Engström, Per; Forsell, Eskil

Working Paper

Demand effects of consumers' stated and revealed preferences

Working Paper, No. 2013:6

Provided in Cooperation with:
Department of Economics, Uppsala University

Suggested Citation: Engström, Per; Forsell, Eskil (2013) : Demand effects of consumers' stated and revealed preferences, Working Paper, No. 2013:6, Uppsala University, Department of Economics, Uppsala,
https://nbn-resolving.de/urn:nbn:se:uu:diva-198662

This Version is available at:
https://hdl.handle.net/10419/82550

UPPSALA
UNIVERSITET

Department of Economics
Working Paper 2013:6

Demand effects of consumers' stated and revealed preferences

Per Engström and Eskil Forsell

Department of Economics
Uppsala University
P.O. Box 513
SE-751 20 Uppsala
Sweden
Fax: +46 18 471 14 78

Working paper 2013:6
April 2013
ISSN 1653-6975

Demand effects of consumers' stated and revealed preferences

Per engström and Eskil Forsell

Papers in the Working Paper Series are published on internet in PDF formats.
Download from http://www.nek.uu.se or from S-WoPEC http://swopec.hhs.se/uunewp/

Demand effects of consumers' stated and revealed preferences

Per Engström*, Eskil Forsell*

April 19, 2013

Knowledge of how consumers react to different quality signals is fundamental for understanding how markets work. We study the online marketplace for Android apps where we compare the causal effects on demand from two quality related signals; other consumers' stated and revealed preferences toward an app. Our main result is that consumers are much more responsive to other consumers' revealed preferences, compared to others' stated preferences. A 10 percentile increase in displayed average rating only increases downloads by about 3 percent, while a 10 percentile increase in displayed number of downloads increases downloads by about 20 percent.

Keywords: peer effects, observational learning, stated preferences, revealed preferences, eWOM, Google play, Android apps, regression discontinuity design, instrumental variable analysis.

JEL: C21, C26, D83, M31

*Department of Economics, Uppsala University, Box 513 , SE-751 20 Uppsala, Sweden. e-mail: (per.engstrom@nek.uu.se / eskil.forsell@nek.uu.se). We thank *Droidmeter* for providing the data; Tommy Andersson, Sebastian Axbard, Niklas Bengtsson, Mikael Elinder, Che-Yuan Liang, Eva Mörk, Olle Nilsson, Jonas Poulsen, Fredrik Åhlund and seminar participants at the economics departments at Lund University and Uppsala University for valuable comments.

1 Introduction

The existing economic literature on consumer peer effects and electronic word of mouth (eWOM) has to a large extent focused either on what other consumers (or experts) say about products, or to what extent other consumers bought the products. Economists tend to think that "talk is cheap" and that true preferences are reflected by actions rather than words. From an economist's perspective, a natural presumption is therefore that the effects of other consumers' actual choices should have larger effects than other consumers' stated preferences. In this study we will focus on both signals: what people say about a product and whether a product is in high demand. In particular, we will ask which signal that matters the most to prospective consumers: other consumers' stated (eWOM) or revealed preferences (observational learning).[1]

The online revolution has facilitated a dramatic increase in the commercial use of user feedback systems. For almost any type of product – e.g. a vacation, a smart phone or a movie – it is now possible to easily access a host of relevant measures of the popularity of a product or service such as user reviews or average ratings. Economists and market researchers are becoming increasingly interested in the workings of eWOM.[2] The central economic question is to what extent electronic user feedback and user statistics actually influence the decisions of potential buyers.

Identification of such causal effects is tricky since measures such as ratings and reviews from a specific site is potentially not the only information that consumers have gathered before making their purchasing decisions. Information about products may come from other sources; e.g. friends, experts and bloggers may also have provided useful signals of the product's quality. High quality of a product may hence induce both high demand, and high user rating, which confounds a causal interpretation of the association between user statistics and demand indicators. A large number of studies have therefore used experimental and quasi-experimental setups to measure the causal effects of eWOM and other related peer effects.[3]

[1] Chen et al. (2011) and Li & Wu (2012) study similar questions. However, their identification strategies are very different from ours.

[2] For a seminal theoretical (economics) contribution see Avery et al. (1999) and for a recent survey see Chan & Ngai (2011).

[3] For the effects of expert reviews see: Reinstein & Snyder (2005) (movies) and Hilger et al. (2011) (wine). For general peer effects see: Duflo & Saez (2002) and Sorensen (2006) (retirement decisions/health plan); Hesselius et al. (2009) (sickness absence); Moretti (2011) (movies). And for average user rating's effect on demand see: Melnik & Alm (2003); Jin & Kato (2006); Resnick et al. (2006); Lucking-Reiley et al. (2007); Cabral & Hortaçsu (2010) (ebay auctions) and Chevalier & Mayzlin (2006) (books).

For the effects of observational learning, see Bandura (1977) for a seminal psychological contribution, and Banerjee (1992) and Bikhchandani et al. (1992) for seminal economics contributions. Prominent empirical studies that isolate the causal effects of other's choices include, Cai et al. (2009), Salganik et al. (2006) and Tucker & Zhang (2007).

The first part of our analysis concerns the effect of others' stated preferences. In order to identify the effect of average user ratings, we utilize the specific way that *Google play* presents the ratings of its apps. In the list view the consumer only observes the average user rating rounded off to the closest half star. A single rating ranges from 1 to 5 stars, in whole star increments, which means that the presented average rating only takes on values in the range: $(1.0; 1.5; 2.0; \ldots; 4.5; 5.0)$. An app that has the exact rating 3.249 will thus be rounded down to 3.0, while an app with exact rating 3.251 will be rounded up to 3.5. Standard results from regression discontinuity design (RD) states that if the app distribution around the thresholds $(1.25; 1.75; \ldots; 4.75)$ is continuous, the causal effect of presented average ratings will be identified. Intuitively we may think of each threshold as a local randomized experiment since apps slightly below and slightly above are, under some conditions, almost identical in expectation. This specific application of RD was pioneered in two recent studies that, independently, applied the same technique on the restaurant review site, Yelp.com. This site presents average user ratings in a way that is similar to that of Google play. The two previous studies find that higher user ratings leads to higher reservation rates (Anderson & Magruder, 2012) and higher revenue (Luca, 2011).

The RD strategy has a particular advantage in this type of setting, where the discontinuity thresholds appear repeatedly throughout the whole range of the forcing variable (average rating). Formally an RD estimate has a local average treatment effect (LATE) interpretation around the identifying threshold. If the threshold appears in a region where mostly "unrepresentativ" elements locate, a LATE estimate may not be a good proxy for the average treatment effect (ATE), which is often more relevant. The fact that we use 10 different thresholds, spread out evenly throughout the whole range of average rating, makes the estimated effect more generalizable.

This first part of our study, where we study stated preference, makes several contributions to the literature on eWOM. First of all, our analysis is based on a much larger dataset than the two previous studies that use a similar identification strategy. Luca (2011) uses 854 restaurants (quarterly data) in his main RD specification while Anderson & Magruder (2012) have access to data from 328 restaurants (daily data) for their main outcome variable (reservations). We use a dataset from *Droidmeter* that includes daily scraped information for all apps on Google play. Our main RD specification is based on data for 42 consecutive

days covering about 50,000 apps.[4] Our large dataset allows for a richer set of RD specifications and robustness checks.

An additional contribution is that we study a novel market that is interesting in itself. The app-market is global and is growing very rapidly: from Google play's introduction in 2008 more than 25 billion apps have been downloaded from the store (as of the 26 September 2012). The one billion threshold was reached as late as in 2010 which indicates a very rapid increase since then.[5] Furthermore, given Google play's position as the official Android marketplace, the ratings on Google play are a very important source of information for potential consumers.

The last contribution relates to the details of how Google play presents information. When apps are listed, the average rating is displayed at half-star precision level. However, when a specific app is chosen, so that more detailed information is displayed, the total rating distribution and the average rating on 0.1 star precision level is displayed. A positive effect of being rounded up to the closest half star therefore implies that users are inattentive to easily accessible information.[6] The RD framework can thus be applied also to the more fine-grained, less saliently displayed, 0.1 star roundoffs. However, we find no indication that consumers react to the average rating displayed at 0.1 star level at all. This indicates that consumers mostly respond to the more salient average rating that is rounded off to the closest half star.

