Billy Beane and Outcomes: What Can Baseball Tell the Nonprofit World About Measures and Measurement?

Ken Berger & Robert M. Penna 04 August 2010

While the notion of outcomes seems to be here to stay (and the jargon of outcomes is certainly everywhere), what is oddly missing in our sector is much evidence of the *practice* of outcomes. Although individual promising examples certainly exist, for the most part the social sector is talking about outcomes much more than it is actually doing much with outcomes, and much of the conversation centers on three questions:

1. What is the "value" of outcomes?
2. What do outcomes tell us; why are they (or why should they be) important?
3. Should they be applied to everyone in the sector?

To get to an answer, perhaps the whole thing ought to be posed in a different way: *How valid is the "knowledge" upon which individual and organizational giving decisions are traditionally and largely based?*

This is an especially important question for donors, because the decisions they make very often determine which efforts will be implemented, and which will survive. Every day, in countless boardrooms, meeting rooms, and living rooms across the county, organizations and individuals make the decision to invest in the work of a given nonprofit. Sometimes they have at least some reliable information upon which to base this decision; very often they have little. So how are these decisions being made?

If the traditional way the sector has been arriving at these decisions for well over fifty years is valid, then the sector really has no need of outcomes. But if outcomes truly do hold the key to making these decisions in a rational, objective way, then virtually all the other considerations we have been using must be wrong—or at least inadequate. A contemporary and accessible example we might consider in thinking about this issue (and moreover one with strong parallels to the situation in the social sector) can be found in, of all places, professional baseball.

Page 2

Organized professional baseball, as even nonfans generally know, is a game that has compiled a staggering number of statistics throughout its long history. The most prevalent of these are, for hitters, the batting average, and for pitchers, the earned run average. Both of these

measures (along with several others) were established by Henry Chadwick in 1859, and have been used ever since to assess performance, to rank players, and, most crucially, to judge the potential of young hopefuls and to justify an investment in them.

Meanwhile in the social sector, it was only in the late 1800s that the first concerted efforts were made at collecting programmatic information regarding the effectiveness of the fairly few public policies enacted at the time in the United States. But the effort was decidedly unsystematized, and early practitioners lurched from one concept of inherent program value to another without much apparent direction.

By the early 20th century, the sector saw a heavy reliance upon legislative fiat or executive order. From temperance to health, various efforts sought to establish rules for general behavior, rather than seeking to change individual behavior. Programs were designed to address *problems*, not necessarily their causes: Outlaw alcohol, so the thinking of the time went, and all social problems stemming from its abuse will disappear. But there also existed a faith in the effectiveness of these programs that required little or no "proof" once the argument regarding the worthiness of the effort had been established (Suchman 1967: 4). There was an implicit assumption that armed with an understanding of social problems, those working in the field would have no trouble changing negative conditions through social policies, social action,

and direct intervention (Zimbalist 1977). The impacts of social programs, to the extent that anyone actually thought in such terms, were assumed to be readily apparent in the number of people touched by an effort or the number of dollars expended. Further investigation was not thought to be necessary (Zimbalist 1977: 4). Later, concepts such as service units and compliance were added to notions of commitment, caring, and effort as the hallmarks and tests of effective programs and organizations. This said, the sector remains in large measure entrenched in the counting of activities, rather than assessing effectiveness.

So what do the respective yardsticks of performance, effectiveness, and potential in baseball and the social sector have in common? Only one thing: They have both proven to be decidedly misleading, and very often simply wrong.

In his best-selling book *Moneyball,* Michael Lewis chronicles the efforts of Oakland Athletics general manager Billy Beane to compensate for the morosely low payroll he had at hand to compete for talent with the deep pockets of teams such as the Yankees, Mets, and Red Sox (Lewis 2004). In a revolutionary epiphany, he realized that the reliance of those well-funded teams on traditional measures of player value not only usually led them to overvalue (and consistently overpay) star players, but also blinded them to the true value of usually overlooked journeymen and minor league players.

But if the traditional formulas, of batting averages (or earned run averages) combined with "the look" and then filtered through the scouts' sage instincts, were as accurate as everyone supposed them to be, there should have been no truly bad teams. More to the point, very expensive yet really awful teams should not exist at all if high payrolls reflected the best talent available. Yet they did, with examples too numerous to mention.

