MANAGEMENT SCIENCE
Vol. 55, No. 1, January 2009, pp. 4–15
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5501 | 0004

DOI 10.1287/mnsc.1080.0907
© 2009 INFORMS

Metrics—When and Why Nonaveraging Statistics Work

Steven M. Shugan, Debanjan Mitra
Warrington College of Business Administration, University of Florida, Gainesville, Florida 32611
{steven.shugan@cba.ufl.edu, deb.mitra@cba.ufl.edu}

Good metrics are well-defined formulae (often involving averaging) that transmute multiple measures of raw numerical performance (e.g., dollar sales, referrals, number of customers) to create informative summary statistics (e.g., average share of wallet, average customer tenure). Despite myriad uses (benchmarking, monitoring, allocating resources, diagnosing problems, explanatory variables), most uses require metrics that contain information summarizing multiple observations. On this criterion, we show empirically (with people data) that although averaging has remarkable theoretical properties, supposedly inferior nonaveraging metrics (e.g., maximum, variance) are often better. We explain theoretically (with exact proofs) and numerically (with simulations) when and why. For example, when the environment causes a correlation between observed sample sizes (e.g., number of past purchases, projects, observations) and latent underlying parameters (e.g., the likelihood of favorable outcomes), the maximum statistic is a better metric than the mean. We refer to this environmental effect as the Muth effect, which occurs when rational markets provide more opportunities (i.e., more observations) to individuals and organizations with greater innate ability. Moreover, when environments are adverse (e.g., failure-rich), nonaveraging metrics correctly overweight favorable outcomes. We refer to this environmental effect as the Anna Karenina effect, which occurs when less-favorable outcomes convey less information. These environmental effects impact metric construction, selection, and employment.

Key words: metrics; metric selection; metric evaluation; summary statistics; environmental effects; natural correlations; forecasting; benchmarking; monitoring; statistical biases; choosing explanatory variables
History: Accepted by Jagmohan S. Raju, marketing; received June 5, 2006. This paper was with the authors $10\frac{1}{2}$ months for 2 revisions. Published online in *Articles in Advance* September 30, 2008.

1. Our Objectives

Metrics are widely used by both practitioners and academics (Gupta et al. 2004, Gupta and Zeithaml 2006). Unlike models, metrics are well-defined formulae (often involving averaging) that transmute multiple measures of raw numerical performance to create informative summary statistics. For example, typical marketing metrics create informative firm-level summary statistics (e.g., average share of wallet, average attrition rate, average customer tenure, average segment share, average preference share, average time to market, average customer retention, average market penetration, average percent awareness, average distribution intensity) from raw performance measures (dollar sales, time to market, number of customers) using well-defined undisputed formulae. Hence, metrics summarize data consisting of observations on some units (individuals, firms, customers), with multiple observations per unit (earnings, citations, referrals, dividends, wins).

Metrics have many uses. Sometimes, metrics are explanatory variables in comprehensive and diverse quantitative models (Hauser 1998, 2001; Cohen et al. 2000; Rust et al. 2004). Other times, we use metrics for benchmarking, monitoring, improving processes, selecting options, creating accountability, allocating resources, diagnosing problems, and so on. For example, we might evaluate cars with an objective reliability metric intending to capture information on relative durability, quality, expected customer satisfaction, and brand reputation. Good metrics should capture the information in multiple observations, i.e., explain variance across observations. For example, when assessing people, metrics computed using observed past performance should explain variance in performance across people. That information has many uses. For example, we can use good people metrics (e.g., Scholastic Assessment Test scores) to select people (e.g., admission decisions) and to predict individual future outcomes (e.g., college grades, starting salaries).

No prior research has developed the necessary theory for constructing, selecting, and employing metrics, particularly for nonaveraging metrics. We do that. Our theory predicts when particular metrics are superior. Moreover, we provide empirical evidence to support our theory. Finally, despite the remarkable properties of ubiquitous averaging statistics, we discover conditions when supposedly inferior nonaveraging statistics are better metrics and argue why most researchers have underestimated their usefulness.

2. Some Interesting Empirical Findings

2.1. Measuring the Information in a Metric

Despite the intuitive appeal of proximity, the statistical yardstick for information is correlation or, more generally, statistical dependence. Correlation (i.e., shared variance) measures the information in one variable about another regardless of the absolute difference (proximity) between the two variables. For example, although the mean metric might be proximal to an unknown parameter of interest, proximity fails to guarantee more information. Only correlation allows prediction of one variable from another. Averaging appears intuitive, but is not necessarily related to correlation. Unfortunately, correlations involve more mathematically complex expressions than proximity (i.e., bias).

2.2. Averaging and Nonaveraging Metrics

The vast majority of metrics employ some averaging. This tendency and reverence for averaging is natural given the remarkable properties and the pivotal role of the sample mean in the development of modern statistics (Ostasiewicz and Ostasiewicz 2000). These properties include being nearly normally distributed (Gibbons and Chakraborti 2003, p. 2), a consistent estimator, a minimal sufficient statistic (at least when the variance is known) for the entire natural exponential family of distributions (e.g., normal, exponential, gamma, inverse Gaussian, negative binomial, binominal, logarithmic, Poisson, gamma, Tweedie, etc.), and being an unbiased estimator of the population mean (regardless of the distribution or the sample size). Moreover, regardless of the population's distribution, the sample mean's distribution (i.e., the sampling distribution) approaches a normal distribution as the sample size increases. The sample mean is the method-of-moments estimator, the least-squares and maximum-likelihood estimator for the population mean for the normal distribution and exponential distribution (Lindsey 1997).

An erroneous reason for preferring averaging metrics is the incorrect belief that many nonaveraging metrics fail to consider all available information. However, for example, although the maximum appears to be a single observation, in fact, computing a maximum requires information from all the observations.

Despite the desirable properties and overwhelming popularity of averaging statistics as metrics, we evaluate four nonaveraging statistics: the maximum, minimum, variance, and count (i.e., number of observations). Common data consist of a recent outcome (earnings, citations, referrals, dividends, wins) for a unit of analysis (firm, author, product) and a background consisting of multiple past outcomes. Our unit of analysis is an individual. For example, an individual author might have a current publication and a background consisting of multiple past publications.

We ask which metric best captures the information in those backgrounds based on how well each metric explains recent outcomes across people. Beyond people, our analysis is relevant for any data containing multiple observations for each unit of analysis (e.g., firms, departments, campaigns, products, cities, areas, and so on).

2.3. Empirical Findings for Three Industries

We use publicly available people data from three very different occupations—baseball batters, movie cast members, and academic authors. Our data capture different units of analysis (games, movies, articles), different outcomes (number of hits in a game, box office, citations), different industries (professional sports, motion pictures, scholarly publishing), different organizations (individual, team, a mixture), different types of activities (athletic, artistic, academic), different marketing efforts (none, large, intermediate), and different competition (between batters and pitchers, between studios, between journals).

2.3.1. Baseball. For baseball, outcomes are the number of hits in each of the 2006 postseason games played by each St. Louis Cardinals batter, reported by http://www.mlb.com, the official website of major league baseball. Our data include 113 hits by the 14 St. Louis Cardinals batters who could play in any or all of the 16 postseason games.

