Statistical infarction: A postmortem of the Cornell Food and Brand Lab pizza publications

Jordan Anaya[1], Tim van der Zee[2], Nicholas J. L. Brown[3]

[1] Omnes Res, Charlottesville, Virginia, omnesres.com

email: omnesresnetwork@gmail.com

twitter: @omnesresnetwork

[2] Graduate School of Teaching (ICLON), Leiden University, Leiden, The Netherlands

email: t.van.der.zee@iclon.leidenuniv.nl

twitter: @Research_Tim

[3] University Medical Center, University of Groningen, The Netherlands

email: nick.brown@free.fr

twitter: @sTeamTraen

Corresponding Author:

Jordan Anaya[1]

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Statistical infarction: A postmortem of the Cornell Food and Brand Lab pizza publications

Jordan Anaya[1], Tim van der Zee[2], and Nicholas J. L. Brown[3]

[1]**Omnes Res, Charlottesville, Virginia**
[2]**Graduate School of Teaching (ICLON), Leiden University, Leiden, The Netherlands**
[3]**University Medical Center, University of Groningen, The Netherlands**

Corresponding author:
Jordan Anaya[1]

Email address: omnesresnetwork@gmail.com

ABSTRACT

We previously reported over 150 inconsistencies in a series of four articles (the "pizza papers") from the Cornell Food and Brand Lab that described a study of eating habits at an all-you-can-eat pizza buffet. The lab's initial response led us to investigate more of their work, and our investigation has now identified issues with at least 45 publications from this lab. Perhaps because of the growing media attention, Cornell and the lab have released a statement concerning the pizza papers, which included a response to the inconsistencies, along with data and code. Many of the inconsistencies were identified with the new technique of granularity testing, and this case has the highest density of granularity inconsistencies that we know of. This is also the first time a data set has been made public after granularity concerns were raised, making it a highly suitable case study for showing the accuracy and potential of this technique. It is also important that a third party audit the lab's response, given the continuing investigation of misconduct and presumably future reports and data releases. Our careful inspection of the data set suggests no evidence of fabrication, but we found the lab's report confusing, incomplete, and error prone. In addition, we found the number of missing, unusual, and logically impossible responses in the data set highly concerning. Unfortunately, given the unsound theory, poor methodology, questionable data, and countless errors, we find it remarkable that these four papers were published and recommend retraction of all four papers.

Keywords: Statistics, Reproducibility, Replication, Reanalysis

INTRODUCTION

> "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." -Ronald Fisher

Concerned about the quality of the scientific literature (O'Grady, 2017) we have developed tools to help readers easily identify inconsistencies in papers (Brown and Heathers, 2016; Anaya, 2016). Using these tools we previously identified around 150 inconsistencies in four[1] articles (the "pizza papers") from the Cornell Food and Brand Lab (henceforth 'lab') after being alerted to their existence in a blog post by the senior author (van der Zee et al., 2017; Wansink, 2016). We were interested in the articles because the description provided in the blog post suggested they were the product of severe *p*-hacking, and seemed to be a clear case of salami slicing as they are based on the same data set.

Upon our initial inspection of the four articles we immediately noticed large numbers of granularity inconsistencies. This inspired us to take a closer look at the articles and more inconsistencies were immediately apparent such as impossible degrees of freedom. A careful reading of all four articles together revealed differing accounts of the methods, conflicts

[1]In that report we labeled the articles as Article 1 (Just et al., 2014), Article 2 (Just et al., 2015), Article 3 (Kniffin et al., 2016), and Article 4 (Sığırcı and Wansink, 2015).

in the sample sizes both within and between the articles, and a host of other problems such as impossibly large standard deviations. To try and identify the source of these problems we emailed the corresponding authors of the articles to inquire about getting access to the data set, but only received a response from the lab's "Communications Specialist". The first response told us that if we were interested in the study we should replicate it and that all the methods to do so were present in the publication. After we made it clear that we simply wanted to check some inconsistencies we received no further responses.

We felt that the problems we identified were severe enough to make public, and that this was a good demonstration for how granularity testing can be used to scrutinize the literature, so we made our findings public with a preprint on January 25th (van der Zee et al., 2017). After posting our preprint, the senior author provided several explanations for why the data could not be shared. On his blog he explained that the data were sensitive to the participants (Wansink, 2016), in an interview with *Retraction Watch* he explained that the data was "tremendously proprietary" (McCook, 2017a), and in an interview with *The Chronicle* admitted to thinking about sharing the data but decided against it because our request appeared "not intended toward helping move science forward" (Bartlett, 2017).

When errors in a publication are identified there is always the possibility they are simply typos or due to other innocent explanations such as unreported missing values. In other words, we can expect a certain base rate of errors which may appear in even the most carefully considered paper. However, given the range and number of problems in the pizza papers, it was difficult to understand how the problems could be explained solely by missing data, or a poor typist. With problems at this scale falsification and/or fabrication comes to mind, but given the seemingly random nature of the problems we found this to be an unlikely explanation. Having received no resolution on this matter, we performed a systematic investigation into other work by this lab hoping to find answers. We found other work from this lab to also contain granularity problems, inconsistent sample sizes, or numbers which simply did not add up. Unfortunately, we also found a large amount of text recycling, a few cases of data recycling, and some very unlikely coincidences and data distributions. Our findings to date are summarized at van der Zee (2017).

At some point during our ongoing investigation the lab decided it was necessary to perform an internal review of the pizza publications, and committed to releasing the underlying data once the review was complete (Wansink, 2017a). True to this statement, Cornell and the lab released a response to our critique along with the underlying data and analysis scripts (Wansink, 2017b; Carberry, 2017). Cornell came to the conclusion that the alleged errors "did not constitute scientific misconduct", and the lab's internal review found the errors did not alter the conclusions of the study. Along with the data and STATA scripts, the lab released Response Tables summarizing the fixed statistics and a List of Inconsistencies detailing each error and marking whether it was "unique and valid". It appears the STATA scripts were independently reviewed by Mathematica Policy Research, along with the errata sent to the journals, but we have not seen these errata and it is not clear if the Response Tables were independently reviewed. In this report we will try to clearly present the underlying problems that caused the 150 errors and describe additional issues that are apparent now that we have access to the data.

We hope that our criticisms of the released data do not discourage Cornell or others from releasing data in the future. While releasing data may open one up to more criticism, reluctance to share data or provide explanations for anomalies is suggestive of serious scientific misconduct. Indeed, after excuses for why data could not be shared the worst was confirmed in the recent cases of Michael Lacour and Oona Lönnstedt (Oransky, 2015; Enserink, 2017). Perhaps Andrew Gelman in his commentary on pizzagate said it best (Gelman, 2017):

> "It seems pretty simple to me: Wansink has no obligation whatsoever to share his data, and we have no obligation to believe anything in his papers."

This should really apply to all research, but is especially applicable for papers that have serious statistical issues.

A POSTMORTEM OF OUR METHODS

We would like to thank the lab for releasing the underlying data. The data release allows us to see what we got right, and what we got wrong. It also allows us the opportunity to show the scientific community what types of problems granularity testing can uncover. In the seminal paper on granularity testing, the authors emailed the corresponding authors of 21 papers

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

with inconsistencies (Brown and Heathers, 2016). The data sets they received revealed the errors to be the result of typos, misreported sample sizes, incorrect rounding/double rounding, or miscalculations. While these sound like small mistakes, because the identities of the papers were protected, it was not made apparent exactly how large or small these mistakes were, and whether they raised doubts about the conclusions of the studies. Our critique of the pizza papers was posted publicly (van der Zee et al., 2017), and the data release was also made public (Wansink and Payne, 2007), which allows members of the scientific community to see for themselves what types of problems granularity testing can reveal, and judge for themselves the seriousness of these errors.

