International Journal of Hospitality Management 69 (2018) 65–74

Contents lists available at ScienceDirect

International Journal of Hospitality Management

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

Discussion paper

Are published techniques for increasing service-gratuities/tips effective? P-curving and R-indexing the evidence

Michael Lynn

School of Hotel Administration, Cornell University, 552 Statler Hall, Cornell University, Ithaca, NY 14853-6902, United States

ARTICLE INFO

Keywords:
Tip income
Field experiments
Restaurants

ABSTRACT

Recently developed statistical tools are used to assess the evidential value and replicability of the published experimental literature on ways to increase tips. Significantly right-skewed full and half p-curves indicate that the literature is more than a collection of Type 1 errors – it provides evidence of real effects. Moreover, those real effects are scattered across both replicated and non-replicated effects as well as across the work of each of the major contributors to this literature. An overall r-index of 0.55 indicates that over half of the reported effects would likely be replicated if the studies were repeated. More research is need to ascertain the reliability of specific effects – especially those reported by Gueguen, because lower power makes his effects less replicable than others in the literature. Nevertheless, readers can be reasonably confident that most of the techniques for increasing tips in this literature will work.

1. Introduction

Consumers around the world often give voluntary sums of money (called "tips") to the hospitality, tourism and other service workers who have served them. The sizes of these individual payments vary across consumers and contexts, but often amount to 10% or more of the costs of the services being tipped (Lynn and Lynn, 2004) and can reach thousands of dollars (Denison, 2016). There are no good records of aggregate tips within or across nations, but Azar (2011) estimates that approximately $45 billion per year is tipped to U.S. waiters and waitresses alone, so worldwide tipping totals across all services are undoubtedly enormous.

Tipping impacts consumers' wallets (Lynn, 2017b) and dining experiences (Lynn, in press), but arguably its biggest impact is on service workers' incomes. For example, Payscale (2015) reports that the percentage of total income from tips is 12% for baristas, 16% for restaurant hosts/hostesses, 24% for busboys/girls, 31% for bartenders, 42% for banquet captains, and 62% for waiters/waitresses. Given this impact on service workers' incomes, many scholars across diverse academic disciplines have conducted field experiments and quasi-experiments testing ways that servers or their managers can increase the tips consumers leave. They have found that servers can earn larger tips if they:

- use makeup (for waitresses) (Guéguen and Jacob, 2011; Jacob et al., 2009),
- make their hair blond (for waitresses) Guéguen, 2012; Jiang and Galm, 2014),
- wear something unusual in their hair (for waitresses) (Jacob et al., 2012; Stillman and Hensely, 1980),
- wear red shirts or lipstick (for waitresses) (guen and Jacob, 2012, 2014; guen and Jacob, 2012, 2014),
- introduce themselves by name (Garrity and Degelman, 1990),
- use customers' names (Adams and Pettijohn, 2016; Seiter and Givens, 2016; Seiter and Weger, 2013),
- squat next to or sit down at the table (Davis et al., 1998; Leodoro and Lynn, 2007; Lynn and Mynier, 1993),
- stand physically close to customers (Jacob and Guguen, 2012),
- touch customers (Crusco and Wetzel, 1984; Guéguen and Jacob, 2005; Hornik, 1992; Hubbard et al., 2003; Lynn et al., 1998),
- smile (Tidd and Lockard, 1978),
- compliment customers (Seiter, 2007; Seiter and Dutson, 2007; Seiter and Weger, 2010),
- mimic customers' verbal behavior (van Baaren et al., 2003; Jacob and Guéguen, 2013),
- entertain guests with puzzles or jokes (Guéguen, 2002; Rind and Strohmetz, 2001b),
- forecast good weather to customers (Rind, 1996),
- write various messages or draw various pictures on the check (Guéguen and Logeherel, 2000; Jacob, et al., 2013; Rind and Bordia, 1995, 1996; Seiter and Gass, 2005),
- use tip trays with credit cardo logos on them (McCall and Belmont, 1996), and
- give customers free candies (Strohmetz et al., 2002).

E-mail address: wml3@cornell.edu.

http://dx.doi.org/10.1016/j.ijhm.2017.10.004
Received 7 August 2017; Received in revised form 6 October 2017; Accepted 7 October 2017

These studies have been reviewed in a free e-book for servers, called *MegaTips 2: Scientifically Tested Ways to Increase Your Tips* (Lynn, 2011), and have received a lot of attention in the press and online (e.g., Gillman, 2017; Shin, 2014).

Unfortunately, the social sciences are undergoing a replication crisis that calls into question the reliability of this published tipping literature. Scholars have discovered that questionable research practices (collectively called "p-hacking") combine with a bias against publishing null results to make Type 1 errors far more common than typically believed (Simmons et al., 2011; Sterling et al., 1995). In fact, even frequently studied and apparently well-established phenomena such as the effects of ego depletion (Carter and McCullough, 2014), power posing (Simmons and Simonsohn, 2017), and money priming (Vadillo et al., 2016) appear to be smaller and less reliable than assumed. The effects of different tip enhancing techniques, which are less frequently studied and replicated, may also be unreliable. This possibility seems particularly plausible given the subtle and/or transitory nature of many of the purported tip enhancing behaviors and the fact that tipping has been found to be only weakly related to service quality (Lynn and McCall, 2000). Further support for skepticism regarding at least one of these tip enhancing techniques comes from Lynn et al's. (2016) recently reported failure to conceptually replicate Guéguen and Jacob's (2014) finding that waitresses received more and larger tips when they wore red shirts than when their shirts were another color. Given these reasons for skepticism, the collective body of field experiments and quasi-experiments on ways service workers and managers can increase consumer tipping is evaluated in the paper below. Recently developed statistical tools – called p-curves and the replicability-index – are used to assess the evidential value and replicability of the published literature on ways to increase tips.