We find that the causal effect of an additional half star corresponds to a $0.015-0.020$ percentage point increase in the daily percentage change in number of ratings (daily rating rate). We use the daily rating rate as a proxy for the daily download rate as we do not have exact data for the latter measure. The baseline daily increase in number of ratings is 0.425 percent. Given this, the percentage point increase translates into a relative effect of a $4-5$ percent increase in the daily rating rate as a result of an additional half star.

As a robustness check we also perform the analysis with actual downloads, measured in intervals (more on this below), as the outcome variable. The estimated effect is similar (but less precisely estimated) which indicates that the daily rating rate is a good proxy for the daily dowload rate.

The estimated effect is stable for a wide range of bandwidths and is statistically unaffected by the inclusion of a number of control variables. It is, however, larger for less downloaded apps (apps with fewer ratings). One explanation could be that information about more popular apps is also spread through many other channels, such as blogs, forums and off-line peer interaction.

[4] The total number of apps available at Google play is more than 10 times larger than this. However, most of the apps have very few downloads, and even fewer ratings, which makes them unsuitable to include in the RD analysis.

[5] Source: the official android blog (http://officialandroid.blogspot.se/)

[6] For a seminal contribution see Simon (1955). Later applications include: Chetty et al. (2009), Gabaix & Laibson (2006), Finkelstein (2009), Lacetera et al. (2011), Pope (2009).

The second part of our analysis complements the above analysis by using number of downloads to measure the effect of others' revealed preferences. We specifically estimate the causal effect of increased number of displayed downloads, on future download rate (as above, approximated by the rating rate).

Google play displays the number of downloads for each app in rough intervals. If we had information on exact downloads, we could in principle identify the causal effect of switching download interval through the above RD strategy. However, we only have access to the publicly available information on Google play, which makes an RD strategy infeasible. Instead we use the particular way the download intervals are specified to identify the effect. We specifically use the fact that the (relative) interval lengths oscillate back and forth; a short interval will be followed by a long interval, and vice versa. The first few intervals will describe the logic: download interval 1, $(1-5$ downloads); interval 2, $(5-10$ downloads); interval 3, $(10-50$ downloads); interval 4, $(50-100$ downloads) and so on. This means that an app starting at the lowest level of an odd (long) interval will have to fivefold its number of downloads to enter the next interval, while an app starting in an even (short) interval will only have to double its downloads to increase interval. An app starting in a short interval the first day we observe it, will be much more likely, than an app starting in long intervals, to have switched to a higher interval by the last day we observe it. In section 4. we exploit this variation as an instrument to identify the effect of switching to a higher download interval.

We find that the daily rating rate increases by somewhere between 0.16 and 0.45 percentage points when an app switches to a higher download interval. In comparison the baseline daily rating rate for apps that switch download intervals is around 1.3 percent – it is natural that the switching apps have a higher baseline rating rate compared to the unconditional average rating rate (which is 0.4 percent). This translates to approximately a $20-30$ percent increase in the daily rating rate, as a result of changing download interval.

To get an idea of which of the two quality signals affects consumers the most, we need to make the estimates from the two signals comparable. For stated preferences, we calculate the average percentile increase in displayed average rating distribution an additional half star would induce. For revealed preferences, we calculate the corresponding average percentile increase in the download interval distribution a higher interval would induce. By weighing our estimated effects against these averages we can compare the relative importance of the two types of quality signals.

The interpretation of our normalized effects are as follows. For stated preferences, we find that a hypothetical app which increases 10 percentiles in the displayed average rating distribution, will approximately experience a 3 percent increase in the daily rating rate. For revealed preferences, the effect on the daily

rating rate of increasing 10 percentiles in the displayed download distribution is instead about 20 percent. Our result thus strongly supports the notion that, in a consumer's view, talk is relatively cheap compared to actions. That consumers respond more to others' revealed preferences is in line with the typical economist's notion that true preferences are reflected in peoples actions rather than words.

The rest of the paper is organized as follows: section 2. describes our data; section 3. describes our method and results when analyzing stated preferences; section 4. does the same for revealed preferences and section 5. concludes.

2 Data

Our data was collected from the publicly available app information on the US version of Google play, which contains approximately 500 000 apps. We collected daily information for 42 consecutive days ("sessions") in October and November 2012. The data consists of the following information for each app and day: exact rating distribution (and hence average rating as well as rounded measures), price, size, category, time since last update, ranking on a leader board (if any) and minimum and maximum number of downloads (download interval).

Google play is available across different types of devices running the Android operating system, such as smart phones and tablets, as well as on desktop computers. The starting page contains top charts where apps that stand out in some way are placed (the algorithm determining which apps are placed on these charts is known only to Google).

When searching for apps, the page of search results contains a brief description of each app along with the option to install each app. For apps that have one rating or more, the store displays a meter of 1 to 5 stars in half star intervals calculated from the user supplied 1 to 5 ratings (see figure 1a). When clicking an app the consumer reaches the "app page" where more detailed information about the app is available such as exact rating distribution, an interval indicator for number of downloads (but not exact number of downloads), a graph of download history for a number of days and written user reviews (see figure 1b). In addition to the previously described star meter and precise rating distribution, the app page also displays average rating rounded to the closest 0.1 decimal. Due to the smaller screen size of many mobile devices the user is usually required to scroll down in order to see much of this information when using the Android version of Google play, which makes the more detailed rating information much less salient than the average rating rounded off to the closest half star.

3 Stated preferences

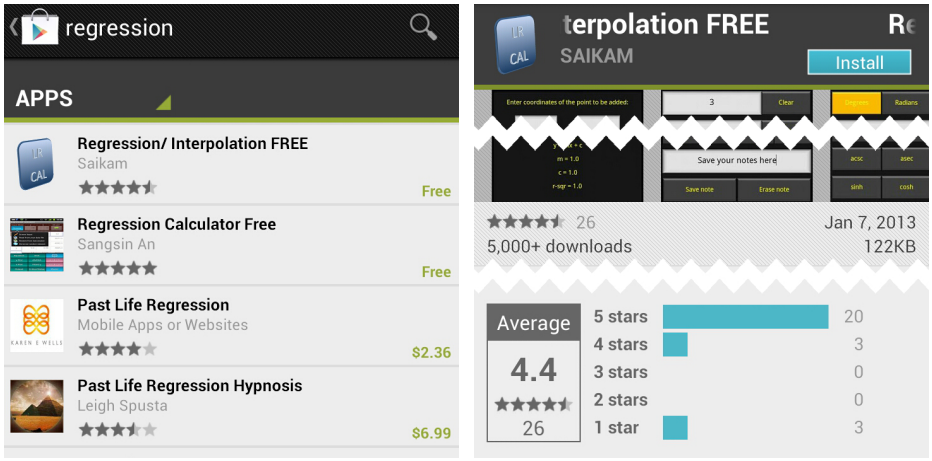
3.1 Method

In order to identify the causal effect of average rating on download rate we use a regression discontinuity design (RD). The intuition is that apps with similar average ratings resemble each other closely in unobservable (and observable) characteristics, such as quality. The apps that are located slightly above and below a given round-off-threshold are likely to be very similar in expectation, while the displayed, rounded off, average rating differ a lot. In the limit, when the bandwidth around the thresholds tends to zero, the RD design will reproduce a local randomized experiment, provided that there is some randomness in the average ratings.

The RD strategy can be specified as follows:

$$y_{i,t+1} = \alpha + \tau \cdot D_{it} + f(r_{it} - c_{it}) + \boldsymbol{W_{it}\gamma} + \varepsilon_{it} \tag{1}$$

where $y_{i,t+1}$ is the logarithm of the quota of number of ratings[7] in period $t+1$ and t (i.e. the approximate percentage change, henceforth denoted as daily rating rate), r_{it} is the actual average rating for app i in period t, c_{it} is the half star round-off-threshold closest to r_{it}, $D_{it} = \mathbf{1}_{\{r_{it} \geq c_{it}\}}$ and $\boldsymbol{W_{it}}$ is a potential vector of covariates. From here on normalized average rating will be denoted $\tilde{z}_{it} = r_{it} - c_{it}$.



(a) A search for the word "regression" among Android apps. **(b)** Information presented on the app page.

Figure 1: Screenshots from the mobile version of the Google play store.