We compute the averaging and nonaveraging metrics for each individual batter's background, i.e., hits in each observed game in the postseason except the most recent outcome (i.e., the last game), which becomes our holdout game. Finally, using the holdout game, we compute the correlation ρ for each of the metrics, across all batters. To illustrate, suppose a batter plays three postseason games with hits of 0, 2, and 1, respectively. The most recent (holdout) outcome is 1 hit. For the two background games, with 0 and 3 hits, the metrics for this batter are mean $= 1.5$, maximum $= 3$, minimum $= 0$, variance $= 2.25$, and count $= 2$. We then compute, across all batters, the correlation of each batter's metric with the holdout outcome (i.e., 1 for this batter). For the mean metric,

$$\rho_{\text{mean}} = \frac{(14(1.5\cdot 1+\cdots)-(1.5+\cdots)(1+\cdots))}{\sqrt{(14(1.5^2+\cdots)-(1.5+\cdots)^2)(14(1^2+\cdots)-(1+\cdots)^2)}},$$

where the additional numbers come from other batters and 14 is the number of St. Louis Cardinals batters. Larger $|\rho|$ implies more information (variance explained) in the metric.

2.3.2. Movies. According to the Internet Movie Database at http://www.imdb.com, there were 200 unique lead cast members (actors) who had at least one movie in 2004. We chose 2004 to ensure complete movie life cycles. For these actors, individual backgrounds are the total gross revenues (i.e., box office) for each movie in their career, a total of 2,932 prior movies. We compute the averaging and nonaveraging metrics for each actor from all movies made by that actor (i.e., their background) excluding the holdout movie. Finally, using the holdout movie, we compute the correlation ρ for each of the metrics, across all actors. For example, suppose an actor makes one \$24 million movie in 2000, two \$6 million movies in 2003, and one \$30 million movie in 2004. Excluding the holdout outcome, \$30 million, the metrics using the remaining 3 movies (with \$24, \$6, \$6 box office) are mean = 12, maximum = 24, minimum = 6, variance = 72, and count = 3. We then compute, across all actors, the correlation ρ of each metric with the holdout outcome.

2.3.3. Publishing. Authors who publish in scholarly journals have backgrounds consisting of observed citations for each of their past articles. Specifically, we use the 190 articles published by the University of Pennsylvania Wharton Marketing Department faculty in five prestigious journals—*Journal of Marketing*, *Journal of Marketing Research*, *Marketing Science*, *Journal of Consumer Research*, and *Management Science* through 2000, giving each article an adequate citation window (Price 1970, Peritz 1982, Stremersch and Verhoef 2005). Google Scholar (http://scholar.google.com) reveals each article's citations and year of publication. We compute each metric from each faculty member's background (i.e., citations for each of their published articles except the last article). For example, suppose a faculty member publishes two articles in year 1997 (in May and June), one article in 1999, and one in 2000. Citations are 20, 110, 80, and 108, respectively. The metrics for this faculty member, computed for the 3 background articles (with 20, 100, 80 citations), are mean = 70, maximum = 110, minimum = 20, variance = 1,400, and count = 3. We compute, across all faculty members, ρ for each metric with the holdout observation (108 citations for this individual).

2.4. Correlations

Table 1 reports each metric's observed correlation ρ computed from background observations with the holdout observation, for our batters, movie actors, and authors. The next section reveals that most of these correlations are highly significant and often much larger than expected given random sampling variation.

Table 1 Correlations of Different Metrics with Holdout Outcome

Metric	Baseball batters	Movie actors	Academic authors
Mean	0.55	0.29	0.35
Maximum	0.62	0.39	0.46
Minimum	N.A.	−0.01	−0.12
Variance	0.57	0.41	0.33
Count	0.44	0.13	0.13

2.5. A Monte Carlo Simulation

Using a standard Monte Carlo Simulation (MCS), we randomly sample data from a normal distribution with the observed characteristics in the data set (i.e., actual observed means and variances). For the movie data, MCS randomly assigns each actor a mean from a normal distribution with the same mean and variance as the observed box-office distribution in the data set. Using each actor's assigned mean, MCS then randomly generates that actor's outcomes using a normal distribution and a random error term (with the observed mean and variance in the data set). Finally, MCS randomly determines the number of observations in each actor's background from a normal distribution with the mean and variance of the observed number of observations in the data set. One hundred replications reveal the sampling distributions for testing purposes. See Table 2.

Unsurprisingly, Table 2 shows that the mean metric's expected (by chance) and actual observed correlations are similar, i.e., 29%. However, most of the nonaveraging metrics (maximum, variance, and count) have dramatically better (at the 0.0000 significance level) correlations than expected. The minimum is significantly worse. Moreover, although the nonaveraging metrics should seldom outperform the mean metric, at least one nonaveraging metric does outperform in all three different industries. For example, we expect (based on simulations) that the maximum metric will rarely (6%) outperform the mean metric. If 6% is typical, there is only a 0.06^3 or 0.00022

Table 2 Monte Carlo Simulation of Expected Correlations for Different Metrics for the Movie Data

Metric	Actual correlation (from Table 1)	Average expected correlation	Actual vs. expected (two-tail p-values)	Actual better than mean? (expected incidence)	Actual best? (expected incidence)
Mean	0.29	0.29	0.8060	N.A.	No
Maximum	0.39	0.19	0.0000	Yes (6%)	No
Minimum	−0.01	0.08	0.0000	No	No
Variance	0.41	0.16	0.0000	Yes (8%)	Yes (8%)
Count	0.13	−0.00	0.0000	No	No

probability of finding that the maximum metric outperforms the mean metric in three of three industries. Still, we find that. In fact, simulations reveal the mean metric should almost always (92%) outperform all of our nonaveraging metrics. If 92% is typical, the probability that the mean metric will be best in at least one of the three industries is $1-(1-0.92)^3$, or 0.99949. Still, we do not find that. Moreover, for the movie data, the maximum metric improves the squared correlation by 80.1%, i.e., $(0.39^2-0.29^2)/0.29^2$, over the mean metric.

Parenthetically, further analysis reveals that trend in the data fails to explain these findings because these findings replicate when predicting a randomly chosen observation (rather than the last) as the holdout observation. We now provide a simple theoretical model that helps explain these surprising empirical findings.

3. Theoretical Model

We model individuals who each have a background with possibly multiple observations. As noted earlier, our model is also appropriate for units of analysis beyond individuals (e.g., organizations or categories), where we have possibly multiple observations for each unit. Each observation j is one of $J \geq 1$ possible outcomes, i.e., $j = 0, 1, 2, \ldots, J-1$, ordered in increasing favorability. To clarify, we use a running baseball example. Suppose the most hits in a game for any batter is 3, then 4 outcomes ($j = 0, 1, 2, 3$ hits) are possible and $J = 4$. Let outcomes j be binomially distributed, so the probability of observing outcome j is $\binom{J-1}{j} \cdot q_t^j(1-q_t)^{J-1-j}$, where $0 < q_t < 1$ captures a type t individual's innate ability (the probability of a hit when at bat) and $t = 1, 2, \ldots, T$ denote the different types of individuals (e.g., slap hitters, power hitters). Each background consists of $n \geq 1$ observations, i.e., $n = 1, 2, \ldots, N$, where N is the largest number of observations in any background. For example, if the largest number of games played in any batter's background is 3, then $N = 3$, so backgrounds consist of $n = 1, 2$, or 3 games. Let n be binomially distributed, so the probability of observing a background of n observations is $\binom{N-1}{n-1} p_t^{n-1}(1-p_t)^{N-n}$, where $0 < p_t < 1$ captures the opportunities for a type t individual (the probability a batter plays in a game). Finally, let $R_t > 0$ denote the fraction of type t individuals where $\sum_{t=1}^{T} R_t = 1$. In sum, we allow for differences in the following: the number of possible outcomes (J), individual types (t), the actual outcome observed (j), the innate ability of a type t individual (q_t), the number of possible observations in a background (N), the actual number of observations in a background (n), the opportunities of a type t individual (p_t), and the fraction of type t individuals (R_t).

