10.5465/AMBPP.2017.242

SENSITIVITY ANALYSIS ON THE RELATION BETWEEN EXTRINSIC REWARDS AND INTRINSIC MOTIVATION

SHEILA K. LIST
School of Business
Virginia Commonwealth University
Richmond, VA 23284-4000
lists@vcu.edu

SVEN KEPES
MICHAEL A. MCDANIEL
Virginia Commonwealth University

ABSTRACT

Despite conflicting support for Cognitive Evaluation Theory (CET), many still believe that extrinsic rewards undermine intrinsic motivation. Results of our comprehensive sensitivity analysis indicate that most meta-analytic mean estimates provided in previously published meta-analyses may be severely misestimated. This suggests that existing meta-analytic support for CET may be untrustworthy.

INTRODUCTION

The idea that extrinsic incentives undermine intrinsic motivation has been one of the most hotly debated topics in the field of management over the past four decades (e.g., Cerasoli, Nicklin, & Nassrelgrgwai, 2016). This debate started when Deci (1971, 1972) reported that tangible rewards decreased subsequent intrinsic motivation. His findings resulted in the development of Cognitive Evaluation Theory (CET; Deci, 1975; Deci & Ryan, 1985). CET suggests that autonomy and competence are central drivers of intrinsic motivation and can be affected by extrinsic rewards. Thus, depending on how extrinsic rewards such as praise or financial rewards affect an individual's feelings of competence and autonomy, they may either undermine or enhance intrinsic motivation (Deci & Ryan, 1985; Deci, Koestner, & Ryan, 1999).

In the years following the initial findings by Deci (1971, 1972) and the development of CET, several primary studies were conducted to test the predicted relations. These studies provided conflicting results; some finding support for CET (e.g., Smith & Pittman, 1978) and some not (e.g., Rosenfield, Folger, & Adelman, 1980). To resolve the resulting controversies, researchers used meta-analytic techniques (e.g., Cameron & Pierce, 1994; Wiersma, 1992). However, the resulting meta-analyses also had conflicting results and, consequently, did nothing to assuage the controversy. Despite this, Ryan and Deci (2000) as well as others (e.g., Gagné & Deci, 2005) claimed that CET was a supported theory. Furthermore, researchers have continued to cite CET in their theory and hypothesis development (e.g., Dyrlund & Wininger, 2006; Matosic, Cox, Amorose, & Jeffrey, 2014), and insist that providing tangible extrinsic rewards is harmful to individuals' intrinsic motivation (e.g., Frey & Osterloh, 2012). This assertion has even made it into the public consciousness (e.g., Pink, 2009).

The question of whether or not extrinsic rewards affect intrinsic motivation is of vital importance to organizations. First, intrinsic motivation is associated with important outcomes

10.5465/AMBPP.2017.242

such as organizational commitment, job satisfaction, burnout, turnover, and job performance (e.g., Cerasoli, Nicklin, & Ford, 2014; Li, Wang, Pyun, & Kee, 2013; Mathieu & Zajac, 1990; Eby, Freeman, Rush, & Lance, 1999). Second, many organizations use extrinsic rewards to incentivize high performance (Heneman & Werner, 2005). Because such reward structures are very common, if extrinsic rewards do undermine intrinsic motivation, it is possible that these systems have a negative impact on individuals, groups, and organizations.

Our current understanding of the relation between extrinsic rewards and intrinsic motivation rests on the assumption that the cumulative knowledge of these relations is trustworthy. Although meta-analytic reviews are typically considered the best way to gain an understanding of our cumulative knowledge in a literature area because they synthesize the results of primary studies in the area (Garg, Hackman, & Tonelli, 2008; Le, Oh, Shaffer, & Schmidt, 2007), the trustworthiness of much of our scientific literature has recently been called into question by researchers and the popular press (e.g., Banks, Kepes, & McDaniel, 2015; Kepes & McDaniel, 2013, Lehrer, 2010). Because results of past meta-analyses on CET have not assessed the robustness of the results, the scientific trustworthiness of this literature is unknown.