What we did wrong

In science it is important to quickly and unconditionally admit to errors so that future researchers do not perpetuate your mistakes or waste time trying to replicate a flawed result or technique. After reviewing the lab's response to our criticism and reanalyzing their data we became aware of a few errors in our methodology. Overall these errors only affect a few minor points in our report, but because we do not want others to make the same mistakes as we did we would like to thoroughly describe these cases.

Checking conversions

In Table 1 of Article 3 the authors reported height in both centimeters and inches, and weight in both kilograms and pounds. We noticed that they had performed these conversions incorrectly, and ambitiously decided to move on to check their BMI calculations. We stated how we recalculated the mean BMI using the ratio of the means provided by the original authors for height and weight, but we are now aware that the method described will not faithfully reproduce the value obtained by the correct method, which is to calculate the BMI for each diner and then take the average of those BMI values.

The problem with the BMI recalculation arises because calculating BMI involves a nonlinear transformation of the data. It is easiest to see this problem when there are only two data points. Figure 1 shows the nonlinear transformation $1/x$ of "original data" to "transformed data". Taking two points on this graph, x_1 and x_2, it is easy to see that the mean of their transformed values does not fall on the curve. Clearly, the transform of the average of the original values of x_1 and x_2 will fall on the curve and be different from the mean of the transforms. This difference is illustrated as "Error" in the figure.

As a result, conversions of means should only be checked if the conversion is a linear transformation such as the conversion between centimeters and inches, or between kilograms and pounds. Interestingly, the response to our critique correctly points out that our BMI recalculations were inappropriate, but it also claims the height and weight recalculations suffered from the same problem. We do not agree with this statement. Calculating an arithmetic mean involves a summation:

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

And summation has the following well established mathematical property:

$$\sum_{i=1}^{n} C \cdot x_i = C \cdot \sum_{i=1}^{n} x_i$$

Multiplying by the constant C is a linear transformation, and it does not matter whether this transformation occurs before or after the summation. In this case, the mean of the transformations is equal to the transformation of the means, and it is unclear why the lab believes our calculations are in error. Unfortunately, because the data release did not contain individual-level height or weight information, we cannot attempt to reproduce their results and identify the source of their error(s).

Two-way ANOVAs

To recalculate the two-way between-subjects ANOVAs in the pizza papers, we used the `rpsychi` package in R (R Core Team, 2016) — which calculates F statistics from means, SDs, and cell sizes — and added our own code to put upper and lower bounds on these test statistics, to take into account the fact that means and SDs reported to two decimal places have been rounded (so that, for example, a mean of 2.52 could correspond to any number between 2.515 and 2.525). The

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

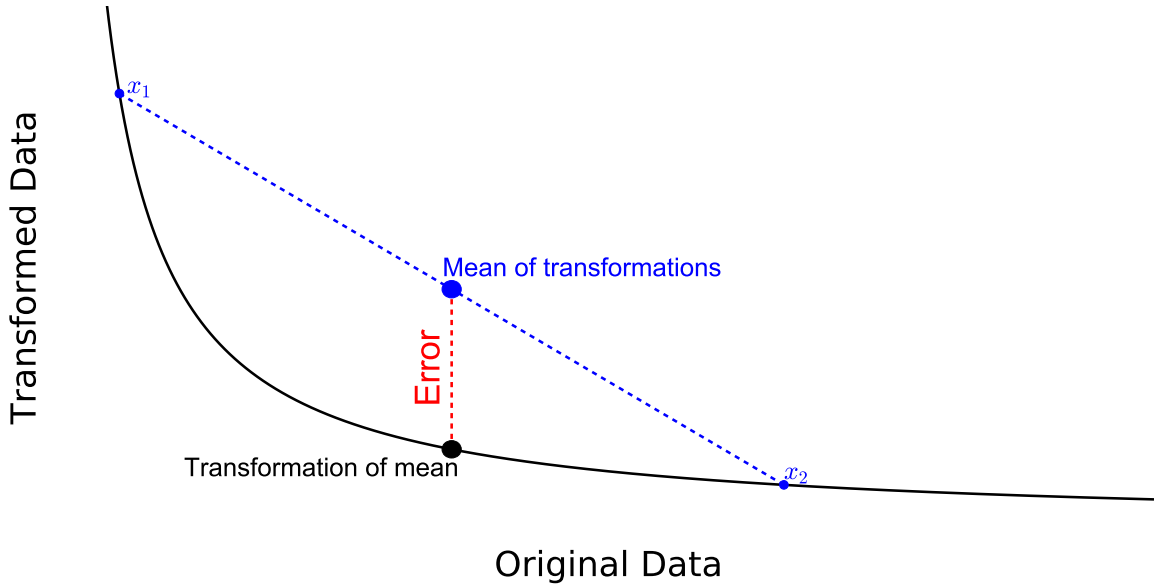

Figure 1. Illustration showing that the mean of transformations does not equal the transformation of the mean if the data underwent a nonlinear transformation. Here x_1 and x_2 are two data points in "original data". The illustration plots their location after having undergone the $1/x$ transformation.

`ind.twoway.second` function in the `rpsychi` package uses the means, SDs, and cell sizes to recalculate the two-way ANOVAs using an unweighted-means solution as described in Cohen (2002). We were unsure how the Food and Brand Lab had calculated their two-way ANOVAs, but since they cited an article about SPSS in Article 1, we speculated that they may have used this software. To check that `rpsychi` would give the same values as SPSS we generated some data sets and checked the results against `rpsychi`; the values were in close agreement in all cases. Furthermore, `rpsychi` successfully reproduced some of the two-way ANOVA calculations in the pizza papers, and therefore we assumed that the discrepancies we observed were due to misreported means, SDs, and/or sample sizes.

The data release provided the opportunity to check whether `rpsychi` was faithfully reproducing the statistics. We confirmed that it had correctly recalculated the unbalanced 2X2 ANOVAs present in Article 3, but not the unbalanced 2X3 ANOVAs present in Article 4. Dr. Barry Cohen (personal communication, April 26, 2017) informed us that the unweighted-means solution will only give the same values as the Type 3 sum of squares solution if the two-way ANOVA is balanced, or each factor only contains two levels. As a result, although some of our estimations for the 2X3 ANOVAs in Article 4 may be accurate, we should not have used `rpsychi` to recalculate those values. It should also be noted that `rpsychi` assumes the use of Type 3 sums of squares; however, this is usually a resonable assumption, as Type 3 is the default in psychology and in most popular statistics software such as STATA, SPSS, and SAS, and can be done in R with the `ezANOVA` function in the package `ez`.

What we did right

Granularity testing can be thought of as a diagnostic test that can give a quick idea of the overall health of a paper. Granularity inconsistencies indicate that a paper has some problems, but cannot on their own reveal the source of the problems (and the absence of granularity problems does not, of course, guarantee a paper's complete health). When multiple inconsistencies are found, a close inspection of the paper is warranted with special attention paid to sample sizes and degrees of freedom, and further diagnostic tests such as the recalculation of test statistics may be needed.

With this strategy we were able to make several correct inferences. For example, we correctly identified that it was impossible for the modal number of pizza slices to be 3 as stated in three of the articles, correctly inferred large discrepancies

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

in the number of diners who ate 2 slices of pizza, and correctly hypothesized that the authors incorrectly used the terms 1st, 2nd, and 3rd to refer to the order in which diners ate slices of pizza in Table 3 of Article 1 (this should have been reported as first, middle, and last). All of these inferences were made by first noticing granularity problems, and then inspecting the sample sizes and degrees of freedom.

