*Article*

*Group Processes & Intergroup Relations*

GPIR

*Group Processes & Intergroup Relations*
2014, Vol. 17(3) 388–403
© The Author(s) 2013
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1368430213502557
gpir.sagepub.com

SAGE

# Broken windows, mediocre methods, and substandard statistics

**Jelte M. Wicherts**[1] **and Marjan Bakker**[2]

**Abstract**

Broken windows theory states that cues of inappropriate behavior like litter or graffiti amplify norm-violating behavior. In a series of quasi-experiments, Keizer, Lindenberg, and Steg altered cues of inappropriate behavior in public places and observed how many passersby subsequently violated norms. They concluded that particular norm violations transgress to other misdemeanors (e.g., graffiti leads to littering or even theft) and that the presence of prohibition signs heightens the saliency of norm violations, thereby aggravating the negative effects of cues such as litter and graffiti. We raise several methodological and statistical issues that cast doubt on Keizer et al.'s results. Problems include confounding factors, observer bias, lacking scoring protocols, a failure to establish interobserver reliabilities, inflated Type I error rates due to dependencies, sequential testing, and multiple testing. We highlight results of a highly similar study that does not support the notion that prohibition signs aggravate the effects of observed norm violations. We discuss potential improvements of the paradigm.

**Keywords**

descriptive norms, littering behavior, methodological critique, statistics, theft

Paper received 24 May 2012; revised version accepted 10 June 2013.

According to broken windows theory, cues of inappropriate behavior like litter, graffiti, or broken windows in public places amplify norm-violating behavior (Cialdini, Reno, & Kallgren, 1990; Kelling & Coles, 1998; Wilson & Kelling, 1982). The empirical support in favor of broken windows theory has been predominantly correlational and so the proposed causal effects of disorder on misdemeanors like littering and petty criminal acts were in need of empirical support (Harcourt & Ludwig, 2006). In two widely publicized papers, Keizer, Lindenberg, and Steg (2008, 2011) aimed to fill this void by conducting thought-provoking field studies in which they varied litter, graffiti, and other norm violations in public places and observed (unobtrusively) subsequent norm violations of passersby. Across eight such field studies, Keizer et al. observed relatively more norm violations (littering, trespassing, and minor theft) when cues of other people's norm

[1] Tilburg University, The Netherlands
[2] University of Amsterdam, The Netherlands

**Corresponding author:**
Jelte M. Wicherts, Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, Tilburg 5000 LE, The Netherlands.
Email: J.M.Wicherts@uvt.nl

violations were made apparent. They found that cues of particular norm violations transgressed to other norms, such that graffiti in an alley or unreturned shopping carts in a parking garage led to a higher prevalence of subsequent littering (i.e., a cross-norm effect). So Keizer et al.'s work went beyond previous findings that showed that people tend to litter more in littered environments (Huffman, Grossnickle, Cope, & Huffman, 1995) by supposedly showing that other norm violations may also lead to more littering. They also claimed to have established that graffiti and litter lead to more minor theft.

Recently, Keizer et al. (2011) studied whether prohibition signs moderate the effect as proposed by broken windows theory (Kelling & Coles, 1998; Wilson & Kelling, 1982) and by theories that distinguish injunctive norms from descriptive norms (Cialdini et al., 2006; Cialdini et al., 1990). Specifically, they studied whether the presence of prohibition signs heightens the saliency of norm violations by others, thereby leading to more misdemeanors. Keizer et al. concluded that the introduction of a sign prohibiting either graffiti (cross-norm) or littering (same-norm) actually led to more littering in an environment in which another norm (i.e., against graffiti) or the same norm (i.e., against littering), was apparently violated by others. And so Keizer et al. (2011) argued against the use of prohibition signs when the prohibitions cannot be effectively enforced (e.g., an antilitter sign in a littered environment).

Here we criticize Keizer et al.'s (2008, 2011) studies on methodological, statistical, and empirical grounds. Methodological problems like confounding factors, a failure to employ standardized observer protocols, and observer bias render Keizer et al.'s empirical support in favor of broken windows theory unconvincing, while proper statistical analyses fail to support the supposedly negative impact of prohibition signs in combination with visible violations of these prohibitions. We discuss results of a similar study that cast doubt on the moderating effect of prohibition signs and offer suggestions for future work.

## Litter and Graffiti

Keizer et al. (2008) reported the results of six field studies in the city of Groningen, the Netherlands. The observed misdemeanors involved littering a flyer (Studies 1, 3, and 4), trespassing (Study 2), and minor theft of an envelope with 5.00 euros (Studies 5 and 6). The first study took place in an alley in a shopping area that is used by many to park their bicycles. Keizer et al. covertly put fake commercial flyers on the bicycles' handlebars after the owners had parked them, and subsequently observed whether the owners took the flyer with them or disposed it when collecting their bicycle. Because there were no trashcans in the alley, disposing either meant littering or putting the flyer on another bicycle. On one day, the walls of the alley were covered with graffiti, while on the other day the walls were clean. On both days, a clearly visible sign pointed out that graffiti was prohibited. A confederate observed the behavior unobtrusively and removed the littered flyers. Results showed that 53 out of 77 bicycle owners littered the flyer (or put it on another bicycle) on the day when the graffiti was present, while only 25 out of 77 did so on the day when graffiti was absent. The other studies in the *Science* paper (2008) involved adaptations of this paradigm both involving littering and other types of misdemeanors (trespassing and minor theft), while those in Keizer et al. (2011) involved the same paradigm and the same dependent variable (littering a flyer or not) and location.