[7] As mentioned in the introduction, we do not know the exact downloads. The rating rate will serve as our proxy for download rate. The validity of this proxy will be tested in the end of this section

The function $f(.)$ captures the underlying continuous relationship between the average rating and the app's rating rate. In particular, if higher quality of an app leads to both higher average rating and higher rating rate, this confounding factor will be picked up by $f(.)$. If this function is discontinuous at the threshold, τ will not be identified (the derivative is however allowed to be discontinuous). As in any RD setting, this assumption is not directly testable but we can perform various robustness tests to validate the assumption.

Our forcing variable, $\tilde{z}_{it}$, ranges from -0.25 to 0.25 as the star discontinuities are distributed in even 0.5 intervals along average rating. In the following analysis we choose to pool our data for all cutoffs. This allows for a more general interpretation of our estimates as the LATE will be estimated on different discontinuities along the whole range of average ratings.

The fact that we use publicly available information means that the consumer downloading an app could potentially know as much about an app as we do. In principle the consumers may be separated into three different levels of sophistication: i) the first type only reacts to the most salient information, the 0.5 star average ratings; ii) the second type instead reacts on the 0.1 decimal precision level that is displayed in the detailed app view; iii) the meticulous third type of consumer does not react to the rounded off figures at all since the whole rating distribution is also available in the detailed app view. Our main empirical strategy only isolates the type i) reaction to average ratings. However, the results from a similar analysis on the 0.1 decimal level give no indication that consumers react to the 0.1 decimal averages at all.[8] This makes the extreme type iii) reaction even more implausible. This suggests that the empirically relevant reaction is due to the most saliently displayed figure, i.e. the 0.5 star ratings, and that most consumers are of type i).

As mentioned above we do not have access to the exact number of downloads of each app. Instead, the daily rating rate will serve as a proxy for the daily download rate. This strategy implicitly assumes that the round-off-indicator, D_{it}, is orthogonal to the rating ratio, i.e. the number of ratings divided by the number of downloads. If there is a psychological effect that makes users more inclined to rate apps that are rounded up, this assumption is violated. However, since we have access to downloads by interval we can perform a robustness check in which the outcome variable instead is an indicator for daily increase in download interval.

Related to this, there is most likely a certain lag between the downloads and the ratings of an app. Furthermore, our data is sampled daily. While this gives

[8] The statistically insignificant point estimates are actually negative. This analysis follows the one detailed below in everything but cutoff thresholds and the fact that we exclude the eight thresholds that coincide with the half star thresholds. The latter modification is in order to avoid attributing the effect of an additional half star to the 0.1 decimal averages.

very good precision, ideally we would have had access to event based data for every new download, rating etc. We cannot be sure if an app that switched rounded value between two days did so just after the first or just before the second – we cannot know for how large fraction of the day it was "treated". These issues both lead to a potential attenuation of the effect of ratings on downloads, and will bias our estimates towards zero.

A minor additional problem with the number of ratings is that this number may fall between two days. This peculiarity is due to the fact that each rating is tied to a user account. Each user can only rate an app once but can if they wish edit, or even remove their rating. Fortunately, the empirical relevance of this issue is minor. Out of a total of about 16.5 million app-days (for which the apps have one rating or more) there are slightly less than 33 000 app-days for which $y_{i,t} < 0$ and the exclusion of these makes a negligible impact on the results.

3.2 Results

The main identifying assumption for our analysis is that the only variable exhibiting a discontinuous behavior around our cutoff is the number of half stars shown. Precise manipulation of the forcing variable, normalized average rating ($\tilde{z}_{it}$), could invalidate this assumption.

Figure 2. shows the frequency distribution of normalized average rating and the unmodified version for apps with more than 50 ratings. These figures give no indication of manipulation and the distribution looks smooth across the cutoff. The highest and lowest bins are the only outliers. The highest bins consists of apps that have been rounded up just slightly (e.g. 3.499), while the lowest bin consists of apps that have been rounded down just slightly (e.g. 3.501). This striking pattern is a bit of a mystery and persists when excluding

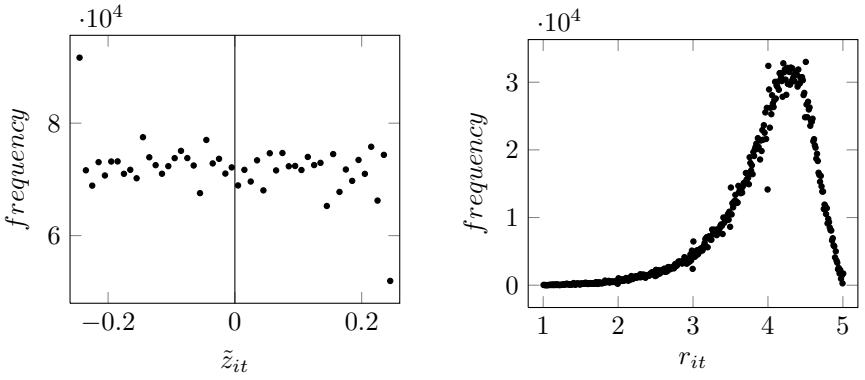


Figure 2: Frequency distribution of the above variables in 0.01-bins, for apps where the number of ratings are greater or equal to 50.

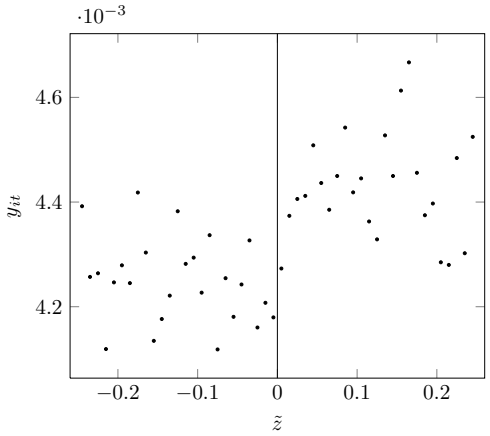


Figure 3: Average daily rating rate by 0.01 bins in normalized average rating. Number of ratings ≥ 50.

apps with ratings corresponding exactly to half a star (eg. 3.500). We have no explanation for why apps that are just above the rounded off values are much more common than apps that locate just below the rounded off values. Fortunately, these regions are located the furthest away from the rounding off thresholds used for identification, and will only be included in the maximum bandwidth sample.

Apart from there being no visual indication of manipulation related to the rounding off thresholds, the rating mechanism in the store makes manipulation of app ratings difficult (and is probably intended to do just that). In order to rate an app the user has to have installed the app on a device, ratings are tied to each user's account and each user can only vote on an app once (Google, 2012). Precise manipulation of ratings would thus require a large number of votes coming from multiple accounts and installations and would become increasingly difficult as the number of ratings grows.

Figure 3. shows the average daily rating rate, by normalized average rating. Overall there is a clear indication that the rating rate is higher for the apps that are rounded up to the closest half star.

As mentioned above we have no reason to believe that our forcing variable is being precisely manipulated. However, as ratings are given as discrete numbers we have a potential problem of artificial discontinuities in the variable. To avoid such confounding discontinuities we have to restrict our analysis to apps with a certain number of minimum ratings. So far we have conditioned the sample to apps with at least 50 ratings. In an attempt to formalize the selection of such a threshold we perform a common test of discontinuities in the frequency of the forcing variable around the cutoff due to McCrary (2008). To implement this test for all days at once we would have to modify it to take the panel structure

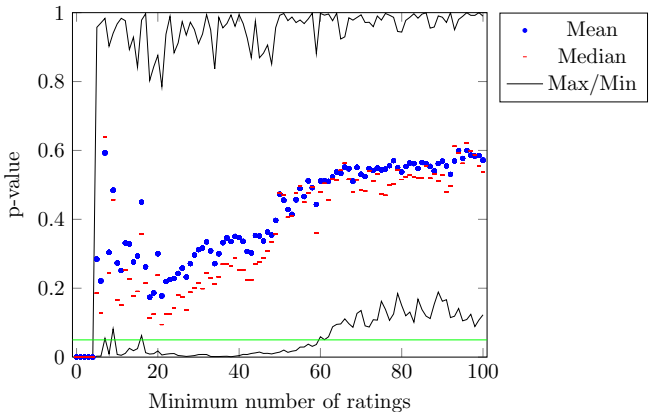


Figure 4: Summary of p-statistics from McCrary density tests performed seperately for each day over a range of minimum number of ratings. The horizontal line represents a p-value of 5 %.

of our data into account (in order to get unbiased standard errors). However, we choose a simpler approach and instead perform the test separately for each day over a range of minimum number of ratings. Figure 4. summarizes these estimates.