Overall, the large discrepancies in the sample sizes, combined with the granularity inconsistencies and test statistic errors, gave us serious concerns about the underlying data and the data handling, which have now been confirmed with the data release.

Granularity Testing

Granularity testing is fairly foolproof, although care must be taken that the underlying data elements are integers and not means from a composite measure (or, if they are, that appropriate compensation has been made). For instance, data for height (in inches), weight (in pounds), and age (in years) are fairly likely to be whole numbers, but could possibly contain decimal values if someone decides to mention, say, an extra half inch. Another potential issue is different rounding conventions, especially at sample sizes that are multiples of 8 when rounding to two decimals (e.g., 105 / 40 = 2.625, which could be reported as 2.62 or 2.63 depending on the software being used). This last issue can largely be avoided by considering both the rounded-up and rounded-down values.

When one or two granularity inconsistencies are identified, it is reasonable to assume that they are typos or due to missing data for that row of the table. When several granularity inconsistencies are present the pattern of the inconsistencies can be used to infer the underlying cause. Inconsistencies that are all in the same column can be due to the sample size at the top of the column being listed incorrectly, and inconsistencies in the same row could be indicative of a question with a poor response rate. When there are inconsistencies spread randomly throughout a table further techniques can be used to help judge their severity. We are not aware of any mistakes in our granularity testing.

Test statistics

The main limitation of granularity testing is that detected inconsistencies are just that: they are only inconsistencies and may not be errors. For example, if the sample sizes for each statistic are different from what is reported then these may show up as granularity inconsistencies. Whether this inaccurate reporting is meaningful is up for debate. Imagine there are 100 participants who give responses to 14 items each, but one person's response to one item was lost. Is that a big deal? Probably not. But what if half of the responses are lost? With granularity testing we don't know how severe the problem is, only that there is a problem, and recalculating test statistics can help reveal the severity of the problem.

If the granularity inconsistencies are due solely to small errors in rounding or a few dropped participants, the test statistics shouldn't change much. Large discrepancies in the test statistics suggest large problems with the means, SDs, or sample sizes. Although this additional information is valuable, recalculating test statistics is more involved than granularity testing, with several potential pitfalls such as between vs within design or equal vs unequal variance[2].

If there are test statistics reported in the text an automated check that the DFs and statistics match the p values can be performed with statcheck (Epskamp and Nuijten, 2015). Statistics and p values for one-way and two-way ANOVAs can be recalculated when only the means, SDs, and sample sizes are known using `rpsychi` or the Python functions provided in our GitHub repository. With the exception of the unbalanced 2X3 ANOVAs in Article 4, we believe our recalculations of test statistics to be accurate and reliable.

Detective work

If granularity issues have been identified, and the recaclulated test statistics also show substantial discrepancies, a detailed investigation is likely warranted. The easiest thing to do is to contact the authors to ask for the data set, and use that to try and understand the inconsistencies you have found. If the authors are unresponsive or uncooperative it is then up to you to decide if the problems are worth pursuing further. The DFs in the paper can reveal issues with sample sizes, and careful attention should be paid to unusually large or small standard deviations. The standard deviations and means can be used to

[2]Use of Welch's ANOVA is generally reported in the text of an article. It can also generally be detected by visual inspection, since the value for the denominator degrees of freedom typically has a fractional component.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

reconstruct the range of data sets which can result in that particular mean and variation, and may reveal a highly unusual distribution (Heathers, 2017).

A POSTMORTEM OF THE DATA SET

As expected from the granularity inconsistencies, the released data set contains a large number of missing responses. On its own, this need not be a major problem, assuming that the missing data points are indicated somehow in the publication—ideally by providing the correct cell sizes for each response. However, given the discrepancies in sample sizes within and between publications we expected the issues to go deeper than missing responses. Indeed, the data release shows that almost every single sample size was reported incorrectly, with many sample sizes being **larger** than what was originally reported. The data also contains numerous logically impossible responses which were included in the original analyses, and for some statistics it is unclear how to reproduce them given the data.

Who, what, when, where, how, why?

Based on the four different papers describing the methods, news coverage of the study, public statements about the study by the senior author, and the data release, we would expect to have clear answers to the basic questions about this study. However, the who, what, when, where, how, and why all remain unclear to a large degree.

When?

All four articles claim the study was conducted over a 2-week period, however the senior author's blog post described the study as taking one month (Wansink, 2016), the senior author told *Retraction Watch* it was a two-month study (McCook, 2017b), a news article indicated the study was at least 3 weeks long (Lazarz, 2007), and the data release states the study took place from October 18 to December 8, 2007 (Wansink and Payne, 2007). Why the articles claimed the study only took two weeks when all the other reports indicate otherwise is a mystery.

Furthermore, articles 1, 2, and 4 all claim that the study took place in spring. For the Northern Hemisphere spring is defined as the months March, April, and May. However, the news report was dated November 18, 2007, and the data release states the study took place between October and December. It is unclear why the articles included this inaccurate statement.

How?

All four articles state that the diners completed a survey after having finished their meals. However, in his *Retraction Watch* interview the senior author stated (McCook, 2017b):

> "These people were eating lunch and they could skip any question they wanted – maybe they were eating with their best buddy or girlfriend and didn't want to be distracted."

This suggests that the diners actually filled out the surveys while they ate. In addition, the authors' response to our preprint states:

> "Field study surveys are notoriously incomplete because people skip questions. This is especially true when people are eating – some do not want to write down their weight and others do not want to write down what they ate."

Again this suggests that the surveys were filled out while the diners were eating.

Article 1 states that the diners were asked to estimate how much they ate, while Article 3 states that the amount of pizza and salad eaten was unobtrusively observed, going so far as to say that appropriate subtractions were made for uneaten pizza and salad. Adding to the confusion Article 2 states:

> "Unfortunately, given the field setting, we were not able to accurately measure consumption of non-pizza food items."

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

In Article 3 the tables included data for salad consumed, so this statement was clearly inaccurate. It was still an open question however whether diners self-reported how much they ate or were carefully observed. In his *Retraction Watch* interview the senior author stated:

> "Also, we realized we asked people how much pizza they ate in two different ways – once, by asking them to provide an integer of how many pieces they ate, like 0, 1, 2, 3 and so on. Another time we asked them to put an 'X' on a scale that just had a '0' and '12' at either end, with no integer mark in between."

A recent Editorial Note from the journal in which Article 3 was published (Shackelford, 2017) confirmed that the pizza and salad consumption data for this article were indeed collected by self-report. Yet, as noted previously, when we first contacted the lab they suggested we perform a replication and that all the information needed to do so was in the papers. This clearly is not the case, and this Editorial Note confirms that fact. We have expressed our concerns in an open letter to the editor of the journal in question (Brown et al., 2017). Simply put, we consider that this avowed discrepancy between what was reported and what actually took place constitutes falsification of the scientific record and should be considered scientific misconduct.

Who?

Articles 1 and 2 claim there were 139 participants, while Article 3 states there were 133 participants. Article 1 mentions 6 diners younger than 18 were eliminated, and we assumed this was the reason for the 6 diner discrepancy between Article 3 and Articles 1 and 2. However, the data release contains 139 diners with no indication that any diners were eliminated because of age. There are 6 diners with missing group information however, which is likely the explanation for the sample size of 133 in Article 3, whose analysis relies upon group information.

This raises another problem. It is unclear why Article 1 states 6 diners were removed because of their age. From our reanalysis and the released STATA scripts this clearly didn't happen. Perhaps there was something in the IRB approval about the age of the participants and the authors intended to remove these participants but didn't, or perhaps this is simply another inaccurate statement.

Articles 1, 2, and 3 state that 8 diners ate alone, and Articles 1 and 2 state that 52 diners ate in groups of 2. However, the data set clearly indicates that only 6 diners ate alone and 54 diners ate in groups of 2. Presumably this was a typo that was subsequently propagated to 3 different papers.