Note that (1) q_t, p_t, and R_t are not observed, i.e., latent; (2) $N \geq 1$ because $n \geq 1$; (3) the duration between background observations could vary, e.g., actors could have three movies in one year and a fourth movie two years later; and (4) we restrict the duration window for the computation (e.g., observations during a two-year period) so that individuals retain the same type. This restriction allows the metric to capture differences across individuals rather than changes within individuals, e.g., innate abilities. However, we advocate moving duration windows. For example, with nine years of data, we could compute the maximum metric using seven moving three-year windows to capture potential changes within individuals or other units of analysis.

3.1. Specific Cases

We begin by providing specific conditions when nonaveraging metrics outperform averaging metrics. These specific conditions will inspire more general intuition associated with the more general, albeit more complex, conditions derived in subsequent sections.

Consider the case of 2 possible outcomes ($J = 2$) called F and S, where $F < S$. As proved later, our two-outcome analysis is independent of the units of S and F because covariance is bilinear, making our correlations unitless. For example, if $F = \alpha_1$ for outcomes under \$X and $S = \alpha_1 + \alpha_2$ for outcomes over \$X, then our equations are exact for any arbitrary constants $\alpha_1, \alpha_2 > 0$. Let the population be equally divided between two individual types where $q_1 = 1/4$ and $q_2 = 3/4$ are the probabilities of outcome S for individual types 1 and 2, respectively. Backgrounds have 1 or 2 observations, i.e., $N = 2$. Let p_1 and p_2 be the probabilities of a background having two background observations (versus one) for individual types 1 and 2, respectively. We compare ρ^2, the squared correlation of the metric with the holdout observation, for the mean metric (given its prevalence and remarkable statistical properties) with the ρ^2 for the maximum metric (an exemplar of nonaveraging metrics). Remember, however, the mean metric's reputation involves proximity rather than correlation.

For the mean metric, as shown later (Table 4), $\rho^2_{\text{mean}} = 1/(16-3p_1-3p_2)$ and $\rho^2_{\max} = (3p_1-3p_2-8)^2/(4(16+3p_1+3p_2)(16-3p_1-3p_2))$ for the maximum metric, so $(\rho^2_{\text{mean}}-\rho^2_{\max})/\rho^2_{\max} = (20p_1-12p_2+6p_1p_2-3p_1^2-3p_2^2)/((8-3p_1+3p_2)^2/3) = Q$. Note that Q is independent of the units of S and F. We now prove that the maximum metric outperforms the mean metric (i.e., $\rho^2_{\max} > \rho^2_{\text{mean}}$) when $p_1 < p_1^*$, where $p_1^* = p_2 + ((10-2\sqrt{25+6p_2})/3)$. The proof follows: (1) $\rho^2_{\max} > \rho^2_{\text{mean}}$, if and only if the numerator of $Q < 0$. (2) There is only one admissible p_1 making the numerator zero, i.e., $p_1 = p_1^*$, because $0 < p_t < 1$. (3) The first derivative of the numerator with respect to p_1 is positive, i.e., $20-6p_1+6p_2>0$. (4) Therefore, the numerator is strictly increasing in p_1, decreasing in p_2 (with analogous reasoning), p_1^* is the only solution, and $p_1<p_1^*$ implies $\rho^2_{\text{max}}>\rho^2_{\text{mean}}$.

Unlike traditional statistical analysis, our underlying parameters influence the sample size, so the sample size conveys information about q_t. The number of observations in an individual's background n could, for example, convey information about that individual's innate ability. The number of observed new products from a firm might convey information about the management's ability to innovate. Similarly the number of brands in a product category might convey information about the category's potential. In each case, nonaveraging metrics better capture this information.

However, even when the sample size n is independent of the likelihood of favorable outcomes or type (e.g., $p_1=p_2$), the maximum metric can still outperform the mean metric. We prove this proposition with a simple example. Suppose that for both individual types, there is an equal likelihood of F and S, i.e., $p_1=p_2=1/2$. If q_1 is slightly less than q_2, say, $q_1=q_2-0.01$, then we can derive $\rho^2_{\text{mean}}-\rho^2_{\text{max}}$. We find that $\rho^2_{\text{mean}}-\rho^2_{\text{max}}$ is increasing in q_2 within the admissible region and zero at $q_2^*=0.21$. Hence, for $q_1<q_2<q_2^*$, the maximum metric outperforms the mean metric when types are equally frequent.

Traditional statistical analysis fails to recognize the nature of the environment. When we know that S is rare (e.g., $q_1<0.2, q_2<0.21$), then we use that knowledge to better summarize that information with a nonaveraging metric. When record albums, for example, reach the very rare RIAA certification of diamond, that certification conveys a great deal of information about the artist's innate ability without knowledge of that artist's other outcomes. The rare blockbuster drug (>$1 billion) might provide important information about the parent pharmaceutical company's innate ability regardless of possible failures. Similarly, a Clio award provides information about an ad agency's innate ability. The maximum metric recognizes these important signals within a noisy failure-rich environment.

3.2. The State Space

The last section provided two specific cases. One case had specific probabilities for different outcomes, but general probabilities for observing different sample sizes (i.e., more observations in a background). The other case had specific probabilities for different sample sizes, but general probabilities for different outcomes. We now consider general probabilities for both sample sizes and outcomes. Our subsequent theorems reveal that two individual types are sufficient to reveal general reasons when and why different metrics perform better. Our theorems also show that two possible outcomes (e.g., less than X and greater than X) are sufficient to produce all of our key findings regarding why and when particular metrics are better. Our findings reveal both intuition and testable predictions applicable to many situations. Nevertheless, we later explore generality through simulations.

We now explain how to compute ρ^2 for different metrics. Although we are concerned with individual metrics, our analysis must define the state space of observations by the possible joint events consisting of a background (one or two observations) and holdout outcome for a particular type of individual. With $J=2$, $N=2$, and $T=2$, Table 3 provides the 24 possible joint events. Remember, q_1 and q_2 are the probability of an S outcome (versus an F outcome) for type 1 and type 2 individuals, respectively. Moreover, p_1 and p_2 are the probability of a background having two observations (versus one observation) for type 1 and type 2 individuals, respectively. Finally, R_1 and $R_2=1-R_1$ are the fraction of type 1 and type 2 individuals, respectively. Note that $q_1\neq q_2$ so that different types have different probabilities of an S outcome. Table 3 provides the probability of observing each of the 24 joint events. Here are some clarifying computational specifics for several rows in Table 3.

Row 1 (*Event* 1). Event 1 is a type 1 individual with a background of one observation S and a holdout outcome S. This event's probability is the probability the individual is type 1 (R_1) times the probability of having one background observation $(1-p_1)$ times the probability that observation is an S outcome (q_1) times the probability the holdout observation is an S outcome (q_1), so the probability of event 1 is $R_1(1-p_1)q_1q_1$.

Row 10 (*Event* 10). Event 10 is a type 1 individual with a background of two observed outcomes (S and F) and a holdout outcome F. This event's probability is the probability the individual is type 1 (R_1) times the probability of having two background observations p_1 times the probability those observations are S and F outcomes $(q_1(1-q_1))$ times the probability the holdout observation is an F outcome $(1-q_1)$, so the probability of event 10 is $R_1p_1q_1(1-q_1)(1-q_1)$.

