7-2013

Rationalizing Organizational Change: A Need for Comparative Testing

Arkadiy Sakhartov
University of Pennsylvania

Timothy B. Folta

Recommended Citation

Sakhartov, A., & Folta, T. B. (2013). Rationalizing Organizational Change: A Need for Comparative Testing. *Organization Science, 24* (4), 1140-1156. http://dx.doi.org/10.1287/orsc.1120.0785

Rationalizing Organizational Change: A Need for Comparative Testing

Abstract
Behavioral theory explains that organizational change is prompted by performance relative to a firm-specific aspiration. Although this explanation has been empirically confirmed, it has not been tested comparatively alongside other explanations, most notably rational choice. This lack of comparative study implies that prior research may be committing Type I errors—confirming aspiration-level decision making when it is not actually occurring. This paper contributes to behavioral theory in two specific ways. First, we show that several foundational studies purporting to provide empirical support for aspiration-level decision making may actually represent maximizing behavior. To consider this potential, we simulate a sample of subjectively rational agents who choose strategies by maximizing expectations. We show that it is possible and highly probable to diagnose satisficing when agents are, in fact, maximizing. Second, we develop and implement recommendations for comparative testing to demonstrate reliability. Analysis shows that the recommendations are effective at reducing Type I and II errors for both behavioral theory and rational choice. This paper is meant to inspire the design of future studies on aspirations and, indeed, all studies of organizational change.

Keywords
behavioral theory, rational choice, aspirations, aspiration levels, expectations, comparative testing

Disciplines
Business Administration, Management, and Operations | Organizational Behavior and Theory

RATIONALIZING ORGANIZATIONAL CHANGE: A NEED FOR COMPARATIVE TESTING

Arkadiy V. Sakhartov
Krannert School of Management
Purdue University
403 West State Street
West Lafayette, IN 47907-2056
asakhart@purdue.edu

and

Timothy B. Folta*
Krannert School of Management
Purdue University
403 West State Street
West Lafayette, IN 47907-2056
foltat@purdue.edu

Draft 3.0

* Corresponding author

We are grateful to Shivaram Devarakonda, Pied Haunsberg, Moren Lévesque, Brian McCann, Kent Miller, Elizabeth Reusch, Kenneth Train, Diemo Urbig, seminar participants at Purdue University, Nick Argyres, and three anonymous reviewers for their valuable comments.

Abstract

Behavioral Theory explains that organizational change is prompted by performance relative to a firm-specific aspiration. While this explanation has been empirically confirmed, it has not been tested comparatively alongside other explanations, most notably rational choice. This lack of comparative study implies that prior research may be committing Type I errors – confirming aspiration-level decision-making when it is not actually occurring. The paper contributes to behavioral theory in two specific ways. First, it is shown that several foundational studies purporting to provide empirical support for aspiration level decision-making may actually represent maximizing behavior. To consider this potential, a sample of subjectively rational agents is simulated who choose strategies by maximizing expectations. It is shown that they can be perceived as satisficing, even when aspirations are absent from the data. Second, recommendations for comparative testing are developed and implemented to demonstrate their reliability. Analysis of simulated data shows the recommendations are effective at reducing Type I and Type II errors for both behavioral theory and rational choice. The paper has strong implications for the design of future studies on aspirations, and indeed, all studies of organizational change.

1. Introduction

It has been frequently argued that managers *satisfice* in their decision-making. In contrast to most models of rational choice where *the best* alternative is selected, Cyert and March (1992: 134) emphasize that their theory of organizational decision-making assumes:

> … an approximate sequential consideration of alternatives. The first satisfactory alternative evoked is accepted . . .

where firm-specific performance aspirations determine what is "satisfactory". This view suggests firms use aspirations to dictate change behavior, and a growing body of empirical work claims to find evidence of this theory. However, the vast majority of confirmations fail to consider alternative explanations. The danger, of course, is Type I error, also known as a false positive, which occurs when a theory is confirmed yet is not true.[1] To avoid this grave danger, the method of multiple working hypotheses is advocated, where theories are tested comparatively.

[1] The rate of Type I error is usually denoted by the Greek letter α, and usually equals the significance level of a test.

Comparative testing refers to the systematic consideration of alternative explanations to a phenomena, as Kuhn (1970: 146) emphasizes: ". . . confirmation or verification is not a relation between a theory and evidence, but a process of selection from amongst rival candidates," and as Stinchcombe (1968:20) argues, stating that a *strong* test of theory requires researchers to "consider the alternative theories which might be explanations" for the observed phenomena. When implemented rigorously, comparative testing disciplines researchers not to unduly fasten their affections (Chamberlin, 1931; Kuhn, 1970). While multiple theories of organizational change exist beyond behavioral theory, one obvious alternative to managers' *satisficing around aspirations* is rationally *maximizing around expectations.*

One should not be dissuaded from comparative testing even though "it is difficult to differentiate maximizing from satisficing" because "most decisions are interpretable in either way" (March and Heath,1994: 20-21). In this sense, this work is different from Lant (1992: 641) who is interpreted as being unconcerned about her conclusion that "there may be a *close relationship* (emphasis added) between expectations and aspirations." If there is a likelihood of collinearity between the constructs then it is premature to assume that either perspective is well-supported unless empiricists simultaneously test both hypotheses. Without comparatively testing one explanation against the other, it is unclear whether one perspective dominates the other, or if both add explanatory power. A review suggests that "a sizeable empirical research stream" (Argote and Greve, 2007: 343) on aspirations ignores comparative testing against the expectation hypothesis. This disregard is clearly at odds with philosophies of Kuhn (1970), Lakatos (1974), and Stinchcombe (1968), who emphasize the power in being comparative.[2] It also contrasts with

[2] Some schools of philosophy disagree on the necessity of comparative testing. For example, relativists believe alternative theories can peacefully co-exist if they describe a phenomenon equally well, but have different causal mechanisms. Lakatos (1974) has criticized this position because it requires certainty about the independence of those explanations.

the founders of the satisficing view (Cyert and March, 1963; Simon 1997b) who recognize theoretical uniqueness hinges on whether empiricists can discern if maximizing or satisficing best describe managerial decision-making.[3]

In response to this premise, this study contributes to the literature in two important ways. First, it is demonstrated that several foundational studies purporting to provide empirical support for aspiration level decision-making--those of Lant and Montgomery (1987), Lant (1992), Greve, (1998), and Mezias et al. (2002)--may actually represent maximizing behavior. To consider this potential, simulation is used to develop a sample of agents who choose strategies by maximizing expectations. It is shown that the agents may be perceived as satisficing, even when aspirations are absent from the data. A plethora of robustness tests, confirm it is remarkably easy to confirm satisficing behavior when the true causality is rational choice. These findings clearly do not invalidate prior research, but strongly imply that empiricists who study organizational change through the lens of aspirations need more explicitly to rule out maximization as an alternative explanation for their findings.

Second, recommendations for comparative testing are offered and implemented to demonstrate their reliability. Again, simulation is valuable because the true causality is known, making it possible to ascertain whether the recommendations are effective at the elimination of Type I and Type II errors for both behavioral theory and rational choice. To accomplish this, the recommendations are examined across three scenarios--when causality is maximizing, when it is satisficing, and when a mixture of maximizing and satisficing is occurring. While this focus on simulated data does not illuminate evidence in favor of any one theory, it does demonstrate

[3] Simon (1997b: 22) argues ". . . empirical evidence . . . would determine which theory is the correct one." In relation to the behavioral theory they develop, Cyert and March (1992: 177) explain "the adequacy of the theory is tested by using it as a basis for decisions and comparing the result with the result obtained from decisions derived from alternatives rules."

sophisticated techniques built soundly upon theory which then can be practically implemented for comparative testing in non-simulated, real world contexts. This final effort provides a roadmap for future research that seeks to understand the boundary conditions of both theories.

To be clear, this paper is not challenging the theoretical uniqueness of behavioral theory. Rather, theoretical uniqueness is assumed, and the focus is on empirical discrimination. However, it is important to emphasize that the theoretical implications of comparative testing are consequential. Chamberlin (1931: 163-164) advocates several benefits to comparative testing beyond empirical discrimination. They "whet the discriminative edge of each" theory, which tends to sharpen the analytic process by specifying more closely the criteria differentiating the hypotheses. They promote "fertility" in research processes, as "each hypothesis suggests its own criteria, its own means of proof, its own method of developing the truth; and if a group of hypotheses encompass the subject on all sides, the total outcome of means and of methods is full and rich." Finally, they urge discipline in not emotionally fixing one's attention or devotion unduly on a single theoretical explanation. Thus, comparatively testing whether change is guided by satisficing or maximizing behavior will amplify and clarify the scope of each hypothesis. While some find offensive our targeting of satisficing behavior, this focus is consistent with the aim of contributing to a target audience: scholars who research organizations.

2. Discriminating Between Alternative Rationalizations of Organizational Change

Argote and Greve (2007: 343) argue that the theory in Cyert and March (1963):

> most directly predicts organizational search and change. Problemistic search implies that organizational aspiration levels adapt to the past experience of the focal organization and those of

> comparable organizations. Once organizational performance falls below the aspiration level, search for solutions will occur and organizational changes become more likely.

This theory has become a leading theory to explain organizational change and its empirical predictions have received "good support" (Argote and Greve, 2007: 343).[4] The literature yields findings nearly universally consistent with the theoretical propositions of behavioral theory (BT) and suggests that past performance, past aspirations, and past performance of comparable firms determine their aspirations; furthermore, organizational change is predicted by the relationship of performance to historical or social aspirations. Moreover, evidence suggests the likelihood of change is not symmetric around the deviation of performance from aspirations—performance above aspirations induces a greater likelihood of change.

A meaningful test of BT, and specifically the role of aspirations in organizational change, requires that it be considered against alternative explanations of the phenomena. The most likely alternative is that organizational change is guided by rational choice (RC)—firms viewed as unitary actors selecting alternatives that maximize expected utility.[5] Comparative testing between BT and RC is not a trivial exercise, partly because considerable theoretical overlap may exist between them. Simon (1978: 503) emphasizes that "In long-run equilibrium it might even be the case that choice with dynamically adapting aspiration levels would be equivalent to optimal choice, taking the costs of search into account." Schwartz et al. (2002: 1178) argue "A satisficer thus often moves in the direction of maximization without ever having it as a deliberate

[4] Reviews of empirical studies on organizational aspirations can be found in Greve (2003) and Nickel and Rodriguez (2002).

[5] It is doubtful that anyone would disagree with this claim. Cyert and March (1992) criticize the RC approach, and explicitly challenge the maximization hypothesis. That the challenge is without potential for compromise is manifested in two ideas they express: (1) an assumption of "an approximate sequential consideration of alternatives" (p. 134) rather than a simultaneous consideration of multiple alternatives, and (2) their conclusion that decision processes are "dominated in large part by non-expectational factors" (p. 75). Note that the decision heuristic described by Cyert and March (1963) differs from that offered by Simon (1997a) and March and Simon (1959). As indicated, an entire body of theoretical and empirical work has followed with the decision heuristic developed by Cyert and March (1963), so it is this heuristic on which this paper is focused.

goal." Cyert and DeGroot (1974: 522) raise concerns about the theoretical and empirical uniqueness of expectations from social aspirations. "Because formulating the alternative theories and deriving their consequences is preeminently a theoretical task," (Stinchcombe, 1968: 28) it is important to clarify what constitutes maximizing behavior in RC and satisficing behavior in BT. Subsequent to the clarification is an evaluation of how the empirical literature distinguishes between satisficing and maximizing.