Why?

Perhaps the most important question is why did this study take place? In the blog post the senior author did mention having a "Plan A" (Wansink, 2016), and in a *Retraction Watch* interview revealed that the original hypothesis was that people would eat more pizza if they paid more (McCook, 2017a). The origin of this "hypothesis" is likely a previous study from this lab, at a different pizza buffet, with nearly identical study design (Just and Wansink, 2011). In that study they found diners who paid more ate significantly more pizza, but the released data set for the present study actually suggests the opposite, that diners who paid less ate more. So was the goal of this study to replicate their earlier findings? And if so, did they find it concerning that not only did they not replicate their earlier result, but found the exact opposite? Did they not think this was worth reporting?

Another similarity between the two pizza studies is the focus on taste of the pizza. Article 1 specifically states:

> "Our reading of the literature leads us to hypothesize that one would rate pizza from an $8 pizza buffet as tasting better than the same pizza at a $4 buffet."

Either they did not read their own previous pizza buffet study, or they do not consider it to be part of the literature, because in that paper they found ratings for overall taste, taste of first slice, and taste of last slice to all be higher in the lower price group, albeit with different levels of significance (Just and Wansink, 2011). However, in the later study they again found the exact opposite, but did not comment on the discrepancy.

Of course, there is a parsimonious explanation for these contradictory results in two apparently similar studies, namely that one or both sets of results are the consequence of modeling noise. Given the poor quality of the released data from the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

more recent articles (see below), it seems quite likely that this is the correct explanation for the second set of studies, at least.

What?

Even with the Editorial Note for Article 3 clarifying that pizza and salad consumption were self-reported, with diners reporting the number of slices they ate and marking on a 13 point scale how much salad they ate, there are still unanswered questions. Figure 2 below shows the distribution of salad responses.

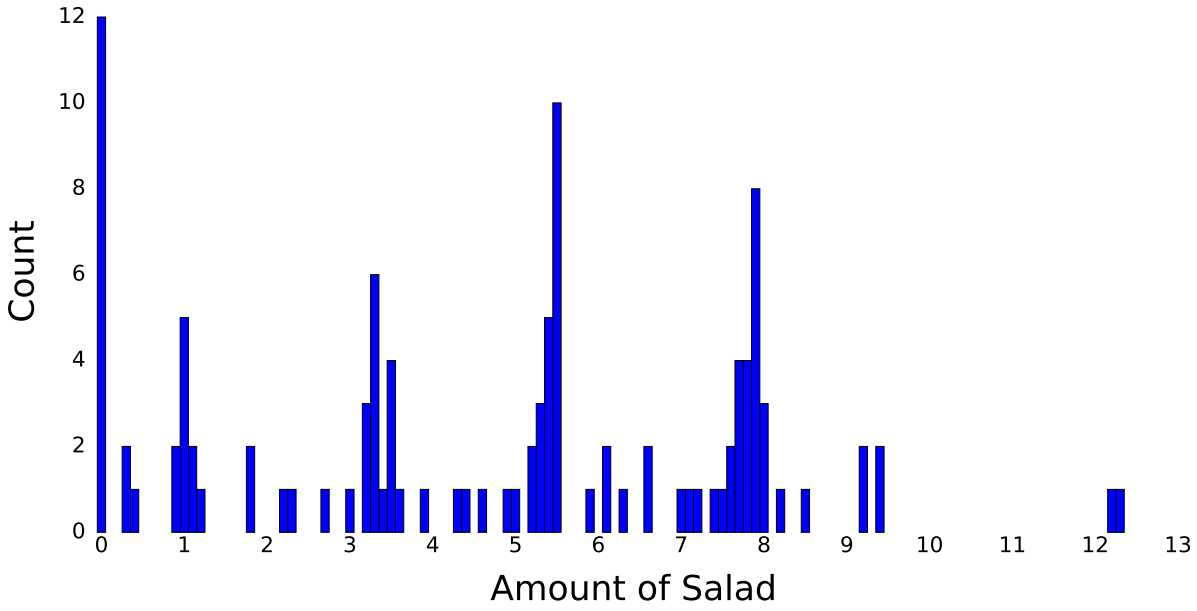

Figure 2. Distribution of salad responses.

One thing to notice is that there are responses above 12. If the scale was a 13 point scale from 0 to 12 how are these values possible? In addition, why are responses such as 5.5 or 7.9 so popular? If the scale contains tick marks you might expect most responses to use those marks and be whole numbers, but it is unclear what type of scale would produce the frequencies observed.

In addition, although the Editorial Note makes it clear that diners self-reported the number of pieces they ate, there are 3 responses that are non-integer values. Perhaps these values were due to the continuous scale mentioned in the *Retraction Watch* interview (but not the Editorial Note), or maybe the diners misunderstood the question (see below for more instances). Furthermore, Articles 1, 2, and 4 state the modal number of pieces was 3. However, Figure 3 clearly shows that this is not the case, as the modal number is either 1 or 2 depending on how fractional responses are handled.

Impossible responses

It is understandable for a field study to be messier than a laboratory study, but at some point we must question whether the data can be trusted. This entire study relies upon diners' responses, however many of these responses are logically impossible, suggesting the diners either did not understand the questionnaire or did not take it seriously. Either way, none of us would have found this data set reliable enough to publish a paper, let alone four that received large amounts of media attention.

Impossible pizza consumption

The researchers asked diners to evaluate their first piece of pizza, middle piece of pizza, and last piece of pizza. If a diner ate 1 piece of pizza they should have a response for the first piece, but not a middle piece or last piece. If a diner ate 2

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

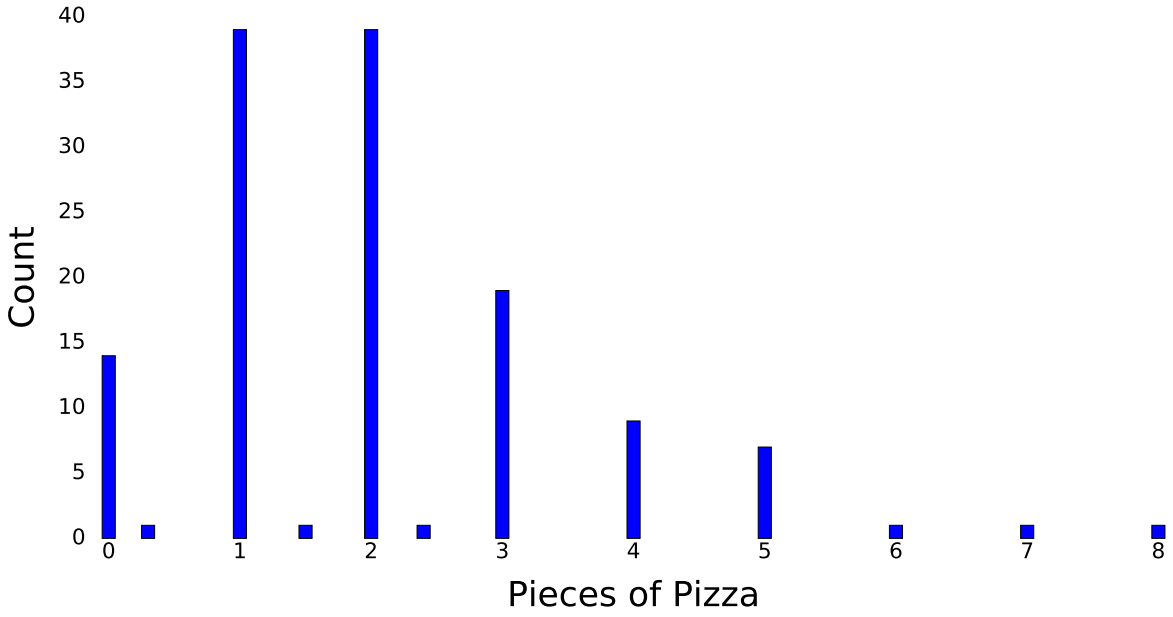

Figure 3. Distribution of pizza responses.

pieces they should have a response for a first piece and last piece, but not a middle piece. Only diners who ate 3 or more pieces should have responses for first, middle, last. With this in mind, the data can be checked for inconsistencies.