The figure shows that for a minimum number of ratings above four, the tests are on average non-significant at the five percent level, while for some days the tests are highly significant. Since one in 20 independent tests is significant by chance under the null, it would be overly conservative to require that the test must be insignificant for all days. The graph shows that the average significances of the tests decrease as the minimum number of ratings increases. In particular, the average significance decreases drastically in the region between 40 and 60 minimum ratings. In fact, for a minimum number of ratings above 60, the test is not significant at the 5 % level for a single day out of our total of 42. Based on these tests it seems that the distribution is quite smooth already at a minimum number of ratings around 10. However, to err on the safe side, and due to the large drop in average significance around 50 minimum ratings, we choose this as the number of minimum ratings in our preferred specification.

Table 1. shows the estimate of τ in model 1. across a number of different bandwidths and specifications of $f(.)$, all excluding covariates. To examine which specification of $f(.)$ that best fits the data we follow the recommendation of Lee & Lemieux (2010) and perform a test of joint significance on a set of bin dummies included in a modified version of each model. The p-values from these tests are presented in square brackets.

As is to be expected, the joint significance of these dummies is reduced as

Table 1: RD estimates of the effect of being rounded up to nearest half star, on the daily rating rate.

Polynomial order:	Bandwidth: 0.25	0.20	0.15	0.10	0.05	0.03
Zero	.0001749	.0001927	.000182	.0001969	.0001708	.0001692
	(.000042)	(.0000473)	(.0000508)	(.0000549)	(.0000627)	(.00007)
	[.2623477]	[.4923134]	[.7909614]	[.750983]	[.592381]	[.7790087]
One	.0002113	.0001818	.0001786	.0001509	.0001249	.0000825
	(.0000676)	(.0000712)	(.0000748)	(.0000816)	(.0001004)	(.0001228)
	[.2541608]	[.4531965]	[.7071857]	[.7299795]	[.8442352]	[.8190711]
Two	.0001324	.0001389	.000162	.0001105	.0001158	.0001811
	(.0000845)	(.0000886)	(.0000959)	(.0001092)	(.0001423)	(.0001736)
	[.1953995]	[.5211518]	[.7757419]	[.7513663]	[.8426831]	[.7831818]
Three	.0001591	.0001774	.0000539	.0001375	.0001719	.0003979
	(.0001007)	(.0001075)	(.0001174)	(.0001347)	(.0001808)	(.000228)
	[.1551933]	[.5049997]	[.8119239]	[.8096395]	[.6809329]	[.6027584]
Four	.0001296	.0000399	.0001594	.0001093	.0003439	.000686
	(.0001175)	(.0001257)	(.0001395)	(.0001638)	(.0002213)	(.0002714)
	[.1412722]	[.6607629]	[.85897]	[.8002244]	[.2987714]	[.4112361]
Five	.0000688	.0001151	.0001622	.0002312	.0006298	.0007534
	(.0001332)	(.0001447)	(.0001606)	(.0001936)	(.0002584)	(.0003066)
	[.1640295]	[.6701925]	[.8489573]	[.6398752]	[.5329843]	[.330277]
	BIC preferred polynomial order:					
	0	0	0	0	0	0

Note: Standard errors in parenthesis. P-values from goodness-of-fit test in square brackets. The goodness of-fit-test is obtained by jointly testing the significance of a set of bin dummies included as additional regressors in the model. The bin width used to construct the bin dummies is 0.01. The optimal order of the polynomial is chosen using the Bayesian information criterion. Number of ratings ≥ 50.

the bandwidth decreases. However, they are insignificant at conventional levels for all bandwidths in the zero order polynomial specification. This implies that a plain comparison of means fits our data reasonably well even at maximum bandwidth. While this may seem surprising at first, it is due to the fact that by pooling data over all thresholds, the mechanical relationship between the forcing variable (distance to closest rounding off threshold) and average rating is heavily attenuated – in fact, depending on the average rating distribution, the correlation may even be negative. That the zero order polynomial is optimal is also indicated by the Bayesian information criterion; the preferred specification of $f(.)$ is the zero degree polynomial for all bandwidths.

The second striking observation that can be made from table 1. is the similarity in magnitude of the estimated effects. For the lower degree polynomial functions, the estimates are mostly in the same quite narrow range. The only

exceptions are the linear regressions for quite narrow bandwidths. This is also reflected in the p-values from the dummy tests in that they are slightly lower (but far from significant) at lower bandwidths with a zero degree polynomial. Returning to figure 3. it can be noted that the average daily rating rate just above the cutoff seems to be more in line with the average to the left of the cutoff, which explains the patterns in table 1. We believe that this is caused by the data collection and proxy issues detailed in the previous section rather than a misspecification of the forcing variable or a failure of our identifying assumption. The fact that the outlying estimates in table 1. are mainly observed at very narrow bandwidths instead strengthens the argument that there is no correlation between normalized average rating and average rating. The pattern is instead likely due to a slight attenuation of the treatment status very close to the threshold. This would lead to an artificial trend in a narrow region around the threshold – when zooming in close enough, any discontinuity would look like a trend, given that an ever so tiny measurement error is present.

Using the information in table 1. and figure 3. it seems natural to exclude the function $f(.)$ altogether and simply make a comparison of means estimate for some bandwidth. However, as treatment is only "randomized" *at* the cutoff such a method will give estimates with a linear bias in the choice of bandwidth. Imbens & Lemieux (2008) show that this bias (notation is modified to fit the current situation) will be:

$$\text{bias}(\hat{\tau}) = \frac{h}{2} \cdot \left(\lim_{\tilde{z}_{it} \downarrow 0} \frac{\partial}{\partial \tilde{z}_{it}} \mu(\tilde{z}_{it}) + \lim_{\tilde{z}_{it} \uparrow 0} \frac{\partial}{\partial \tilde{z}_{it}} \mu(\tilde{z}_{it}) \right) + O(h^2)$$

where h is the bandwidth and $\mu(.)$ is the conditional mean of y_{it}. In order to reduce this bias it is common to include a first degree polynomial which will generally give a bias of order h^2 instead. As is obvious from the specification of the bias the linearity in h stems from the fact that the conditional mean of the dependent variable is often related to the forcing variable. On the other hand, if the relationship between the forcing variable and the conditional mean is nonexistent on both sides of the cutoff (so that $\lim_{\tilde{z}_{it} \downarrow / \uparrow 0} \frac{\partial}{\partial \tilde{z}_{it}} \mu(\tilde{z}_{it}) = 0$) the estimates from a simple comparison of means will exhibit a bias of a similar order as those from a local linear regression.

We have argued above that pooling our data for many cutoffs gives no reason to believe that there truly is a relation between our forcing variable (normalized average rating) and the conditional mean of our the outcome variable (daily rating rate). As we have no reason to believe that the bias properties of the local linear regression are better than those of a standard comparison of means, we use the latter approach when presenting our results below.

Figure 5. shows the estimates of τ at a 0.1 bandwidth for an increasing minimum number of app ratings. As we exclude apps with few ratings (and

hence downloads) the point estimates steadily decrease. This is in line with the intuition that the effect of average ratings on downloads would be largest for apps where few other signals of quality are available.

Figure 6a. shows the estimate of τ across a large number of bandwidths. As in table 1. the point estimates are stable across bandwidths (around 0.00019) but drop (along with the significance of the estimates) at smaller bandwidths. When comparing to the average baseline daily rating rate among apps to the left of the cutoff (0.00425), an estimate of 0.00019 translates into approximately a 4.5 % increase in the daily rating rate.

Figure 6b. shows the same estimates when including covariates.[9] In general the estimates are of the same magnitude as when not including covariates. This is reassuring as it suggests that the covariates are balanced over the cutoff and that our identifying assumption is not violated.