2.1 Subjective Rationality as an Alternative to Behavioral Theory

The decision heuristic common to every characterization of RC is maximization – agents prefer the alternative providing the highest utility, and will change to that alternative from status quo if doing so provides a higher expected utility. Satisficers are unconcerned about comparing utilities across alternatives. They focus on whether current performance is satisfactory relative to aspired performance, and if it is not they change to the first available alternative. This paper emphasizes the need for comparative testing of whether these heuristics lead to fundamentally different outcomes. As described in the previous paragraph, it is not obvious that they should.

The challenge of comparatively testing RC and BT is enhanced by our inability to know with certainty how expectations or aspirations are determined by agents. It is therefore necessary to make instrumental assumptions about the formation of expectations and aspirations. For example, Simon (1997a: 84) recognizes that we might assume that agents are *objectively rational* in that they maximize based on complete knowledge and perfect processing capabilities, or we might assume them to be *subjectively rational* by maximizing attainment relative to their actual knowledge.[6] Similarly, we might assume what performance is deemed "satisfactory" to be based on a complete history of performance or weighted more heavily on recent histories, or it might

[6] Objectively rational agents might also consider of their rival's likely responses, as epitomized in game theory.

be based on historical or social comparisons. Since it is necessary to make instrumental assumptions to implement a comparative test, one should be careful to clearly denote the assumptions underlying their characterization of the comparative theories to demarcate the applicability of their test. We choose to implement comparative tests between BT and subjective rationality, as defined by Simon (1997a), and consistent with Arrow (1951) and Sen (1997).[7] This choice was made partly because it is preferable to comparatively test a more general interpretation of theory (Stinchcombe, 1968), and partly because of the extraordinary computational complexities associated with imparting perfect knowledge and processing capabilities on a sample of numerous and heterogeneous agents. It is essential to emphasize that subjectively rational agents maximize with the information they possess – they do not intentionally satisfice. The specific ways in which subjectively rational behavior and satisficing behavior are instrumented is described more carefully in later sections.

2.2 An Absence of Comparative Tests Between Satisficing and Maximizing

Graham Allison (1969) was one of the first to explicitly compare how organizational change might be driven by either changes in expectations or changes in performance relative to aspirations. He emphasized how BT and RC differentially explain events around the Cuban Missile Crisis, but seemed uninterested in comparatively testing which theory best explained the events.[8] There is no evidence of any serious challenges to aspiration-based reasoning with RC reasoning, such as would be the case in the presence of comparative testing. Some work has

[7] Simon (1997a: 84) specifically refers to a decision being "subjectively rational if it maximizes attainment relative to the actual knowledge of the subject." Arrow (1951: 406) described "rational behavior simply means behavior in accordance with some ordering of alternatives in terms of relative desirability," and Sen (1997: 746) suggests that maximization "only requires choosing an alternative that is not judged to be worse than any other." Note that these three Nobel laureates place maximization at the center of their descriptions of rationality. There is no requirement that agents have complete knowledge.

[8] It is noteworthy that Allison (1969: 694) also recognized that maximization was the fundamental RC proposition: "An increase [/a decrease] in the cost of an alternative, i.e., a reduction [/an increase] in the value of the set of consequences which will follow from that action, or a reduction [/an increase] in the probability of attaining fixed consequences, reduces [/increases] the likelihood of that alternative being chosen."

acknowledged the potential for RC to explain the observed outcome, but these studies fall considerably short of a comparative test. For example, Lant and Montgomery (1987), Greve (1998), and Iyer and Miller (2008), admit organizational change can be described through the lens of RC, but they ultimately reason it is unlikely for their empirical contexts (*Markstrat* simulations, radio broadcasting, and acquisitions, respectively). Pertinent is the advice of Camerer and Weber (1999: 62), who explain that comparative tests are most effectively made "with data that clearly rule out an alternative hypothesis than with mere reason and argument." The most direct attempt to consider RC alongside BT is the work of Lant (1992). She was interested in determining whether sales targets in the *Markstrat* game are better predicted by rational expectations or by attainment discrepancy (performance relative to aspirations). She considered two attainment discrepancy models: that of Lewin et al. (1944), which focuses on aspirations of individuals, and that of Levinthal and March (1981), which characterizes organizational aspirations. Lant (1992) found that the model of Lewin et al. (1944) predicted sales targets better than the model of Levinthal and March (1981) or the RC model. Interestingly, she did find that under certain conditions, the attainment discrepancy models and the RC model converged in their predictive ability, suggesting that, under those circumstances, a strong correlation may exist between aspirations and expectations. Despite her contributions, for three reasons this work fails to conduct a comparative test between BT and RC. First, it does not test the key proposition that attainment discrepancy predicts organizational change, and instead focuses on predicting sales targets because they are believed to represent aspirations. Second, if whether or not sales targets actually represent aspirations or expectations is ambiguous, then the finding that attainment discrepancy models best fit the data does not necessarily mean that sales targets represent aspirations. The correct conclusion ultimately hinges on how accurately the models characterize the theory. Lant (1992) implicitly adopted an assumption that expectations could not be adapted based on performance relative to prior expectations. While scholars sometimes disallow adaptive expectations to facilitate model convergence, it is important to recognize that adaptive expectations might easily be instrumented in RC. Nobel laureate Kenneth Arrow (1971: 171) states that the "observer of the outcome of an activity can be supposed to form new probability judgements," and Cyert and DeGroot (1974) provide one example of how expectations are adapted by new information. Surprisingly, Lant (1992:627) recognizes that "organization theorists have adopted models of adaptive expectations from the economic literature," but she does not incorporate such adaptation in her own test of RC. Finally, while her *Markstrat* players have two alternative brands around which she predicts sales targets, she estimates separate models for each brand. By doing so, her estimation treats the sales targets of the brands as independent of one another, a conclusion in clear opposition of the maximization behavior based on comparative expectations. Consequently, the participants of the game might have maximized in their choice over relative quantities of the two goods.

In light of Lant's (1992) finding that predictions using aspirations and expectations may converge, the fact that no paper explicitly rules out maximizing behavior is surprising. The closest attempt at doing so seems to be the work of Iyer and Miller (2008) who study how aspirations affect the hazard of acquisition. They include the acquirer's market-to-book ratio to proxy for its growth options, which might be interpreted as an attempt to control for available alternatives. Curiously, their finding that the probability of acquisition decreases with performance below aspirations contradicts the view that change behavior is driven by problemistic search, advocated by BT.

To summarize, despite recognition that BT and RC are competing explanations for change, the empirical literature fails to explicitly compare BT against its most probable theoretical competitor (RC). This oversight raises concern as to whether confirmations of aspiration-level decision-making are false positive confirmations, also known as Type I errors, resulting from a spurious relationship where the true causality may be maximizing. In Section 3 this concern is empirically examined. For this purpose, simulation has one key advantage relative to studies of real data–the true causality is known because it is determined *ex ante*.

3. Is it Possible to Diagnose Satisficing When Maximizing is Occurring?

This section investigates the potential for committing a Type I error by falsely confirming BT when it is not true. We do so by:

- simulating a sample of subjectively rational agents (firms) that maximize in their decision-making by estimating performance expectations for each alternative;
- inferring aspirations from the subjectively rational agents–it is common to empirically derive aspirations and identify them to be behavioral, but when using these same methods to identify aspirations it is known there is nothing behavioral about them because agents are predetermined to be maximizers;
- fitting the simulated data to models purporting to provide foundational support for aspiration-level decision-making to observe whether Type I errors obtain in the simulated data, indicating that BT is confirmed when the true causality is RC; and
- ascertaining how robust Type I errors appear in the simulated data.

This process reveals that Type I errors in research confirming BT may be consequential. Section 4 develops recommendations for overcoming Type I and Type II errors in BT and RC.

3.1. Simulating a Sample of Subjectively Rational Agents Who Maximize

Simulation is used to predict *change* among a population of agents. *Change* is simulated through rational choice, implemented via *expected utility maximization* as specified by Arrow (1971) and Pratt (1964), where utility is a function of wealth. One redeeming quality of this approach is that enables quantification of utilities across a wide range of risk attitudes. Since expected outcomes matter differently to agents with different risk attitudes, quantification is valuable for ascertaining change behavior. One key element of the model is agents are randomly endowed with one of five risk attitudes as noted in Appendix A, three types of risk averse utilities and two types of risk seeking utilities. The other two key elements to the model are the calculation of actual and expected utilities. Expected utilities are important since they dictate strategy choice.

3.1.1 Calculating Actual Utilities

In the model, each agent i gets utility, U_{ijt}, for participating in strategy j at time t, where

$$U_{ijt} = U(W_{ijt}) = U(W_{it-1} - S + I_{ijt}) \quad (1),$$

indicating that utility is a function of wealth, W_{ijt}. W_{ijt} is determined by wealth at the beginning of a period (W_{it-1}), the cost of switching strategies (S) when a switch happens, and an agent's income under a strategy (I_{ijt}). All agents are assigned zero wealth in the first period. S is randomly drawn from the distribution $\mathcal{U}(0, 0.01)$ for each run of the simulation, but within a run is invariant across time or strategy. I_{ijt} is determined by random variables D_{1t}, D_{2t}, D_{3t}, D_{4t}, and D_{5t} representing the population's income for each of the five strategies. D_{jt} is shared equally by all n_{jt} agents choosing the strategy based on expectations (equation 2 below), such that I_{ijt}=I_{jt}= D_{jt}/n_{jt}. In period 1, each D_{j1} is randomly drawn from the uniform distribution $\mathcal{U}(0, 5)$. Every period thereafter, each D_{jt} is adjusted by uncertainty δ_{jt}, drawn for each strategy from distribution $\mathcal{U}(-\lambda, \lambda)$, and λ is a parameter drawn once in the beginning of each run from distribution $\mathcal{U}(0,10)$.

Thus, I_{ijt} is determined by the initial draw for a strategy (D_{j1}), the exogenous uncertainty for a strategy each period (δ_{jt}), and the number of agents choosing a strategy (n_{jt}). If agents were risk neutral, U_{ijt} would be a linear function of W_{ijt}. However, since they have different risk attitudes, they calculate U_{ijt} by different formulas, specified in Appendix A.