BACKGROUND

There are several reasons why a body of literature may be considered untrustworthy. For instance, although replications are an essential part of the scientific method (Nosek et al., 2015), they are difficult to publish in the fields of management and organizational behavior because of our fields' emphasis on creating new theory (Hambrick, 2007). The lack of replications makes it possible that some effect sizes observed in the literature are due to chance or questionable research practices (QRPs; e.g., Banks et al., 2016; McDaniel, Kepes, Hartman, & List, 2017). QRPs tend to be poor scientific practices (Banks et al., 2016; Bedeian, Taylor, & Miller, 2010; O'Boyle, Banks, & Gonzalez-Mulé, 2017). For instance, HARKing (hypothesizing after results are known; Kerr, 1998), occurs when researchers alter their hypotheses to fit their findings. Bedeian et al. (2010) reported that roughly 92% of the surveyed faculty know someone who had engaged in HARKing. Selective hypothesis reporting, or choosing to remove hypotheses that were not supported from papers prior to publication (e.g., Kepes, Banks, McDaniel, & Whetzel, 2012) is also a very common practice (Banks et al., 2016). Because journals tend to publish studies with statistically significant results (Fanelli, 2012), authors may decide to eliminate hypotheses that were not supported. Furthermore, some studies may never be published because the results are inconsistent with a theory (e.g., CET) and are thus not submitted by the author or the study was rejected by multiple journals. Taken together, the lack of replications, HARKing, selective hypothesis reporting, and non-publishing of studies with "nil" results can each impact the trustworthiness of meta-analytic results because potentially viable data, typically associated with non-significant results (Kepes & McDaniel, 2013), is being suppressed from the publicly available literature.

One hint of potential publication bias in the CET literature is that primary studies included in the meta-analyses assessing the effect of extrinsic rewards on intrinsic motivation are likely to be underpowered. To demonstrate this, assuming a "moderate" effect size of $d = .50$ (which is larger than the estimated meta-analytic effect in the most recent meta-analyses assessing the relation between extrinsic rewards and intrinsic motivation [Cameron, Banko, & Pierce, 2001; Deci et al., 1999]), a power analysis indicates that 102 individuals are necessary to detect the effect (assuming a power of .80 using G*Power version 3.1.9.2 [Faul, Erdfelder,

10.5465/AMBPP.2017.242

Buchner & Lang, 2009]). However, examining the studies included in the largest and most recent meta-analysis (Cameron et al., 2001), fewer than 6% of the included effect sizes were associated with a sample size of 102 or greater. This suggests that the effect sizes from the published studies may overestimate the population effect size because underpowered studies without statistically effect sizes are unlikely to get published in our journals (e.g., Fanelli, 2010). In brief, when the effect sizes of these available underpowered studies are synthesized meta-analytically, the resulting meta-analytic effect size is most likely to be an overestimate of the "true" effect size of the relation between extrinsic rewards and intrinsic motivation (e.g., LeLorier, Gregoire, Benhaddad, Lapierre, & Derderian, 1997; Maxwell, 2004).