Observers might be forgiven for seeing in this parallels to the social sector, where the best minds have sought out the best programs for over forty years, and yet in the words of one long-time activist, "We've been fighting this 'war on poverty' since Johnson was in the White House. But not only haven't we won it, I'm not sure for all we've spent that we even have anything to show for it." Could it be that our "traditional formulas," like those in baseball, have been simply and irredeemably wrong?

In Lewis's account of events, it was a rabid fan and complete outsider, a crusty gentleman named Bill James, who was the first to ask, "If we can't tell who the good [players] are from the record books...how *can* we tell?" (Lewis 2004: 69). The answer, James concluded, was by *counting* things. But not the things baseball was already counting, because they told, at best, an incomplete story.

James began with what his eyes told him he was seeing in various players and with what other people said was there

in terms of talent and performance. But he then asked himself, "Is any of it true? Can you validate it? Can you measure it?" (Lewis 2004: 75). What he learned upon examination was that you could *not* validate the traditional wisdom regarding talent or performance by using baseball's traditional measures. In fact, his new measures proved most conventional wisdom to be dead wrong.

In similar fashion, the social sector has also been counting. By 1907 the New York City Bureau of Municipal Research was collecting data on social conditions, and helped to launch a broad tradition of counting. Over the years, a staggering number of governmental and private agencies have counted numbers of children, numbers of residents per dwelling, numbers of people living in poverty, and a host of other things. Over the last several decades, our sector itself has maintained the practice. We counted service units, clients, graduates, and even processes and interventions themselves. Yet for all this, for the most part no one was providing compelling evidence that any of our programmatic efforts were actually working effectively (Zimbalist 1977).[1]

Over time, however, just as Bill James realized that "runs batted in" was not really an accurate assessment of a baseball player's value,[2] some thinkers in the social sector slowly began to realize that service units and compliance measures were a poor proxy for actual performance and effectiveness. The early voices calling for an objective measure of program (and by implication, organizational)

effectiveness were lonely and few. For example, Knutson (1955) argued the importance of defining program objectives more specifically. Ciocco (1960), perhaps mindful of the growing reliance upon strict measurement in the corporate field of management theory, followed five years later by stressing the importance of using indices of "known reliability and validity."

Yet, in 1969, the Urban Institute completed an extensive study of the performance evaluation of federally sponsored programs and concluded that such assessments of effectiveness were "almost non-existent" (Horst et al. 1974: 300). Worse still, that same year, Suchman observed that "what passes for evaluative research [today] is...a mixed bag at best, and chaos at worst." Looking at program objectives themselves, he concluded that "far too many [program goals]...are grandiose-but-usually-vague statements of intent and procedure...based upon largely untested or even unsound assumptions whose validity rests primarily upon tradition or 'common sense,' and not upon proven effectiveness" (Suchman 1967: 16). In other words, we were still largely taking it on faith that we were producing meaningful results, and in the relatively few cases where we were attempting any measurement at all, we were looking at essentially meaningless information.

But if this was true, upon what "knowledge" were programs being funded for all these years by philanthropy and government? The answer was a combination of the

tried and true measures of head counts and compliance, mixed with the judgment of the funding agencies regarding what successful programs (and the organizations that ran them) "looked like." Oddly enough, *just* like baseball.

Beginning with the foundations laid by Donald Campbell, and continuing through the work of Claude Bennett, Joseph Wholey, Martha Taylor Greenway, Williams, Phillips and Webb, and others, a powerful argument began to form that suggested that measuring performance—outcomes—was a far more accurate way to assess, rank, and (ultimately) fund charities than were the old measures that had been in use for so long. Yet just as Bill James' insights on baseball mostly fell on deaf ears, the logically inarguable idea of outcomes met resistance: they were a fine intellectual concept, the critics murmured in staid and profound tones, but could never work in practice. Indeed, many critics still make such arguments, but the consensus is slowly shifting away from such ideas.

In the case of baseball, Bill James' work went largely ignored by the professional establishment until utter and dire necessity forced Billy Beane to look beyond the stats that had everyone else's attention, and to try to identify the diamonds in the rough or the underrated gems that the Big Boys had overlooked in their rush to sign superstars. Billy's team was impoverished and likely to remain so. The only way he could compete was to play smarter. Just as he could not afford superstars, neither

could he afford to allow the extremely limited resources he had on hand to be squandered by ignorance, bad decision-making, or hidebound thinking among his scouts or managerial staff. Success called for a new way of thinking; in fact, the entire game, in Beane's new view, needed to be rethought right down to the fundamentals. And so, as but one example, he willfully ignored baseball's hallowed devotion to runs scored (and the players who scored them) and focused instead on limiting outs, specifically seeking players who made the fewest.[3]

The situation in the social sector is not all that different.

