.018	0.17
Age	0.006	1.43
Male	0.056	0.58
SamePrize	0.758	6.52***
PrevChallenges	0.096	2.03**
Challenge Control Variables	Yes	
N (Part.)	8,686	(1,855)
R^2	0.042	
Area under the ROC curve	0.609	

Panel C: Descriptive Statistics after Match

	Gift Card = 1		Gift Card = 0			
Variable	Mean	Std Dev	Mean	Std Dev	Abs. Mean Std. Difference	Variance Ratio
Employee	71.7%	45.5%	61.0%	48.8%	0.21	1.15
Age	47.7	10.4	43.8	11.1	0.36	1.15
Male	46.3%	49.9%	34.4%	47.5%	0.24	0.91
SamePrize	73.0%	44.4%	71.3%	45.2%	0.04	1.04
PrevChallenges	3.07	3.01	3.07	3.00	0.00	0.98
Complete	86.4%	34.3%	81.5%	38.8%		
N (Part.)	2,599 (855)		2,599 (1,232)			

***, **, * represent p-values < 0.01, 0.05, and 0.10, respectively. The propensity score matching (Panel B) matched using one-to-one matches without replacement and a caliper distance of 0.2 times the pooled estimate of the common standard deviation of the logits of the propensity scores (Austin, 2011). In Panel C, the standardized mean differences and variance ratios are presented. Values less than absolute value of 0.25 and between 0.5 and 2 suggest effective balance, respectively (Rubin, 2001, p. 174; Stuart, 2010, p. 11). Variables are defined in Table 1. We note that the number of unique participants identified in Panel A (i.e., the (Part.) number) exceeds the total number of participants because some participants are counted in each group (that is, this is the unique number of people in that group, but people can be in both groups).

Table 5
The Association Between Challenge Completion and Weight Loss.

Parameter	Full Sample		PSM-Matched Sample	
	Estimate	t-value	Estimate	t-value
Intercept	1.124	1.91*	0.829	0.84
Complete	−0.877	−5.54***	−1.087	−4.69***
Employee	0.047	0.34	−0.088	−0.33
Age	0.008	1.70*	−0.002	−0.25
Male	0.599	4.32***	0.522	1.97**
SamePrize	−0.032	−0.27	−0.013	−3.96***
PrevChallenges	0.233	3.25***	0.207	0.78
WeightStart	−0.016	−8.11***	0.33	2.59***
Challenge Control Variables	Yes		Yes	
Standard errors Clustered by Participant	Yes		Yes	
R-squared	0.074		0.112	
N (Part.)	3,125 (1,477)		644 (484)	

***, **, * represent p-values < 0.01, 0.05, and 0.10, respectively. Panel A shows a regression using all observations whereas Panel B only includes observations that were successfully matched using propensity-score matching. For details of the PSM-matched sample see Table 3. Variables are defined in Table 1.

Panel B provides a more robust analysis and test, and shows that participants who selected the gift card incentive are more likely than those who selected the tangible reward or cash to complete the challenge. This result holds in both the full sample and the propensity score matched sample, and provides support for our hypothesis.

Panel B tests the effect of gift card selection against the joint selection of cash and tangible rewards. We perform separate comparisons of gift card selection against each of these other incentives in Panel C.[15] The analyses show that gift card participants achieve higher completion percentages than participants choosing either of the other incentives, while cash and tangible reward participants do not differ in their completion percentages (see f-test at the bottom of the table). These analyses provide further support for our prediction, and suggest that

[15] In Panel C we are unable to report propensity score matched results because there are three comparison groups. As an alternative analysis, we create two different propensity score matched comparisons: (1) *GiftCard* and *Tangible* and (2) *GiftCard* and *Cash*. When we create the propensity scores and run the analysis for these separate groups, we find that for group (1) the results are in the same direction but statistically significant at the p-value = 0.105 level (the lack of significance is likely because of the significant reduction in sample size in creating the matched sample). The results for group (2) are in the same direction and at the same level of significance.

Table 6
Wellness Program Performance Based on Reward Type.