To determine the trustworthiness of our cumulative knowledge of the relation between extrinsic rewards and intrinsic motivation, we use a comprehensive battery of sensitivity analyses. Sensitivity analyses involve asking a series of questions aimed at answering the question, "What happens if some aspect of the data or the analysis is changed?" (Greenhouse & Iyengar, 2009: 418). In the meta-analytic context, sensitivity analyses may involve using different meta-analytic methods, assessing the presence and impact of publication bias (PB), and identifying and removing outliers (Greenhouse & Iyengar, 2009). PB occurs when the primary studies included in a meta-analysis are not representative of all studies examining the relation of interest (Banks et al., 2015; Kepes & McDaniel, 2015; Rothstein, Sutton & Borenstein, 2005). Ferguson and Brannick (2012) found that roughly 40% of published meta-analyses in psychology are affected by PB, yet proper assessments of PB are rarely conducted in our literature (e.g., around 4% according to Banks et al., 2012). Outliers also have the potential to affect the accuracy of meta-analytic conclusions by distorting the mean and/or variance of the distribution of effect sizes to be analyzed. They may also affect the accuracy of PB assessments (e.g., Stanley & Doucouliagos, 2014; Terrin et al., 2003). However, only about 3% of meta-analytic studies assess the presence of outliers (Aguinis et al., 2011). As of yet, no studies have assessed the trustworthiness (e.g., the impact of PB and outliers) of the literature examining the relation between extrinsic rewards and intrinsic motivation. In this study, we aim to fill this gap by using outlier detection techniques and assessing the degree of PB in the meta-analyses in this literature area. Furthermore, we aim to determine whether the greater threat to trustworthiness stems from PB or from outliers and whether outliers affect the results of the PB assessments.

METHODS

Meta-Analytic Datasets and Analytic Approach

Meta-analyses assessing the relation between extrinsic rewards and intrinsic motivation were identified by searching online databases (i.e., PsycInfo, ABI/INFORM, Academic Search Complete) using keywords such as *financial reward, incentive, merit pay, meta-analysis, synthesis, quantitative review*, etc. A total of eight relevant meta-analyses were identified, however, due to space considerations, only the two largest and most recent meta-analyses (i.e., Cameron et al., 2001; Deci et al., 1999) were selected for inclusion in this study.

We identified 56 distributions and subdistributions from these studies. However, because distributions with less than 10 effect sizes can suffer from low statistical power (Sterne et al., 2011) and second-order sampling error (Schmidt & Hunter, 2015), we attempted to recreate 40 of the 56 (71%) distributions. Unfortunately, due to lack of clarity in the original meta-analyses, meta-analytic results could not be reproduced for several distributions.

10.5465/AMBPP.2017.242

We used analyses in the Hedges and Olkin's (1985; Hedges & Vevea, 1988) tradition for conducting meta-analyses in this study, as many of the sensitivity analysis techniques have not been adapted for psychometric meta-analyses (Schmidt & Hunter, 2015). For the re-analysis of Cameron et al.'s (2001) meta-analysis, we first converted the provided effect sizes (Hedge's *g*) to Cohen's *d,* using the formulas provided by Hedges and Olkin (1985). All analyses were conducted in *R* using the "metafor" package (Viechtbauer, 2015) and the DL estimation method (DerSimonian & Laird, 1986), with the exception of the selection model analyses, which were conducted using the "meta" package (Schwarzer, 2007).

Our comprehensive battery of sensitivity analyses comprised several methods, including PB methods such as contour-enhanced funnel plots, trim and fill, cumulative meta-analysis by precision, a priori selection models, and PET-PEESE (e.g., Duval, 2005; Kepes et al., 2012; Kepes & McDaniel, 2015; Peters et al., 2008; Stanley & Doucouliagos, 2014; Stanley, Jarrell, & Doucouliagos, 2010). The presence and influence of outliers was assessed using the one-sample removed analysis (osr; Borenstein, Hedges, Higgins, & Rothstein, 2009) and Viechtbauer and Cheung's (2010) comprehensive multivariate influence diagnostics. Detailed explanations of these methods can be found in the references provided.

For each meta-analytic distribution, we first assessed how the meta-analytic results are affected by PB. Then, we removed the outliers identified by Viechtbauer and Cheung's (2010) influence diagnostics. After the identified outliers were removed, if the resulting *k* was 10 or greater, the meta- and PB analyses were re-run. This approach allowed for a comparison of the effects of PB and outliers on meta-analytic results; it also provided information about how outliers affect PB assessments.