The main goal of the studies in *Science* (2008) was to lend support to the effects of cues of disorder on violations of another norm (a cross-norm effect), while the main hypothesis of both studies in *Group Processes & Intergroup Relations* was that "making a norm more salient by means of a prohibition sign in a setting with cues signaling that other people did not conform to the norm … will actually increase the number of people violating that norm" (Keizer et al., 2011, p. 682). Keizer et al. (2011) studied the moderating effect of prohibition signs in two quasi-experiments in the same alley that featured alterations in terms of the presence or absence of littering (Study 1) or graffiti (Study 2) and in terms of the presence or absence of signs related to either littering (Study 1) or graffiti (Study 2). According to Keizer et al. (2011), their studies showed that the presence of the prohibition sign led to a higher prevalence of littering in both an environment with graffiti and an environment with litter. They argued that this is so because the signs increase the salience of the norm and the violation thereof by others. Specifically, they contended that the signs led a descriptive norm (e.g., "many people ignore the sign against littering") to supersede an injunctive norm (e.g., "one ought not litter"), which leads to more misdemeanors (littering; Cialdini et al., 2006; Cialdini et al., 1990).

It is well established that people litter more in littered environments (Cialdini et al., 1990; Durdan, Reeder, & Hecht, 1985; Ernest-Jones, Nettle, & Bateson, 2011; Finnie, 1973; Huffman et al., 1995; Krauss, Freeman, & Whitcup, 1978; Reiter & Samuel, 1980; Schultz, Bator, Large, Bruni, & Tabanico, 2011). Previous studies of littering behavior also found relations of littering with for example, the proximity or presence of trash cans (Huffman et al., 1995), number of bystanders (Ernest-Jones et al., 2011), and personal characteristics such as age and sex (Schultz et al., 2011).

## Mediocre Methods

In this section,[1] we criticize Keizer et al.'s (2008, 2011) studies on methodological grounds. The eight studies reported by Keizer and colleagues (2008, 2011) are quasi-experiments because they did not involve random assignment of participants to conditions. Quasi-experiments in relatively uncontrolled environments are methodologically challenging (Shadish, Cook, & Campbell, 2002), because of uncontrolled confounding factors. Although the conditions within each of Keizer et al.'s (2008, 2011) quasi-experiment involved the same location and the researchers controlled for potential confounding factors related to time of day and weather conditions, it is quite likely that participant characteristics differed systematically between the conditions. Several variables like age and sex are associated with littering behavior (e.g., Schultz et al., 2011) and could be scored readily (at least by reasonable approximation). Days of the week differed necessarily between conditions, which may have had a number of relevant effects such as the number of passersby. For example, in shopping areas in Groningen it is more crowded on Thursdays because of this day's expanded opening hours, while fewer people go shopping on Mondays because of its short opening hours. In addition, the alley that featured in Study 1 of Keizer et al. (2008) and all studies in Keizer et al. (2011) is located next to a market square (the Vismarkt) with a food market on particular weekdays but not on others, which may affect relevant characteristics of the passersby (including age and sex). Both day of the week and the number of people present have been found to affect littering behavior in another field study involving a more homogeneous sample (Ramos & Torgler, 2010). In yet another field study in a cafeteria, the number of people present also clearly diminished the prevalence of littering (Ernest-Jones et al., 2011; cf. Meeker, 1997). We requested dates of data collection from Keizer et al. in order to check for potential day of the week effects, but this was to no avail. We gave up the effort to gather data after 4 months (based on earlier experience; Wicherts, Borsboom, Kats, & Molenaar, 2006).

Keizer et al.'s studies lacked rigorous standardization of the observations of behavior. Only a few details were given concerning how behavior was observed and observers did not appear to have employed any scoring protocol (cf. Endnote 1). This is particularly problematic for the studies involving littering, because many people who litter a flyer in a public setting do this inconspicuously. For instance, no guidelines were stipulated to unambiguously distinguish between (a) littering when taking the bicycle (i.e., at the spot where it was parked), (b) littering when riding the bicycle, or (c) littering the flyer when (or slightly after) leaving the alley. Notably, the latter behavior could be scored either as "taking the flyer" or "littering" depending on the observer's viewpoint and location in the alley (which may or may not have been controlled). Interobserver reliabilities could shed light on the degree to which misdemeanors could be scored unambiguously, but such reliabilities appear not to have been determined in any of the studies by Keizer et al. (2008, 2011). More importantly, the observers were not blind to the study's hypothesis or to the condition under which behavior was observed. Observer biases and observer effects are particularly potent when observations are not standardized.[2]

Although field studies are often characterized by a lack of control on participant characteristics and external influences, several methodological and statistical controls could have improved Keizer et al.'s studies considerably. Controls for observer biases and observer effects are quite standard in behavioral research and are easily implemented by blinding procedures. Moreover, a counterbalanced design in which days of the week function in each condition could avert potential confounding of crowding in the area, the number of parked bicycles, and participant characteristics. In addition, participants' characteristics can be controlled either statistically or by design. Taken together, the studies by Keizer et al. (2008, 2011) are characterized by a lack of methodological controls, many of which could be applied readily in future research. The lack of the detail offered by Keizer et al. (2008, 2011) and their reluctance to share information by email impedes a thorough assessment of the severity of these methodological problems and obstructs independent replications of their work.

## Substandard Statistics

In this section we address several problems with the statistical results reported by Keizer et al. (2008, 2011). These include (a) dependencies of data points, (b) the use of intermediate tests, and (c) the use of one-sided testing after much of the data had already been observed. We provide several reanalyses of their data and a data simulation to come to grips with the severity of these problems.

Analyses conducted by Keizer et al. (2008, 2011) assume that data points are independent, but this assumption is unlikely to be tenable in their data on littering because people often shop (e.g., Jazwinski & Walcheski, 2011) and/or travel together. In the televised replication of the first experiment in the *Science* (2008) paper (Folkersma & Rademaker, 2009), approximately a fourth of the bicycle owners were seen collecting their bicycles in the presence of others. If your travel partner does not litter his or her flyer, you are unlikely to litter yours, and vice versa. Keizer et al. did not indicate how they dealt with such potentially important social factors (e.g., Raafat, Chater, & Frith, 2009) in their studies of littering.[3] Because of the expected positive correlations between littering behavior of groups of participants, such dependencies in the data lead to inflated $\chi^2$s and hence overly positive significance levels (e.g., Hedges, 2007). The effects of dependencies between observations on the Type I error rate depend on the number of grouped observations and the correlation between the behaviors within groups.