Row 17 (*Event* 17). Event 17 is a type 2 individual with a background of two outcomes (F and S) and a holdout outcome S. This event's probability is the probability the individual is type 2 $(1-R_1)$ times the probability of having two background observations p_2 times the probability those observations are F and S outcomes $((1-q_2)q_2)$ times the probability the holdout observation is an S outcome (q_2), so, the probability of event 17 is $(1-R_1)p_2(1-q_2)q_2q_2$.

Table 3 Events, Likelihoods, and Computation of Metrics

Event	Type	Background	Mean	Maximum	Minimum	Variance	Count	Holdout outcome	Probability (P_{event})
1	1	$[S]$	S	S	S	0	1	S	$R_1(1-p_1)q_1q_1$
2	1	$[F]$	F	F	F	0	1	S	$R_1(1-p_1)(1-q_1)q_1$
3	1	$[S,S]$	$(S+S)/2$	S	S	0	2	S	$R_1p_1q_1q_1q_1$
4	1	$[S,F]$	$(S+F)/2$	S	F	$(S-F)^2/4$	2	S	$R_1p_1q_1(1-q_1)q_1$
5	1	$[F,S]$	$(F+S)/2$	S	F	$(S-F)^2/4$	2	S	$R_1p_1(1-q_1)q_1q_1$
6	1	$[F,F]$	$(F+F)/2$	F	F	0	2	S	$R_1p_1(1-q_1)(1-q_1)q_1$
7	1	$[S]$	S	S	S	0	1	F	$R_1(1-p_1)q_1(1-q_1)$
8	1	$[F]$	F	F	F	0	1	F	$R_1(1-p_1)(1-q_1)(1-q_1)$
9	1	$[S,S]$	$(S+S)/2$	S	S	0	2	F	$R_1p_1q_1q_1(1-q_1)$
10	1	$[S,F]$	$(S+F)/2$	S	F	$(S-F)^2/4$	2	F	$R_1p_1q_1(1-q_1)(1-q_1)$
11	1	$[F,S]$	$(F+S)/2$	S	F	$(S-F)^2/4$	2	F	$R_1p_1(1-q_1)q_1(1-q_1)$
12	1	$[F,F]$	$(F+F)/2$	F	F	0	2	F	$R_1p_1(1-q_1)(1-q_1)(1-q_1)$
13	2	$[S]$	S	S	S	0	1	S	$(1-R_1)(1-p_2)q_2q_2$
14	2	$[F]$	F	F	F	0	1	S	$(1-R_1)(1-p_2)(1-q_2)q_2$
15	2	$[S,S]$	$(S+S)/2$	S	S	0	2	S	$(1-R_1)p_2q_2q_2q_2$
16	2	$[S,F]$	$(S+F)/2$	S	F	$(S-F)^2/4$	2	S	$(1-R_1)p_2q_2(1-q_2)q_2$
17	2	$[F,S]$	$(F+S)/2$	S	F	$(S-F)^2/4$	2	S	$(1-R_1)p_2(1-q_2)q_2q_2$
18	2	$[F,F]$	$(F+F)/2$	F	F	0	2	S	$(1-R_1)p_2(1-q_2)(1-q_2)q_2$
19	2	$[S]$	S	S	S	0	1	F	$(1-R_1)(1-p_2)q_2(1-q_2)$
20	2	$[F]$	F	F	F	0	1	F	$(1-R_1)(1-p_2)(1-q_2)(1-q_2)$
21	2	$[S,S]$	$(S+S)/2$	S	S	0	2	F	$(1-R_1)p_2q_2q_2(1-q_2)$
22	2	$[S,F]$	$(S+F)/2$	S	F	$(S-F)^2/4$	2	F	$(1-R_1)p_2q_2(1-q_2)(1-q_2)$
23	2	$[F,S]$	$(F+S)/2$	S	F	$(S-F)^2/4$	2	F	$(1-R_1)p_2(1-q_2)q_2(1-q_2)$
24	2	$[F,F]$	$(F+F)/2$	F	F	0	2	F	$(1-R_1)p_2(1-q_2)(1-q_2)(1-q_2)$

Note that, in actual data, some events are indistinguishable because observed outcomes are identical. Also, the component probabilities for each event depend on the unobserved individual type. Also, note that, although these computations are for independent events, many forms of dependency (e.g., upward and toward trends) would only strengthen our results and favor nonaveraging metrics. We now compute the different metrics.

3.3. Computing the Metrics

Using the event probabilities in Table 3, we compute each metric across all 24 events assuming (without loss in generality) that $F < S$. For example, event 1 consists of a background of S. Therefore, the mean metric is S, the count metric is 1, the maximum metric is S, and so on. For event 16, with background $[S, F]$, the mean metric is $(S + F)/2$, the maximum metric is S, the minimum metric is F, the variance metric is $(S - F)^2/4$, and the count metric is 2. Here, bilinear covariances make our analysis *independent* of the units of F and S.

3.4. Deriving the Squared Correlations for Each Metric

This section shows how to compute the correlation ρ^2_{mean} between the mean metric and the holdout outcome. Analogous steps produce $\rho^2_{\max}$, ρ^2_{variance}, and ρ^2_{count}. Let E[·] denote the expectations operator, i.e., the expected value of a metric across all events. Then,

$$\rho^2_{\text{metric}} = \left(\frac{\mathrm{E}[\text{metric}\times\text{outcome}]-\mathrm{E}[\text{metric}]\mathrm{E}[\text{outcome}]}{\sqrt{\mathrm{E}[\text{metric}^2]-\mathrm{E}[\text{metric}]^2}\sqrt{\mathrm{E}[\text{outcome}^2]-\mathrm{E}[\text{outcome}]^2}}\right)^2.$$

We now compute each term in ρ^2_{mean}. The expected mean metric follows. See column 4 in Table 3.

$$\begin{aligned}\mathrm{E}[\text{mean metric}] &= P_1S + P_2F + P_3\left(\frac{S+S}{2}\right) + P_4\left(\frac{S+F}{2}\right) + P_5\left(\frac{F+S}{2}\right)\\ &\quad + P_6\left(\frac{F+F}{2}\right) + P_7S + P_8F + P_9\left(\frac{S+S}{2}\right) + \cdots .\end{aligned}$$

Note that $P_1, P_2, P_3, \ldots$ are the event probabilities in the last column of Table 3. We can also compute the expected squared mean metric across all events.

$$\begin{aligned} &\mathrm{E}[(\text{mean metric})^2] \\ &\quad = P_1 S^2 + P_2 F^2 + P_3\left(\frac{S+S}{2}\right)^2 + P_4\left(\frac{S+F}{2}\right)^2 \\ &\qquad + P_5\left(\frac{F+S}{2}\right)^2 + P_6\left(\frac{F+F}{2}\right)^2 + P_7 S^2 + P_8 F^2 + \cdots. \end{aligned}$$

Next, compute the expected cross products of the mean metric with the holdout outcome (columns 4 and 9 in Table 3).

$$\begin{aligned} &\mathrm{E}[(\text{mean metric}) \cdot \text{outcome}] \\ &\quad = P_1 SS + P_2 FS + P_3\left(\frac{S+S}{2}\right)S + P_4\left(\frac{S+F}{2}\right)S \\ &\qquad + P_5\left(\frac{F+S}{2}\right)S + P_6\left(\frac{F+F}{2}\right)S + P_7 SF + P_8 FF + \cdots. \end{aligned}$$

Let $A = R_1 q_1 + (1-R_1) q_2$ denote the expected probability of an S outcome for a randomly chosen individual. Substituting all of these terms into ρ^2_{mean} yields the following equation for the squared correlation between the mean metric and the holdout outcome. Note that ρ^2_{mean} is independent of the specific values of S and F.

$$\begin{aligned} \rho^2_{\text{mean}} = {} & \big(2(A-q_2)^2(A-q_1)^2(q_2-q_1)\big) \\ & \cdot \big(A(1-A)(q_1q_2(p_2-p_1)+2A(1-A)(q_2-q_1) \\ & - (A+q_1q_2)(p_2q_2-p_1q_1)+A(p_2q_2^2-p_1q_1^2))\big)^{-1}. \end{aligned}$$

Following these steps for each metric yields the squared correlations shown in Table 4.