A careful inspection of the responses to items evaluating the quality of pizza consumed shows that in many cases these responses are inappropriate given the number of pieces that were eaten. As can be seen in Table 1, a large percentage of diners apparently responded to questions that they shouldn't have. For example, there are 15 diners who reported eating 0 slices, and yet 5 of them responded to at least one question about the taste, satisfaction, or enjoyment of the pizza. Perhaps they tasted with their noses. Even more concerning, out of the 40 diners who ate 2 pieces of pizza, half of them responded to questions about the middle piece of pizza. These diners don't have a middle piece, only diners who ate 3 or more pieces should have a middle piece. If half of the diners are providing clearly impossible responses it raises questions about whether the diners understood the survey or took it seriously, and calls into question the validity of all the responses, even those that are not logically impossible.

Another concern is the four diners who did not provide a response for how many pieces of pizza they ate, but yet provided responses about the overall quality of the pizza. It is unclear to us what the appropriate action is here, but we note that two of these diners gave the same rating for every question, suggesting that they were either rushed or did not take the survey seriously. Despite this, these responses were included in the original analyses and even in some of the reanalyses by the authors (see below).

Table 1. Number of diners with impossible responses

Pieces eaten[3]	0	1	2
Total number of diners	15	40	40
Problem diners	5	10	20
Ratio problem/total	33%	25%	50%

[3]Because of the 3 fractional responses these are technically the diners who ate 0<=pieces<1, 1<=pieces<2, and 2<=pieces<3.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Impossible calorie consumption

Another set of responses to the questionnaire that can easily be checked for logically impossible responses is the number of calories that diners estimated that they had consumed. For example, 11 diners reported eating either 0 or 1 calories, and of these 11 diners only 1 reported eating 0 pizza and 0 salad. There were other dining options besides pizza and salad which weren't recorded, but maybe this patron really ate nothing at an all-you-can-eat buffet. The other responses are difficult to explain, since most of them include clear indications that pizza and/or salad were consumed. Again, we see here how the presence of strange responses raises questions about the relevance of the data set.

A POSTMORTEM OF THE STATISTICS

With the data release, Response Tables, and STATA scripts, we are now able to understand the basis for many of the impossible values we detected in the publications.

Response Tables

Here we will go over each Response Table in detail. However, we should state from the start that it is unclear to us exactly how to interpret the Response Tables. Are they meant to show the reason for the inconsistencies or are they meant to show the correct statistics? These two possibilities may seem the same, but they are not. For most original statistics, to explain the problems that we detected the authors simply needed to provide the correct cell sizes or fix any typos or rounding errors. The provided STATA scripts correctly exclude impossible responses (most of the time) that were included in the original analyses, but the results from the STATA scripts are only sometimes included in the Response Tables. It is thus unclear what the authors intend to send to the journals as errata.

The Response Tables provided by the authors note "supplements/corrections appear in blue bold print". However, sometimes numbers are changed which are neither blue nor bold, and sometimes numbers are bold, but are not blue. To make things as clear as possible we have provided 3 different tables. The "Original Table" is what appeared in our critique, where red indicated numbers which were mathematically impossible. The "Response Table" is the table from the authors' response, with numbers that were either blue and bold, or just bold, in blue bold print. For numbers which changed but were not colored blue or bold, we changed the color to purple to highlight the apparent oversight. Numbers colored purple may or may not be mathematically possible. When a blue or bold number was mathematically impossible we colored those red. We then present our responses to the Response Tables, which we refer to as "Correct Tables". These tables remove impossible responses in the analyses, like the STATA scripts, and should be seen as a gold standard for this data set. Any numbers which needed to be corrected relative to the Response Table are shown in gold and bold.

To make matters even worse there are a few cases where a number is blue and bold, but did not change from the original table. For simplicity, we decided to leave these numbers unchanged (i.e., keeping the blue and bold format) instead of creating yet another category.

Of course our code to reproduce these analyses is publicly available. It is perhaps instructive to show how these types of analyses can be done with a free, open source language instead of proprietary statistical software. The results were checked against the STATA output (which was not provided by the authors but was kindly provided to us by colleagues with access to STATA).

Article 1, Table 1

As can be seen in the original table there are mathematically impossible means, percents, and test statistics. The multiple impossible means along with the incorrect test statistics in rows without granularity problems indicates extensive problems with sample sizes.

The data release and Response Table confirms the issues with the cell sizes, and most of the values can be reproduced with the released data. Although there are missing responses, surprisingly the cell sizes are actually larger than originally reported. Article 1 clearly states that 122 diners were included in the study, with 6 diners removed because of their age and

[4]Table presented as in the original publication. Numbers in red are inconsistent values. This note applies to all "Original" tables.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Original[4] Table 1

	$4 buffet (N = 62)	$8 buffet (N = 60)	*F* test (*p* value)
Age	44.16 (18.99)	46.08 (14.46)	0.42 (0.52)
Gender (male percent)	57.4	47.9	
Height	68.52 (3.95)	67.91 (3.93)	0.76 (0.37)
Weight	180.84 (48.37)	182.31 (48.41)	0.03 (0.87)
Number in group	3.00 (1.55)	3.28 (1.29)	1.34 (0.25)
I was hungry when I came in	6.62 (1.85)	6.64 (2.06)	0.00 (0.95)
I am hungry now	1.88 (1.34)	1.85 (1.75)	0.01 (0.91)

11 others removed because of nonresponses. The sample sizes of the original Table 1 are consistent with this statement. Despite this, data from all 139 diners is used in this table.

Interestingly, the authors did not change any of the statistics, except for a furtive change of the Age SD. We cannot check the Age, Height, and Weight means and SDs since demographic data were not released. However, with the new cell sizes we could check the test statistics and found the Height *p* value to be incorrect.

Response[5] Table 1

	$4 buffet	$8 buffet	*F* test (*p* value)
Age	44.16 (19.00)	46.08 (14.46)	0.42 (0.52)
N	**64**	**65**	
Gender (male percent)	**60**	**51.5**	
N	**65**	**68**	
Height	68.52 (3.95)	67.91 (3.93)	0.76 (0.37)
N	**64**	**63**	
Weight	180.84 (48.37)	182.31 (48.41)	0.03 (0.87)
N	**62**	**54**	
Number in group	3.00 (1.55)	3.28 (1.29)	1.34 (0.25)
N	**65**	**68**	
I was hungry when I came in	6.62 (1.85)	6.64 (2.06)	0.00 (0.95)
N	**66**	**70**	
I am hungry now	1.88 (1.34)	1.85 (1.75)	0.01 (0.91)
N	**67**	**66**	

The data release is consistent with the original statistics and the Response Table's new cell sizes, with the exception of the mean and SD for "Number in group", $4, and the corresponding test statistic for that row. Our code, along with the released STATA code, identifies a problem with these values. Why the Response Table did not reflect the output of the STATA code is unclear.