3.1.2 Calculating Expected Utilities

Since agents make choices before U_{ijt}'s are observed, change is determined by their expectations, $E[U_{ijt}]$'s.[9] In the first period ($t=1$), strategies are randomly assigned, but for each $t>1$, agents choose the strategy, j^*_t, providing the highest expected utility:

$$\text{Choose } j^*_t,\ where \quad E[U_{ij^*t}] = \max\{E[U_{ij(1)t}];E[U_{ij(2)t}];E[U_{ij(3)t}];E[U_{ij(4)t}];E[U_{ij(5)t}]\} \qquad (2).$$

Since W_{it-1} and S are invariant to strategy choice, the primary variable affecting expected utility is the income for each strategy, I_{ijt}. Agents cannot perfectly anticipate I_{ijt} because of exogenous uncertainty (δ_{jt}) and their inability to observe n_{jt} until the end of the period. Instead, they estimate a discrete probability distribution for each I_{ijt} to use in the calculation of $E[U_{ijt}]$'s.[10] Specifically, for each strategy j and period t, agents calculate $E[U_{ijt}]$'s by:

a) calculating the probability of achieving income, by observing all prior incomes in strategy j that fall within discrete intervals s $(s=1,\ldots,12)$, and estimating the historical frequency (p_{jts}) that income falls within discrete interval s (Note that since all agents have the same information, they form the same probabilities, p_{jts}'s);

[9] "Expectations" refer to the expected value of a random variable, defined in probability theory as the integral of the random variable with respect to its probability measure.

[10] The model is too complex to represent the cumulative distribution analytically. It is common to represent the cumulative discretely when analytical solutions are infeasible. The choice of twelve discrete intervals was the authors, and involves a tradeoff between precision and computation intensity. Each interval has an identical range, and the complete range over all intervals is the same for all strategies and determined by observing the range of all I_{ijt} across all prior periods.

b) calculating utility under the midpoint incomes (I_{ijts}) of each interval s, such that $U_{ijts} = W_{ijts} = W_{it-1} - S + I_{ijts}$, and based on formulas specified in Appendix A;

c) multiplying p_{jts} and U_{ijts} for each s; and

d) summing the twelve products derived in (c).

The heuristic dictates that *change* occurs when $E[U_{ij^*t}] > E[U_{ij^*t-1}]$, unless agents are constrained because $W_{it-1}<S$, where they choose strategy $j_t=j^*_{t-1}$. It is entirely consistent with the principle of maximization underlying choice for subjectively rational agents, whose actual knowledge we specify. Satisficing behavior is not present.

The simulation is run 200 times, each involving 100 agents over thirty periods, resulting in 600,000 observations.[11] 200 runs were estimated to eliminate the consequences of random assignments of initial parameters. Table 1 illustrates one agent's decisions.

(Insert Table 1 about here)

3.2 Inferring "Aspirations" from the Sample of Maximizers

Even though satisficing behavior is absent from the sample, historical and social aspirations can be derived using methods common to BT. To infer historical aspirations (A) for each agent, the model of Levinthal and March (1981) is used:

$$A_{it} = \alpha P_{it-1} + (1-\alpha) A_{it-1} \quad (3),$$

where α is a weight coefficient, $0 \leq \alpha \leq 1$, dictating the emphasis placed on an agent's prior performance, P_{it-1}, versus aspirations in the prior period, A_{it-1}. In the simulated model, P_{it-1} is

[11] All agents persist throughout the 30 periods because the benchmarks experiments of Lant (1992) and Lant and Montgomery (1987) do not allow for entry and exit of subjects.

equivalent to I_{ij^*t-1}, the prior income under the optimal strategy j^*. Like in Greve (1998), the choice of α was determined by which value provided the best model fit, which was $\alpha=0.5$.

To infer social aspirations (B), Greve's (1998) is used:

$$B_{mt} = \frac{\sum_{1}^{k_m} P_{it-1}}{k_m} \tag{4}$$

B_{mt} represents the mean income of all k_m agents in social reference group m in period $t-1$, where m is dictated by risk attitude.[12] Again, P_{it-1} is the income derived in the prior period, I_{ij^*t-1}.

The derivation of historical (A) and social (B) aspirations from the simulated data is helpful for several reasons. First, one can be completely confident that the derived aspirations are not behavioral because agents are predetermined to be rational. Second, since they were derived from methods common to BT implies that it is possible to infer aspirations in real data that are merely a manifestation of maximizing behavior. Indeed, this represents the concern about committing Type I error when falsely confirming BT. Third, it is now possible to test how potent this concern might be by using the data to replicate prior research confirming support for BT. If replications using aspirations derived from the simulation create similar results revealed in these foundational empirical analyses, then the concern about Type I error in these prior studies is accentuated, underscoring the imperative necessity of comparative testing.

[12] Festinger (1954) developed the concept of social reference groups for individuals, not organizations. The behavioral literature has adapted the concept to organizations and uses different criteria for identifying social reference groups, such as firm size or the 2-digit SIC code. In the simulated data, risk attitude represents the most pertinent basis for social comparisons.

3.3 Fitting the Simulated Data to Foundational Models Purporting to Support Aspiration-based Decision-Making

Concerns about spurious support for BT emanate from (a) the correlation between expectations and aspirations, and (b) the potential for expectations, and not aspirations, to be the true causal factor. Historical aspirations (A) and social aspirations (B) are correlated ($p<0.001$) with expectations ($E[U_{ij*t}]$) at 0.33 and 0.19, respectively.[13] Correlations in real data might be equally high because both aspirations and expectations derive from past performance. However, these constructs are unique in that historical aspirations are based on an agent's own history of performance, social aspirations are based on the immediate past performance of related others, and expectations are based on maximization over all historical performance of all firms across alternative strategies. Correlation does not imply causation.

The first examination into whether expectations might be the true causal factor underlying studies confirming BT involves a replication of three studies that have investigated aspiration formation in real data. In contrast to the empirically derived aspirations of Levinthal and March (1981) and Greve (1998), Lant (1992), Lant and Montgomery (1987), and Mezias et al. (2002) all assume that aspirations are represented by sales targets (T) specified in questionnaires, and predict them in the following way:

$$T_{it} = \beta_0 + \beta_1 T_{it-1} + \beta_2 (P_{it-1} - T_{it-1}) + \beta_3 (P_{it-1} - B_{it-1}) + e_{it} \quad (5),$$

where P_{it-1} is an agent's performance in the prior period, ($P_{it-1}–T_{it-1}$) is an agent's attainment discrepancy relative to historical aspirations, and ($P_{it-1}–B_{it-1}$) is attainment discrepancy relative to social aspirations. Lant (1992) and Lant and Montgomery (1987) assume $\beta_3=0$, implying

[13] A and B are correlated at 0.75, which is similar to the 0.65 correlation Greve (1998) reported for attainment discrepancies involving historical and social aspiration. Unreported analysis suggests that significant correlations between aspirations and expectations are robust across time periods and in Spearman and Kendall Tau rank tests.

aspirations are not determined by social comparison, but only historical aspirations. The three studies claim confirmation that sales targets represent aspirations because in each case T_{it} is positively affected by T_{it-1} and negatively affected by attainment discrepancy. However, the confirmations are spurious if sales targets truly represent expectations.

While it may never be known whether sales targets are expectations or aspirations, the true causality is known in the simulated data, enabling an assessment of whether expectations might be predicted in the same way as aspirations. The data are fit to the models used in those prior studies, except that T_{it} is replaced with $E[I_{ij^*t}]$ and P_{it-1} is represented by I_{ij^*t-1}.

(Insert Table 2 about here)

Table 2 illustrates the results alongside those of the three papers noted above. Note that each replicated model produces coefficients that are significant and identical in direction. These results show complete convergence–the same factors that have influenced sales targets in prior studies influenced expected income in this study. This result has several implications. First, sales targets may represent expectations and not aspirations. While some may dismiss this logic because expectations may derive from aspirations, in the simulated data it is certain that expectations do not emanate from aspirations. Second, prior support for the effect of social comparison may be spurious, because an identical result is replicated when social aspirations do not exist in the data. Third, aspiration formation model tested in prior research may be too accommodating and unable to discriminate between expectations and aspirations.

A much less safe BT model is replicated in the second examination of whether expectations might be the true causal factor underlying studies confirming BT. Greve's (1998) study of organizational change offers a more specific prediction--that the negative relationship for attainment discrepancy is stronger when a firm performs above its aspiration than when it performs below its aspiration.[14] Such a prediction might have been lauded by Popper (1963: 36) who advocated that "confirmations should count only if they are the result of risky predictions" rather than "safe" ones. The fact that Greve (1998) found support for his hypotheses is compelling evidence in favor of BT. He did not, however, consider that expectations might be driving his result. His pooled logit model is replicated with the simulated data:

$$\log \frac{\Pr(Y_{it}=1)}{1-\Pr(Y_{it}=1)} = \tau_0 + \tau_1[I_{it-1} - A_{it-1}]_{>0} + \tau_2[I_{it-1} - A_{it-1}]_{<0} + \tau_3[I_{it-1} - B_{it-1}]_{>0} + \tau_4[I_{it-1} - B_{it-1}]_{<0} + \tau_5 X_{i5} + \ldots + \tau_n X_{in} + \varepsilon \quad (6),$$

where $Y_{it}=1$ if a strategy change occurred for agent i and $Y_{it}=0$ otherwise; A and B represent historical and social aspirations, respectively; P_{it-1} is again represented by I_{ij^*t-1}; N_{it-1} is a negative inconsistency variable explicating the cases when performance falls between the two aspirations; and X_{in} represents the remaining vector of variables, which are ignored because of the controlled environment. Greve's (1998) two main hypotheses predict that $\tau_1, \tau_2, \tau_3,$ and τ_4 should all be negative, and $\tau_1<\tau_2$ and $\tau_3<\tau_4$.

(Insert Table 3 about here)

Table 3 shows results alongside those of Greve (1998). Note that these results mirror Greve's (1998), confirming the risky prediction noted immediately above. The main difference is that Greve (1998) estimated his model on real data in the radio broadcasting industry and cannot be sure of causality, while there is certainty that satisficing is not causing this result. Again, it seems that prior empirical support for BT may be subject to Type I error.

3.4 Ascertaining how robust Type I errors appear in the simulated data

To assure that the results are not endemic to the original specifications, a plethora of robustness tests were undertaken. The replications in Tables 2 and 3 demonstrated remarkable robustness:

---

[14] See Appendix B for Greve's hypotheses.

(a) after removing the budget constraint ($W_{ijt}<S$), because some may argue that capital is readily available; (b) across ten deciles of switching costs (S); (c) across five different risk attitudes (m); (d) across ten deciles of uncertainty (λ); and (e) across all values of the weight coefficient (α) for aspiration formation. Replications were confirmed across many of these robustness checks. For example, the signs of coefficients were robust, as was the replication of Lant and Montgomery (1987) and Lant (1989). The replication of Mezias et al. (2002) was confirmed across all specifications for coefficients related to aspirations and attainment discrepancy, although the coefficient for social comparison was not supported in a few specifications. In almost every case, Greve's (1998) hypotheses for a negative relationship for attainment discrepancy are confirmed. While failing to always simultaneously confirm his hypotheses 1B and 2B, either hypothesis 1B or 2B (see Table 9) are always confirmed. Finally, to consider whether the confirmations were due to the fact that five strategies were considered, the data were re-simulated with alternative numbers of strategies (i.e., 1-10, 30, 100), and tests continue to commit Type I error. These robustness checks demonstrate that, across a full array of model specifications, it is possible to diagnose aspirations when they are absent. These results should be clear evidence that comparative testing is needed to rule out alternative explanations and raise suspicion about prior support for BT.