3.5. Findings

The correlations in Table 4 quantify the amount of information each metric contains by computing how much variance each metric explains across individuals (or other units of analysis). Consequently, the correlations also quantify each metric's ability to predict the holdout outcomes for each individual. Knowing that allows us to construct and evaluate how well various metrics will perform in different environments regardless of their use (e.g., decision making, diagnosing, monitoring, evaluating, forecasting, estimating, etc.).

Table 4 reveals necessary and sufficient conditions when any metric outperforms any other metric (e.g., specific parameter values when each metric is best). Although specific values are latent, knowledge about the nature of the environment might still reveal when particular metrics should do well. The following theorems provide sufficient conditions in a form that reveals general qualitative insights linked to the environment. Our theorems compare the ρ^2 for each nonaveraging metric with the ρ^2_{mean} for the mean metric to reveal when known environments (e.g., efficient, adverse, etc.) favor nonaveraging metrics over the mean metric. Although the mean metric is our benchmark for reasons noted earlier, additional comparisons are doable and might produce insightful future research.

Theorem 1. *The maximum metric contains more information than the mean metric* ($\rho^2_{\max} > \rho^2_{\text{mean}}$) *when*

$$p_2 > 1 - q_1 > 1 - q_2 > p_1 \quad (\textit{the Muth environmental effect}) \tag{1}$$

$$\textit{or} \quad \frac{1-p_1}{4} = \frac{1-p_2}{4} > q_2 > q_1 \quad (\textit{the Anna Karenina environmental effect}). \tag{2}$$

See the appendix for the proof.

Table 4 **Squared Correlation of Different Metrics with Holdout Outcome** (ρ^2)

$$\rho^2_{\text{mean}} = \frac{2(A-q_2)^2(A-q_1)^2(q_2-q_1)}{A\bar{A}(q_1q_2(p_2-p_1)+2A\bar{A}(q_2-q_1)-(A+q_1q_2)(p_2q_2-p_1q_1)+A(p_2q_2^2-p_1q_1^2))}$$

$$\rho^2_{\max} = \frac{((A-q_2)^2(A-q_1)^2((q_2-q_1)+(p_2q_2-p_1q_1)-(p_2q_2^2-p_1q_1^2))^2)}{(A\bar{A}(q_1q_2(p_2-p_1)+\bar{A}(q_2-q_1)-(A+q_1q_2)(p_2q_2-p_1q_1)+A(p_2q_2^2-p_1q_1^2))(-q_1q_2(p_2-p_1)+A(q_2-q_1)+(A+q_1q_2)(p_2q_2-p_1q_1)-A(p_2q_2^2-p_1q_1^2)))}$$

$$\rho^2_{\min} = \frac{((A-q_2)^2(A-q_1)^2((q_2-q_1)-(p_2q_2-p_1q_1)+(p_2q_2^2-p_1q_1^2))^2)}{(A\bar{A}(q_1q_2(p_2-p_1)-\bar{A}(q_2-q_1)-(A+q_1q_2)(p_2q_2-p_1q_1)+A(p_2q_2^2-p_1q_1^2))(-q_1q_2(p_2-p_1)-A(q_2-q_1)+(A+q_1q_2)(p_2q_2-p_1q_1)-A(p_2q_2^2-p_1q_1^2)))}$$

$$\rho^2_{\text{variance}} = \frac{((A-q_2)^2(A-q_1)^2((p_2q_2-p_1q_1)-(p_2q_2^2-p_1q_1^2))^2)}{(A\bar{A}(2q_1q_2(p_2-p_1)+(q_2-q_1)-2(A+q_1q_2)(p_2q_2-p_1q_1)+2A(p_2q_2^2-p_1q_1^2))(-q_1q_2(p_2-p_1)+(A+q_1q_2)(p_2q_2-p_1q_1)-A(p_2q_2^2-p_1q_1^2)))}$$

$$\rho^2_{\text{count}} = \frac{(A-q_2)^2(A-q_1)^2(p_2-p_1)^2}{A\bar{A}(A(p_2-p_1)-(p_2q_1-p_1q_2))(-A(p_2-p_1)+(q_2-q_1)+(p_2q_1-p_1q_2))}$$

Where $A = R_1q_1 + (1-R_1)q_2$, $\bar{A} = 1 - A$

The conditions in Theorem 1 reveal when and why the maximum metric conveys more information than the mean metric. Condition (1) holds when the environment provides a sufficiently positive relationship between p_t and q_t, so that individuals with greater innate ability (i.e., a greater likelihood of more favorable outcomes) tend to have a greater number of observed outcomes in their background. Conversely, less innate ability (smaller q_t) tends toward fewer observations (smaller p_t). Hence, the number of observed outcomes conveys information about underlying parameters (e.g., the individual's latent innate ability). Many efficient environments tend to satisfy this condition (1) because of what we refer to as the Muth effect, i.e., rational markets do not waste information as argued by the eminent economist and the father of rational expectations John F. Muth (Muth 1961). Here, by using all available information, rational market environments efficiently recognize higher innate ability individuals (or organizations or categories) and reward them with more opportunities. Of course, beyond our Muth reasoning, other environmental forces might magnify this relationship, ensuring that condition (1) is met.

Statistical theory virtually ignores this situation because most statistical procedures assume either that exogenous sample sizes are fixed or that sample sizes only reflect issues related to accuracy. Hence, by assumption, sample sizes contain no information. Despite that tradition, samples sizes do reveal information. For example, product categories with more products usually differ from those with fewer products. Product categories with more price promotions usually differ from those with fewer price promotions. Hence, the number of observations is revealing.

Now consider why the maximum metric can exploit information in the number of observations whereas the mean metric cannot. The reason is that the distribution of the maximum statistic depends on the number of observations, whereas the mean does not. As the sample size increases, the expected value of the mean remains the same, but the maximum metric increases because, by definition, the maximum metric is the most favorable observation. For example, when we go from 1 to 2 to 3 to 4 observations, the probability of observing the most favorable outcome increases from $1-(1-q_t)^1$ to $1-(1-q_t)^2$ to $1-(1-q_t)^3$ to $1-(1-q_t)^4$. When $q_t = 10\%$, this probability increases from 10% to 19% to 27% to 34%. In general, the expected value of the maximum metric captures sample size information, whereas the mean metric does not. However, in our situation, the mean metric displays a small positive correlation with sample size because the mean of larger samples is slightly larger when larger samples contain more type 2 individuals. Note that in actual applications the maximum metric can increase and decrease when we employ a moving duration window. For example, with eight years of data, we could compute the maximum metric for six three-year moving windows.