In Figure 1, we provide the results of a data simulation (10,000 iterations per condition) under the null hypothesis (i.e., no relation between the two variables of interest like presence or absence of graffiti and whether people litter or not) under various scenarios of dependence between observations. In keeping with the typical littering rates and cell sizes in Keizer et al. (2011), we fixed the chance of littering at 50% for both conditions, while the sample size of each condition was set at 74. We varied the percentage of pairs among these 2*74 = 148 observations and the degree to which these pairs showed overlapping behavior (see Appendix B for the R code). Results show a clear increase of the Type I error rate of up to over 10%, with stronger inflations when a higher proportion of observations are paired and when participants' behaviors are more strongly overlapping. In other words, the test statistic becomes overly liberal due to dependencies. Insofar that observed passersby in Keizer et al.'s studies were together and acted similarly, the Type I error rate in their analyses was well above 5%.

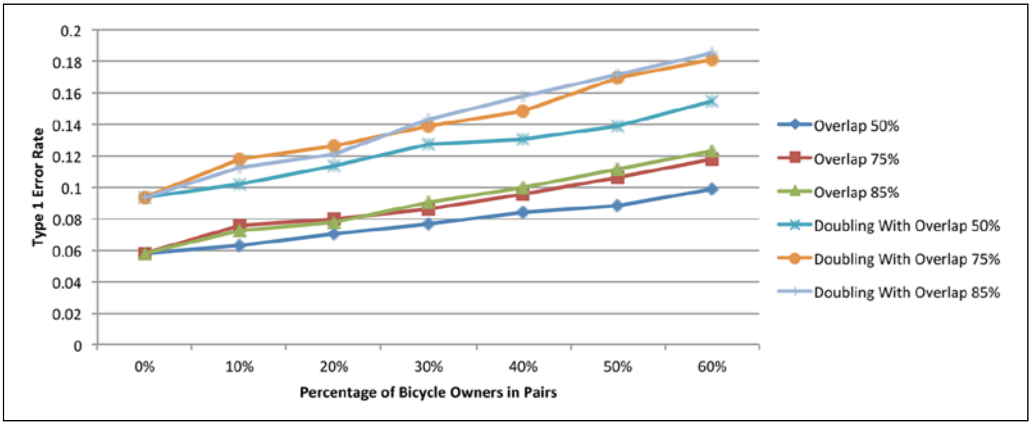

**Figure 1.** Simulated effects of dependencies and doubling of sample size after intermediate testing on the Type I error rate of Pearson's $\chi^2$ test of independence in a 2 x 2 table. Results show inflations of Type I error rate well above the nominal level.

It is noteworthy that in their recent paper they chose to collect more data (doubling the sample size in several conditions; cf. Table A1) after an earlier analysis failed to show significant results of pair-wise tests between conditions (cf. Table 1). It has been long known that such intermediate (or sequential) testing directly violates the rationale of null hypothesis significance testing (Anscombe, 1954; Armitage, McPherson, & Rowe, 1969) and that it leads to potentially severe inflations of the Type I error rate (Bakker, van Dijk, & Wicherts, 2012; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, 2007).

We ran additional simulations to come to grips with the effect of a doubling of the sample size after an intermediate test rendered a nonsignificant result (at α .05) on the actual Type I error in the same set-up as mentioned in the previous lines. In our simulation without any dependent observations, we found that this type of intermediate testing results in a Type I error rate of 9.4%. If this strategy of intermediate testing and doubling the sample size is combined with the dependencies in the data, the Type I error rates go up even further. Figure 1 gives the results of a simulation in which the dependencies due the overlapping behaviors of pairs of observations are combined with the intermediate testing. When both practices are combined, Type I rates can be as high as 18.5%. Thus, combining intermediate testing with dependencies in the data leads to biased outcomes of Pearson's $\chi^2$ tests in which the chances of Type I errors are unacceptably high.

On the basis of 95th percentile of the 10,000 $\chi^2$s in each simulated condition, we determined the threshold for significance (α = .05) that could be used as a threshold that effectively controls for the bias as simulated. The average 95th percentile was 4.82 for dependencies only, 4.58 for doubling the sample size (intermediate testing) only, and 5.62 for when dependencies were combined with doubling the sample size. So when controlled for either of these biases as simulated, the results of Studies 5 and 6 in the *Science* (2008) paper (related to minor theft and not involving dependencies because only single passersby were included) and most within-study pair-wise comparisons reported in the 2011 paper are no longer significant. The outcomes of Studies 1–4 in the *Science* paper were quite dramatic (log odds ratios of 1.52, 2.47, 1.18, 1.33, respectively) and do meet these more stringent thresholds for significance (all $\chi^2$s > 8.5).

Keizer et al. (2011) reported a total of 12 $\chi^2$ tests of independence of 2 x 2 cross-tables. The authors indicated their use of a one-tailed test of independence and have divided the *p* values