Panel A: Descriptive Statistics for % who Completed the Task

Variable	N	Part.	Mean	Std
Gift Card	2,599	855	86.4%	34.3%
Tangible Reward	800	470	81.1%	39.2%
Cash	5,287	1,395	86.1%	34.6%

Panel B: Statistical Comparison of Conditions (Tangible and Cash are included in the intercept)

	Full Sample		PSM-Matched Sample	
Parameter	Estimate	t-value	Estimate	t-value
Intercept	0.012	0.04	0.449	6.92***
GiftCard	0.245	2.81***	0.037	3.18***
Employee	−0.277	−2.51**	−0.041	−2.64***
Age	0.017	3.90***	0.002	2.64***
Male	0.324	3.25***	0.050	3.42***
SamePrize	−1.258	−10.04***	−0.100	−8.25***
PrevChallenges	0.460	10.42***	0.081	8.22***
Challenge Control Variables	Yes		Yes	
Standard errors Clustered by Participant	Yes		Yes	
R-squared	0.073		0.070	
N (Part.)	8,686 (1,855)		5,198 (1,639)	

Panel C: Statistical Comparison of Conditions (GiftCard is included in the intercept)

	Full Sample	
Parameter	Estimate	t-value
Intercept	0.258	0.76
Tangible	−0.244	−1.85**
Cash	−0.246	−2.67***
Employee	−0.277	−2.51**
Age	0.017	3.88***
Male	0.324	3.26***
SamePrize	−1.258	−9.83***
PrevChallenges	0.460	10.39***
Challenge Control Variables	Yes	
Standard errors Clustered by Participant	Yes	
R-squared	0.073	
N (Part.)	8,686 (1,855)	

F-test of Tangible = Cash; F-value = 0.00, p-value = 0.990.
***, **, * represent p-values < 0.01, 0.05, and 0.10, respectively. Panel A provides descriptive statistics of the percentage of participants who completed each challenge by reward type. Panel B and Panel C provides a logistic regression with Complete as the dependent variable. Panel C does not include a propensity-score matched evaluation since there are three different groups (GiftCard, Tangible, and Cash) and so a first stage prediction of one group is not possible. All variables are defined in Table 1.

gift card incentives are associated with greater wellness program success than either cash or tangible reward incentives.

6.1. Additional analyses

6.1.1. Differences between employees and spouses

We have argued that the motivational power of the challenge reward is likely to be a function of both its fungibility and the extent to which it can be perceived as separate from salary. Given that the gift card reward in our setting takes the form of money added to the participant's university ID card (to be used at campus stores and restaurants), it is likely that the perceived fungibility of the gift card is likely to differ between employees (who are regularly on campus where the gift card money can be spent) and spouses (who are not regularly on campus, making the gift card more difficult to use). All else equal, then, an employee is more likely than a spouse to perceive the gift card as a fungible reward. Given this difference, we would expect that the predicted results predicted are more likely to be observed among employees than among spouses.

To examine this possibility, we conduct our analyses separately for employees and for spouses. The results are presented in Table 7 and show that the overall results are driven primarily by employee participants—those choosing the gift card incentive achieve a significantly higher completion percentage than those choosing either the cash or tangible reward incentive. In contrast, and consistent with expectations, spouse participants who select the gift card incentive perform marginally better than those who receive the tangible reward or the cash incentive but only when compared to both groups (Panel B) and not when compared to each group separately.[16] Collectively, these results provide support for our theoretical argument about reward fungibility as a factor in the relationship between reward types and wellness program performance.

6.1.2. Goal completion difficulty

Presslee et al. (2013) find that individuals being rewarded with tangible rewards choose easier goals and perform worse (because their goals don't challenge them to do better). It is possible that participants selecting tangible rewards in our sample performed worse because they selected easier personal goals for the task. We view this as unlikely to be

[16] This result may be explained by a small sample of spouse participants, limiting the power of statistical tests.

Table 7
Comparison of Wellness Program Performance For Employees and Spouses.

Panel A: Descriptive Statistics for % who Completed the Challenge

	Employee				Spouse			
Variable	N	Part.	Mean	Std	N	Part	Mean	Std
Gift Card	1,837	594	86.0%	34.8%	762	261	87.5%	33.1%
Tangible Reward	593	340	80.9%	39.3%	207	130	81.6%	38.8%
Cash	3,608	968	85.8%	34.9%	1,679	427	86.8%	33.9%

Panel B: Statistical Comparison with Full Sample of Conditions (Tangible and Cash are included in the intercept)*

	Employee		Spouse	
Parameter	Estimate	t-value	Estimate	t-value
Intercept	0.340	4.98***	0.543	6.32***
GiftCard	0.031	2.55***	0.027	1.58*
Age	0.002	3.74***	0.002	1.67*
Male	0.048	3.62***	−0.006	−0.26
SamePrize	−0.116	−9.61***	−0.099	−5.53***
PrevChallenges	0.087	8.22***	0.065	4.75***
Challenge Control Variables	Yes		Yes	
Standard errors Clustered by Participant	Yes		Yes	
R-squared	0.086		0.063	
N (Part.)	6,038 (1,290)		2,648 (565)	

*Similar results are observed using a PSM-constructed sample for this analysis.