Condition (2) remarkably reveals that the maximum metric can outperform the mean metric even when the Muth effect is entirely absent, and sample size is unrelated to the underlying latent ability (i.e., $p_1 = p_2$). This condition's key feature is the extremely adverse or failure-rich environment, i.e., $q_t < (1-p_t)/4$. Many market environments might satisfy condition (2) because of the Anna Karenina (AK) effect (Shugan 2007). The AK effect, an eponym for the main character and namesake title of Count Lev Nikolayevich Tolstoy's famous novel (Tolstoy 1998), reflects the widely quoted first line of Tolstoy's novel, roughly translated to "Happy families are all alike; every unhappy family is unhappy in its own way" (Diamond 1997, p. 157). One implication is that when passively observing survivors, survivors show little disparity on the variables necessary for survival, making these variables appear as constants (Shugan 2007). Another implication is that favorable outcomes require every detail to be right, whereas an unfavorable outcome only requires one wrong detail, making most environments adverse. Consequently, favorable outcomes are rare and more informative than unfavorable outcomes. For example, individuals, new products, or organizations could fail for many environmental reasons beyond their control, such as lack of support, bad timing, incompetent distributors, unexpected competitive moves, economic turns, new regulations, or bad luck. In sum, in adverse or failure-rich environments, failures provide less information because there are many causes for failure. However, favorable outcomes often reflect the presence of many propitious conditions including, for example, a high-innate-ability individual.

The maximum metric accentuates more favorable outcomes, whereas the mean metric tempers more and less favorable outcomes. Accentuating favorable outcomes allows the maximum metric to exploit the information in a mixed background (i.e., having both more- and less-favorable observations) better than the mean metric. For example, the maximum metric equates an individual with a mixed background with an individual with only observed favorable outcomes. Consequently, the maximum metric extracts more information about outcomes when the likelihood of at least one favorable outcome is sensitive to the latent individual type. This situation occurs in extremely adverse or failure-rich environments. To understand, let x denote the difference between the two individual types, i.e., $x = q_2 - q_1$. For two observations, the difference in the probability of observing at least one favorable outcome is

$[(1-(1-q_2)^2)]-[1-(1-q_1)^2]=x(2-2q_2+x)$, which is larger for smaller values of q_2 holding x constant. Hence, the maximum metric better distinguishes the two types when q_t is small (i.e., a failure-rich environment). Parenthetically, the AK effect fosters a failure-rich environment where successes S are rare and contain more information as implied by Shannon's well-known self-information measure in information theory (Shannon 1948; Cover and Thomas 1991, p. 107). The maximum identifies informative signals within a noisy environment, whereas the mean fails to discount the noise.

Of course, beyond the AK environment, other forces might magnify this relationship, ensuring that condition (2) is met. In sum, either environmental effect (AK or Muth) can cause the maximum metric to convey more information about outcomes than the mean metric. This reasoning for the maximum metric suggests a symmetric result for the minimum metric. Theorem 2 confirms that reasoning.

Theorem 2. *The minimum metric contains more information than the mean metric* ($\rho^2_{\min} > \rho^2_{\text{mean}}$) *when*

$$p_1 > q_2 > q_1 > p_2 \tag{3}$$

$$\text{or} \quad q_2 > q_1 > \frac{p_1+3}{4} = \frac{p_2+3}{4}. \tag{4}$$

The two conditions in Theorem 2 reveal when and why the minimum metric conveys more information than the mean metric. Condition (3) implies a sufficiently negative relationship between p_t and q_t, i.e., individuals with greater innate ability have a smaller expected number of observations. Hence, there is an inverse relationship between individual ability (i.e., the probability of more favorable outcomes) and the number of observations (i.e., the sample size). For example, as we go from 1 to 2 to 3 to 4 observations, the probability of at least one failure increases from $1-q_t$ to $1-q_t^2$ to $1-q_t^3$ to $1-q_t^4$. Consequently, unlike the mean metric, the minimum tends to decrease as the number of observed outcomes increases, thereby capturing information in the sample size. When $p_1 > q_2 > q_1 > p_2$, the increased sample size also captures decreased latent innate ability. Like the maximum metric, the minimum metric can increase or decrease when computed over different time intervals (i.e., a moving window).

Although condition (3) is sufficient, it is unnecessary. Minimum metrics can outperform the mean without sample-size information (i.e., $p_1 = p_2$). Condition (4) implies $q_t > (3+p_t)/4$, so the environment is extremely favorable or success (S) rich. Here, the minimum metric benefits from unfavorable outcomes being more informative than favorable outcomes. In other words, successes are noisy while failures provide informative signals. Hence, in success-rich environments, the minimum metric better distinguishes latent type than the mean. Incidentally, Theorem 2 follows from Theorem 1 because (1) $\rho^2_{\max}$ evaluated at $q_1 = q_1^*$ and $q_2 = q_2^*$ equals $\rho^2_{\min}$ evaluated at $q_1 = 1-q_1^*$ and $q_2 = 1-q_2^*$ and (2) $\rho^2_{\max}$ evaluated at $q_1 = q_1^*$, $q_2 = q_2^*$, $p_1 = p_1^*$, $p_2 = p_2^*$ equals $\rho^2_{\min}$ evaluated at $q_1 = q_2^*$, $q_2 = q_1^*$,$p_1 = p_2^*$, $p_2 = p_1^*$ for any $q_1^*, q_2^*, p_1^*, p_2^*$.

Theorem 3 considers the variance metric.

Theorem 3. *The variance metric contains more information than the mean metric* ($\rho^2_{\text{variance}} > \rho^2_{\text{mean}}$) *when*

$$\frac{p_2}{2} > \frac{1}{5} > 2q_1 > q_2 > q_1 = p_1. \tag{5}$$

The condition in Theorem 3 reveals when and why the variance metric conveys more information than the mean metric. Unlike the maximum metric, we can prove that the variance metric never reveals more information about outcomes than the mean metric when $p_2 = p_1$. Hence, Theorem 3 requires both effects. First, $1/5 > 2q_1 > q_2 > q_1 = p_1$ ensures the AK effect, i.e., q_1 and q_2 small. Second, condition (5) ensures the Muth effect, i.e., ($p_2 > p_1$, $q_2 > q_1$). Finally, the condition $p_2 > 2/5$ ensures that the probability of observing multiple outcomes (for type 2 individuals) is sufficiently large to allow the variance metric to extract information about individual types. Only one observation yields no variance (i.e., no information in the metric).

Theorem 4. *The count metric contains more information than the mean metric* ($\rho^2_{\text{count}} > \rho^2_{\text{mean}}$) *when*

$$\frac{p_2}{2} > q_2 > q_1 > p_1. \tag{6}$$

The condition in Theorem 4 reveals when and why the count metric conveys more information than the mean metric. Theorem 4 reveals that the count metric can explain more information across outcomes than the mean metric because, unlike the mean, the count metric exploits information in the sample size. However, to outperform the mean metric, the count metric requires the Muth environmental effect, i.e., ($p_2 > p_1$, $q_2 > q_1$). Moreover, the count metric requires a stronger form of the Muth effect than the maximum because, unlike the count metric, the maximum metric exploits both the number of observations and their magnitude. For example, the count metric fails to distinguish between several favorable outcomes and the same number of mixed outcomes. The count metric only exploits the number of observed outcomes in each background. This finding is consistent with the count metric's observed performance in our data.