To determine the practical impact of any non-robustness, we calculated the baseline range estimate (BRE) and maximum range estimate (MRE) following the guidelines in Kepes and McDaniel (2015). The BRE is the difference between the naïve meta-analytic mean estimate (i.e., the estimate without any adjustments for potential bias; Copas & Shi, 2000) before outlier removal and the sensitivity analysis result that is farthest away from this original naïve meta-analytic mean estimate, either before or after the removal of outliers. The MRE is the difference between the two estimates farthest away from each other, either before or after outlier removal. To calculate the relative difference for the BRE and MRE, the naïve meta-analytic mean estimate before outlier removal was used as the base. For the BRE and MRE, if the relative difference is less than 20%, the practical difference will be considered "negligible." If this relative difference is 20% to 40%, it will be considered "moderate." Relative differences greater than 40% will be considered "large." These ranges are consistent with guidelines provided in Kepes and McDaniel (2015). Overall conclusions regarding the robustness of the results were determined using BRE and MRE. If the two values were not in agreement, the lesser of the two estimates was used to indicate the overall conclusion.

To determine if any bias or non-robustness was due to outliers or PB, a series of decision rules were used. Outliers were considered to be a source of non-robustness if (a) the absolute difference between the original naïve meta-analytic mean prior to the removal of outliers and any of the osr analyses prior to the removal of outliers was greater than 20%, or (b) the absolute difference between the original naïve meta-analytic mean prior to the removal of outliers and any sensitivity analyses after the removal of outliers was greater than 20%. PB was considered to be a source of non-robustness if the absolute difference between the original naïve meta-analytic mean prior to the removal of outliers and results of any sensitivity analyses before or after the removal of outliers was greater than 20%.

10.5465/AMBPP.2017.242

RESULTS

Robustness of Deci et al.'s (1999) Meta-Analytic Results

Based on the guidelines provided in Kepes et al. (2012), none of the meta-analytic results can be considered robust, with all but one distribution indicating a "large" practical difference (our results indicated that the one remaining distribution had a "moderate" practical difference). However, even among those distributions with a "large" practical difference, widely varying degrees of robustness were observed. For instance, although our results indicated the distribution "free-time: all reward: verbal: college students" is potentially only misestimated by up to 42% (.19; BRE) or 53% (.24; MRE), the "self-reported: all reward" distribution is potentially misestimated by up to 3900% (.78; BRE) or 4400% (.88; MRE). With the exception of the "free-time: all reward: verbal" distribution, all BRE and MRE estimates were greater than 40%, indicating that the original naïve meta-analytic mean estimates are likely to be severely non-robust (Kepes et al., 2012; Rothstein et al., 2005). Based on the results, it can also be concluded that both outliers and PB affected the degree of misestimation in the original meta-analytic effect sizes. Of the 19 re-analyzed distributions, 15 were affected by both PB and outliers, whereas four distributions were only affected by PB. Thus, in the meta-analysis by Deci et al. (1999), both PB and outliers impacted the estimates of the effect of various extrinsic rewards on intrinsic motivation.

Robustness of Cameron et al.'s (2001) Meta-Analytic Results

Results of our re-analysis indicated that the naïve meta-analytic mean estimates for all 21 distributions that we re-analyzed appear to be misestimated. For instance, the "free-choice: all rewards: verbal" distribution is potentially misestimated by up to 58% (.19; BRE) or 67% (.22; MRE), which represents a "severe" misestimation (Kepes et al., 2012; Rothstein et al., 2005). The majority of the remaining distributions have even larger BRE and MRE values, indicating that their respective naïve meta-analytic mean effect size estimates are likely to be misestimated by an even greater amount. Based on this assessment, one can conclude that Cameron et al.'s (2001) meta-analytic results are likely to be non-robust. In regards to the source of this non-robustness, 16 of the 21 distributions were affected by the presence of both outliers and PB, while the remaining five distributions were affected only by PB. This suggests that, similar to the Deci et al., (1999) meta-analysis, the results of the Cameron et al. (2001) meta-analysis were affected by both PB and outliers; with PB having the larger adverse effect.