The days of virtually unlimited funding are long gone. Like Billy Beane, the leaders and managers of today's social sector organizations are being forced to try to do more with less. Yet just as the baseball establishment initially spent more effort trying to stymie Beane's success than they did understanding and emulating it, the social sector today seems to be spending more time discussing the challenges presented by outcomes than actually implementing them.

1. ^ In his survey of over 60 years of social science writing, Zimbalist did not find evidence that there existed within early research efforts *any* question about the effectiveness of social work's interventions.
2. ^ James pointed out that a run batted in, or RBI, was in fact not a measure of an individual player's

prowess, because he needed *someone* else to get on base before him so that he might, through his own hit, allow that person to score. The fallacy of this statistic, James noted, was that a runner who played it too cautiously on third base could rob a hitter of an RBI, just as an audacious one, sprinting from second base and taking a chance on getting home safely, might unduly reward the hitter of a mere blooper single by stretching an expected one-base advance to a two-base score. Either way, the hitter actually had nothing to do with the result. More to the point, James' analysis led to the inescapable conclusion that hitters on teams with poor offenses had fewer opportunities to rack up RBIs because their teammates were lousy hitters. The RBI, in other words, was more of a reflection of a *team's* combined hitting than that of any individual player.

3. ^ Beane's calculation was that runs are potentially unlimited, but that the out was the game's most precious commodity. Teams are allowed only 27 of the latter, and the faster they accumulate them, the quicker the chance to score runs diminishes. His thinking was that by limiting the number of avoidable outs made during a game, even a team of (less expensive) average hitters would have ample opportunities to score sufficient runs to win—especially if the other side was squandering its outs. He similarly realized that the hallowed batting average was far less a reliable predictor of team

scoring success than was a player's on-base percentage, which takes into consideration bases awarded on walks: the patient hitter who drew more walks would usually score more often than the power hitter who also swung at bad pitches and struck out more frequently.

Page 3

What are the lessons for the social sector?

We suggest at least four:

1. The conventional wisdom about what a successful program (or organization) looks like is misleading at best and dead wrong at worst. Just like the baseball establishment, we have been looking at (and responding to!) the wrong things.
2. The measures the sector has used in the past—activity counts and compliance, commitment and passion, and even management savvy and financial strength—are inadequate. They tell us no more about the contribution an organization or program is actually making to meaningful, sustained social change than batting average tells about a hitter's actual value to a team's victories, or a pitcher's ERA tells us about the quality of the defense arrayed around the bases behind him.[4]
3. We need better measures, and outcomes are the

best we are likely to have for a long time. We need, as a guide to our individual and institutional giving decisions, to begin asking several basic but essential questions. Among them are the following:

1. Does an organization utilize a program management system that includes measurable outcome indicators?
2. Does it do so in a systematized way?
3. Do an organization's targeted outcomes have any intrinsic value—are they, in other words, *meaningful* outcomes?
4. Are outcomes being used to enhance performance? In other words, does the organization change *its* behavior if targeted outcomes are not met?
5. Are the organization's outcome reports available to funders and the public?

4. We need to begin using, throughout the sector, the outcome-based tools that have already been developed. There are now tools and formats available to practitioner organizations for virtually every stage of programmatic planning, management, and reporting. There are tools for identifying and establishing outcomes, for tracking and managing toward them, and for capturing and presenting them.[5] These need to be used by more than the relatively few organizations now doing so; they need to be utilized across the sector. In order to make this possible, however, donors need to begin making a

concerted effort at providing the wherewithal that practitioner organizations need in order to learn about, master, and adopt these tools.

For the purposes of informed social investing—giving with the right information—the failure of the predictive power of our sector's traditional measures has been evident for a long while. Like Billy Beane, we must look elsewhere if we are to intelligently allocate the scarce resources we have available. There is really no excuse not to do so.

Yet the questions remains, what will we do?

The ultimate choice is up to our sector's institutional and individual donors. When they begin demanding better, more reliable information upon which to base their giving decisions, the sector will have no choice but to comply.