2. Overview of P-curve and R-index analyses

2.1. P-curve analysis

P-curve analysis uses the frequency distribution of p-values (aka, p-curve) associated with significant effects in a set of studies to assess the likelihood that selective reporting is the sole explanation for the set of effects (Simonsohn et al., 2014; Simonsohn et al., 2015). If selective reporting cannot explain a set of significant effects, that set is said to contain evidential value. Evidential value can be inferred from the shape of the p-curve as described below.

A uniform frequency distribution of significant p-values is indicative of no evidential value – in other words, the significant effects are likely to be selectively reported Type 1 errors. This follows from the fact that p-values reflect the probability of getting a result that large or larger by chance alone. When the null hypothesis is true, there is a 5% chance of getting a p-value of 0.05, a 4% chance of getting a *p*-value of 0.04, a 3% chance of getting a p-value of 0.03, etc… This means that *p*-values between 0.05 and 0.04 should occur 1% of the time, as should p-values between 0.04 and 0.03 as well as p-values between 0.03 and 0.02. In other words, when the null hypothesis is true, significant p-values should be uniformly distributed across values less than 0.05, 0.04, 0.03, 0.02 and.01. Thus, a uniform frequency distribution of significant p-values is expected when the null hypothesis is true and such a distribution provides no evidence of a real effect.

A right-skewed frequency distribution of significant p-values (in other words, with more p-values of 0.01 than 0.05) is indicative of some evidential value – of a true effect underlying at least some of the findings in the set. The shape of p-curves is a function only of effect size and sample size. For any true non-zero effects, p-curves associated with unbiased tests of those effects are right-skewed with the amount of skewness increasing with true effect sizes and/or sample sizes. Thus, right-skewed p-curves are consistent with expectations when non-zero effects are tested and such a distribution indicates that at least some of the tested effects are real.

A left-skewed frequency distribution of significant p-values (in other words, with more p-values of 0.05 than.01) is indicative of intense p-hacking. P-hacking is the use of questionable research practices (such as making post-hoc decisions about how much data to collect and which data points and/or measures to retain in the analyses) to produce "significant effects" and increase the odds of publication. P-hacking tends to produce left-skewed frequency distributions of p-values because the p-hacking effort required to produce significant tests when true effects are zero increases exponentially as the alpha-level (or target p-value) decreases and alpha/p-levels of 0.05 are typically sufficient to get published. Thus, left-skewed p-curves are consistent with expectations when p-hacking is responsible for a set of significant effects and such a distribution indicates that the set of findings provide no evidence of a real effect.

P-curve analysts assess the statistical significance of a p-curve's right skewness using two tests.[1] First, for each p-value less than 0.05 (the "full p-curve"), they compute the probability (called the "pp-value") of getting a value that extreme conditioned on having a value of at least 0.05 and convert those pp-values to z-scores, which are then combined using Stouffer's method. This is the most powerful test of evidential value, but it is potentially biased by ambitious p-hacking that targets alpha-levels below 0.05. To address this concern, a second test is typically done to see if the p-values less than 0.025 (the "half p-curve") are right skewed. For each of these p-values, the probability of getting a value that extreme conditioned on having a value of at least 0.025 is calculated and those pp-values are converted to z-scores, which are then combined using Stouffer's method. This latter test uses less information and has less power than the full p-curve test, but is also less biased by ambitious p-hacking. These two tests are used together to get both of the benefits that each provides – to get both statistical power and resistance to ambitious p-hacking. If the half p-curve's skewness is reliable at the 0.05 level or both the full and half p-curves' skewness is reliable at 0.10 level, then a set of studies is judged to contain evidential value.

2.2. R-index analysis

R-index analysis assesses the replicability of a set of studies by comparing the proportion of reported results that are statistically significant (called the "success rate") with expectations given the median observed (or post-hoc) power of the studies (Schimmack, 2016). Specifically, the replicability-index is calculated using the following formula:

R-Index = Median Observed Power − (Success Rate − Median Observed Power)

This index does not reflect the average probability of replication for a set of studies (a R-Index of 0.22 does not imply an average replicability of 22%), but it is monotonically related to the average probability of replication in the set of studies. Thus, comparisons of R-indices across authors, journals, schools, etc… will reflect their rankings in terms of replicability. Furthermore, the R-index over-estimates average true power that is less than 50% and under-estimates average true power that is more than 50%, so an R-index below 0.50 indicates that the average replicability of the set of studies is below 50% and an R-index above 0.50 indicates that the average replicability of the set of studies exceeds 50%.

3. Identification of studies and effects to be analyzed

This paper presents p-curve and r-index analyses of the body of published field experiments and quasi-experiments testing viable ways

[1] If a p-curve is not reliably right skewed, p-curve analysts also typically test whether the p-curve is significantly flatter than expected if the studies had a power of 33%. This test is a slightly more complicated analog to the right-skewness test and is used to assess whether or not a set of studies contains sufficient information to assess evidential value when the p-curve is not significantly right skewed. If this flatness test is significant, then the set of studies does contain sufficient information to assess evidential value and the absence of right-skewness is meaningful – i.e., indicative of no evidential value in the studies. If it is not significant, then the set of studies does not contain enough information to judge whether it does or does not contain evidential value.