4. Generalizing via a Numerical Simulation

To investigate the generality of condition (1), we examine 10,000 cases. Each case consists of a random N distributed between 3 and 10, a random J

Table 5 **Metric Correlations Based on Random Parameters With and Without Condition (1)**

Metrics	Expected correlation (no restriction)	Expected correlation (condition (1) restriction)	Better than mean (no restriction) (%)	Better than mean (condition (1) restriction) (%)	Best (no restriction) (%)	Best (condition (1) restriction) (%)
Mean	0.27	0.12	N. A.	N. A.	82	0
Maximum	0.18	0.22	6	100	3	41
Minimum	0.17	0.03	9	2	4	0
Variance	0.10	0.17	10	47	7	13
Count	0.04	0.21	5	88	4	45

distributed between 3 and 10, and probabilities (q_t, p_t, and R_t) randomly distributed between 0.01 and 0.99. For each case, we generate a background and holdout observations using the binomial distributions described earlier. We then compute the correlations for each metric (i.e., between the metric and holdout observation) for all the cases and, next, for only cases with parameters satisfying condition (1).

Table 5 provides the results. Table 5 reveals that the maximum metric outperforms the mean metric in every case when condition (1) holds. Hence, it is likely that condition (1) generalizes to $N > 2$ and $J > 2$ because we would only expect (with no conditions) the maximum metric to outperform the mean metric 6% of the time (see Table 5). Finally, Table 5 also suggests that condition (1) favors the variance and count metrics, but fails to be sufficient. The count metric does particularly well because it specifically isolates the sample size. Of course, simulations only reveal probabilistic evidence for underlying relationships, whereas our theorems provide exact proofs.

5. Conclusions and Implications

Good metrics are well-defined formulae (often involving averaging) that transmute multiple measures of raw numerical performance (e.g., dollar sales, referrals, number of customers) to create informative summary statistics (e.g., average share of wallet, average customer tenure). Despite myriad uses (benchmarking, monitoring, allocating resources, diagnosing problems, explanatory variables in forecasting models), most uses require metrics that contain information (explained variance) about observed outcomes. On this criterion, we show empirically (with people data), theoretically (with exact proofs), and numerically (with simulations) that, although averaging has remarkable theoretical properties, supposedly inferior nonaveraging statistics are often better. We find:

- Good metrics capture information (explained variance) across observations rather than proximity (minimal bias). These metrics often become explanatory variables in formal models that achieve proximity.
- Empirical analysis of the performance of individuals (baseball batters, movie actors, and authors) all reveal that seemingly inferior nonaveraging metrics (e.g., the maximum metric) do much better than expected by chance, and often outperform the mean metric.
- The Muth effect provides an environmental explanation—markets provide more opportunities to people and organizations with greater innate ability causing larger samples sizes (e.g., number of observations in individual backgrounds). The implications are vast. The number of observed new products from a firm might convey information about the firm's innate ability to innovate. The number of brands in a product category might convey information about the category potential. Decision makers can exploit knowledge about their environment by choosing specific nonaveraging metrics.
- The Anna Karenina (AK) effect is another environmental explanations—in adverse environments, favorable outcomes convey more information than unfavorable outcomes. Similarly, in propitious environments, favorable outcomes convey less information. The implications are far-reaching. When record albums reach the very rare RIAA certification of Diamond, for example, that certification conveys a great deal of information about the artist's innate ability, irrespective of other outcomes. The rare blockbuster drug (>$1 billion) might provide important information about the parent pharmaceutical company's innate ability regardless of other outcomes. Similarly, a Clio award provides information about an ad agency's innate creativity. Nonaveraging metrics exploit this environment. The maximum metric recognizes these important observations within a noisy failure-rich environment. Decision makers can exploit knowledge about their environment by choosing specific nonaveraging metrics.
- Specific cases, general theoretical developments and more general simulations all show that these environmental effects (Muth and AK) do favor nonaveraging metrics.

Our research provides many practical implications. First, academics and practitioners should more seriously consider (if not immediately adopt) nonaveraging metrics (e.g., the maximum, the top decile, the bottom quartile, the standard deviation). Second, despite traditional statistical theory, empirically observed sample sizes (e.g., number of products in a category, number of returned questionnaires, number of firms in an industry) might convey important information about underlying parameters. For example, the number of reviews of a product might reflect the quality of the product. Third, the environment itself often conveys valuable information. For example, we might learn more from surviving firms in hostile environments than from much more successful firms in friendly environments. Fourth, our findings might help understand the logic behind some apparent anomalies in consumer behavior involving premature termination of search, selective perception, repetition bias, escalation of commitment, and biased recall. For example, the AK effect explains the famous taxicab problem (Tversky and Kahneman 1982, p. 156). Respondents might believe that rare events are more informative (more easily remembered) and conclude the witness is likely correct because the probability of remembering an uncommonly colored taxi is greater. Fifth, our findings could explain why individuals with one very successful outcome (e.g., a highly cited article, a blockbuster movie, a creative new strategy) remain coveted despite many subsequent less-successful outcomes (e.g., articles with less impact, movies that flop, strategies that fail). For example, consider Robert L. Nardelli, who had failures after his success at General Electric, and was still hired as chairman and chief executive office of Chrysler. Sixth, in many situations, only nonaveraging data (e.g., the best, worst, or truncated) are available. For example, we might only observe successful products, selective résumés might only report more favorable accomplishments, advertising agencies might only promote successful ads, and athletic record books report only the best (e.g., minimum) times. These incomplete data might still provide adequate information to predict future outcomes. Seventh, given cross-sectional data with multiple observations for every unit of analysis (individuals, firms, categories, etc.), we can estimate the relationship between any nonaveraging metric (e.g., maximum, minimum, etc.) with any predicted outcome (sales, market share, penetration, profits, etc.) to exploit the desirable properties of nonaveraging metrics in particular environments. For example, empirical analysis of movie data shows that the maximum metric better predicts box-office outcomes than the mean metric. Eighth, when summarizing the characteristics of a market, nonaveraging metrics (e.g., top quartile) might contain more information than standard averaging metrics from accounting. For example, this is an explanation for the popularity of metrics such as price to peak earnings and gross positive fair value. Ninth, unlike models that seek to produce accurate (proximal) forecasts, we should evaluate metrics on their information content (e.g., variance explained, statistical dependence) and not necessarily on proximity or bias. For example, the identity of the developer of a new product might be a better predictor of success than the average failure rate in the product category. Tenth, perhaps the disappointing correlation between market outcomes (e.g., share, sales) and common marketing metrics (e.g., average customer satisfaction, average intent to purchase) would be improved by using nonaveraging metrics. For example, the AK effect explains why "top box" is a better metric for predicting eventual purchase. Eleventh, computing nonaveraging metrics from moving duration windows could capture both intertemporal and cross-sectional information for the best predictions. For example, with eight years of data, we could compute the maximum metric for six three-year moving windows and predict future trends.

Obviously, myriad topics await future research. Generality could be explored in still more settings. Theoretical research could compare our nonaveraging metrics and prospect new ones, including combinations of extant metrics. Sufficient conditions could be weakened. Events could be statistically dependent (e.g., people could change types).

Acknowledgments

The authors appreciate the helpful comments of Yubo Chen and the seminar attendees at Yale University, Singapore Management University, Santa Clara University, University of California at Berkeley, University of Utah, Texas A&M University, Washington University, Northwestern University, North Carolina State University, University of Texas at Dallas, University of Southern California, and University of Florida.