Figure 7. shows the estimates of a comparison of means just as the one in figure 6a. with the available covariates as dependent variables. Overall there seem to be little evidence of the covariates being unbalanced. Only two estimates are marginally significant for a bandwidth of 0.15, "share of 4 star ratings" and "months since update". While the significance of the former estimate quickly disappears when the bandwidth is reduced, the latter is more persistent. That the number of months since an app was updated would in general be lower for apps with higher average ratings seems intuitive, but as mentioned the correlation between $\tilde{z}_{it}$ and r_{it} is fairly low. Even if there was such a trend it would be expected to be a continuously decreasing function in $\tilde{z}_{it}$ and not exhibit any discontinuity around the cutoff. We are unable to come up with a reasonable explanation for this marginally significant discontinuity and we simply attribute

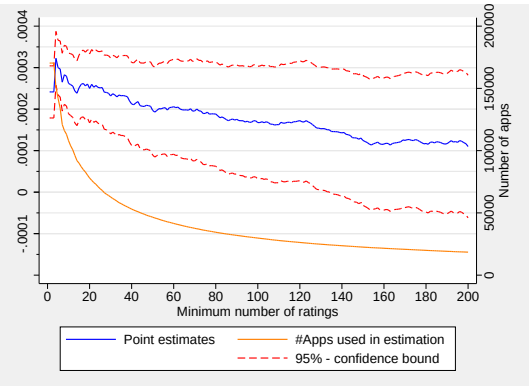
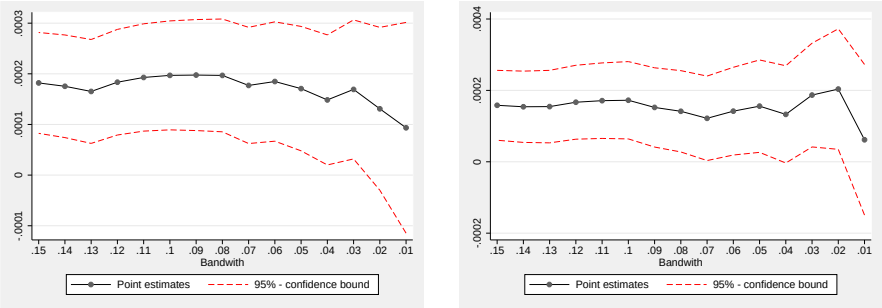


Figure 5: Estimates of being rounded up on daily rating rate, and number of apps included in these estimations, across different number of minimum ratings. Covariates excluded, bandwidth is 0.10.

[9]Dummies for: months since update; day; app category; leaderboard; price and size.



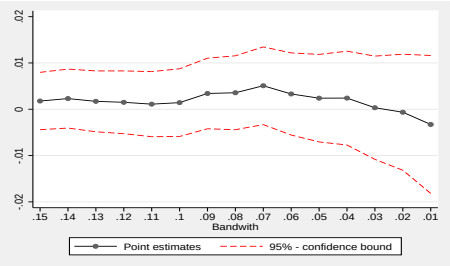
(a) No covariates. **(b)** Covariates.

Figure 6: Non parametric estimates of being rounded up, on daily rating rate, across different bandwidths. Number of ratings ≥ 50.

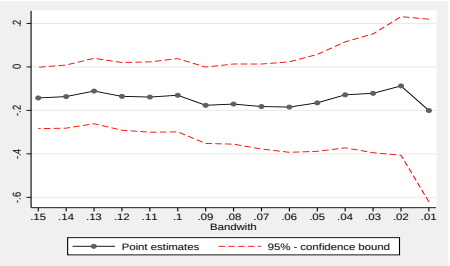
this to chance.

Worth noting is the bottom right sub-figure of figure 7. which shows the comparison of means estimates of an extra half star's effect on average rating. Note that if we only used one rounding off threshold, there would be a mechanical one-to-one positive relationship between normalized average rating and average rating. However, when pooling the data we actually get a negative insignificant relation between normalized average rating and average rating. This takes the RD closer to a proper randomized experiment and further strengthens the case for covariates related to average rating being balanced around the cutoff. In addition, it also lends support to excluding the polynomial when estimating the effects.

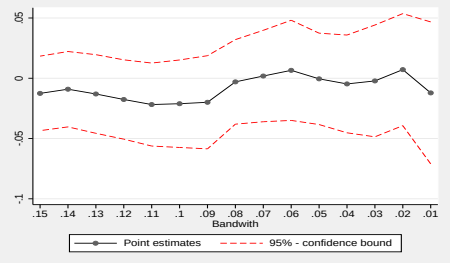
In all of the analysis above we have used the daily rating rate (percentage change in the number of ratings between one day and the following) as a proxy for the daily download rate. It might be argued that finding an effect on this proxy variable does not necessarily mean that the download rate is affected. A skeptic may instead attribute the effect to some psychological reaction related to the rating inclination of app users. If the consumers' inclinations to rate an app are affected by the difference between an app's average rating and the rounded off average rating, the relationship between our preferred outcome variable (download rate) and our proxy (rating rate) is not stable over the rounding off threshold. To investigate this we run the same comparison of means regressions as above on an indicator variable that is one in the day that an app changes download interval; e.g. if it switches from being classified as having $(5-10)$ downloads to having $(10-50)$ downloads this indicator will be one for the first day in the higher interval. Among our total of 3 508 814 app days for apps with a minimum number of ratings of 50 there are only 14 004 days in which an app changes download interval. This means that a substantial amount of variation is thrown away when using the indicator variable instead of the proxy variable.



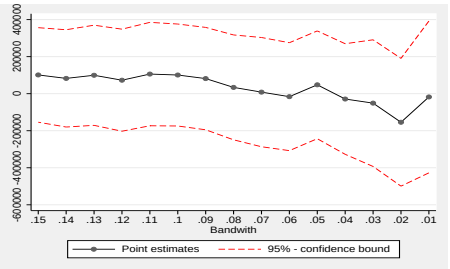
Leaderboard dummy



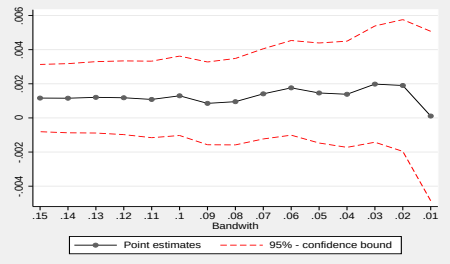
Months since update



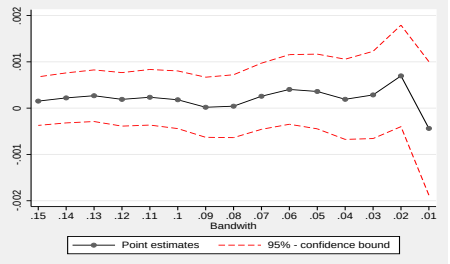
Price in US dollars



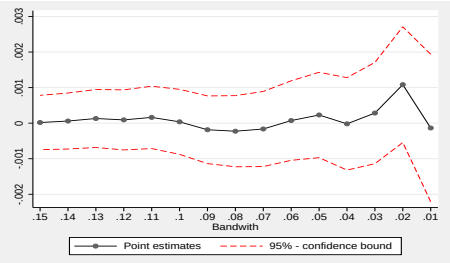
Size in bytes



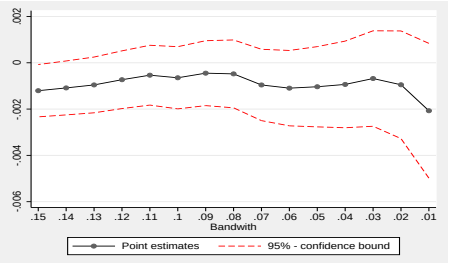
Share of 1 star ratings



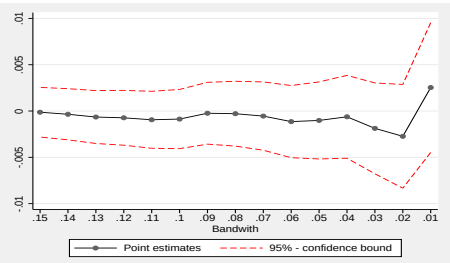
Share of 2 star ratings



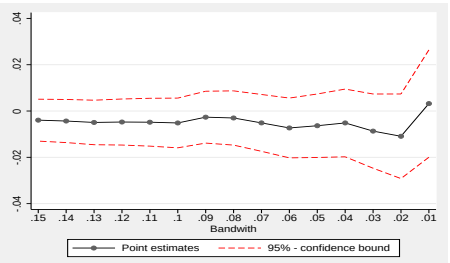
Share of 3 star ratings



Share of 4 star ratings



Share of 5 star ratings



Average rating, r

Figure 7: Non parametric estimates of being rounded up on different covariates across different bandwidths. Number of ratings ≥ 50.