4.0 Recommendations for Comparative Testing

This section offers recommendations for scholars interested in designing a study suitable for comparative testing between RC and BT. These recommendations focus on operationalizing the unique theoretical constructs and evaluating statistical tests capable of comparative testing. More generally, distinguishing between BT and RC is challenging because it is not desirable to reduce the risk of Type I error while excessively enhancing the risk of Type II error--where BT (or RC) is falsely rejected when it is present. For this reason, it is necessary to examine the recommendations across different contexts–where the investigated causality (e.g., BT) is present, and where it is not. Specifically, three plausible scenarios in an empirical setting are investigated: true causality is pure maximizing behavior (as explained in section 3.1); true causality is pure satisficing behavior (explained in Appendix B); and true causality is a mixture of maximizing and satisficing behavior (explained in Appendix C, 60% of agents are satisficers and 40% are maximizers). This third scenario is invoked because it may be empirically plausible, and it enables a check for robustness in the most challenging empirical scenario–the context of mixed behavior. Table 4 clarifies the types of errors that might exist across these different contexts. Given that the true causality is known in each context, it is possible to test the effectiveness of the recommendations in simultaneously coping with Type I and II error. Since larger samples are the only way to simultaneously reduce the risks of both errors (Neter, Wasserman, and Whitmore, 1988), the effectiveness of our recommendations is examined on samples with nearly 600,000 observations. In practice, however, large samples may not always be feasible, so we report how robust our recommendations are in smaller samples.

(Insert Table 4 about here)

Recommendations are presented in increasing order of technical sophistication. The recommendation researchers choose and adopt is best guided by their relative advantages and disadvantages, discussed below.

4.1 Recommendation 1: Deliberate Controls

The first recommendation follows directly from Stinchcombe (1968), who recommends the deliberate control of the value of possible spurious variables. Critical to this recommendation are accurate operationalizations of maximizing/satisficing behavior and appropriate modeling techniques.

4.1.1 Operationalizing Maximization

Section 3 has already detailed Greve's (1989) operationalization of satisficing behavior. March and Heath (1994: 20) advocate accounting for maximizing behavior through "the relative position of alternatives." This statement is interpreted to refer to the relative expected payoff to alternatives, which can be estimated in many cases. If alternatives can be enumerated, along with a reliable proxy for their expected payoffs, it is possible to identify whether firms are choosing strategies with the highest expected outcome (i.e., maximizing). For example, when studying whether attainment discrepancy affects the decision to acquire a target, as was done by Iyer and Miller (2008), expected payoffs for acquiring alternative targets can be estimated with indicators readily available in Compustat or CRSP, such as the price-to-earnings multiple, which DePamphlis (2003) suggested bidders actually use.[15] When studying format changes in the radio broadcasting industry, Greve (1998) might have been able to estimate the "relative position of alternatives" through a comparison of expected sizes of target audiences for alternative formats.

One way to the measure the relative value of alternatives is the Fechner Index (Fechner, 1966), which is the standardized difference between the expected utilities of two alternatives:

$$\text{Fechner Index} = \frac{E[U_1] - E[U_2]}{\eta_1 - \eta_2} \qquad (7),$$

where η_1 and η_2 are noise parameters denoting errors made by agents in estimating E[U_1] and E[U_2], respectively.[16] Equation 7 is modified in three ways. First, it is adapted to the context of organizational change, where one alternative is status quo (*sq*) and the other alternative is the

[15] Alternative candidates for acquisition might be determined by identifying firms in the same market.

[16] In empirical setting, the denominator in equation 8 is to be estimated (see Harrison, 2008).

maximum of alternative strategies. Second, the denominator is eliminated because the simulated agents do not make mistakes. The denominator should be retained in contexts involving non-simulated data. Third, since in real data utility is an unobservable function of wealth depending upon latent risk preferences, the index is simplified by assuming risk-neutral utility functions. Consequently, the modified index considers differences in expected wealth, where $W_{jt} = I_{jt}+W_{jt-1}-S$.[17] Note that while this simplification enhances empirical applicability, it sacrifices rigor in estimating utility maximization, which may make it more difficult to obtain support for RC. The modified index can be interpreted as the opportunity cost of remaining in the status quo, and the name is revised accordingly:

$$\text{Opportunity Cost}_{\text{it}} = \max\{E[W_{i,j\neq sq,t}]\} - E[W_{i,j=sq,t}] \qquad (8).$$

Since maximizing agents change strategies whenever *Opportunity Cost*$_{jt}$ > 0, opportunity cost should be operationalized as a dichotomous variable, such that *Opportunity Cost Dummy*$_{jt}$ = “1” if *Opportunity Cost*$_{jt}$ > *0*, and “0” otherwise.

4.1.2 Modeling Organizational Change with Deliberate Controls

Having derived a proxy for opportunity cost, we use Tables 5-7 to show how *Opportunity Cost Dummy* and Greve’s (1998) operationalization of aspirations impact the testing of the null hypothesis that a theory does not explain organizational change. Two models are used to estimate organizational change: pooled logit as in Greve (1998) and random effects logit to account for time and cross-sectional dependence. The implementation of a random effects model may be needed to appropriately account for path dependence emphasized in both theories, while pooled models assume temporal independence of observations. Failure to apply appropriate estimation techniques may result in misdiagnosing support for a theory.

---

[17] After eliminating the denominator, the Fechner Index becomes E[$U(W_{j\neq sq,t})$]-E[$U(W_{j=sq,t})$]. *S* is not in the second term because under the status quo alternative, switching cost is not relevant.

4.1.3 Evaluating the Effectiveness of Recommendation 1

Table 5 illustrates models fitted to the data where the true causality is RC, and helps in assessing whether the recommendation is useful in ruling out a false positive (Type I error) for BT and a false negative (Type II error) for RC (cells A and B in Table 4). Column 1 replicates column 2 of Table 3, which was already demonstrated to be a Type I error. Column 2 of Table 5 fits an alternative model aimed at diagnosing RC, using *Opportunity Cost Dummy*. As expected, *Opportunity Cost Dummy* is positive, suggesting agents are more likely to prefer change when the opportunity cost of the status quo is positive. Column 3 implements the first recommendation, by combining the BT variables with *Opportunity Cost Dummy* in a pooled logit model.[18] This combined model yields a positive and significant *Opportunity Cost Dummy*, implying the recommendation rules out Type II error for RC. It also shows some ability to eliminate (Type I error) false confirmation of BT because robust support for BT is lost: social aspirations are rejected, while spurious support for historical aspirations persists. Replicating the comparative model in a random effects logit in column 5 significantly improves model fit and completely eradicates (Type I error) the spurious support for BT found in column 3, and avoids Type II error by revealing a positive coefficient for *Opportunity Cost Dummy*.

(Insert Tables 5 ,6, and 7 about here)

Table 6 illustrates models fitted to the data where the true causality is BT. This table helps assess whether the recommendation is useful in ruling out a false positive (Type I error) for RC and a false negative (Type II error) for BT (cells C and D in Table 4). Column 1 fits a behavioral model and demonstrates that agents change strategies based on aspirations as

[18] The combined model is an improvement over models 1, as expected. Interestingly, it is also an improvement over model 2, which is attributed to the fact that a relatively primitive model is used relying on the assumption that the coefficient is the same for all agents. Since there are five distinct risk attitudes present in the data, this should not be the case.

predicted by Greve (1998). This confirmation of BT creates confidence that Greve's (1998) formalization of Cyert and March (1963) is implemented correctly in the simulation.[19] Column 2 fits an alternative model aimed at diagnosing RC, but the model is not significant. Column 3 implements the first recommendation in a pooled model. This comparative model yields robust estimates for both social and historical aspirations, ruling out Type II error with respect to BT. It also rules out Type I error because *Opportunity Cost Dummy* is insignificant, it does not falsely confirm RC. These results hold up in a random effects logit model.

The results reported in Tables 5 and 6 suggest that recommendation 1 is effective at eliminating Type I and II errors when the data generating processes are pure – all agents use either RC or BT. Table 7 illustrates models fitted to data where the true causality is a mixture of BT and RC. Since both heuristics are present, this scenario will help assess whether the recommendation is useful in ruling out false negatives (Type II error) for both RC and BT. Models of mixed data will not help in ruling out false positives (Type I errors) because they cannot occur if both behaviors are present. *Opportunity Cost Dummy* is insignificant across all models in the table, suggesting that we commit a Type II error, because we know RC is used by 40 percent of the sample. BT is largely supported in our comparative tests (models 3 and 5), although it should be noted that support for social aspirations is lost in the random effects logit. These findings do not embolded great confidence that recommendation 1 is effective in discriminating between theories in the presence of mixed behavior.

To assure that the recommendation was capable of discriminating between theories in smaller samples, we also tested the effectiveness of recommendation 1 across forty samples of 14,000 observations, where agents were randomly drawn without replacement. Note that each of

[19] Also, note that there is a close relationship between the estimated coefficients in column 1 of table 6 with the parameters in the simulation explained in Appendix B.

these subsamples included fewer observations than Greve's (1989) sample. When the true causality is RC, Type I and Type II errors are avoided in all forty samples. When the true causality is BT, Type I error was avoided in all forty samples, and Type II error occurred in one of forty samples (we falsely confirmed RC 5 percent of the time). When the true causality is mixed, Type II error for BT was avoided in all 40 cases, but Type II error for RC occurred in 95 percent of the cases. In short, these results on smaller samples show an identical pattern to the pattern of results reported in Tables 5-7.

In summary, a key advantage of the first recommendation is the practicality in developing a proxy for opportunity cost and modeling it alongside BT. A disadvantage is the assumption of a simplified utility function, although we found no evidence of that this simplification placed the recommendation in jeopardy when data generating processes was pure. Another potential shortcoming is that combining RC and BT in a single model assumes their independence, and will lead to biased estimates if this assumption is false. Despite these concerns, this first recommendation prevented Type I and II errors across all pure scenarios, and was always effective at preventing Type II error for BT in mixed behavior when 60 percent of the agents were satisficers, but in that scenario was relatively inefficient in preventing Type II error for RC.