Panel C: Statistical Comparison of Conditions (GiftCard is included in the intercept)

	Employee		Spouse	
Parameter	Estimate	t-value	Estimate	t-value
Intercept	−0.259	−0.60	0.720	1.28
Tangible	−0.235	−1.52*	−0.263	−1.04
Cash	−0.259	−2.33***	−0.245	−1.49
Age	0.019	3.77***	0.014	1.56
Male	0.405	3.64***	−0.038	−0.2
SamePrize	−1.315	−8.37***	−1.122	−5.03***
PrevChallenges	0.480	8.85***	0.410	5.34***
Challenge Control Variables	Yes		Yes	
Standard errors Clustered by Participant	Yes		Yes	
R-squared	0.143		0.112	
N (Part.)	6,038 (1,290)		2,648 (565)	
F-test of Tangible = Cash	F = 0.03; p = 0.871		F = 0.01; p = 0.939	

***, **, * represent p-values < 0.01, 0.05, and 0.10, respectively. Panel A provides descriptive statistics of the percentage of participants who completed each challenge by reward type and employee status. Panel B and Panel C provides a logistic regression with Complete as the dependent variable and either results for only employees (Panel B) or only spouses (Panel C). All variables are defined in Table 1.

the principal driver of our results for two reasons. First, the correlation between gaining points from exercising and from personal goals is 0.71, suggesting these goals are highly correlated. Since participants cannot choose an easier exercise goal (as that is set by the program), our results are unlikely due to the selection of easier goals. Second, if we only analyze the number of points participants gained from exercising to meet the program-set goal (i.e., if we exclude any points gained from achieving self-set personal goals), we find that gift card participants still perform the best.[17] Thus, it appears that the results cannot solely be attributed to self-set goals of varying difficulty.

6.1.3. Timing differences across reward types

In the setting we examine, cash rewards were paid out on a different timeline (i.e., at the end of the year) than that of the other incentives (which were distributed at the end of each challenge period). One concern about this difference is that the cash incentive may have had reduced motivational power. To examine this possibility, we analyze challenge performance during the first three challenges of the year. The results (untabulated) are consistent with those presented in Panel B of Table 6. That is, the variable *GiftCard* is in the same direction and is statistically significant and the magnitude of the coefficient is similar to the analysis if we look at just the first three periods.[18] The results are similar for Panel C for the variable *Cash,* but the results for *Tangible*, while in the same direction, are no longer statistically significant due to a lack of power.

6.1.4. Testing alternative PSM models

PSM-created samples can differ based on modeling choices. To

[17] We perform this analysis in various ways including using the entire sample without fixed effects, including fixed-effects, and only using the propensity-score matched sample for completion (we cannot create a propensity score matched sample for just exercising since there is not a dichotomous variable to use in the first stage of the analysis). In all of these samples, we find the same results.

[18] We collapse to three challenge periods to have sufficient data for testing the relation. If we collapse just the two first challenge periods, the results are in the same direction and marginally statistically significant (p-value < 0.10). If we look only at the first period, the results are directionally consistent but not statistically significant.

verify that our results are robust to various PSM modeling techniques, we conduct several sensitivity analyses of the results in Tables 5 and 6. First, we re-perform all PSM testing using various caliper distances ranging from 0.01 to 1, and find that the results are similar to those reported (i.e., the key variable of interest is in the same direction and statistically significant at $p < 0.05$).[19] Second, we re-perform the PSM testing using one-to-many matches (i.e., one treatment can match as many other responses that are within the caliper distance), and find that the results are similar to those reported. Third, we re-perform the PSM testing allowing for replacement and find similar results to those reported. Finally, we test various formulations of including different variable combinations in our first-stage model. In all instances, the results are robust to these variations. Importantly, for most iterations the covariate balance was reasonable as assessed by standardized mean differences and variance ratios (Rubin, 2001; Stuart, 2010; Austin, 2011). In a few instances for the results in Table 6, the standard mean difference scores are greater than the absolute value of 0.25 for some covariates; nevertheless, these differences never cause different inferences for the variables of interest and the same variables are not always above the acceptable threshold for all iterations. Taken together, our results appear to be robust to many different constructions of the PSM sample.