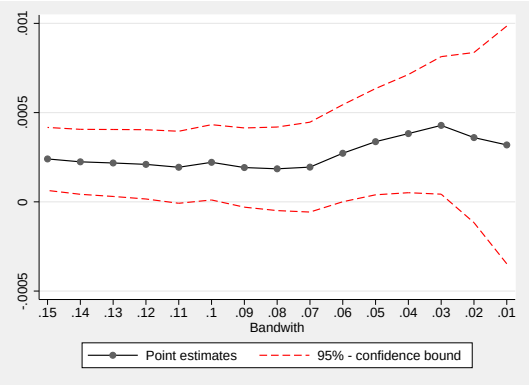


Figure 8: Non parametric estimates of being rounded up on changing download interval. Number of ratings ≥ 50.

Figure 8. shows the estimates over a range of bandwidths. While they are mostly just marginally significant at the 5 % level, this marginal significance prevails until the bandwidth is very low (in fact for some smaller bandwidths the estimates are significant at the 5 % level). Overall, figure 8. indicates that the rating rate is a good proxy for the download rate. The baseline probability of changing download category for apps that are rounded down is 0.00383 (for a 0.10 bandwidth) which means that an estimate of 0.00022 translates into an approximately 5.7 % higher chance of changing download category for apps with one extra half star. This is very close to the estimated effect on the the daily rating rate (4.5 %).

4 Revealed preferences

4.1 Method

The purpose of this section is to estimate the causal effect of the displayed number of downloads on future download rate (as above, approximated by the rating rate). The presumption is that a large number of downloads is an important quality signal that will have a positive effect on future downloads. The problems of causation are obvious since we essentially want to measure the stock variable's (number of downloads) effect on the corresponding flow variable (download rate).

However, the details of what information Google play reveals to the users, and what remains unobserved, provide a potential solution for identification, but not without some caveats. As mentioned above, the Google play store does not reveal the exact number of downloads for an app. This is good since the signaled number of downloads remains constant for a long time while the true number of

Ratings			Interval
Min	Max	$^{\text{Max}}/_{\text{Min}}$	type
1	5	5	Long
5	10	2	Short
10	50	5	Long
50	100	2	Short
...	...	...	...

Table 2: Structure of download intervals on the Google play store.

downloads changes. If we had access to the true number of downloads, we could use the same type of regression discontinuity approach as above to estimate the causal effect of signaled number of downloads on download rate. Unfortunately, as we only have access to publicly available data we do not have the required running variable, exact number of downloads, and hence cannot use a regression discontinuity design to analyze how daily downloads change in response to the displayed download intervals.

Another way to analyze the effect in question using our data would be to look at apps that switch download intervals and compare the rating rate before and after that switch. However, this method would lead to biased estimates since temporary positive shocks to the download rate will be mechanically overrepresented just prior to a switch in download interval. An instrument for switching download interval could solve the identification problem through the ordinary IV-approach. Such a variable needs to be highly correlated with switching download interval but plausibly uncorrelated with the unobservable characteristics, such as underlying quality or temporal shocks to the download rate.

Due to a nice feature of how download intervals are defined on Google play we have access to an instrumental variable that potentially satisfies these requirements. Table 2. shows the minimum and maximum number of downloads for the first four download intervals. It is clear that for every other download interval a fivefold increase in the number of downloads is required to advance to the next interval, while for the other intervals a doubling is enough. This structure is repeated for all intervals. We label these two types of download intervals long and short respectively.

The type of interval an app initially belongs to is somewhat correlated with the previous rating rate (apps with a higher rating rate are more likely to be found in a long interval). However, this difference in means will be absorbed by app fixed effects in our specification. Apps starting in short intervals should be more likely to have switched to a higher download interval for any given time, compared to apps starting in long intervals. This is the variation that we use for identification. Technically our instrument consists of the interaction between

initial download interval type and time (day dummies).

In the following analysis we thus run two stage least squares regressions using the interactions between initial type of download interval, and dummies for each day as instruments for whether an app has switched download interval. This instrument specification translates into the cumulative probability of an app, in each respective interval type, having switched download interval in a given day. That apps in short intervals will have a higher probability of switching download interval allows us to disentangle the effect of this from a general time trend. Formally our model is specified as:

$$y_{i,t+1} = \alpha_i + \kappa \cdot I_{it} + \boldsymbol{S_t\delta} + \boldsymbol{W_{it}\gamma} + \varepsilon_{it} \tag{2}$$

$$I_{it} = \alpha_i' + T_i\boldsymbol{S}_t\boldsymbol{\phi} + \boldsymbol{S_t\delta'} + \boldsymbol{W_{it}\gamma'} + \upsilon_{it} \tag{3}$$

where y_{it} (the rating rate) is specified as above, α_i and α_i' are app specific constants, I_{it} is an indicator for whether app i has switched intervals by day t, $\boldsymbol{S_t}$ is a vector of day dummies, $\boldsymbol{W_{it}}$ is a vector of covariates and T_i is an indicator for the app being in a short download interval in its initial session. Equation 2. represents the main equation and equation 3. the first stage.

4.2 Results

The first question is whether apps in short download intervals really do have a higher probability of switching intervals – this question thus relates to the strength of the first stage. Figure 9. shows the daily average of I_{it} for apps initially belonging to each respective download interval type. As suspected, the probability of an app having switched download interval is always higher

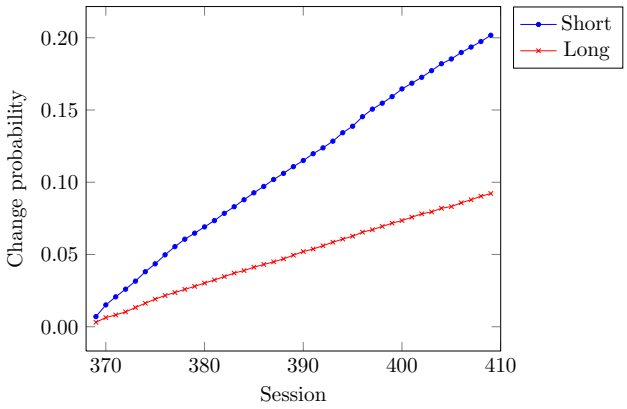
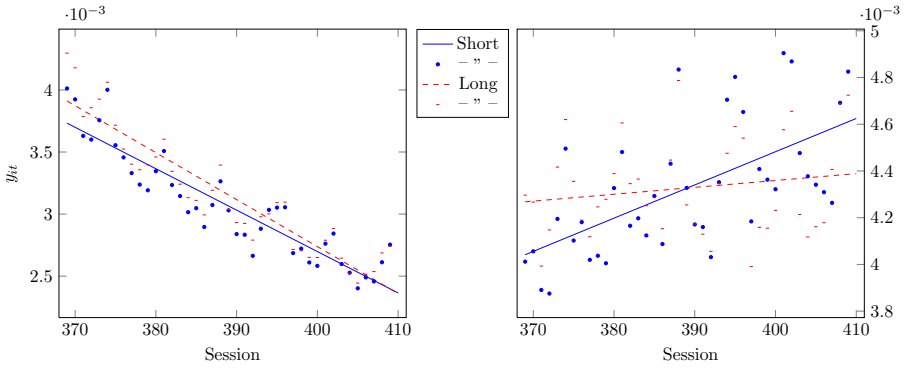


Figure 9: Cumulative probability of having switched download interval over days. Long (short) signifies that the app was in a long (short) interval in its first session. Number of ratings ≥ 50.



(a) First day. **(b)** Any day.

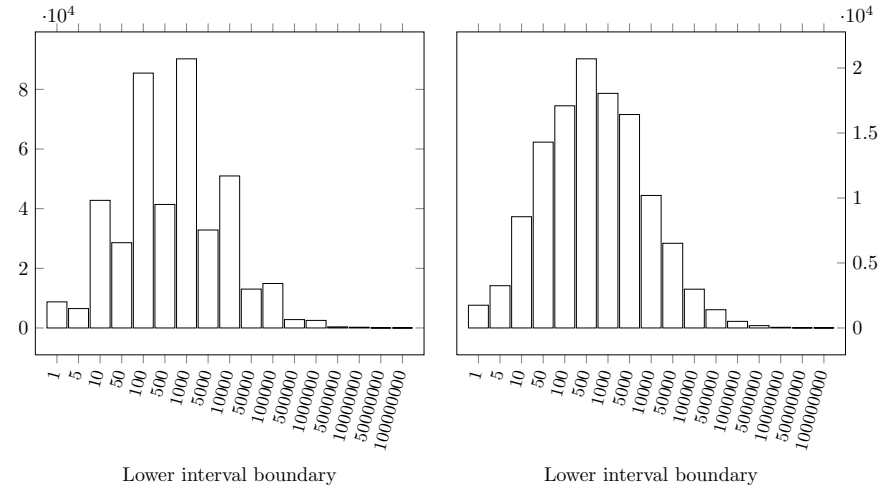
Figure 10: Daily rating rate by initial interval type. "First day" and "Any day" signifies which day an app must first appear in our data in order to be included in the analysis. Number of ratings ≥ 50.

for apps initially belonging to a short download interval, and this difference is increasing in the number of days that pass.