Correct[6] Table 1

	$4 buffet	$8 buffet	*F* test (*p* value)
Age	44.16 (19.00)	46.08 (14.46)	0.42 (0.52)
N	64	65	
Gender (male percent)	60.0	51.5	
N	65	68	
Height	68.52 (3.95)	67.91 (3.93)	0.76 (**0.38-0.39**)
N	64	63	
Weight	180.84 (48.37)	182.31 (48.41)	0.03 (0.87)
N	62	54	
Number in group	**3.03 (1.52)**	3.28 (1.29)	**1.04 (0.31)**
N	65	68	
I was hungry when I came in	6.62 (1.85)	6.64 (2.06)	0.00 (0.95)
N	66	70	
I am hungry now	1.88 (1.34)	1.85 (1.75)	0.01 (0.91)
N	67	66	

Article 1, Table 2

This table had several serious problems. First, the table uses the terms "first", "middle", "last", but those terms were not defined in Article 1. Second, readers of Article 1 would be forgiven if they assumed the sample sizes at the top of the table

[5]Corrected table from the lab's reanalysis. Numbers in blue are values corrected by the lab, and in purple are numbers which were also changed by the lab but without them reporting as such. This note applies to all "Response" tables.

[6]Corrected table based on our reanalysis. Numbers in **gold** are values which have been corrected based on our reanalysis. This note applies to all "Correct" tables.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

were applicable for every row. However, from Article 2 it was clear that for the $4 buffet the cell sizes were 62, 41, 47, and for the $8 buffet the cell sizes were 60, 26, 38. Even using these inferred cell sizes there were still numerous granularity problems and issues with the test statistics, indicating that we still did not have the correct cell sizes.

Original Table 2

	$4 buffet (N = 62)	$8 buffet (N = 60)	*F* test (*p* value)
The pizza, in general, tasted really great	6.89 (1.39)	7.44 (1.60)	4.24 (0.04)
The first piece of pizza I ate tasted really great	7.08 (1.30)	7.45 (1.60)	1.97 (0.16)
The first piece of pizza I ate was very satisfying	7.08 (1.37)	7.34 (1.70)	0.82 (0.37)
The first piece of pizza I ate was very enjoyable	7.05 (1.40)	7.47 (1.55)	2.40 (0.12)
The middle piece of pizza I ate tasted really great	6.68 (1.49)	7.97 (1.21)	15.42 (0.00)
The middle piece of pizza I ate was very satisfying	6.68 (1.49)	7.97 (1.21)	14.69 (0.00)
The middle piece of pizza I ate was very enjoyable	6.64 (1.48)	7.81 (1.22)	12.48 (0.00)
The last piece of pizza I ate tasted really great	6.15 (1.89)	7.58 (1.39)	15.16 (0.00)
The last piece of pizza I ate was very satisfying	6.16 (1.87)	7.41 (1.55)	10.99 (0.00)
The last piece of pizza I ate was very enjoyable	5.98 (1.86)	7.45 (1.52)	15.60 (0.00)

The data release and Response Table again showed an interesting mix of cell sizes that were either larger or smaller than those originally reported. The data release is consistent with the original statistics and Response Table cell sizes, with the exception of a row which was surreptitiously corrected, and incorrectly corrected at that (SD should be 1.10 not 1.11). The curious thing about the Response Table is that it is not consistent with the STATA output. The STATA code correctly removes impossible responses, for example a response for a middle piece when the diner only ate 1 piece of pizza. However, even the STATA code is not completely correct. The STATA code does not remove diners who did not report how many pieces of pizza they ate. It is unclear what to do with these diners, but the STATA code removes these diners in the Table 3 analysis. Altering inclusion criteria depending on the analysis is an example of a researcher degree of freedom (Simmons et al., 2011), can be considered a form of *p*-hacking, and should be avoided.

Response Table 2

	$4 buffet (N = 62)	$8 buffet (N = 60)	*F* test (*p* value)
The pizza, in general, tasted really great	6.89 (1.39)	7.44 (1.60)	4.24 (0.04)
N	**63**	**61**	
The first piece of pizza I ate tasted really great	7.08 (1.30)	7.45 (1.60)	1.97 (0.16)
N	**62**	**60**	
The first piece of pizza I ate was very satisfying	7.08 (1.37)	7.34 (1.70)	0.82 (0.37)
N	**60**	**59**	
The first piece of pizza I ate was very enjoyable	7.05 (1.40)	7.47 (1.55)	2.40 (0.12)
N	**60**	**60**	
The middle piece of pizza I ate tasted really great	6.72 (1.50)	8.00 (1.11)	15.42 (0.00)
N	**43**	**29**	
The middle piece of pizza I ate was very satisfying	6.68 (1.49)	7.97 (1.21)	14.69 (0.00)
N	**40**	**29**	
The middle piece of pizza I ate was very enjoyable	6.64 (1.48)	7.81 (1.22)	12.48 (0.00)
N	**39**	**31**	
The last piece of pizza I ate tasted really great	6.15 (1.89)	7.58 (1.39)	15.16 (0.00)
N	**47**	**38**	
The last piece of pizza I ate was very satisfying	6.16 (1.87)	7.41 (1.55)	10.99 (0.00)
N	**45**	**39**	
The last piece of pizza I ate was very enjoyable	5.98 (1.86)	7.45 (1.52)	15.60 (0.00)
N	**44**	**40**	

Our analysis of the data removes impossible responses and diners who did not report how much pizza they ate, and as a result our values differ from the original tables and the STATA output.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Correct Table 2

	$4 buffet	$8 buffet	*F* test (*p* value)
The pizza, in general, tasted really great	6.92 (1.37)	7.55 (1.42)	6.11 (0.01)
N	59	58	
The first piece of pizza I ate tasted really great	7.05 (1.31)	7.56 (1.44)	4.01 (0.05)
N	59	57	
The first piece of pizza I ate was very satisfying	7.09 (1.35)	7.45 (1.49)	1.80 (0.18)
N	57	56	
The first piece of pizza I ate was very enjoyable	7.09 (1.30)	7.51 (1.50)	2.56 (0.11)
N	57	57	
The middle piece of pizza I ate tasted really great	6.81 (1.57)	8.00 (1.15)	6.51 (0.02)
N	21	16	
The middle piece of pizza I ate was very satisfying	6.84 (1.61)	7.88 (1.20)	4.48 (0.04)
N	19	16	
The middle piece of pizza I ate was very enjoyable	6.74 (1.59)	7.82 (1.19)	5.29 (0.03)
N	19	17	
The last piece of pizza I ate tasted really great	6.11 (1.98)	7.56 (1.40)	12.34 (0.00)
N	36	34	
The last piece of pizza I ate was very satisfying	6.09 (1.96)	7.47 (1.48)	10.76 (0.00)
N	34	34	
The last piece of pizza I ate was very enjoyable	5.88 (1.93)	7.36 (1.55)	12.44 (0.00)
N	33	36	

Article 3, Table 1

For this table we did not check the statistics for granularity issues given that we were unsure if the data for Age, Height, and Weight were whole numbers. However, we were able to identify that every single *t* statistic was wrong, indicating issues with the means, SDs, and/or sample sizes.

Original Table 1

	Males eating with females (N = 40)	Males eating with males (N = 20)	*t*	Females eating with males (N = 35)	Females eating with females (N = 10)	*t*
Age (years)	44 (18.86)	43 (11.19)	0.42	44.52 (17.09)	48.18 (16.49)	0.64
Height (cm)	178.02 (7.72)	181.11 (7.32)	1.59	165.83 (7.71)	164.82 (5.88)	0.37
Weight (kg)	86.35 (17.92)	100.80 (21.33)	2.87	64.63 (10.95)	75.54 (12.42)	2.38
BMI	27.20 (5.13)	30.96 (6.62)	2.52	23.46 (3.53)	27.77 (3.68)	2.96

The Response Table again shows a combination of larger sample sizes than those that were originally reported along with some nonresponses. The demographic data were not released so we cannot check the cell sizes, means, or SDs, but we can check the *t* statistics. Interestingly, 2 of the *t* statistics are clearly wrong. Also interesting is the fact that the STATA script performs Welch's *t* test while the statistics in the table are consistent with a Student's *t* test. As a result, it is unclear what code was used to reproduce these values, and the source of the errors.