4.2 Recommendation 2: Test of Non-Nested Alternative Hypotheses

Recommendation 1 encourages ascribing variables from different theoretical models into a single empirical model. This is perhaps the most common approach to testing competing theories. It is equivalent to saying that the two theories are nested in a combined model, which implies that each of them "can be reduced to the other [combined] model" (Clarke, 2001: 727). To clarify, the following equations correspond to columns 1-3 in tables 5-7:

$$
\begin{aligned}
H_{BT}: &\quad Y = \omega X_1 + \theta && (9a)\\
H_{RC}: &\quad Y = \psi X_2 + \rho && (9b)\\
H_{Combined}: &\quad Y = \omega X_1 + \psi X_2 + \pi && (9c)
\end{aligned}
$$

where Y is organizational change; X_1 and X_2 are vectors of variables; ω and ψ are coefficient vectors; and θ, ρ, and π are errors. The models 3 and 5, fit in Tables 5-7, rely on the assumption that X_1 and X_2 are not co-determined. If X_1 and X_2 are co-determined then estimates in the combined model (9c) will be biased. While one approach would be to estimate the combined model (9c) after controlling for endogeneity via instrumental variable probit, when there are no strong instruments for potentially endogenous parameters, this approach may reduce the reliability of estimation compared to models excluding instruments.[20] Since there exist no strong instruments for aspirations and expectations a combined model may not be reliable. Recommendation 2 overcomes the problem noted above through a non-nested test, which helps in assessing the relative predictive power of model (9a) versus (9b). This test is particularly useful when there is a suspicion that one of the models (either 9a or 9b) provides a spurious correlation between dependent and independent variables.

4.2.1 Evaluating the Effectiveness of Recommendation 2

In our analysis we use Vuong's (1989) test, which generates likelihood-ratio based statistics for testing the null hypothesis that the competing models are equally close to the true data generating process against the alternative hypothesis that one model is closer. Vuong's test is noteworthy for its generality, allowing for theories to "be nested, non-nested, or overlapping, and that both, only one, or neither of the competing models may contain the true law generating the

[20] Hahn and Hausman (2003: 118) discuss the implications of applying weak instruments: "…a researcher may estimate "bad results" and not be aware of the outcome." While their work focused on linear models, Adkins (2009) finds that in probit models weak instruments enhance the bias.

observations" (p. 307).[21] This test can be implemented with or without an adjustment in the likelihood ratio statistic for the number of parameters in the respective models. Since adjusted statistics penalize BT models more than RC models because they employ more variables, we use both adjusted and unadjusted statistics. Table 8 reveals the results of the non-nested test across our three scenarios. Under each scenario, the test effectively avoids Type I and Type II errors.

(Insert Table 8 about here)

To assure this recommendation is practical in smaller samples, we also tested the effectiveness of recommendation 2 on 40 samples of 14,000 observations. The non-nested test proved 100 percent effective at avoiding Type II error when the data generating processes were pure. In all cases of pure behavior, the adjusted and unadjusted likelihood-ratio statistics were significant. The non-nested test using the unadjusted likelihood-ratio test also proved 100 percent effective at avoiding Type II error in mixed behavior. However, we found the adjusted likelihood-ratio statistic to be 100 percent ineffective at determining BT was the dominant theory in the data. The fact that adjustment imposes a greater penalty in smaller samples suggests that a more conservative test should use the unadjusted statistic.

In summary, a key advantage of this recommendation is that it avoids concerns about endogeneity when theoretical constructs may be co-determined. It is easy to implement and reliably predicts the dominant causality in the data. Like recommendation 1, it relies on a simplified utility function. Note that passing such test does not mean that the theory in question is true, it simply means that the theory is better than the available alternative. Finally, this

[21] Other notable non-nested tests are offered by Akaike (1973) and Cox (1961). Compared to the Akaike test, the Vuong test does not require that a "best" model be chosen if the competing models are statistically equivalent. Compared to the Cox test, the Vuong test allows for competing models to be either strictly non-nested or overlapping. In the presence of mixed behavior, allowing for both of these is important.

recommendation effectively identifies the theory most robustly present in the data, and therefore more credible, but it does not imply that the losing theory is absent in the data.

4.3 Recommendation 3: Comparative Robustness of Alternative Theories

The third recommendation is similar to recommendation 2 in its aversion to a combined model (e.g., equation 9c). It is similar to recommendation 1 in testing the null hypothesis that a theory does not explain organizational change. The emphasis is on the relative robustness of each separate model (e.g., equations 9a and 9b). Here, a sort of competition, or test of comparative robustness, is encouraged for how persistent confirmations are to empirical assumptions used to instrument the theory. While it is well-known that models robust to alterations in empirical assumptions are less likely to generate Type I error, we advocate that the winner of the competition is the theory that is more robust to alterations in assumptions, implying it is more likely to represent the true causality.[22] Of course, a number of types of robustness tests may be reasonable candidates. Below, two specific tests are recommended for a competition between RC and BT. Each test is described separately, and the results of the competition are summarized.

4.3.1 Robustness Test of RC model

In recommendations 1 and 2, the assumption was made that all firms have the same utility function. Empirically, utility was estimated through a single coefficient for the *Opportunity Cost Dummy*. While this assumption is pragmatic because utilities are latent, firms may have different utility functions, and correspondingly, may be differentially sensitive to wealth. If the

[22] The idea that robustness reduces concerns about Type I error is noted by Wimsatt (1981: 140): "If a result is robust over a range of parameter values in a given model or over a variety of models making different assumptions, this gives us some independence of knowledge of the exact structure and parameter values of the system under study: a prediction of this result will remain true under a variety of such conditions and parameter values. This is particularly important in scientific areas where it may be difficult to determine the parameter values and conditions exactly."

assumption of homogeneity of utilities across firms is relaxed, it is possible to test the robustness of the initial findings (recommendations 1 and 2) when accounting for heterogeneous utilities.

Very recent developments in econometrics of choice identify latent utility classes using non-parametric techniques. A non-parametric mixed logit developed by Train (2008) is estimated under the assumption that an agent's utility for a particular strategy is a function of wealth under the strategy, as indicated below:

$$U_{ijt} = \lambda_i W_{ijt} + \nu_{ijt} \qquad (10),$$

where λ is a coefficient, and υ represents error.[23] Note that because coefficient λ is specific to each agent *i*, the model allows heterogeneous utility, and therefore, generates superior fit relative to models assuming homogeneity. Because the model is designed to achieve superior fit, the most reliable indicator of the type of behavior dominating the sample is $\bar{\lambda}$, the population mean coefficient (Train, 2008). $\bar{\lambda}$ is the sum of the products of each class's coefficient and proportion in the population. Accordingly, to ascertain whether RC is robustly present in the data, the sign and significance of $\bar{\lambda}$ is analyzed. The test involves three steps: (1) deriving the number of latent classes affecting utility; (2) estimating the model given the number of latent classes; and (3) bootstrapping $\bar{\lambda}$ to ascertain significance.

The test generated the following results across the three scenarios. When the true causality is RC, $\bar{\lambda}$=1997.9 and significantly greater than zero (*p=0.001)*. That $\bar{\lambda}$ is positive suggests that RC is robustly present because, on average, agents prefer choices that give them more wealth. When the true causality is BT, $\bar{\lambda}$ = -0.9 and significantly less than zero (*p=0.001*). That $\bar{\lambda}$ is negative suggests that RC is not robustly present because wealth is not a positive predictor of choice, on average. When the true causality is mixed, $\bar{\lambda}$=12.9 and significantly

[23] The term W_{ijt} is used in this estimation because it is unique to each alternative *j*, whereas Opportunity Cost$_{it}$ is a property of a pair of alternatives.

greater than zero ($p < 0.001$), suggesting that RC is robustly present. Thus, in contrast to the prior recommendations, recommendation 3 identifies RC in mixed behavior when rational behavior is in the minority. This same pattern of results was obtained on a random sample of 14,000 observations across each of the three scenarios, suggesting that the Train model is suitable for discriminating the presence of RC in smaller samples, a finding already established by Train (2008), who used a sample of 508.[24]

4.3.2 Robustness Test of BT Model

In recommendations 1 and 2, when testing for BT with Attainment Discrepancy for historical aspirations, the assumption was made that estimation of the model should involve the unique value of the weight coefficient, α, providing the best fit. This assumption is consistent with current approaches in the literature but is problematic in two respects. First, it does not simultaneously estimate α with other parameters in the model. Simultaneous estimation is appropriate if α is endogenous to aspirations, which is stated to be the case within BT (see equation 3). Second, it does not provide test statistics about whether α contributes meaningfully to the prediction, which makes it impossible to ascertain whether α is statistically different from zero or any other values of α. This motivates consideration of how sensitive the support for BT is to the choice of α. If the true causality in the population is BT, it is expected that the diagnosis of BT will be robust across more specifications of α. Whereas, if the true causality is RC, or a mixture of RC and BT, it is expected that less robustness would obtain across values of α.[25]

---

[24] We did not replicate this test across forty randomly generated subsamples in three scenarios because based on Train's (2008) insight, we calculate the computational time to be two years on a single-core machine.

[25] So, like Baum, et al. (2005), tests are encouraged for robustness of findings to the choice of α, but suggest this is only one of two components in the competition for comparative robustness between RC and BT. An unbiased comparative test would simultaneously compare the robustness of α, and robustness of maximization described in the preceding section.

The model reported in column 1 of Tables 5-7 was fit for different values of α. Table 9 reports the number of times BT was confirmed across all three scenarios. When the true causality is RC, it was impossible to simultaneously confirm both types of aspirations. When the true causality is BT, it was possible to simultaneously confirm both types of aspirations for 12 out of 14 values of α. When the behavior is mixed, simultaneous confirmation of both types of aspirations obtained for 11 out of 14 values of α. The analysis illustrates that the recommendation is effective in diagnosing robust support for BT when it is present, while avoiding robust support for RC when it is not present. An inference for researchers is that the more robust findings are to variations in α, the more reliable are confirmations of BT.

(Insert Table 9 about here)

To demonstrate that this robustness check is feasible on smaller samples, we replicated the tests in forty random subsamples of smaller size (n=14,000), constituting 560 attempts for each scenario (14α x 40) . When the true causality was RC, we never simultaneously confirmed both types of aspirations. When the true causality was BT, we simultaneously confirmed both types of aspirations 75.1 percent of the time. In mixed behavior, we simultaneously confirmed both types of aspirations 64.1 percent of the time. Note that historical aspirations were confirmed 100 percent of the time in pure BT and mixed scenarios. These results suggest that this robustness check can be reliably performed in samples of 14,000.

4.3.3 Evaluating the Effectiveness of Recommendation 3

The comparative robustness of confirmations to RC and BT can be examined across the three scenarios, which should point researchers unambiguously toward the true causality. When the true causality is RC, the two tests outlined above demonstrate robust confirmation of RC and a lack of robustness for BT. When the true causality is BT, the tests demonstrate non-robustness of

RC and robust support for BT. When the true causality is a mixture of RC and BT, the tests demonstrate confirmation of RC and relatively strong robustness of BT. In each scenario these comparative tests help to successfully diagnose the true causality present in the data.

Finally, this third recommendation effectively diagnoses true causality. Relative to the other two recommendations, it is more rigorous in estimating utility, less reliant on specific operationalizations of theory, and avoids concerns about endogeneity. A limitation is that it does not involve a statistical test for comparative robustness.