Figure 10. plots the average daily rating rate for the two download interval types and shows a simple linear fit of the values. The downward sloping pattern in figure 10a. is due to the sample being restricted to apps that entered our sample already the first day (session 369) and stayed in the sample until the final day (session 410). It is thus based on a balanced panel. In figure 10b. we also include apps that enter our sample later on. The initial download interval type is in this case based on the first session in which we observe the app in the sample. In general, new apps have a higher daily rating rate than old apps. This explains why there is a general negative trend in the left graph, while the general trend is positive in the right graph (the whole market is growing). The relative pattern for "short apps" (solid) and "long apps" (dashed) is consistent with a positive causal effect on downloads from switching to a higher interval. As time passes more of the apps initially in the short download intervals switch interval and experience an increase in the daily rating rate.

Figure 11a. shows the raw frequency distribution of apps over different download intervals. It is clear that there is a large number of apps in the long download interval "$1\,000 - 5\,000$ downloads". Figure 11b. shows the normalized frequency distribution; the normalization simply consists of dividing the frequencies in short intervals by 2, and the frequencies in long intervals by 5, in order to compensate for the interval lengths. The normalized frequencies closely resembles a normal distribution. This gives little reason to suspect any problems with apps being strangely distributed across download intervals.

While a perfect instrumental variable will always give asymptotically unbi-



(a) Raw frequency. **(b)** Normalized frequency.

Figure 11: Distribution of apps in their first day over different download intervals. Number of ratings ≥ 0.

ased estimates, this is not true if the exclusion restriction does not hold. Then the bias will be increasing in the correlation between the instrument and the error term in the main equation but decreasing in correlation between the instrument and the endogenous variable. With no formal tests of whether the exclusion restriction holds it is common to check the latter correlation, the strength of the first stage. The stronger the instruments are, the smaller any problems caused by violation of the exclusion restriction will be.

Figure 9. suggests that the statistical relationship between our endogenous variable and instrument is large. Table 3. presents the F-statistic from a joint test of our instrumental variables being non-zero. Across all specifications this statistic is very large which reassures that the instrument is indeed strong.

Table 3. summarizes the estimates of κ in model 2. from a number of different specifications. Focusing on the first two columns, the point estimates are lower when the sample is restricted to only include apps with a higher number of minimum ratings but do not seem to vary much when including covariates. The third column of table 3. shows the point estimates when restricting our sample to apps for which we have information for all days in our samples (the balanced panel). The point estimates in these specifications are smaller and do not vary much across the different minimum number of ratings specifications.

The point estimates on their own give little information of the economic significance of switching download interval. Instead it is the relative effects, when compared against a baseline, that is of interest. As apps that change interval during the period under observation might be systematically different

Table 3: Estimates of the effect of switching download interval on daily rating rate

2SLS, # Ratings $\geq$ 0			
Cumul. Increase	0.00444***	0.00412***	0.00155**
	(0.000735)	(0.000765)	(0.000611)
FS F stat.	3110.9	3149.1	3553.3
N	16037658	16037658	15196186
# Apps	422772	422772	379666
2SLS, # Ratings $\geq$ 20			
Cumul. Increase	0.00402***	0.00378***	0.00196***
	(0.000897)	(0.000873)	(0.000627)
FS F stat.	1099.3	1106.3	1213.3
N	5593767	5593767	5272857
# Apps	147351	147351	131416
2SLS, # Ratings $\geq$ 50			
Cumul. Increase.	0.00215**	0.00219**	0.00159**
	(0.00108)	(0.00106)	(0.000748)
FS F stat.	786.8	788.1	785.2
N	3508987	3508987	3283771
# Apps	92775	92775	81881
Session FE	Yes	Yes	Yes
App FE	Yes	Yes	Yes
Rating dummies	No	Yes	Yes
Board dummy	No	Yes	Yes
Startsession 369	No	No	Yes

Notes: Clustered standard errors in parentheses, p-values: * - 10 %, ** - 5 %, *** - 1 %. "FS F Stat." presents F-values from a test of the excluded instruments.

from apps that do not change we are estimating an average treatment effect on the treated (ATT). Thus our baseline consists of apps that will switch download interval, but has not yet done so.

Focusing on the first two columns this baseline average daily rating rate is around 0.013 when letting apps enter our dataset in any day (it does not vary much with the different minimum number of rating specifications). The point estimates of around 0.0042 when using our full sample then translates into approximately a 31 % increase in daily rating rate, when switching to a higher download interval. When restricting analysis to 20 and 50 minimum ratings the point estimates of around 0.0039 and 0.0022 translate into approximately a 28 % and 17 % increase respectively. When restricting attention to the estimates in the third column, where we exclude new apps, the point estimates are all around 0.0017 which translates roughly into a 22 % increase in daily rating rate, as the baselines for these specifications are all around 0.008.

5 Conclusion

In this paper we have investigated how consumer behavior responds to information about others' stated and revealed preferences. We found, based on daily data from the Google play store, that an extra half star leads to approximately a 4.5 % increase in daily download rate (approximated by daily rating rate). In the same manner, an increase in displayed download category increased daily download rate by approximately 25 %. In order to compare the sizes of these effects, we need to normalize them.

We have chosen to do this by calculating the percentile increase in the underlying distribution that an increase in the respective measure induces; i.e. an extra half star's impact on position in the average rating distribution, and a higher download interval's impact on position in the download interval distribution. The average percentile increase from one extra half star is approximately 16 percentiles. The corresponding average percentile increase from switching to a higher download category is about 12 percentiles.

The normalized effects can then be expressed as follows: a 10 percentile increase in the average rating distribution increases daily download rate by approximately 3 percent, while a 10 percentile increase in the download distribution increases the daily download rate by approximately 20 percent. Our results thus indicate that future downloads are much more responsive to other consumers' revealed preferences than other consumers' stated preferences – in other words, observational learning seems to be much more important than eWOM.

This finding is in line with the standard economist's notion that talk is cheap and that actions speak louder than words. However, there are a number of competing explanations that are left to be explored in future studies. One such explanation for the relatively modest consumer response to others' stated preferences is that the average rating in the app market is rather high – slightly above 4. This might be an indication of rating inflation taking place, which could attenuate the informational value of an app's average rating.

Furthermore, the attribution of the revealed preferences effect to observational learning is not without caveats. Pure observational learning is caused exclusively by an updated estimate of a product's true quality. In many cases the true quality of a product is endogenous to the number of users – many products carry positive (and sometimes negative) network externalities. Unfortunately, we cannot isolate what share of the revealed preference effect that should be attributed to pure observational learning and what is instead due to positive network externalities.

Finally, a limitation of the study is that the costly apps (i.e. the non free apps) are too few, and have too few ratings, to be analyzed separately. Arguably,

the external validity of the study would increase if the relative sizes of the effects were to prevail in an separate analysis based exclusively on costly apps. However, the specific market that we analyze – and the software market in general – is quite special in that there is often a connection between specific costly and free apps. For each costly app, there is generally a "sister app" that is free of charge. The purpose of the free sister app is usually to promote the costly version, to collect additional charges within the app and to generate advertising revenue. One may therefore hypothesize that the most important quality information about a costly app comes from using the free version of the same app. In that case, the average rating and download statistics of a costly app may have very limited effects on demand as it is redundant for someone who has already experienced the free version of the same app. Instead, the download statistics and average rating of the free sister app could be the most important influences on demand for the costly app. This relates to the effectiveness of using a free app to advertise the costly app. In a future study we plan to analyze this issue further by matching the apps into sibling pairs.

References

Anderson, M., & Magruder, J. (2012). Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *The Economic Journal*, *122*, 957–989.

Avery, C., Resnick, P., & Zeckhauser, R. (1999). The market for evaluations. *American Economic Review*, *89*(3), 564–584.