**Table 1.** Results of our reanalysis of studies by Keizer et al. (2008, 2011) and *p* values reported by them.

| Study | Condition | Viol. | Not viol. | % Violating | $\chi^2$ ($df$ = 1) | *p* | *p* as reported |
|---|---|---|---|---|---|---|---|
| S1 | No graffiti & sign | 25 | 52 | 32 | | | |
| | Graffiti & sign | 53 | 24 | 69 | 20.37 | <.001 | <.001 |
| S2 | Correctly parked bicycles | 12 | 32 | 27 | | | |
| | Incorrectly parked bicycles | 40 | 9 | 82 | 27.79 | <.001 | <.001 |
| S3 | No unreturned shopping carts | 18 | 42 | 30 | | | |
| | Unreturned shopping carts | 35 | 25 | 58 | 9.77 | .002 | .002 |
| S4 | No firecrackers | 26 | 24 | 52 | | | |
| | Firecrackers | 37 | 9 | 80 | 8.59 | .003 | .003 |
| S5/S6 | No graffiti & no litter | 9 | 62 | 13 | | | |
| | Graffiti[1] | 16 | 44 | 27 | 4.12 | .042 | .035[5] |
| | Litter[1] | 18 | 54 | 25 | 3.55 | .060 | .047[5] |
| 1.1 | Baseline[2] | 36 | 41 | 47 | | | |
| | Littered | 46 | 29 | 61 | 3.25 | .071 | .036 |
| 1.3 | Littered[3] | 91 | 59 | 61 | | | |
| | Littered & sign | 105 | 45 | 70 | 2.89 | .089 | .045 |
| 1.4 | Litter-free & sign | 29 | 45 | 39 | | | |
| | Baseline[2] | 36 | 41 | 47 | 0.88 | .348 | .174 |
| 2.1 | Baseline[2] | 36 | 41 | 47 | | | |
| | Graffiti | 48 | 29 | 62 | 3.77 | .052 | .026 |
| 2.3 | Graffiti[3] | 93 | 55 | 63 | | | |
| | Graffiti & sign[4] | 108 | 42 | 72 | 2.85 | .091 | .046 |
| 2.4 | Graffiti-free & sign | 25 | 52 | 32 | | | |
| | Baseline[2] | 36 | 41 | 47 | 3.29 | .070 | .035 |

*Note.* [1]Tested against the same baseline condition; [2]represents same baseline group (no litter & no graffiti, & no sign); [3]involves the use of intermediate testing and data from 1.1/2.1; [4]includes data from S1; viol. versus not viol.: was norm violated by participant or not?; [5]Keizer et al. (2008) used a one-tailed Fisher exact test without indicating this in their paper.

of Pearson's $\chi^2$ tests by 2. This amounts to a directional proportion test without a continuity correction. Table 1 gives the results of the six core analyses of the 2011 paper and shows that in none of these, the null hypothesis of independence could be rejected at an alpha of .05 when the alternative hypothesis was bidirectional. Table 1 did not include all 12 tests because many tests reported by Keizer et al. (2011) were redundant or involved overlapping data (see Table 1). Given the overlap of samples, it is imperative that some correction for multiple testing is applied. Keizer et al. (2011) used one-tailed tests after much of the data had already been analyzed, which is widely considered to be problematic (e.g., Agresti & Franklin, 2007). We see no cogent argument to employ one-tailed tests in these analyses (Giner-Sorolla, 2012).

Analyses presented in Appendix A raise additional problems with data presented by Keizer et al. (2008, 2011). The analysis shows that their results are more consistent across seven sets of independent samples than is to be expected from standard sampling ($p$ = .006). When asked about the potential reuse of data, Dr. Keizer acknowledged that data from one condition in the 2008 paper were reused without proper acknowledgment in the 2011 paper (see Table 1). The set of six remaining replications also showed overly consistent results ($p$ = .017), which may indicate that Keizer et al. reused additional data without mention. However, other explanations are that observer bias, dependencies between data points, sequential testing, or publication bias have rendered the sampling scheme nonstandard. For instance, the reported replications may have been selected from a larger set of replications, some of

**Table 2.** Frequencies of conditions in both studies of Keizer et al. (2011) that could be submitted to a full analysis.

| Study | Sign | Disorder | Littered | Did not litter | % Did not litter |
|---|---|---|---|---|---|
| 1[1] | – | – | 36 | 41 | 53 |
| 1[2] | – | + | 91 | 59 | 39 |
| 1 | + | – | 29 | 45 | 61 |
| 1[2] | + | + | 105 | 45 | 30 |
| 2[1] | – | – | 36 | 41 | 53 |
| 2[2] | – | + | 93 | 55 | 37 |
| 2 | + | – | 25 | 52 | 68 |
| 2[2,3] | + | + | 108 | 42 | 28 |

*Note.* [1]Represents same baseline group; [2]involves the use of sequential testing; [3]includes data from the *Science* paper; Disorder: was environment prelittered or not?; Littered/did not litter: did participant litter or not?; Sign: was there a prohibition sign or not?

which may have showed less desirable results (see Pires & Branco, 2010, for a discussion). We contend that biases we discussed here may well explain the overly consistent results presented by Keizer et al. (2011) and the minor theft studies presented by Keizer et al. (2008).

## Another Analysis

Study 1 of Keizer et al. (2011) does not involve the reuse of data from the 2008 paper. This study ($N$ = 451) entails a full 2 x 2 factorial with factors disorder (littering) and prohibition sign (see Table 2). These data could be submitted to a standard log-linear analysis or logistic regression to see whether there is (a) a main effect of the prohibition sign, (b) a main effect of disorder (littered environment), and (c) an interaction between disorder and the prohibition sign on the prevalence of observer littering. The latter interaction enables a statistical test of the reversal effect that is superior to the six semidependent pair-wise tests conducted (after intermediate testing) by Keizer et al. (2011).

In a standard logistic regression (the standard log-linear analysis provided nearly identical results), the interaction between disorder and the prohibition sign was not significant: $\chi^2$ ($df$ = 1) = 3.10, $p$ = .078.[4] The main effect of disorder (litter) from the model was significant: $\chi^2$ ($df$ = 1) = 19.97, $p$ < .001, which corroborates earlier findings that people litter more in littered environments. The main effect of the sign failed to reach significance: $\chi^2$ ($df$ = 1) = 0.66, $p$ = .418. So notwithstanding potential inflations of significance levels in these analyses, the data do not lend much support to the negative effect of prohibition signs or its interaction with litter in the environment. We consider the results relating to prohibition signs in the 2011 paper inconclusive.

To summarize, our reanalysis casts doubt on results from Studies 5 and 6 of Keizer et al. (2008), the core interaction in Keizer et al. (2011), and the core pair-wise comparisons in Keizer et al. (2011). The statistical results from Studies 1 and 4 of Keizer et al. (2008) and both studies of Keizer et al. (2011) may be inflated due to statistical dependencies. Finally, results of the latter two studies are hard to interpret because of the authors' use of intermediate (sequential) testing.