5. Conclusion

It has been nearly fifty years since Cyert and March (1963) considered that organizational change may be guided not by rational behavior, but by boundedly rational behavior. While their thesis is decidedly comparative in nature, the empirical research seeking to confirm aspiration-level decision-making has not been as comparative, which raises concerns about erroneous confirmations. While comparative tests between behavioral theory and rational choice have not been implemented in prior research, we use simulated data to demonstrate that it is possible to do so effectively, and offer a template for how they might be conducted with real (non-simulated) data. Analysis of simulated data shows that rational decision-makers behave in a statistically similar manner as predicted in several seminal projects claiming support for behavioral theory. The simulated agents appear to be satisficing, when they are clearly not. Results are remarkably robust, and imply that until future work implements comparative testing, support that organizations change based on aspiration is tenuous. Three recommendations for comparatively testing behavioral theory and rational choice are shown to be remarkably effective at controlling Type I and Type II error in simulated data. If implemented together, they overcome the bounds of any singular test.

The recommendations for comparative testing are designed to improve empirical discrimination between behavioral theory and rational choice, but the very process of implementing them will amplify and clarify the scope of behavioral theory. Comparative testing may lead to three potential outcomes. First, it may confirm behavioral theory as the dominant explanation in real data, which will serve to accentuate the credibility of the theory and empirical approximations of the theory. Second, it may disconfirm behavioral theory as the dominant explanation in real data. This finding is a signal for researchers to look for theoretical and empirical causes of disconfirmation, which may result from applying the theory beyond the original boundary conditions of the theory. For example, since Cyert and March (1963) developed their theory for large, multi-product firms, where information efficiencies and political processes are at play, it may be inappropriate to map the theory unaltered onto entrepreneurial firms. Disconfirmation may also lead scholars to reconsider the original boundary conditions established by Cyert and March (1963). Finally, disconfirmation may help deduce imprecise operationalization of theoretical concepts and the relationships they imply, or inspire research designs better enabling scholars to distinguish aspirations from expectations.

A third possible outcome from implementing the recommendations could simultaneously confirm some in the population satisfice while others maximize. While the non-nested test developed by Vuong (1989) explicitly allows for theoretical overlap in samples, it only allows one to discern whether one theory better explains the data, and not the extent to which *each* explains the data. *Finite mixture models* have been recently developed to tackle this issue. They "estimate the parameters of each theory while simultaneously estimating the probability that each theory applies to the sample" (Harrison and Rutström, 2009: 133). A challenge with mixture models of this kind is that convergence may not occur in maximum likelihood estimations

(Harrison and Rutström, 2009: 137), as was the case with the data simulated for this paper. Nevertheless, such problems are being actively addressed in econometrics of choice, suggesting an interesting opportunity is eminent. Transparent attempts at comparative testing in mixed data will strengthen the support of behavioral theory.

Some may criticize this study's replication of prior empirical work as mere manipulation of data and initial conditions. This critique is impotent for several reasons. First, the replications are robust across a wide range of specifications. While these replications are not "globally" robust because the considered range does not reflect the universe of potential specifications, they are "locally" robust. It is infeasible to show global robustness because the potential set of initial parameters is unbounded. Moreover, doing so would suggest that behavioral theory reduces to rational choice, which is certainly not the intent of this paper. Moreover, the fact that recommendations effectively discriminate between behavioral theory and rational choice should exhibit evidence that the replications are not a product of initial specifications.

Others may criticize as unrealistic the characterization of pure satisficing or maximizing behavior, instead preferring some hybrid representation. As demonstrated, however, these pure scenarios are necessary for distinguishing between Type I and Type II error. Some may prefer comparatively testing behavioral theory against a rational choice that is instrumented with Simon's objective rationality, incorporating game-theoretic reasoning. Since agents would not rely as heavily on prior performance as they do in our model of subjective rationality, such an approach may make it more difficult to replicate empirical confirmations of behavioral theory, which may be important to those fastening affections to behavioral theory. Comparative tests between the theories should focus on the maximizing and satisficing heuristics, not on assumptions about knowledge that are merely used to instrument both theories. Indeed, it is advisable to adopt the most general interpretation of a theory, not a more strict interpretation.

The striking similarity observed between aspirations and expectations should not be interpreted as a condemnation of behavioral theory or satisficing behavior. Clearly, organizational aspirations are important. Rather, this study is a call for greater insight into when aspirations are important and when they are not. One could take issue with the targeted attack on empirical work around aspirations, claiming that empirical work around rational choice suffers equal limitations. Clearly the principles of comparative testing apply equally to both theories. The readers of this journal, however, tend to be predominantly organizational. Moreover, the recommendations do not favor either behavioral theory or rational choice. The use of three scenarios demonstrates agnosticism about true causality.

One limitation is that behavioral theory and rational choice are analyzed as the only alternatives predicting change. Other explanations are plausible, and in the presence of a third unobserved theory, comparatively testing two theories is not sufficient for diagnosing causality. Recall the purpose was not to comparatively test in real data, but demonstrate the consequences of failing to do so and offer reliable solutions for comparative testing. This purpose is most effectively accomplished by focusing on a single theory and its most likely alternative.

The explicit focus on behavioral theory must not blind scholars who are studying organizational change and who have other affections. Affections can endanger research by straining the integrity of the intellectual process. Comparative testing is needed to discipline all inquiry of organizational phenomena. Of course, the rigor with which one implements comparative testing is the ultimate discipline.

Appendix A: Utility Functions Assumed in Section 3.1

Risk attitude (*m*)	**Utility Function (*U*)**	**Type of risk attitude**
1	$U(W) = e^{0.60W}-1$	risk-seeking
2	$U(W) = e^{0.50W}-1$	risk-seeking
3	$U(W) = -e^{-1.00W}+1$	risk-averse
4	$U(W) = -e^{-1.05W}+1$	risk-averse
5	$U(W) = -e^{-1.10W}+1$	risk-averse

Appendix B: **Description of the Simulation of Change Under Satisficing Behavior**

Like in Greve (2002), *change* is simulated through satisficing behavior, but this simulation is altered to make the behavior entirely consistent with the benchmark model of Greve (1998), who specifies a decision heuristic in hypotheses 1a, 1b, 2a, and 2b, listed below.[26]

> Hypothesis la (2a): When performance relative to the social (historical) aspiration level increases, the probability of change decreases.
>
> Hypothesis lb (2b): The decrease in the probability of change is greater for performance increases above the social (historical) aspiration level.

In accord with BT, change behavior is simulated in the following way:

1. Historical aspirations in the first period (A_{i1}) are randomly drawn from $\mathcal{U}(0, 0.25)$.
2. Groups are determined to be reference groups for social aspirations.
3. Beginning period 2, agents calculate attainment discrepancies for historical (I_{it-1}-A_{it-1}) and social aspirations (I_{it-1}-B_{it-1}).
4. In the first period, choice of strategy is random to generate the first record of performance.
5. In subsequent periods, probability of change q_{it} for each agent is determined the following logistic function:

$$\log\frac{q_{it}}{1-q_{it}} = \omega_0 + \omega_1\left[I_{it-1} - A_{it-1}\right]_{>0} + \omega_2\left[I_{it-1} - A_{it-1}\right]_{<0} + \omega_3\left[I_{it-1} - B_{it-1}\right]_{>0} + \omega_4\left[I_{it-1} - B_{it-1}\right]_{<0} + \omega_5 N_{it-1}$$

   where coefficients ω_1, ω_2, ω_3, ω_4 are arbitrarily assigned values -10, -5, -8, -4, respectively, consistent with Greve's (1998) prediction that the probability of change is more negative when performance is above aspirations than below; coefficient ω_5 is arbitrarily assigned a value of 20, consistent with Greve's (1998) expectation that it should be positive; and intercept ω_0 is assigned a negative value, as found in Greve's (1998) model.

6. q_{it} is then used to predict *change* (Y_{it}), a value of zero or one, through the Bernoulli distribution.
7. If $Y_{it}=1$, agents randomly move to one of the four remaining strategies (Greve, 2002).

All parameters not mentioned above are unchanged from section 3.1. 200 runs result in 600,000 observations. Note that the simulation is entirely consistent with Cyert and March (1963).

---

[26] Greve (1998) uses the *negative inconsistency parameter* to resolve the ambiguity when performance lies between social and historical aspirations, while Greve (2002) resolves the ambiguity by assigning importance weights to historical and social aspirations.

Appendix C: **Description of the Simulation of Change Under Mixed Behavior**

Change is simulated through a mixture of satisficing and maximizing behaviors. There are infinite ways to mix behavior. For this sample, mixture is established in the following way:

1. Agents are randomly assigned to five groups.
2. Agents in groups 3-5 behave in accord with Appendix B.
3. Agents in group 1 have the same utility function as specified for risk attitude $m=3$, and behave in accord with section 3.1.
4. Agents in group 2 have the same utility function as specified for risk attitude $m=5$, and behave in accord with section 3.1.

All parameters not mentioned above remain unchanged from section 3.1. 200 runs result in 600,000 observations, approximately 60 percent of which represent satisficing behavior and 40 percent represent maximizing behavior. Note that the simulation is entirely consistent with both BT and RC for separate parts of the sample. Recognize the simulation does not just merge data sets generated separately by maximizers and satisficers. They act simultaneously in the same setting, resulting in interactions between agents with different heuristics. One subtlety of this mixed behavior is that because agents in groups 3-5 randomly choose strategies if dissatisfied, the landscape is more effectively revealed to maximizing agents, who otherwise would not have recognized superior alternative that were unexplored.

References

Adkins, L. C. 2009. *Testing Parameter Significance in Instrumental Variable Probit Estimators: Some Simulation Results*. Working Paper.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory* 267-281.

Allison, G. T. 1969. Conceptual models and the Cuban missile crisis. *Amer. Political Science Rev*. **63**(5) 689-718.

Argote, L., H. R. Greve. 2007. A behavioral theory of the firm—40 years and counting: Introduction and impact. *Organ. Sci.* **18**(3) 337–349.

Arrow, K. J. 1951. Alternative approaches to the theory of choice in risk-taking situations. *Econometrica*. **19**(4) 404-437.

Arrow, K. J. 1971. *Essays in the Theory of Risk-Bearing*. Markham, Chicago, IL.

Baum, J. A. C., T. J. Rowley, A. V. Shipilov, Y.-T. Chuang. 2005. Dancing with strangers: Aspiration performance and the search for underwriting syndicate partners. *Admin. Sci. Quart*. **50**(4) 536-575.

Camerer, C., R. Weber. 1999. The econometrics and behavioral economics of escalation of commitment: A re-examination of Staw and Hoang's NBA data. *J. Econom. Behav. Organ*. **39**(1) 59-82.

Clarke, K.A. 2001. Testing nonnested models of international relations: reevaluating realism. *Amer. J. Political Sci.* **45**(3) 724-744.

Chamberlin, T. C., 1931. The method of multiple working hypotheses. *J. Geology.* **39**(2) 155-165.

Cochrane, D, Orcutt, G.H. 1949. Application of least squares regression to relationships containing auto-correlated error terms *J. Amer. Statist. Association* **44** 32–61.

Cox, D. 1961. Test of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1** 105-123.

Cyert, R. M., M. H. DeGroot. 1974. Rational expectations and Bayesian analysis. *J. Political Econom*. **82**(3) 521-536.

Cyert, R. M. , J. G. March. 1963. *A Behavioral Theory of the Firm*. Wiley-Blackwell, London.