6.1.5. Testing using a fixed-effects model

An alternative approach to using regression and clustering the standard errors to control for lack of independence is to analyze a fixed-effects model.[20] To do this, we remove all control variables that do not vary within participant (e.g., age, gender, etc.) and then estimate the model including fixed effects. In untabulated results, we observe the same direction and level of significance for the results in Table 5. For the results in Table 6, the results are in the same direction but are significant at the $p < 0.10$ level for Panel B and not the $p < 0.01$ level. For the results in Panel C of Table 6, *Tangible* is no longer statistically significant but *Cash* is at the $p < 0.05$ level. The difference between *Tangible* and *Cash* remains insignificant. Overall, we interpret our results to be largely robust to this alternative modeling technique.

7. Conclusion

This study examines reward type choices and the associated performance in organizational wellness programs. We analyze data from a natural field setting; thus the choice data in our study represent actual reward choices made by participants, rather than survey or post-experimental expressions of reward preferences. We find that wellness program participants are most likely to choose cash rewards, followed by gift cards and tangible rewards, respectively. However, we find that employee participants who choose to be rewarded with gift cards perform at a higher level than those who choose either of the other reward types, while cash participants and tangible reward participants perform relatively similarly.

Wellness programs represent a unique organizational setting in which to examine reward type associations. First, they formalize efforts to motivate behaviors and outcomes that many employees are likely to want for themselves even in the absence of extrinsic motivations (i.e., improved health and well-being). Second, the reward selection feature of our setting increases the likelihood that each participant is equally motivated (at least extrinsically), having chosen the reward they most prefer. Finally, the program we study is completely voluntary, further increasing the presumed similarity of participant motivation.[21] Despite these features, we find significant variation in performance, consistent with theory that has previously been invoked to predict performance effects in more traditional incentive programs (Jeffrey, 2009; Choi and Presslee, 2017). Our results provide additional and compelling evidence that individual choices are not always congruent with the factors that truly influence motivation and performance.

With respect to wellness programs in particular, our results suggest that firms may benefit from offering reward choice, but that some choices are associated with better performance than others. For employees, tangible rewards and cash are not associated as highly with performance as gift cards. Gift cards appear to offer a balance between fungibility and mental separateness from salary, and it may therefore be in an organization's interest to merely offer a wide selection of gift card options as wellness participation rewards. Indeed, a 2013 survey indicates that 88 percent of respondent companies use gift cards in their non-sales employee incentive programs (Incentive Federation, 2013). It may also be useful for organizations to inform participants about these results in an effort to help them make distal decisions that are actually more aligned with proximal motivations. Participants who learn about the evidence we document here may make different and more optimal decisions at the time of reward selection.

More generally, our results support previous evidence of an incongruence between reward preferences and effectiveness (Jeffrey, 2009). Specifically, our evidence confirms that individuals have strong preferences for economic fungibility when choosing rewards; however, their individual incentive choices are not always associated with higher performance. In light of this, managers may benefit from considering how employees think about (and are motivated to engage in) desired behaviors when selecting rewards versus performing job tasks when writing contracts and designing incentive systems.

The existing literature on incentive reward type and performance has produced a variety of mixed results. Most surveys of participant preferences find that cash is the most preferred reward type, and our results using actual choices confirm that preference. However, the results on the performance effects of different rewards have not demonstrated a consistent pattern. We add to this literature by showing that in our study the benefits and costs of cash versus tangible rewards seem to offset each other, while gift cards seem to offer an optimal balance of fungibility and separateness from salary that results in the highest performance (for employees). Thus we see our results as informing the development of theory in this area (see Choi and Presslee, 2017), as well as encouraging further empirical research.

While our study provides strong external validity, the natural setting, along with certain data restrictions, pose a number of important limitations that present valuable opportunities for future research. For example, the absence of random reward type assignment means that we cannot perfectly disentangle motivational versus self-selection effects (or those arising from other correlated omitted variables) that may be operating in our setting. In addition, there is a timing difference across reward types; cash rewards are received at the end of the year, while tangible and gift card rewards are received at the end of each successful challenge. Thus it is difficult to provide clear evidence about the causal mechanisms underlying our results. Future research could provide

[19] The choice of caliper distances is an arbitrary decision (Austin, 2011). We relied on the findings of Austin (2011) who found that 0.2 was optimal, and that the "MSE was minimized by using calipers that were equal to a width of between 0.20 and 0.55 times the standard deviation of the logit of the propensity score when at least one of the covariates were continuous. Furthermore, the use of calipers of these widths tended to result in confidence intervals with approximately correct coverage rates and significance tests with approximately correct empirical type I error rates." Every iteration of the results we ran using different caliper distances produced similar results.