## Retrospective Replication

Thirty-five years ago, Reiter and Samuel (1980)[5] conducted a study that can, retrospectively, be considered a fairly close replication of Keizer et al.'s (2011) first study. Reiter and Samuel (1980) used a Californian parking garage with seven stories and put flyers on windshields of cars when the parking garage was full. In a counterbalanced design the floors of the parking garage were randomly assigned to six conditions in a 2 x 3

**Table 3.** Frequencies of littering for prohibition sign and no-sign conditions of Reiter and Samuel (1980).

| Sign | Disorder | Littered | Did not litter | % Did not litter |
|---|---|---|---|---|
| − | − | 43 | 85 | 66 |
| − | + | 60 | 58 | 49 |
| + | − | 51 | 152 | 75 |
| + | + | 79 | 125 | 61 |

*Note.* Disorder: was environment prelittered or not?; littered/did not litter: did participant litter or not?; sign: was there a prohibition sign or not?

factorial design. The first factor concerned prelittering of the floors or not, and the second factor involved three levels: no sign, a prohibition sign (with the text "Littering is unlawful and subject to a $10 fine"), and a cooperative sign depicting a man throwing a flyer in a trashcan (and the text "Pitch-in"). Littering was objectively measured by counting at the end of the day the number of littered flyers on each floor. The top floor was not involved in the study to preclude weather effects.

Unlike Keizer et al.'s (2011) first study, this study did not suffer from potential confounding due to day-of the-week effects. Reiter and Samuel's use of counting of littered flyers can be seen as a relatively objective measure that diminishes the chances of observer bias. However, the failure to remove littered flyers and the repeating of the measurement after a week may have led to dependencies in the data. Such potential dependencies can be expected to inflate significance levels (i.e., they cannot explain nonsignificant findings).

Results for the prohibition sign are given in Table 3 (pooled for both days of testing).[6] We ran a logistic regression (a log-linear analysis provided similar conclusions) involving data from Study 1 of Keizer et al. (2011) and from Reiter and Samuel (1980) with backward selection of prediction on the basis of likelihood ratio tests. Results showed a higher prevalence of littering in the Keizer et al. (2011) study ($\chi^2[df = 1] = 46.95$, $p < .001$), a main effect for disorder (littered environment leads to more littering; ($\chi^2[df = 1] = 35.69, p < .001$) and an interaction between study and the presence or absence of the sign ($\chi^2[df = 1] = 7.28$, $p = .007$). More importantly, results show no interaction between prohibition signs and litter in the environment; $\chi^2(df = 1) = 0.89, p = .346$. So combined, the data do not lend support to Keizer et al.'s (2011) assertion that prohibition signs reverse the effect of a littered environment on littering. A random effects meta-analysis on the basis of differences between odds ratios with and without the sign corroborated this lack of support for the proposed interaction between signs and littered environments (see Appendix C).

## Conclusion

We consider the paradigm from Keizer et al. (2008, 2011) interesting and potentially viable to test the relevance of broken windows theory for littering and other misdemeanors in real-life circumstances, but feel that their paradigm is in need of methodological improvement. We have argued that the results of studies on the broken windows theory by Keizer et al. (2008, 2011) are unconvincing because none of these studies involved proper and feasible methodological controls for confounding factors and observer bias. The lack of detail in the description of the days of the week on which conditions were run, observations, crowding, scoring protocols, and information on interobserver reliabilities renders the observations ambiguous and impedes independent replications. Objective protocols could be developed (e.g., the counting of littered flyers) and/or behaviors could be videotaped in a way that enables scoring by independent observers who are blind to conditions. In addition, even approximate data on potentially relevant personal characteristics like sex and age could be collected and used as controls. The number of bystanders, (double-)parked bicycles, and passersby in the alley could be observed and used as controls in either the design or the analysis. Days of the week could be chosen as fixed or dealt with in a counterbalanced design. We note that many journals including *Science* and *Group Processes and Intergroup Relations* allow authors to upload large files of documentation and even raw data as supplementary online material and so issues relating to journal space no longer impede comprehensive reporting of studies that deal with topics of clear societal relevance.

We highlighted several problems in the statistical analyses reported by Keizer et al. (2008, 2011) and showed that a proper analysis of the data fails to support the negative effect of prohibition signs for which they claimed to have found clear support. We also showed in a simulation that dependencies in the data and the choice to double the sample size after intermediate testing severely inflate the Type I error rates. We have shown that the results of replications of the same behaviors under the same conditions by Keizer et al. are significantly more consistent than is to be expected under standard sampling, which raises questions about potential reuse of data, observer bias, violations of independence, sequential testing, and publication bias. Regardless of what explains the consistency, it casts doubt on the validity of the data presented by Keizer et al.

A previous study with a larger sample and several methodological controls failed to show any negative impact of prohibition sings on littering in a littered environment (Reiter & Samuel, 1980). The effects of litter and/or graffiti in Studies 5 and 6 of Keizer et al. (2008) on minor theft do not appear robust when analyzed correctly. In a study that involved several controls for confounding, Cialdini et al. (2006) found that prohibition signs that invoked the message that others violated the prohibition led to more theft of pieces of petrified wood from a natural park. Unfortunately, violations of independence of data points may also have led to inflated significance levels in that study.[7] The norm violation spoken about in prohibition signs of Cialdini et al. (2006) was identical to later norm violations. The cross-norm effect of graffiti, unreturned shopping carts, illegally parked bicycles, and the illegal firing of firecrackers on other misdemeanors (Keizer et al., 2008) remains to be replicated independently in studies with proper methodological controls. We are familiar with only one conceptual replication in a controlled and randomized experiment (Austrup, 2011). The results showed that participants in a littered room littered more but did not lie more for financial profit than did participants in a tidy room. Clearly, more research under controlled circumstances with observations that are rigorously scored and blinded, and proper statistical analyses are needed to shed light on the notion that evidence of other peoples' norm violations transgress to other norm violations. It also remains to be seen whether indeed prohibition signs highlight other people's norm-violating behaviors, thereby aggravating the effects of these norm violations on the same or other norm violations. To be convincing, future studies of the spreading of disorder should involve established methods and sound statistical analyses. Preferably, such studies are preregistered to preclude effects of publication bias and sequential testing (Open Science Collaboration, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