-------, -------. 1992. *A Behavioral Theory of the Firm*. 2nd ed. Prentice Hall, Englewood Cliffs, New York.

DePamphlis, D. M. 2003. *Mergers, Acquisitions, and Other Restructuring Activities: An Integrated Approach to Process, Tools, Cases, and Solutions*. 2nd ed. Academic Press, Amsterdam; Boston.

Fechner, G. 1966. *Elements of Psychophysics*. Holt, Rinehart and Winston: New York.

Festinger, L. 1954. *A theory of social comparison processes*. *Human Relations*. **7**(2) 117-140.

Fuller, W.A., Battese, G.E. 1974. Estimation of linear models with crossed-error structure. *J. Econometrics* **2** 67-78.

Greve, H. R. 1998. Performance, aspirations, and risky organizational change. *Admin. Sci. Quart*. **43**(1) 58-86.

Greve, H. R. 2002. Sticky aspirations: Organizational time perspective and competitiveness. *Organ. Sci.***13(1)** 1-17.

Greve, H. R. 2003. *Organizational Learning from Performance Feedback: A Behavioral Perspective on Innovation and Change*. Cambridge University Press, Cambridge, UK.

Hahn, J., J. Hausman. 2003. Weak instruments: Diagnosis and cures in empirical econometrics. *Amer. Econ. Rev.* **93**(2) 118-125.

Harrison, G. W. 2008. *Maximum Likelihood Estimation of Utility Functions Using Stata*. Working Paper 06-12, Department of Economics, College of Business Administration, University of Central Florida.

Harrison, G. W., E. E. Rutström. 2009. Expected utility theory and prospect theory: One wedding and a decent funeral. *Exp. Econ.* **12** 133-158.

Iyer, D. N., K. D. Miller. 2008. Performance feedback, slack, and the timing of acquisitions. *Acad. Management J*. **51**(4) 808-822.

Kuhn, T. S. 1970. *The Structure of Scientific Revolutions*. 2nd ed. University of Chicago Press, Chicago, IL.

Lakatos, I. 1974. The role of crucial experiments in science. *Stud. Hist. Philos. Sci*. **4**(4) 309-325.

Lant, T. K. 1992. Aspiration level adaptation: An empirical exploration. *Management Sci*. **38**(5) 623-644.

Lant, T. K., D. B. Montgomery. 1987. Learning from strategic success and failure. *J. Bus. Res*. **15**(6) 503-517.

Levinthal, D. A., J. G. March. 1981. A model of adaptive organizational search. *J. Econom. Behav. Organ*. **2**(4) 307-333.

Lewin, K., T. Dembo, L. Festinger, P. S. Sears. 1944. Level of aspiration. J. M. Hunt, ed. *Perssonality and the Behavioral Disorders*, Vol. 1. 333-378. Ronald Press, New York.

March, J. G., C. Heath. 1994. *A Primer on Decision Making: How Decisions Happen*, Free Press, New York.

March, J. G., H. A. Simon. 1959. *Organizations*. 2nd printing. John Wiley & Sons, Inc., New York.

Mezias, S. J., Y.-R. Chen, P. R. Murphy. 2002. Aspiration-level adaptation in an American financial services organization: A field study. *Management Sci*. **48**(10) 1285-1300.

Neter, J., Wasserman, W., Witmore, C.A. 1988. *Applied Statistics*. 3rd ed. Allyn and Bacon, Boston.

Nickel, M. N., M. C. Rodriguez. 2002. A review of research on the negative accounting relationship between risk and return: Bowman's paradox. *Omega* **30**(1) 1–18.

Popper, K. 1963. *Conjectures and Refutations*. London: Routledge and Keagan Paul.

Pratt, J. W. 1964. Risk aversion in the small and the large. *Econometrica*. **32**(1/2) 122-136.

Schwartz, B., A. Ward, S. Lyubomirsky, J. Monterosso, K. White, D. R. Lehman. 2002. Maximizing versus satisficing: Happiness is a matter of choice. *J. Personality and Soc. Psych.* **83**(5) 1178-1197.

Schwartz, G. 1978. Estimation the Dimension of a Model. *The Annals of Statistics*. **6** 461-464.

Sen, A. 1997. Maximization and the act of choice. *Econometrica*. **65**(4) 745-779.

Simon, H. A. 1978. Rational decision making in business organizations. *Amer. Econom. Rev*. **69**(4) 493-513.

Simon, H. A. 1997a. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations.* Free Press, New York.

Simon, H. A. 1997b. *An Empirically Based Microeconomics*. Cambridge University Press, Cambridge, UK.

Stinchcombe, A. L. 1968. *Constructing Social Theories*. Harcourt, Brace and World, New York.

Train, K. E. 2008. EM algorithms for nonparametric estimation of mixing distributions. *J. Choice Modelling*. **1**(1) 40-69.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. **57**(2) 307-333.

Wimsatt, W. Robustness, reliability, and overdetermination. Brewer, M.B., Collins, B.E. eds. *Scientific Inquiry and the Social Sciences*, 124-163. Jossey-Bass, San Francisco.

Table 1. Representation of a History for a Single Agent from a Run for All 30 periods[27]

Period (t)	Wealth at the beginning of the period (W_{it-1})	Past strategy (j_{t-1}^*)	Costs of switching (S)	Expected utility ($E[U_{ijt}]$)					Chosen strategy (j_t^*)
				j=1	j=2	j=3	j=4	j=5	
1	0.0000000	-	0.0022996	-	-	-	-	-	4
2	0.1282474	4	0.0022996	0.20705	0.31545	0.14657	0.20896	0.26324	2
3	0.1170309	2	0.0022996	0.19765	0.16454	0.13646	0.19765	0.25451	5
4	0.1170295	5	0.0022996	0.19765	0.19268	0.13646	0.19765	0.16580	1
5	0.2243357	1	0.0022996	0.28487	0.29133	0.22847	0.28314	0.25289	2
6	0.2775171	2	0.0022996	0.35263	0.31332	0.27037	0.32208	0.29346	1
7	0.2398448	1	0.0022996	0.25693	0.28388	0.24093	0.29472	0.26495	4
8	0.2760066	4	0.0022996	0.28288	0.31056	0.26921	0.28029	0.29234	2
9	0.3917541	2	0.0022996	0.36495	0.39309	0.35285	0.36111	0.36106	2
10	0.3842933	2	0.0022996	0.34810	0.38092	0.34776	0.35609	0.35604	2
11	0.4086289	2	0.0022996	0.36454	0.39163	0.36421	0.37234	0.36384	2
12	0.4169403	2	0.0022996	0.37006	0.39339	0.36973	0.39018	0.36936	2
13	0.3978081	2	0.0022996	0.35728	0.37244	0.35695	0.38698	0.35657	4
14	0.4792343	4	0.0022996	0.40995	0.42247	0.40964	0.44388	0.40929	4
15	0.5640773	4	0.0022996	0.46024	0.47169	0.45996	0.49350	0.45964	4
16	0.5901301	4	0.0022996	0.47480	0.48595	0.47453	0.50046	0.47422	4
17	0.6458557	4	0.0022996	0.50465	0.51516	0.50439	0.52476	0.50410	4
18	0.7339408	4	0.0022996	0.54841	0.55799	0.54818	0.56891	0.54791	4
19	0.7625536	4	0.0022996	0.56178	0.57108	0.56155	0.57918	0.56129	4
20	0.8038235	4	0.0022996	0.58036	0.58927	0.58014	0.59520	0.57989	4
21	0.8913667	4	0.0022996	0.61721	0.62534	0.61701	0.63222	0.61679	4
22	0.8777467	4	0.0022996	0.61170	0.62097	0.61150	0.62295	0.61127	4
23	0.9148593	4	0.0022996	0.62654	0.63881	0.62635	0.63660	0.62613	2
24	0.8934351	2	0.0022996	0.61804	0.62595	0.61785	0.62744	0.61762	4
25	0.903338	4	0.0022996	0.62200	0.62856	0.62180	0.63171	0.62158	4
26	0.9760581	4	0.0022996	0.64978	0.65889	0.64960	0.65956	0.64939	4
27	1.004296	4	0.0022996	0.66002	0.67057	0.65984	0.66911	0.65964	2
28	1.017782	2	0.0022996	0.66480	0.67490	0.66462	0.67297	0.66442	2
29	1.116383	2	0.0022996	0.69776	0.70807	0.69760	0.70514	0.69743	2
30	1.111112	2	0.0022996	0.69609	0.70561	0.69593	0.70350	0.69575	2

[27] A sample agent with m=4.

Table 2. Estimating the Determinants of Aspirations: A Comparison of Lant and Montgomery (1987), Lant (1992), and Mezias et al. (2002) with Replications on the Data Where True Causality is RC[28]

	Hypothesized Relationships in the Benchmark Studies	Lant and Montgomery (1987) & Lant (1992)[29]	Replication of (*1*) on simulated data where true causality is RC[30]	Mezias et. al (2002)[31]	Replication of (*3*) on simulated data where true causality is RC
		1	*2*	*3*	*4*
Constant		+ *	+ ***	-	+ ***
Historical Aspiration (A_{it-1})	+	+ **	+ ***	+ ***	+ ***
Attainment Discrepancy for Historical Aspiration (I_{it-1}-A_{it-1})	+	+ **	+ ***	+ ***	+ ***
Attainment Discrepancy for Social Aspiration (I_{it-1}-B_{it-1})	-			- **	- ***
Adjusted R^2		0.839	0.888		
R^2				0.976	0.512
F-statistic		271.72**	2,141,315***		
Hausman test for random effects				n.a.	8554.740***
F-test for addition of attainment discrepancy social				*	***
Number of observations		105	540,000	860	135,000

* $p < 0.05$;
** $p<0.01$;
***$p <0.001$

[28] Only the direction of the relationships is indicated in the table, since the referred studies do not standardize coefficients.

[29] From Table 2 of Lant and Montgomery (1987) and Table 2 of Lant (1992). Lant (1992) estimated the generalized linear models with the Corchrane and Orcutt (1949) transformation. Because the signs of corresponding coefficients are the same across industries considered in her study, we report results only for Industry 1.

[30] Because the Corchrane and Orcutt (1949) transformation dictates omission of the first observation in each series and social aspirations can only be derived starting with *t*=2, only observations starting with *t*=3 were used, resulting in 540,000 observations in this model.

[31] From Table 3 of Mezias et al. (2002). They estimated the Fuller and Battese (1974) variance component models with two alternative proxies for Attainment Discrepancy for Social Aspiration, which they name *Social Comparison*. Due to the fact that the signs of the coefficients are the same, we report results only for their Model 4. Unfortunately, SAS does not estimate the TSCSREG procedure when sample sizes are large, which is the case with the simulated data. Therefore, sample histories of randomly selected 5,000 agents in periods 2-30 were used for analysis.