[20] When estimating the logistic regression results in Table 6, we use conditional maximum likelihood. A full discussion of this choice is available at https://pdfs.semanticscholar.org/84f5/55569662b8c4882b213cd13f75622eaf495e.pdf.

[21] Most wellness programs are voluntary, but not all. Some firms have begun to impose severe non-participation penalties (such as the removal of employee-sponsored health insurance) that render the programs virtually mandatory (Greenfield, 2016).

stronger causal evidence through the benefits of random assignment and more internally valid procedures. Moreover, all wellness challenge behavior and health outcomes are self-reported in our setting. As such, any systematic differences in such reports that are not captured by our control variables and PSM analyses may influence the results we report here. We also note that due to small sample size and lack of data, we are unable to directly test the effects of reward type on health outcomes (e.g., weight, blood pressure, etc). A more direct examination of reward type effects on specific health outcome measures would be a fruitful expansion of this literature. Finally, a worksite wellness program represents a distinctive research setting which may limit the generalizability of our results. Future research will benefit from the consideration of contextual features in the emerging literature on reward type effects across unique and traditional organizational control systems.

References

Alonzo, V., 1996. The trouble with money. Incentive 170 (2), 26–31.

Austin, P.C., 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat. Med. 28, 3083–3107.

Austin, P.C., 2011. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm. Stat. 10 (2), 150–161.

Baicker, K., Cutler, D.D., Song, Z., 2010. Workplace wellness programs can generate savings. Health Aff. 29 (2), 1–8.

Barham, K., West, S., Trief, P., Morrow, C., Wade, M., Weinstock, R.S., 2011. Diabetes prevention and control in the workplace: a pilot project for county employees. J. Public Health Manag. Pract. 17 (3), 233–241.

Berry, L., Mirabito, A.M., Baun, W.B., 2010. What's the hard return on employee wellness programs? Harv. Bus. Rev. 88 (12), 104–112.

Bradler, C., Dur, R., Neckermann, S., Non, A., 2016. Employee recognition and performance: a field experiment. Manage. Sci. 62 (11), 3085–3099.

Brino, A., 2015. Employer Wellness Programs Spend Record $693 Per Worker. (March 26,2015). Available at:. http://www.healthcarepayernews.com/content/employer-wellness-programs-spend-record-693-worker#.VovyH_krIuU.

Butler, K.M., 2014. Wellness program success? It's a long shot. Workforce 93 (3), 17.

Chen, C.X., Nichol, J.E., Zhou, F.H., 2017. The effect of incentive framing and descriptive norms on internal whistleblowing. Contemp. Account. Res. 34 (4), 1757–1778.

Choi, J., Presslee, A., 2017. The Performance Effects of Tangible Versus Cash Rewards: the Mediating Role of Categorization. Working paper. University of Pittsburgh.

Finkelsteing, E.A., Kosa, K.M., 2003. Use of incentives to motivate healthy behaviors among employees. Gender Issues 21 (3), 50–59.

FMR LLC, 2015. Companies Are Spending More on Corporate Wellness Programs but Employees Are Leaving Millions on the Table. (March 26, 2015). Available at:. https://www.businessgrouphealth.org/pressroom/pressRelease.cfm?ID=252.

Gandolfi, F., 2013. Workforce downsizing. J. Manag. Res. 13 (2), 67–76.

Greenfield, R., 2016. Employee Wellness Programs Not so Voluntary Anymore: Take a Blood Test or Lose Your Health Coverage. Bloomberg. http://bloom.bg/1Zp5aRY.

Gubler, T., Larkin, I., Pierce, L., 2018. Doing well by making well: the impact of corporate wellness programs on employee productivity. Manage. Sci. 64 (11), 4967–4987.

Hall, B., 2008. Health incentives: the science and art of motivating healthy behaviors. Benefits Q. 24 (2), 12–22.

Higgins, M., D'Agostino, R., Kannet, W., Cobb, J., 1993. Benefits and adverse effects of weight loss. Ann. Intern. Med. 119 (7), 758–764.

Incentive Federation, 2013. Incentive Market Study: October 2013. Aspect Market Intelligence, Falls Church, VA.

Inside Public Accounting, 2018. PwC Invests $45M in Wellness Perks and New Parents' Benefits. http://blog.insidepublicaccounting.com/2018/04/pwc-invests-45m-in-wellness-perks-and-new-parents-benefits/.