4. ^ Similarly flawed, for that matter, is the NFL's insistence on measuring the value and performance of a quarterback based upon a ratio of touchdowns to interceptions, the latter of which can be caused by a multiplicity of factors, only some having to do with the quarterback.
5. ^ For more information on these tools and formats, contact the authors through Charity Navigator, or at kberger@CharityNavigator.org or rmpc52@aol.com.

For those still reluctant to embrace the shift to outcome thinking and its associated practices, we point out that those courageous-but-few organizations that have gone down the outcomes path have reaped its rewards. One such organization is Summer Search, a national leadership development program that helps low-income young people graduate high school, go to college, gain successful careers, and give back to society. A popular program that had existed for nearly twenty years, it nonetheless had no set curriculum and no set outcome targets. Instead, there was activity, lots of activity. The result was that recruitment was becoming "a challenge," according to its CEO, and retention had become an even bigger problem. There was just too little the program could concretely point to as a way to attract new participants or keep the ones it had.

After a recent move to outcomes, however, the organization not only saw its largest acceptance rate ever, but overall enrollment and retention blossomed. Of particular interest to other nonprofits is the fact that even in the current economic downturn, where so many charities are seeing their funding slashed, Summer Search has seen its funding sustained. One significant donor recently told them that their verifiable outcomes *proved* that they are a "high impact investment."

Equally inspiring has been the success of the Oakland area's First Place for Youth, an organization that supports

youth in their transition from foster care to successful adulthood by promoting choices and strengthening individual and community resources. After a move to utilize outcomes tracking to illustrate the effectiveness of its model, the organization experienced growth from a $1.7 million budget and staff of 25 to an $8 million budget and a staff of 60. Even more amazing, however, has been the organization's success in helping to attract state dollars in these tough economic times. Partnering with the John Burton Foundation and relying upon the proof of effectiveness its outcome information provided, First Place was able to convince lawmakers in Sacramento to increase from 40 percent to 100 percent the state's portion of the Transitional Housing Plus funding stream that supported its work, an increase that translated into an investment that grew from $5 million to $40 million.

This mention of Oakland, of course, brings us back to Billy Beane and his A's. That's where our story began, and where it should end. So how *did* Billy's shift in thinking work out?

Billy abandoned the old school wisdom of the scouts (and his own reliance upon the game's hallowed and traditional measures of potential and success). In their place, he employed a new analysis that led him to a collection of players nobody else wanted. To the stunned amazement of the rest of baseball, and armed only with this collection of castoffs, rejects, and nobodies, he improved the A's third-place standing and 46-36 record on July 1 to first

place and an 87-51 record on September 1. In other words, over the course of 56 games played in July and August, Billy's misfit A's lost only 15—in spite of having the third lowest payroll in the game. People noticed.

In October of that year, Mr. John Henry, the new owner of the Red Sox, decided that 86 years of frustration was enough. He was going to emulate Billy Beane. He was going to hire Billy's inspirational sage Bill James to help him do it. In fact, he was even going to try to hire Billy Beane himself.

In the end, Billy stayed in Oakland, but his methods were transplanted to Boston. Two short years later, the Sox stunned the world by finally shaking off the Curse of the Bambino and winning their first championship since 1918.

Maybe there *is* something to be said for trying something new.

Ken Berger is the President and CEO of Charity Navigator. Robert Penna, Ph.D., is an independent outcomes consultant, author of the forthcoming *Outcomes Toolbox*, and a member of the Charity Navigator Advisory Panel.

References

References

Ciocco, A. (1960, May). On Indices for the Appraisal of

Health Department Activities. *Journal of Chronic Diseases*, 11: 509-522.

Horst, P., J. N. Nay, J. W. Scanlon, and J. S. Wholey. (1974). Program Management and the Federal Evaluator. *Public Administration Review*, 34(4): 300-308.

Knutson, A. (1955, March). Evaluation Program Progress. *Public Health Reports*, 70: 305-310.

Lewis, M. (2004). *Moneyball*. New York: W.W. Norton.

Suchman, E. (1967). *Evaluative Research*. New York: Russell Sage Foundation.

Zimbalist, S. E. (1977). *Historic Themes and Landmarks in Social Welfare Research*. New York: Harper and Row. Cited in E. J. Mullen, Evidence-based Social Work—Theory & Practice: Historical and Reflective Perspective. Fourth International Conference on Evaluation for Practice. University of Tampere, Tampere, Finland, July 4-6, 2002. Available at http://www.uta.fi/laitokset/sospol/eval2002/CampbellContext.PDF.