Appendix

Proof of Theorem 1. First, we make the following transformations: $q_1 = (1+b)/(1+a+b)$, $q_2 = (q_1+c)/(1+c)$, $p_1 = (d/(1+d))(1-q_1)/(1+c)$, $p_2 = (1+a)/(1+a+b)$, $R_1 = 1/(1+R)$, where the temporary constants $a, b, c, R > 0$ differ for each proof. These transformations ensure that $1 > p_2, q_2, q_1, p_1 > 0$. Note, $p_2 + q_1 = ((2+a+b)/(1+a+b)) > 1 \Rightarrow p_2 > (1-q_1)$, $q_2 = (q_1+c)/(1+c) > q_1 \Rightarrow (1-q_1) > (1-q_2)$, and $p_1 = (d/(1+d))(1-((1+c)q_2-c))/(1+c) = (d/(1+d))(1-q_2) \Rightarrow p_1 < (1-q_2)$. Combining these expressions yields $p_2 > (1-q_1) > (1-q_2) > p_1$. Hence, the transforms enforce condition (1). Substituting the transforms into the expression for $\rho^2_{\text{max}} - \rho^2_{\text{mean}}$ from Table 4 reveals that all the terms in the resulting expression are positive for $a, b, c, d, R > 0$, ergo $\rho^2_{\text{max}} > \rho^2_{\text{mean}}$. Q.E.D.

To prove condition (2) is sufficient, make the following transforms: $q_1 = (1-p_1)/(4+b+c)$, $q_2 = (1-p_1)/(4+b)$, $p_1 = 1/(1+a)$, $p_2 = p_1$, $R_1 = 1/(1+R)$. These transforms ensure that $1 > q_1, q_2, p_1, p_2 > 0$ and enforce condition (2), i.e., $(1-p_1)/4 = (1-p_2)/4 > q_2 > q_1$. Substituting these transforms into $(\rho^2_{\text{max}} - \rho^2_{\text{mean}})$ reveals that all the resulting terms are positive for $a, b, c, R > 0$, ergo $\rho^2_{\text{max}} > \rho^2_{\text{mean}}$. Q.E.D.

Proof of Theorem 2. The transformations $q_1 = (p_2 + c)/(1 + c)$, $q_2 = (q_1 + b)/(1 + b)$, $p_1 = (q_2 + a)/(1 + a)$, $p_2 = 1/(1 + d)$, $R_1 = 1/(1 + R)$ ensure condition (3) holds, i.e., $p_1 > q_2 > q_1 > p_2$, and that $1 > q_1, q_2, p_1, p_2 > 0$ for $a, b, c, d, R > 0$. Substituting these transforms into $\rho^2_{\text{min}} - \rho^2_{\text{mean}}$ from Table 4 reveals that all the resulting terms are positive for $a, b, c, d, R > 0$, ergo $\rho^2_{\text{min}} > \rho^2_{\text{mean}}$. Q.E.D.

To prove condition (4) is sufficient, substitute $q_1 = (3 + b)/(4 + b)$, $q_2 = (3 + b + c)/(4 + b + c)$, $p_1 = (4q_1 - 3)/(1 + a)$, $p_2 = p_1$, $R_1 = 1/(1 + R)$ into $\rho^2_{\text{min}} - \rho^2_{\text{mean}}$ from Table 4. These transforms ensure that $1 > q_1, q_2, p_1, p_2 > 0$ and enforce condition (4). All the resulting terms are positive for $a, b, c, d, R > 0$, ergo $\rho^2_{\text{min}} > \rho^2_{\text{mean}}$. Q.E.D.

Proof of Theorem 3. The transformations $q_1 = 1/(10 + a)$, $q_2 = (cq_1 + 2q_1)/(1 + c)$, $p_1 = q_1$, $p_2 = (2 + d)/(5 + d)$, $R_1 = 1/(1 + R)$ ensure $1 > q_1, q_2, p_1, p_2 > 0$ and enforce sufficient condition (5). Substituting these transformations into $\rho^2_{\text{variance}} - \rho^2_{\text{mean}}$ reveals only positive terms for $a, b, c, d, R > 0$, ergo $\rho^2_{\text{variance}} > \rho^2_{\text{mean}}$. Q.E.D.

Proof of Theorem 4. The following transformations $q_1 = q_2/(1 + b)$, $q_2 = p_2/(2 + a)$, $p_1 = q_1/(1 + c)$, $p_2 = 1/(1 + d)$, $R_1 = 1/(1 + R)$ ensure $1 > q_1, q_2, p_1, p_2 > 0$ and enforce sufficient condition (6). Substituting these transforms into $\rho^2_{\text{count}} - \rho^2_{\text{mean}}$ in Table 4 reveals that all the resulting terms are positive for $a, b, c, d, R > 0$, ergo $\rho^2_{\text{count}} > \rho^2_{\text{mean}}$. Q.E.D.

References

Cohen, M. A., J. Eliashberg, T.-H. Ho. 2000. An analysis of several new product performance metrics. *Manufacturing Service Oper. Management* **2**(4) 337–349.

Cover, T. M., J. A. Thomas. 1991. *Elements of Information Theory*. Wiley-Interscience, New York.

Diamond, J. M. 1997. *Guns, Germs, and Steel*: *The Fates of Human Societies*. W. W. Norton, New York.

Gibbons, J. D., S. Chakraborti. 2003. *Nonparametric Statistical Inference*, 4th ed. CRC Press, New York.

Gupta, S., V. Zeithaml. 2006. Customer metrics and their impact on financial performance. *Marketing Sci.* **25**(6) 718–739.

Gupta, S., D. R. Lehmann, J. A. Stuart. 2004. Valuing customers. *J. Marketing Res.* **41**(1) 7–18.

Hauser, J. R. 1998. Research, development, and engineering metrics. *Management Sci.* **44**(12) 1670–1689.

Hauser, J. R. 2001. Metrics thermostat. *J. Product Innovation Management* **18**(3) 134–153.

Lindsey, J. K. 1997. Exact sample size calculations for exponential family models. *The Statistician* **46**(2) 231–237.

Muth, J. F. 1961. Rational expectations and the theory of price movements. *Econometrica* **29**(3) 315–335.

Ostasiewicz, S., W. Ostasiewicz. 2000. Means and their applications. *Ann. Oper. Res.* **97**(1) 337–355.

Peritz, B. C. 1982. Matched case-control studies in citation analysis. *J. Amer. Soc. Inform. Sci.* **33**(5) 333–337.

Price, D. J. D. 1970. Citation measures of hard science, soft science, technology and nonscience. C. E. Nelson, D. K. Pollack, eds. *Communication Among Scientists and Engineers*. D. C. Heath, Lexington, MA, 3–22.

Rust, R. T., T. Ambler, G. S. Carpenter, V. Kumar, R. K. Srivastava. 2004. Measuring marketing productivity: Current knowledge and future directions. *J. Marketing* **68**(4) 76–89.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Tech. J.* **27**(July, October) 379–423, 623–656.

Shugan, S. M. 2007. The Anna Karenina bias: Which variables to observe? *Marketing Sci.* **26**(2) 145–148.

Stremersch, S., P. C. Verhoef. 2005. Globalization of authorship in the marketing discipline: Does it help or hinder the field? *Marketing Sci.* **24**(4) 585–594.

Tolstoy, L. N. 1998. *Anna Karenina* (L. Maude & A. Maude, Trans.). Oxford University Press, Oxford, UK. (Original work published 1875–1877.)

Tversky, A., D. Kahneman. 1982. Evidential impact of base rates. D. Kahneman, P. Slovic, A. Tversky, eds. *Judgment Under Uncertainty*: *Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 153–160.