Table 3. Estimating Organizational Change: A Comparison of Greve (1998) with Replication on the Data Where True Causality is RC

	Hypothesized Relationships in the benchmark study[32]	Model	
		Greve (1998) Pooled logit	Replication of (1) on simulated data where true causality is RC [33]
		1	*2*
Constant		-2.085	-1.049*
Positive Attainment Discrepancy for Historical Aspiration $(I_{it-1}\text{-}A_{it-1})_{>0}$	– –	-0.168*	-11.211*
Negative Attainment Discrepancy for Historical Aspiration $(I_{it-1}\text{-}A_{it-1})_{<0}$	–	-0.054*	-9.971*
Positive Attainment Discrepancy for Social Aspiration $(I_{it-1}\text{-}B_{it-1})_{>0}$	– –	-0.187*	-0.968*
Negative Attainment Discrepancy for Social Aspiration $(I_{it-1}\text{-}B_{it-1})_{<0}$	–	-0.048*	-0.785*
Negative inconsistency, (N_{it-1})		0.020	20.595*
Log Likelihood		-5,832.65*	-316,981.24*
Number of observations		16,294	560,000

*p <0.001

[32] Model 5 in Greve (1998, Table 2) is used as a benchmark for comparison.

[33] 560,000 observations are used for analysis because strategies were assigned in period 1 and the derivation of social aspirations (described in section 3.2) is allowable after period 2.

Table 4. Type I and Type II Errors When Scenarios are Pure Maximizing, Pure Satisficing, or Mixed Behavior

	True Causality		
	Maximizing	**Satisficing**	**Mixed**
Type I error (false positive)	**A** *Confirm BT when it is not true*	**C** *Confirm RC when it is not true*	
Type II error (false negative)	**B** *Reject RC when it is true*	**D** *Reject BT when it is true*	**E** *Reject BT and/or RC when they are true*

Table 5. Estimation of the Data Where True Causality is RC

	Model				
	Pooled logit			Random-effects logit	
	1	*2*	*3*	*4*	*5*
Constant	-1.049**	-1.018**	-1.029**	-1.263**	-1.264**
Opportunity Cost		19.447**	19.837**		25.014**
Positive Attainment Discrepancy for Historical Aspiration $(I_{it-1}-A_{it-1})_{>0}$	-11.211**		-15.555**	-10.244**	-16.087**
Negative Attainment Discrepancy for Historical Aspiration $(I_{it-1}-A_{it-1})_{<0}$	-9.971**		-12.553**	-13.756**	-17.499**
Positive Attainment Discrepancy for Social Aspiration $(I_{it-1}-B_{it-1})_{>0}$	-0.968**		1.090**	0.529*	3.113**
Negative Attainment Discrepancy for Social Aspiration $(I_{it-1}-B_{it-1})_{<0}$	-0.785**		3.482**	-1.041**	4.498**
Negative inconsistency, (N_{it-1})	20.595**		12.078**	28.225**	18.056**
Log Likelihood	-316,981**	-305,593**	-291,920**	-308,671**	-278,882**
Likelihood-ratio test vs. model *1*, χ^2			27,346.90**		
Likelihood-ratio test vs. model *2*, χ^2			50,123.24**		
Likelihood-ratio test vs. model *4*, χ^2					59,578.84**
Likelihood-ratio test vs. respective pooled model, χ^2				17,000**	26,000**
Number of observations	560,000	560,000	560,000	560,000	560,000
Effectiveness in Eliminating Type I Error (false positive) for RC			n.a.		n.a.
Effectiveness in Eliminating Type I Error (false positive) for BT			partially effective		completely effective
Effectiveness in Eliminating Type II Error (false negative) for RC			completely effective		completely effective
Effectiveness in Eliminating Type II Error (false negative) for BT			n.a.		n.a.

* $p < 0.05$; ** $p < 0.001$

n.a. = not applicable

Table 6. Estimation of the Data Where True Causality is BT

	Model				
	Pooled logit			Random-effects logit	
	1	*2*	*3*	*4*	*5*
Constant	-0.934*	-0.678*	-0.942*	-0.934*	-0.942*
Dummy Opportunity Cost		-0.007	0.011		0.011
Positive Attainment Discrepancy for Historical Aspiration $(I_{it-1}-A_{it-1})_{>0}$	-8.982*		-8.982*	-8.985*	-8.986*
Negative Attainment Discrepancy for Historical Aspiration $(I_{it-1}-A_{it-1})_{<0}$	-4.900*		-4.900*	-4.902*	-4.901*
Positive Attainment Discrepancy for Social Aspiration $(I_{it-1}-B_{it-1})_{>0}$	-5.849*		-5.848*	-5.849*	-5.849*
Negative Attainment Discrepancy for Social Aspiration $(I_{it-1}-B_{it-1})_{<0}$	-4.111*		-4.111*	-4.112*	-4.112*
Negative inconsistency, (N_{it-1})	16.482*		16.482*		16.487*
Log Likelihood	-253,292*	-358,151	-253,291*	-253,291*	-253,290*
Likelihood-ratio test vs. model *1*, χ^2			1.85		
Likelihood-ratio test vs. model *4*, χ^2					1.85
Likelihood-ratio test vs. respective pooled model, χ^2				0.52	0.55
Number of observations	560,000	560,000	560,000	560,000	560,000
Effectiveness in Eliminating Type I Error (false positive) for RC			completely effective		completely effective
Effectiveness in Eliminating Type I Error (false positive) for BT			n.a.		n.a.
Effectiveness in Eliminating Type II Error (false negative) for RC			n.a.		n.a.
Effectiveness in Eliminating Type II Error (false negative) for BT			completely effective		completely effective

*p < 0.001;

n.a. = not applicable

Table 7. Estimation of the Data Where True Causality is Mixed Behavior

	Model				
	Pooled logit			Random-effects logit	
	1	*2*	*3*	*4*	*5*
Constant	-0.107*	-0.117*	-0.100*	-0.439*	-0.432*
Dummy Opportunity Cost		-0.000	-0.010		-0.009
Positive Attainment Discrepancy for Historical Aspiration $(I_{it-1}-A_{it-1})_{>0}$	-7.059*		-7.059*	-6.554*	-6.554*
Negative Attainment Discrepancy for Historical Aspiration $(I_{it-1}-A_{it-1})_{<0}$	-3.026*		-3.025*	-4.909*	-4.909*
Positive Attainment Discrepancy for Social Aspiration $(I_{it-1}-B_{it-1})_{>0}$	-3.165*		-3.165*	-2.419*	-2.419*
Negative Attainment Discrepancy for Social Aspiration $(I_{it-1}-B_{it-1})_{<0}$	-2.462*		-2.462*	-3.664*	-3.664*
Negative inconsistency, (N_{it-1})	9.560*		9.561*	12.079*	12.080*
Log Likelihood	-307,280*	-387,195	-307,279*	-288,489*	-288,489*
Likelihood-ratio test vs. model *1*, χ^2			1.62		
Likelihood-ratio test vs. model *6*, χ^2					1.21
Likelihood-ratio test vs. respective pooled model, χ^2				38,000*	38,000*
Number of observations	560,000	560,000	560,000	560,000	560,000
Effectiveness in Eliminating Type I Error (false positive) for RC			n.a.		n.a.
Effectiveness in Eliminating Type I Error (false positive) for BT			n.a.		n.a.
Effectiveness in Eliminating Type II Error (false negative) for RC			ineffective		ineffective
Effectiveness in Eliminating Type II Error (false negative) for BT			completely effective		partially effective

* $p < 0.001$;

n.a. = not applicable

Table 8. Non-nested Test of Alternative Hypotheses[34]

True Causality	Null Hypothesis	Vuong's (1989) Test		Effectiveness in Eliminating			
		z-statistic	z-statistic corrected[35]	Type I Error for RC	Type I Error for BT	Type II Error for RC	Type II Error for BT
RC	RC is not better than BT	34.249*	101.856*	n.a.	n.a.	completely effective	n.a.
	BT is not better than RC	-34.249	-101.856	n.a.	completely effective	n.a.	n.a.
BT	BT is not better than RC[36]	296.371*	240.383*	n.a.	n.a.	n.a.	completely effective
	RC is not better than BT[36]	-296.371	-240.383	completely effective	n.a.	n.a.	n.a.
Mixed	BT is not better than RC [36]	249.542*	187.685*	n.a.	n.a.	n.a.	completely effective
	RC is not better than BT[36]	- 249.542	-187.685	n.a.	n.a.	n.a.	n.a.

*p < 0.001;

n.a. = not applicable

[34] Because the true causality is not known *ex ante* in an empirical setting, two null hypotheses were tested for each of the three scenarios. The two tests performed on the same data generate statistics of the same absolute value but opposite signs and have p-values, summing to one. While testing both null hypotheses is redundant in the case of known causality, this step is exercised in the paper to illustrate its empirical implementation. Based on the p-values, one of the following 3 outcomes may potentially obtain: (1) neither of the theories is diagnosed to be better than another; (2) BT is diagnosed to be better than RC; or (3) RC is diagnosed to be better than BT.

[35] The Schwartz (1978) correction is used.

Table 9. Robustness of Support to BT to the Value of α under Alternative Scenarios[37]

		# of Tested Values of α	# of Complete Confirmations of BT	Proportion of Complete Confirmations of BT	# of Confirmations of the Effects of Social Aspirations	Proportion of Confirmations of the Effects of Social Aspirations	# of Confirmations of the Effects of Historical Aspirations	Proportion of Confirmations of the Effects of Historical Aspirations
True Causality	RC	14	0	0.000	8	0.571	6	0.429
	BT	14	12	0.857	12	0.857	14	1.000
	Mixed	14	11	0.785	12	0.857	13	0.929

[37] Robustness of α is investigated in in the range of [0,0.7] with a step of 0.05. A confirmation to BT is considered complete, if all of the specific predictions on the effects of aspirations on change (Hypotheses 1A, 1B, 2A, and 2B in Greve, 1998) are jointly confirmed. A confirmation of the effects of Social Aspirations obtains, if the predictions on the effects of social aspirations on change (Hypotheses 1A and 1B in Greve, 1998) are jointly confirmed. A confirmation of the effects of Historical Aspirations obtains, if the predictions on the effects of historical aspirations on change (Hypotheses 2A and 2B in Greve, 1998) are jointly confirmed.

Table 10. Summary of Effectiveness of Recommendations

	Effective at Eliminating											
	Type I Error when Agents						Type II Error when Agents					
	Satisfice		Maximize		Mixed		Satisfice		Maximize		Mixed	
Tested Theory	RC	BT	RC	BT	RC	BT	RC	BT	RC	BT	RC	BT
Recommendation 1: Deliberate Control for Maximizing or Satisficing	✔	n.a.	n.a.	✔	n.a.	n.a.	n.a.	✔	✔	n.a.		✔
Recommendation 2: Test of Non-Nested Alternative Hypotheses	✔	n.a.	n.a.	✔	n.a.	n.a.	n.a.	✔	✔	n.a.	✔	✔
Recommendation 3: Comparative Robustness of Alternative Theories	✔	n.a.	n.a.	✔	n.a.	n.a.	n.a.	✔	✔	n.a.	✔	✔

n.a. = not applicable; ✔ denotes instances where recommendations are effective at eliminating errors in tests for the presence of the theory.