Bandura, A. (1977). *Social Learning Theory*. Englewood Cliffs, NJ: Prentice Hall.

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *The Journal of Political Economy*, *100*(5), 992–1026.

Cabral, L., & Hortaçsu, A. (2010). The dynamics of seller reputation: Evidence from eBay. *The journal of Industrial Economics*, *LVIII*(1), 54–78.

Cai, H., Chen, Y., & Fang, H. (2009). Observational Learning: Evidence from a Randomized Natural Field Experiment. *The American Economic Review*, *99*(3), 864–882.

Chan, Y. Y., & Ngai, E. (2011). Conceptualising electronic word of mouth activity: An input-process-output perspective. *Marketing Intelligence & Planning*, *29*(5), 488–516.

Chen, Y., Wang, Q., & Xie, J. (2011). Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning. *Journal of Mar*, *XLVIII*(April), 238–254.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, *99*(4), 1145–1177.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Duflo, E., & Saez, E. (2002). Participation and investment decisions in a retirement plan: the influence of colleagues' choices. *Journal of Public Economics*, *85*(1), 121–148.

Finkelstein, A. (2009). E-ZTax: Tax Salience and Tax Rates. *The Quarterly Journal of Economics*, *124*(3), 969–1010.

Gabaix, X., & Laibson, D. (2006). Shrouded attributes, consumer myopia, and information suppression in competitive markets. *The Quarterly Journal of Economics*, *121*(2), 505–540.

Google (2012). Google Play for Developers: Ratings and Comments.

Hesselius, P., Johansson, P., & Nilsson, J. P. (2009). Sick of Your Colleagues' Absence? *Journal of the European Economic Association*, *7*(2-3), 583–594.

Hilger, J., Rafert, G., & Villas-Boas, S. (2011). Expert opinion and the demand for experience goods: An experimental approach in the retail wine market. *The Review of Economics and Statistics*, *93(4)*(November), 1289–1296.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Jin, G. Z., & Kato, A. (2006). Price , quality , and reputation: Evidence from an online field experiment. *The RAND Journal of Economics*, *37*(4), 983–1004.

Lacetera, N., Pope, D. G., & Sydnor, J. R. (2011). Heuristic Thinking and Limited Attention in the Car Market. Working Paper.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(June), 281–355.

Li, X., & Wu, L. (2012). Measuring Effects of Observational Learning and Social-Network Word-of-Mouth (WOM) on the Sales of Daily-Deal Vouchers. Working Paper.

Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp.com. Working Paper.

Lucking-Reiley, D., Bryant, D., Prasad, N., & Reeves, D. (2007). Pennies from Ebay: The Determinants of Price in Online Auctions. *The Journal of Industrial Economics*, *LV*(2), 223–233.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

Melnik, M. I., & Alm, J. (2003). Does a Seller's eCommerce Reputation Matter? Evidence from eBay Auctions. *The Journal of Industrial Economics*, *50*(3), 337–349.

Moretti, E. (2011). Social Learning and Peer Effects in Consumption: Evidence from Movie Sales. *The Review of Economic Studies*, *78*(1), 356–393.

Pope, D. G. (2009). Reacting to rankings: evidence from "America's Best Hospitals". *Journal of health economics*, *28*(6), 1154–65.

Reinstein, D. A., & Snyder, C. M. (2005). The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics. *The Journal of Industrial Economics*, *53*(1), 27–51.

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics*, *9*(2), 79–101.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of

Inequality and Unpredictability in an Artificial Cultural Market. *Science*, *311*(5762), 854–856.

Simon, H. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(February), 99–118.

Sorensen, A. T. (2006). Social learning and health plan choice. *The Rand journal of economics*, *37*(4), 1–29.

Tucker, C., & Zhang, J. (2007). Long Tail or Steep Tail? A Field Investigation into How Online Popularity Information Affects the Distribution of Customer Choices. Working Paper.

WORKING PAPERS*
Editor: Nils Gottfries

2011:19 Stefan Eriksson and Karolina Stadin, The Determinants of Hiring in Local Labor Markets: The Role of Demand and Supply Factors. 33 pp.

2011:20 Krzysztof Karbownik and Michał Myck, Mommies' Girls Get Dresses, Daddies' Boys Get Toys. Gender Preferences in Poland and their Implications. 49 pp.

2011:21 Hans A Holter, Accounting for Cross-Country Differences in Intergenerational Earnings Persistence: The Impact of Taxation and Public Education Expenditure. 56 pp.

2012:1 Stefan Hochguertel and Henry Ohlsson, Who is at the top? Wealth mobility over the life cycle. 52 pp.

2012:2 Susanne Ek, Unemployment benefits or taxes: How should policy makers redistribute income over the business cycle? 30 pp.

2012:3 Karin Edmark, Che-Yuan Liang, Eva Mörk and Håkan Selin, Evaluation of the Swedish earned income tax credit. 39 pp.

2012:4 Simona Bejenariu and Andreea Mitrut, Save Some, Lose Some: Biological Consequences of an Unexpected Wage Cut. 67 pp.

2012:5 Pedro Carneiro and Rita Ginja, Long Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start. 82 pp.

2012:6 Magnus Carlsson and Stefan Eriksson, Do Reported Attitudes towards Immigrants Predict Ethnic Discrimination? 23 pp.

2012:7 Mikael Bask and Christian R. Proaño, Optimal Monetary Policy under Learning in a New Keynesian Model with Cost Channel and Inflation Inertia. 25 pp.

2012:8 Mikael Elinder and Oscar Erixson, Every man for himself. Gender, Norms and Survival in Maritime Disasters. 78 pp.

2012:9 Bertil Holmlund, Wage and Employment Determination in Volatile Times: Sweden 1913–1939. 43 pp.

2012:10 Indraneel Chakraborty, Hans A. Holter and Serhiy Stepanchuk, Marriage Stability, Taxation and Aggregate Labor Supply in the U.S. vs. Europe. 63 pp.

2012:11 Niklas Bengtsson, Bertil Holmlund and Daniel Waldeström, Lifetime versus Annual Tax Progressivity: Sweden, 1968–2009. 56 pp.

* A list of papers in this series from earlier years will be sent on request by the department.

2012:12 Martin Jacob and Jan Södersten, Mitigating shareholder taxation in small open economies? 16 pp.

2012:13 John P. Conley, Ali Sina Önder and Benno Torgler, Are all High-Skilled Cohorts Created Equal? Unemployment, Gender, and Research Productivity. 19 pp.

2012:14 Che-yan Liang and Mattias Nordin, The Internet, News Consumption, and Political Attitudes. 29 pp.

2012:15 Krzysztof Karbownik and Michal Myck, For some mothers more than others: how children matter for labour market outcomes when both fertility and female employment are low. 28 pp.

2012:16 Karolina Stadin, Vacancy Matching and Labor Market Conditions. 51 pp.

2012:17 Anne Boschini, Jan Pettersson, Jesper Roine, The Resource Curse and its Potential Reversal. 46 pp.

2012:18 Gunnar Du Rietz, Magnus Henrekson and Daniel Waldenström, The Swedish Inheritance and Gift Taxation, 1885–2004. 47pp.

2012:19 Helge Bennmarker, Erik Grönqvist and Björn Öckert, Effects of contracting out employment services: Evidence from a randomized experiment. 55 pp.

2012:20 Pedro Carneiro and Rita Ginja, Partial Insurance and Investments in Children. 32pp.

2013:1 Jan Pettersson and Johan Wikström, Peeing out of poverty? Human fertilizer and the productivity of farming households. 43 pp.

2013:2 Olof Åslund and Mattias Engdahl, The value of earning for learning: Performance bonuses in immigrant language training. 52 pp.

2013:3 Michihito Ando, Estimating the effects of nuclear power facilities on local income levels: A quasi-experimental approach. 44 pp.

2013:4 Matz Dahlberg, Karin Edmak and Heléne Lundqvist, Ethnic Diversity and Preferences for Redistribution: Reply. 23 pp.

2013:5 Ali Sina Önder and Marko Terviö, Is Economics a House Divided? Analysis of Citation Networks. 20 pp.

2013:6 Per Engströma and Eskil Forsell, Demand effects of consumers' stated and revealed preferences. 27 pp.

See also working papers published by the Office of Labour Market Policy Evaluation
http://www.ifau.se/ ISSN 1653-6975