Response Table 1

	Males eating with females (**N = 46**)	Males eating with males (**N = 19**)	*t*	Females eating with males (**N = 41**)	Females eating with females (**N=12**)	*t*
Age (years)	**45.22 (18.72)**	**43.47 (12.95)**	**0.37**	**43.68 (16.49)**	48.18 (16.49)	**0.80**
N	**45**	**19**		**40**	**11**	
Height (cm)	**177.63 (7.90)**	**181.74 (6.71)**	**1.99**	**165.84 (7.26)**	**164.68 (5.96)**	**0.45**
N	**46**	**19**		**41**	**9**	
Weight (kg)	**87.09 (16.88)**	**98.51 (22.23)**	2.75	**64.31 (10.56)**	**76.14 (12.52)**	**2.63**
N	**45**	**18**		**35**	**7**	
BMI	**27.62 (5.20)**	**30.00 (6.40)**	2.13	**23.37 (3.64)**	**28.00 (3.71)**	**3.06**
N	**45**	**18**		**35**	**7**	

Using the provided means, SDs, and cell sizes we provided possible values that the incorrect *t* statistics can take (Student's *t* test was used).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Correct Table 1

	Males eating with females (N = 46)	Males eating with males (N = 19)	*t*	Females eating with males (N = 41)	Females eating with females (N=12)	*t*
Age (years)	45.22 (18.72)	43.47 (12.95)	0.37	43.68 (16.49)	48.18 (16.49)	0.80
N	45	19		40	11	
Height (cm)	177.63 (7.90)	181.74 (6.71)	1.99	165.84 (7.26)	164.68 (5.96)	0.45
N	46	19		41	9	
Weight (kg)	87.09 (16.88)	98.51 (22.23)	2.21	64.31 (10.56)	76.14 (12.52)	2.63
N	45	18		35	7	
BMI	27.62 (5.20)	30.00 (6.40)	1.53-1.54	23.37 (3.64)	28.00 (3.71)	3.06
N	45	18		35	7	

Article 3, Table 2

This table also had several granularity inconsistencies as well as test statistic errors indicating serious problems.

Original Table 2

	Males eating with females (N = 40)	Males eating with males (N = 20)	Females eating with males (N = 35)	Females eating with females (N = 10)	*F* test Effect of gender	*F* test Effect of group type	*F* test Effect of gender×group
Salad consumed	5.00 (2.99)	2.69 (2.57)	4.83 (2.71)	5.54 (1.84)	3.84	1.36	4.83
Pizza slices consumed	2.99 (1.75)	1.55 (1.07)	1.33 (0.83)	1.05 (1.38)	14.58	9.26	4.22
I overate	2.67 (2.04)	2.76 (2.18)	2.73 (2.16)	1.00 (0.00)	3.57	3.33	4.15
I felt rushed	1.46 (1.07)	1.90 (1.48)	2.29 (2.28)	1.18 (0.40)	0.02	0.83	4.53
How many calories of pizza you think you ate?	478.75 (290.67)	397.50 (191.37)	463.61 (264.25)	111.71 (109.57)	5.01	10.39	4.05
I am physically uncomfortable	2.11 (1.54)	2.27 (1.75)	2.20 (1.71)	1.91 (2.12)	0.15	0.03	0.39

Our analysis of the data release did not reproduce the original table. Interestingly, unlike the Response Table for Article 1, Table 2, where the authors simply updated the cell sizes instead of updating the values with the output of the STATA code, in this Response Table the authors did use the STATA output. Perhaps the difference is that the values in Article 1, Table 2, can be reproduced if the impossible responses are included, whereas here it is not clear what has to be done to reproduce the original values, and perhaps the authors also did not know how to reproduce them.

Response Table 2

	Males eating with females (**N = 46**)	Males eating with males (**N = 19**)	Females eating with males (**N = 41**)	Females eating with females (**N = 12**)	*F* test Effect of gender	*F* test Effect of group type	*F* test Effect of gender×group
Salad consumed	**5.27 (3.07)**	**2.44 (2.61)**	**5.23 (2.84)**	5.54 (1.84)	**4.41**	**2.98**	**4.64**
N	**40**	**16**	**33**	**7**			
Pizza slices consumed	**2.89 (1.77)**	**1.37 (1.21)**	**1.54 (0.88)**	**1.25 (0.87)**	**6.43**	**9.87**	**4.52**
N	**46**	**19**	**39**	**12**			
I overate	**3.13 (2.51)**	**2.95 (2.57)**	**2.74 (2.19)**	**1.36 (1.21)**	**3.78**	**2.38**	**1.38**
N	**45**	**19**	**39**	**11**			
I felt rushed	**1.87 (1.67)**	**2.47 (2.22)**	**2.23 (2.31)**	1.18 (0.40)	1.19	**0.27**	**3.78**
N	**45**	**19**	**39**	**11**			
How many calories of pizza you think you ate?	**458.33 (307.25)**	**291.33 (226.05)**	**444.00 (279.94)**	**142.44 (168.37)**	**1.50**	**12.37**	**1.02**
N	**42**	**15**	**35**	**9**			
I am physically uncomfortable	2.15 (1.54)	**2.47 (2.32)**	**2.28 (1.77)**	1.91 (2.12)	**0.31**	**0.00**	**0.74**
N	**45**	**19**	**40**	**11**			

Because the original values cannot be reproduced it is impossible to determine the cause of the granularity inconsistencies. They could be due to sample size issues, miscalculations, typos, etc. In the Response Table, one of the means from the STATA output has been incorrectly rounded and one of the changes was not colored blue.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Correct Table 2

	Males eating with females (N = 46)	Males eating with males (N = 19)	Females eating with males (N = 41)	Females eating with females (N = 12)	*F* test Effect of gender	*F* test Effect of group type	*F* test Effect of gender×group
Salad consumed	5.27 (3.07)	2.44 (2.61)	5.23 (2.84)	5.54 (1.84)	4.41	2.98	4.64
N	40	16	33	7			
Pizza slices consumed	2.89 (1.77)	1.37 (1.21)	1.54 (0.88)	1.25 (0.87)	6.43	9.87	4.52
N	46	19	39	12			
I overate	3.13 (2.51)	2.95 (2.57)	2.74 (2.19)	1.36 (1.21)	3.78	2.38	1.38
N	45	19	39	11			
I felt rushed	1.87 (1.67)	2.47 (2.22)	2.23 (2.31)	1.18 (0.40)	1.19	0.27	3.78
N	45	19	39	11			
How many calories of pizza you think you ate?	458.33 (307.25)	291.33 (226.05)	444.00 (279.94)	142.44 (168.37)	1.50	12.37	1.02
N	42	15	35	9			
I am physically uncomfortable	2.16 (1.54)	2.47 (2.32)	2.28 (1.77)	1.91 (2.12)	0.31	0.00	0.74
N	45	19	40	11			

Article 3, Table 3

This table had granularity inconsistencies and impossible *F* statistic values from the ANOVAs.