Detailed results are available from the first author.

DISCUSSION

Despite conflicting meta-analytic support for CET (e.g., Cameron et al., 2001; Deci et al., 1999), researchers have continued to argue that providing extrinsic rewards can undermine intrinsic motivation (e.g., Dyrlund & Wininger, 2006; Frey & Osterloh, 2012). Yet, the trustworthiness of some of the literature in management and related fields in the social sciences has recently been called into question (e.g., Banks et al., 2015; Kepes & McDaniel, 2013; McDaniel et al., 2017). To address this, and to determine the validity of CET, two meta-analytic datasets were re-examined using a comprehensive sensitivity analysis approach. Following

10.5465/AMBPP.2017.242

guidelines by Kepes et al. (2012) and Kepes and McDaniel (2015), we calculated the BRE and MRE for each of the distributions. We excluded the estimates of the severe selection models from these calculations because they were consistently more extreme than the other for-PB adjusted estimates and typically nonsensical (see Kepes et al., 2012; Vevea & Woods, 2005). The range estimates suggested that of the 40 originally obtained naïve meta-analytic mean estimates, all may be misestimated by at least a moderate degree (i.e., by 20% or more). In fact, the range estimates suggested that 98% of the distributions' originally derived naïve meta-analytic mean estimates were potentially misestimated by over 40%. This degree of misestimation can be considered severe (Kepes et al., 2012). Taken together, our results indicate that the meta-analytic results regarding the relation between extrinsic rewards and intrinsic motivation are non-robust. We also determined that outliers were a source of non-robustness in 78% of the distributions, while PB was deemed to be a source of non-robustness in 100% distributions. A combined effect of outliers and PB on the original naïve meta-analytic mean estimate was observed in 78% of the distributions. This suggests that PB is the greater source of non-robustness to this literature area. However, the precise effect that PB had on the original naïve meta-analytic mean estimates (i.e. upwardly/downwardly biased) cannot be conclusively determined. Outliers, however, generally had an upwardly biased effect on the original naïve estimate. This suggests that outliers are also responsible for the degree of potential misestimation of naïve meta-analytic effect sizes. Furthermore, our results indicate that, with the exception of the estimate by the five most precise samples, PB assessments are influenced by outlier induced heterogeneity.

Implications for Research and Practice

Our results suggest that the meta-analytic mean estimates of the relation between various extrinsic rewards and intrinsic motivation by two meta-analyses (Cameron et al., 2001; Deci et al., 1999) are untrustworthy. This has several implications. First, it means that authors should be wary about using CET as a theoretical justification for their hypotheses, as CET has, based on our analyses, no trustworthy support. Second, we should make every effort to inform practitioners that there is no trustworthy meta-analytic evidence to suggest that financially rewarding employees will be harmful (as long as the desired behaviors are being rewarded [Kerr, 1975]). Instead, practitioners should listen to their employees, whom consistently state that they prefer merit pay systems and, thus, extrinsic rewards (e.g., Heneman & Werner, 2005). Additionally, the obtained results indicate that outliers affected both the meta-analytic mean estimates and PB results. This suggests that, consistent with recommendations by Kepes and McDaniel (2015), sensitivity analyses should include both the assessment of outliers and PB.

As with all studies, this study has a few limitations. First, we were only able to re-analyze 71% of the total distributions included in the two meta-analyses (the remaining distributions were too small for our re-analysis [Schmidt & Hunter, 2015; Kepes et al., 2012]). However, given that small sample studies tend to produce less robust results than large sample studies, our conclusion would likely stay the same. Furthermore, there are currently no best practices in regards to which PB analysis and outlier detection technique is the most appropriate to use in different situations, therefore, consistent with the triangulation approach (Orlitzky, 2012), we used multiple PB methods.

REFERENCES AVAILABLE FROM THE AUTHORS