Jakobson, L., 2013. Making wellness work. Incentive 187 (3), 12–19.

Jeffrey, S.A., 2009. Justifiability and the motivational power of tangible noncash incentives. Hum. Perform. 22, 143–155.

Johnson, S.R., 2014. Working out wellness. Mod. Healthc. 44 (21), 16–18.

Kelly, K., Presslee, A., A, Webb, A., 2015. The effects of tangible rewards versus cash rewards in a sales tournament: a field experiment. Account. Rev. 92 (6), 165–185.

Kocakulah, M., Powers, J., 2015. Saving money through wellness programs. Strategic Finance 97 (6), 22–33.

Kosfeld, M., Neckermann, S., 2011. Getting more work for nothing? Symbolic awards and worker performance. Am. Econ. J. Microecon. 86-99.

Lara, A., Yancey, A.K., Tapia-Conye, R., Flores, Y., Kuri-Morales, P., Mistry, R., Subirats, E., McCarthy, W.J., 2008. Pausa para tu salud: Reduction of weight and waistlines by integrating exercise breaks into workplace organizational routine. Prev. Chronic Dis. 5 (1) A12-A12.

List, J.A., Shogren, J.F., 1998. The deadweight loss of christmas: comment. Am. Econ. Rev. 88 (5), 1350–1355.

Maas, V., van Rinsum, M., 2013. How control system design influences performance misreporting. J. Account. Res. 51 (5), 1159–1186.

Niv, N., Cohen, A.N., Hamilton, A., Reist, C., Young, A.S., 2014. Effectiveness of a psychosocial weight management program for individuals with schizophrenia. J. Behav. Health Serv. Res. 41 (3), 370–380.

Odell, P., 2005. Motivating the masses. Promo 18 (10), 39–50.

Otley, D., 2016. The contingency theory of management accounting and control: 1980-2014. Manag. Account. Res. 31, 45–62.

Presslee, A., Vance, T.W., Webb, R.A., 2013. The effects of reward type on employee goal setting, goal commitment, and performance. Account. Rev. 88 (5), 1805–1831.

Racette, S.B., Deusinger, S.S., Inman, C.L., Burlis, T.L., Highstein, G.R., Buskirk, T.D., Steger-May, K., Peterson, L.R., 2009. Worksite opportunities for wellness (WOW): effects on cardiovascular disease risk factors after 1 year. Prev. Med. 49 (2-3), 108–114.

Rubin, D.B., 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv. Outcomes Res. Methodol. 2, 169–188.

Shaffer, V.A., Arkes, H.R., 2009. Preference reversals in evaluations of cash versus noncash incentives. J. Econ. Psychol. 30, 859–872.

Sheehan, J., J.M, Hassard, J., 2000. Redundancies in Chinese state enterprises: a research report. Ind. Relat. 39 (3), 486–501.

Shields, M.D., 1997. Research in management accounting by North Americans in the 1990s. J. Manag. Account. Res. 9, 3–61.

Shipman, J.E., Swanquist, Q.T., Whited, R.L., 2017. Propensity score matching in accounting research. Account. Rev. 92 (1), 213–244.

Sisaye, S., 2005. Teams and management control systems: a synthesis of three organizational development approaches. Leadersh. Organ. Dev. J. 26 (3/4), 172–185.

Stuart, E.A., 2010. Matching methods for causal inference: a review and a look forward. Stat. Sci. 25, 1–21.

Thaler, R., 1985. Mental accounting and consumer choice. Mark. Sci. 4 (3), 199–214.

Thaler, R., 1999. Mental accounting matters. J. Behav. Decis. Mak. 12 (3), 183–206.

Towers, Watson, 2012. 2013. Global Talent Management and Rewards Study. Available at:. https://www.worldatwork.org/waw/adimLink?id=71255.

Valentin, E.K., Allred, A.T., 2012. Giving and getting gift cards. J. Consum. Mark. 29 (4), 271–279.

Verbeeten, F.H.M., 2008. Performance management practices in public sector organizations: impact on performance. Account. Audit. Account. J. 21 (3), 427–454.

Walfogel, J., 1993. The deadweight loss of Christmas. Am. Econ. Rev. 83 (5), 1328–1336.

Walfogel, J., 1996. The deadweight loss of christmas: reply. Am. Econ. Rev. 86 (5), 1306–1308.

White, R.J., 2006. Format Matters in the Mental Accounting of Funds: the Case of Gift Cards and Cash Gifts. Working Paper. University of Chicago.