## Notes

1. We have sent Dr. Keizer a total of 16 emails requesting additional information, but after four months we only received a short note that provided partial answers. His coauthors either did not reply or suggested we contact Dr. Keizer. We failed to acquire dates of data collection and received no answers to our questions of (a) whether they used any scoring protocols, (b) whether observers were blind to conditions, and (c) whether they had ever determined interobserver reliabilities. We interpret Dr. Keizer's failure to respond to these questions as a negative answer.
2. Another potential source of bias due to observer bias is that data collection is halted ad hoc, such as immediately after witnessing a "hit" (i.e., someone whose behavior is in line with the hypothesis in a

given condition) rather than after having obtained a prespecified sample size.

3. Keizer et al. (2008) only included single passersby in Studies 5 and 6 and they took groups of people as single cases in Study 2, but Keizer et al. (2008, 2011) did not indicate how they dealt with groups of people in the remaining studies. In an email Dr. Keizer later indicated that whenever bicycles were locked to each other owners were not observed, but this was not documented in any of the papers or supplementary materials.
4. The interaction was significant ($p$ < .05) in the second study, but we consider this result ambiguous because of the use of older data in three of the conditions and the possibilities of inflated Type I errors in this analysis.
5. Keizer et al. (2008) described Reiter and Samuel's (1980) results incorrectly as showing that: "a sign drawing attention to the antilitter norm is more influential in reducing littering when placed in a nonlittered setting than when it is placed in a prelittered setting" (p. 1682). In fact the relevant interaction was nonsignificant in Reiter and Samuel's study. Keizer et al. (2011) did not refer to Reiter and Samuel (1980).
6. For logistical reasons, Reiter and Samuel (1980) used the first floor as the no-litter, no-sign condition on both days of testing. This led to an interaction with day of testing that is not of interest here. Also, an analysis involving only the littered environments (prohibition sign vs. no sign) corroborated the main result.
7. Cialdini et al. (2006) analyzed the data at the level of individual pieces of stolen petrified wood. The analyses in Cialdini et al. (2006) may suffer from violations of independence also if people steal more than one piece of petrified wood at the same time.

## References

Agresti, A., & Franklin, C. (2007). *Statistics. The art and science of learning from data.* Upper Saddle River, NJ: Pearson Education.

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, *10*, 89–100.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*, 235–244.

Austrup, S. (2011). *The person behind the "broken window." The influence of the environment and personality on undesired behavior* (Unpublished bachelor's thesis). University of Twente, Enschede, the Netherlands.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*, 3–15. doi:10.1080/15534510500181459

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015–1026. doi:10.1037/0022-3514.58.6.1015

Durdan, C. A., Reeder, G. D., & Hecht, P. R. (1985). Litter in a university cafeteria: Demographic data and the use of prompts as an intervention strategy. *Environment and Behavior*, *17*, 387–404. doi:10.1177/0013916585173007

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: A field experiment. *Evolution and Human Behavior*, *32*, 172–178. doi:10.1016/j.evolhumbehav.2010.10.006

Finnie, W. C. (1973). Field experiments in litter control. *Environment and Behavior*, *5*, 123–144. doi:10.1177/001391657300500201

Folkersma, B., & Rademaker, J. (Producers). (2009). Nieuwslicht [Television documentary]. Hilversum, the Netherlands: Vara.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*, 562–571. doi:10.1177/1745691612457576

Harcourt, B. E., & Ludwig, J. (2006). Broken windows: New evidence from New York City & a five-city social experiment. *University of Chicago Law Review*, *73*, 271–320.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370. doi:10.3102/1076998606298043

Huffman, K. T., Grossnickle, W. F., Cope, J. G., & Huffman, K. P. (1995). Litter reduction: A review and integration of the literature. *Environment and Behavior*, *27*, 153–183. doi:10.1177/0013916595272003

Jazwinski, C. H., & Walcheski, C. H. (2011). At the mall with children: Group size and pedestrian economy of movement. *Environment and Behavior*, *43*, 363–386. doi:10.1177/0013916510364461

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*, 1681–1685. doi:10.1126/science.1161405

Keizer, K., Lindenberg, S., & Steg, L. (2011). The reversal effect of prohibition signs. *Group Processes & Intergroup Relations*, *14*, 681–688. doi:10.1177/1368430211398505

Kelling, G. L., & Coles, C. M. (1998). *Fixing broken windows: Restoring order and reducing crime in our communities*. New York, NY: Free Press.

Krauss, R. M., Freeman, J. L., & Whitcup, M. (1978). Field and laboratory studies of littering. *Journal of Experimental Social Psychology*, *14*, 109–122. doi:10.1016/0022-1031(78)90064-1

Meeker, F. L. (1997). A comparison of table-littering behavior in two settings: A case for a contextual research strategy. *Journal of Environmental Psychology*, *17*, 59–68. doi:10.1006/jevp.1996.0039

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660. doi:10.1177/1745691612462588

Pires, A. M., & Branco, J. A. (2010). A statistical model to explain the Mendel–Fisher controversy. *Statistical Science*, *25*, 545–565. doi:10.1214/10-STS342

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*, 420–428. doi:10.1016/j.tics.2009.08.002

Ramos, J., & Torgler, B. (2010). *Are academics messy? Testing broken windows theory with a field experiment in the work environment.* Fondazione Eni Enrico Mattei Working Paper No. 493.

Reiter, S. M., & Samuel, W. (1980). Littering as a function of prior litter and the presence or absence of prohibitive signs. *Journal of Applied Social Psychology*, *10*, 45–55. doi:10.1111/j.1559-1816.1980.tb00692.x

Schultz, P. W., Bator, R. J., Large, L. B., Bruni, C. M., & Tabanico, J. J. (2011). Littering in context: Personal and environmental predictors of littering behavior. *Environment and Behavior*, *45*, 35–59. doi:10.1177/0013916511412179

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). New York, NY: Wiley.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804. doi:10.3758/BF03194105

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. doi:10.1177/1745691612463078

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. doi:10.1037/0003–066X.61.7.726

Wilson, J. Q., & Kelling, G. L. (1982, March). Broken windows: The police and neighborhood safety. *Atlantic Monthly*. Retrieved from http://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/

## Appendix A

### *Overly Consistent Results*

In seven instances, Keizer et al. (2008, 2011) replicated their own findings by collecting more data at the same location and under the same environmental condition (i.e., the original study only differed from the replication in terms of the day [date] of data collection). These seven sets of studies (conditions) are described in Table A1. Only for Studies 5 and 6 in the *Science* (2008) paper did the original study and its replication differ slightly in terms of environmental cues (litter vs. graffiti), but Keizer et al. (2008) expected these cues to have the same effect on behavior. Table A1 gives frequencies for each pair of original study and its replication. A Fisher's exact test (two-tailed) and Pearson's $\chi^2$ test of independence between the original study and its replication provide checks of consistency under the null hypothesis of an identical underlying proportion of people who violated a norm. Differences in percentages of observed misdemeanors for the seven sets are quite small in light of the expected standard errors (around 5.4%), namely 1.7%, 1.3%, 3.2%, 1.0%, 0.0%, 0.0%, and 6.5% respectively. In two cases results were exactly identical and in four cases Fisher's exact test gives $p = 1.000$. The sum of Pearson's $\chi^2$s represents an omnibus test and equals 1.064, $df = 7$, $p = .994$. Under the assumptions of the model, one would expect to find such consistent results (or more consistent results) in less than one out of 150 sets of replications ($p = .006$).

We asked Dr. Keizer whether data from the *Science* (2008) paper were reused in the *GPIR* paper and after several emails he eventually acknowledged that Set 6 represented the same data. This overlap in data across both papers was not mentioned by Keizer et al. (2011) and represents a violation of *GPIR* author guidelines and casts further doubt on the use of one-sided testing by Keizer et al. (2011). If we take the data from all remaining six replications as independent, the sum of Pearson's $\chi^2$s becomes 1.064, $df = 6$, $p = .983$. This implies that only in 1.7% would results in these six replications be more consistent under the null hypothesis of exact replications ($p = .017$). Given that the null hypothesis is quite unlikely in an uncontrolled field study, this overly high level of consistency casts doubt on the validity of Keizer et al.'s data. Explanations for this finding include sequential testing, violations of dependency, observer biases, and publication bias (i.e., a selection of these replications from a larger set of replications).

**Table A1.** Results of original studies and replications by Keizer et al. (2008, 2011) and tests for consistency.

| Set no. | Conditions | Original | | Replication | | Fisher | Pearson's $\chi^2$ | $p$ |
|---|---|---|---|---|---|---|---|---|
| | | Yes | No | Yes | No | | | |
| 1 | S5-graffiti S6-litter | 16 | 44 | 18 | 54 | .844 | 0.0475 | 0.827 |
| 2 | G1.1lit_nosign G1.3lit_nosign | 46 | 29 | 45 | 30 | 1.000 | 0.0279 | 0.867 |
| 3 | G1.2lit_sign G1.3lit_sign | 53 | 21 | 52 | 24 | .723 | 0.1829 | 0.669 |
| 4 | G2.1gra_nosign G2.3gra_nosign | 48 | 29 | 45 | 26 | 1.000 | 0.0172 | 0.896 |
| 5 | S1nogra_sign G2.4nogra_sign | 25 | 52 | 25 | 52 | 1.000 | 0.0000 | 1.000 |
| 6 | S1gra_sign G2.2gra_sign* | 53 | 24 | 53 | 24 | 1.000 | 0.0000 | 1.000 |
| 7 | G2.2gra_sign G2.3gra_sign | 53 | 24 | 55 | 18 | .467 | 0.7881 | 0.375 |

*Note.* S: study from *Science* (2008) paper; G: study from *GPIR* (2011) paper; lit: littered environment; graf: graffiti in environment; sign: sign in the environment; nosign: no sign in the environment; yes represents misdemeanor (norm violation); *Dr. Keizer acknowledged that the data of S1gra_sign was the same as the data from G2.2gra_sign, while the other sets represented novel data.

## Appendix B

R *Code Used for the Simulations*

```
########################
##### Broken Windows ###
########################
#H0 is true; i.e., independence of littering and condition
pF=.5        #chance of littering
pS=.8        #chance to shop together
pZ=.5        #chance that pairs behave in the same manner
n=74         #sample size within each condition
nsim=10000   #number of simulations
pn=pp=pnb=ppb=rep(NA,nsim)
chin=chip=chinb=chipb=rep(NA,nsim)
for(i in 1:nsim)
{
#normal samples (nominal)
      g1n=rbinom(n,1,pF)
      g2n=rbinom(n,1,pF)
      mn=matrix(c(sum(g1n),sum(g2n),n-sum(g1n),n-sum(g2n)),2,2)
      pn[i]=chisq.test(mn,correct=F)$p.value
      chin[i]=chisq.test(mn,correct=F)$statistic

  #data with paired samples
        #number of sets that are paired
        p1=rbinom(round(n/2),1,pS)
        p2=rbinom(round(n/2),1,pS)
        #number of pairs that show the same behavior
        pZ1=rbinom(sum(p1),1,pZ)
        pZ2=rbinom(sum(p2),1,pZ)
        #data of pairs that show the same behavior
        dp1=rbinom(sum(pZ1),1,pF)
        dp2=rbinom(sum(pZ2),1,pF)
        #twice the data from pairs that show the same behavior,
        #and remaining cases are sampled individually.
        g1p=c(dp1,dp1,rbinom(n-2*length(dp1),1,pF))
        g2p=c(dp2,dp2,rbinom(n-2*length(dp2),1,pF))
        mp=matrix(c(sum(g1p),sum(g2p),n-sum(g1p),n-sum(g2p)),2,2)
        pp[i]=chisq.test(mp,correct=F)$p.value
        chip[i]=chisq.test(mp,correct=F)$statistic

  #intermediate testing
        if(chisq.test(mn,correct=F)$p.value<.05)
        {
            pnb[i]=chisq.test(mn,correct=F)$p.value
            chinb[i]=chisq.test(mn,correct=F)$statistic
        }else{
            g1nb=c(g1n,rbinom(n,1,pF))
```

```
          g2nb=c(g2n,rbinom(n,1,pF))
          mnb=matrix(c(sum(g1nb),sum(g2nb),n*2-sum(g1nb),
          n*2-sum(g2nb)),2,2)
          pnb[i]=chisq.test(mnb,correct=F)$p.value
          chinb[i]=chisq.test(mnb,correct=F)$statistic
    }

#Intermediate testing with dependent data

    if(chisq.test(mp,correct=F)$p.value<.05)
    {
          ppb[i]=chisq.test(mp,correct=F)$p.value
          chipb[i]=chisq.test(mp,correct=F)$statistic
    }else{
          #number of sets that are paired

          p1b=rbinom(round(n/2),1,pS)
          p2b=rbinom(round(n/2),1,pS)
          #number of pairs that show the same behavior

          pZ1b=rbinom(sum(p1b),1,pZ)
          pZ2b=rbinom(sum(p2b),1,pZ)
          #data of pairs that show the same behavior
          dp1b=rbinom(sum(pZ1b),1,pF)
          dp2b=rbinom(sum(pZ2b),1,pF)
    #twice the data from pairs that show the same behavior,
    #and remaining cases are sampled individually.
          g1pb=c(g1p,dp1b,dp1b,rbinom(n-2*length(dp1b),1,pF))
          g2pb=c(g2p,dp2b,dp2b,rbinom(n-2*length(dp2b),1,pF))
          mpb=matrix(c(sum(g1pb),sum(g2pb),2*n-sum(g1pb),
          2*n-sum(g2pb)),2,2)
          ppb[i]=chisq.test(mpb,correct=F)$p.value
          chipb[i]=chisq.test(mpb,correct=F)$statistic
    }
}
type1n=length(which(pn<.05))/nsim
type1p=length(which(pp<.05))/nsim
type1nb=length(which(pnb<.05))/nsim
type1pb=length(which(ppb<.05))/nsim

type1n    #type 1 error rate nominal
type1p    #type 1 error rate data with dependencies
type1nb   #type 1 error rate with intermediate testing
type1pb   #type 1 error rate with intermediate testing &
dependencies

#chin     #chi squares, nominal
#chip     #chi squares, dependencies
#chinb    #chi squares, intermediate testing
```

```
  #chipb    #chi squares, intermediate testing & dependencies
```

## Appendix C

### *A Meta-Analysis of the Reversal Effect of Prohibition Signs*

This meta-analysis on the reversal effect of prohibition includes both studies by Keizer et al. (2011; Table 2 in the main text) and the study by Reiter and Samuel (1980; Table 3 in the main text). We focus on whether the effect litter on littering behavior is moderated by absence or presence of a prohibition sign in that environment.

### *Method*

We first computed within each cell of the design the simple effects that concern littering behavior (yes or no) and the presence or absence of litter in the environment, which represents the typical finding that littered environments lead to more littering. The effect size for these simple effects is a log odds ratio, with positive values indicating more littering in littered environments. Because the reversal effect of prohibition signs concerns an interaction between the presence of litter (with/without) and sign (with/without), we subsequently compared for each study separately the simple effects (log odds ratios) between conditions that either had a prohibition sign or not. To this end, we subtracted the simple effects in the sign condition from the simple effects in the no sign condition. A positive difference term (interaction) would reflect the reversal effect of prohibition signs. Because of the methodological and substantive differences between the three studies we employed a random effects model. We used the corresponding sums of sampling variances as the sampling variance for the difference term (because of independence of data points).

### *Results*

The log odds ratios for the simple effects and the difference term (interaction) are given in Table C1. Note that this analysis also includes data from the second study in Keizer et al. (2011) and that it also involves some overlapping data because of the reuse of the baseline condition (see Table 2 of main text).

The random effects model for the interaction term (difference) provided a mean estimate of 0.52 (*SE* = 0.34). The 95% confidence interval included zero: [−0.15, 1.91]. In addition, a formal test (Sterne & Egger, 2005) indicated funnel plot asymmetry: $Z = 2.15$, $p = .03$, which indicates that larger interaction effects were associated with smaller sample sizes. Such funnel plot asymmetry is widely seen as an indicator of publication bias.

### *Conclusion*

This meta-analysis corroborates the analyses in the main paper by showing that the reversal effect of prohibition signs as proposed by Keizer et al. (2011) is currently insufficiently supported by the data. In addition, the funnel plot asymmetry we found can be seen as an indicator of publication bias or related biases.

**Table C1.** Log odds ratios of littering behavior for the three studies (simple effects) and difference between log odds ratios between conditions with and without the prohibition sign for Keizer et al. (2011) and Reiter and Samuel (1980).

| Study | Log odds ratio (*SE*) | | Difference (*SE*) |
|---|---|---|---|
| | Without sign | With sign | |
| Keizer et al. (2011) Study 1 | 0.56(0.28) | 1.29(0.29) | 0.72(0.41) |
| Keizer et al. (2011) Study 2 | 0.66(0.28) | 1.68(0.30) | 1.02(0.42) |
| Reiter & Samuel (1980) | 0.72(0.26) | 0.63(0.22) | −0.08(0.34) |