Original Table 3

	Only-male groups (*N* = 20)	Only one male in mixed-sex groups (*N* = 21)	More than one male in mixed-sex groups (*N* = 19)	*F* test
Salad consumed	2.69 (2.57)	5.55 (2.66)	4.33 (3.31)	5.16
Pizza slices consumed	1.55 (1.07)	2.79 (1.54)	3.13 (2.18)	4.89
I overate	2.76 (2.19)	2.92 (2.30)	2.53 (1.81)	0.18
I felt rushed	1.90 (1.48)	1.65 (1.34)	1.47 (1.23)	0.49
How many calories of pizza you think you ate?	397.50 (191.38)	409.52 (246.87)	555.26 (321.84)	0.15
I am physically uncomfortable	2.27 (1.75)	2.32 (1.77)	1.95 (1.24)	0.72

Like the previous table, we do not know how to reproduce the statistics given the data. As a result, it is again unclear what the original problems were. Also like the previous table, the authors used the STATA output in the Response Table, but rounded a value incorrectly.

Response Table 3

	Only-male groups (**N = 19**)	Only one male in mixed-sex groups (**N = 23**)	More than one male in mixed-sex groups (**N = 23**)	*F* test
Salad consumed	**2.44 (2.61)**	5.72 (**3.21**)	**4.86 (2.96)**	**5.66**
N	**16**	**19**	**21**	
Pizza slices consumed	**1.37 (1.21)**	**2.91 (1.65)**	**2.87 (1.91)**	**5.80**
N	**19**	**23**	**23**	
I overate	**2.95 (2.57)**	**3.32 (2.77)**	**2.96 (2.29)**	**0.15**
N	**19**	**22**	**23**	
I felt rushed	**2.47 (2.22)**	**2.00 (1.88)**	**1.74 (1.48)**	**0.82**
N	**19**	**22**	**23**	
How many calories of pizza you think you ate?	**291.33 (226.05)**	**384.21 (306.51)**	**519.57 (300.66)**	**3.06**
N	**15**	**19**	**23**	
I am physically uncomfortable	**2.47 (2.32)**	**2.32 (1.86)**	**2.00 (1.17)**	**0.38**
N	**19**	**22**	**23**	

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Correct Table 3

	Only-male groups (N = 19)	Only one male in mixed-sex groups (N = 23)	More than one male in mixed-sex groups (N = 23)	*F* test
Salad consumed	2.44 (2.61)	**5.73** (3.21)	4.86 (2.96)	5.66
N	16	19	21	
Pizza slices consumed	1.37 (1.21)	2.91 (1.65)	2.87 (1.91)	5.80
N	19	23	23	
I overate	2.95 (2.57)	3.32 (2.77)	2.96 (2.29)	0.15
N	19	22	23	
I felt rushed	2.47 (2.22)	2.00 (1.88)	1.74 (1.48)	0.82
N	19	22	23	
How many calories of pizza you think you ate?	291.33 (226.05)	384.21 (306.51)	519.57 (300.66)	3.06
N	15	19	23	
I am physically uncomfortable	2.47 (2.32)	2.32 (1.86)	2.00 (1.17)	0.38
N	19	22	23	

Article 4, Table 1

Again, we previously did not check Age, Height, and Weight for granularity inconsistencies, but did find all of the *t* statistics to be incorrect, suggesting the existence of genuine problems.

Original Table 1

Demographics	$4 (n = 43)	$8 (n = 52)	*t*
Age (years)	43.67 (18.50)	44.55 (14.30)	0.25
Height (inches)	68.65 (3.67)	66.51 (9.44)	1.38
Weight (pounds)	184.83 (63.70)	178.38 (45.71)	0.52

We cannot confirm the cell sizes and summary statistics of the Response Table, but we checked the *t* statistics, and found one of them to be wrong. While the STATA code for Article 3 performed a Welch's *t* test (but the table included Student's *t* test values), the STATA code for Article 4 appears to do a Student's *t* test. The reason for these inconsistencies is unclear.

Response Table 1

Demographics	$4 (n = 43)	$8 (n = 52)	*t*
Age (years)	43.67 (18.50)	44.55 (14.30)	**0.26**
N	**42**	**49**	
Height (inches)	68.65 (3.67)	**67.76** (**3.87**)	1.12
N	**42**	**42**	
Weight (pounds)	**178.20** (**48.11**)	178.38 (45.71)	**0.02**
N	**40**	**40**	

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Correct Table 1

Demographics	$4 (n = 43)	$8 (n = 52)	*t*
Age (years)	43.67 (18.50)	44.55 (14.30)	0.26
N	42	49	
Height (inches)	68.65 (3.67)	67.76 (3.87)	**1.07-1.10**
N	42	42	
Weight (pounds)	178.20 (48.11)	178.38 (45.71)	0.02
N	40	40	

Article 4, Table 2

This table is a good illustration of the limitations of granularity testing. Granularity testing revealed multiple issues, but all of the granularity inconsistencies are explained by incorrect sample sizes or incorrect rounding. The values with the largest problems were not detected as inconsistent by granularity testing. This can happen because even randomly generated numbers have a certain chance of being consistent. This underscores an important point: **granularity inconsistencies represent a lower bound**. For example, if the sample size is 50, the data are reported to two decimals, and half of the values are inconsistent, it is quite reasonable to assume that many more of the values (perhaps all of them) are wrong, since 50% of randomly-generated means will pass the test by chance.

Original Table 2

	$4 (Discounted-price)			$8 (Full-price)			*F* statistics		
	One piece (N = 18)	Two pieces (N = 18)	Three pieces (N = 7)	One piece (N = 17)	Two pieces (N = 19)	Three pieces (N = 10)	Effect of price	Effect of pieces	Effect of price×pieces
I ate more pizza than I should have	2.63 (2.06)	4.82 (2.55)	6.00 (2.00)	1.76 (1.82)	3.53 (2.39)	4.40 (3.24)	5.37	10.77	0.15
I feel guilty about how much I ate	2.39 (1.94)	3.44 (2.47)	3.71 (1.49)	2.26 (1.79)	1.68 (1.42)	2.90 (2.08)	4.28	1.49	1.67
I am physically uncomfortable	2.17 (1.88)	2.94 (2.12)	2.43 (1.51)	1.97 (1.68)	1.45 (0.94)	2.25 (1.81)	4.19	0.25	1.15
I overate	2.11 (1.81)	3.89 (2.59)	3.71 (1.79)	1.67 (1.28)	1.67 (1.24)	3.50 (2.74)	5.02	4.09	2.27
I ate more than I should have	2.50 (2.20)	4.28 (2.44)	4.57 (2.22)	2.00 (1.45)	2.14 (1.77)	3.92 (2.81)	6.20	5.00	1.14

The Response Table correctly changed most of the values, furtively changed a few values, and changed a value incorrectly. Most of the values in the original table are consistent with the data set, except for a few values in the $8 group which have no obvious explanation.

Response Table 2

	$4 (Discounted-price)			$8 (Full-price)			*F* statistics		
	One piece (N = 18)	Two pieces (N = 18)	Three pieces (N = 7)	One piece (**N = 19**)	Two pieces (**N = 21**)	Three pieces (**N = 12**)	Effect of price	Effect of pieces	Effect of price×pieces
I ate more pizza than I should have	2.63 (2.06)	4.82 (2.55)	6.00 (2.00)	1.76 (1.82)	**4.05** (1.82)	**4.92** (**3.23**)	**2.65**	**12.08**	**0.02**
I feel guilty about how much I ate	2.39 (1.94)	3.44 (2.47)	3.71 (1.49)	2.26 (1.79)	**2.19** (**2.18**)	**3.33** (**2.39**)	**1.59**	**1.95**	**0.72**
I am physically uncomfortable	2.17 (1.88)	2.94 (2.12)	2.43 (1.51)	1.95 (1.68)	1.45 (0.94)	2.25 (**1.82**)	**2.81**	**0.17**	**1.60**
I overate	2.11 (1.81)	3.89 (2.59)	3.71 (1.79)	1.67 (1.28)	1.67 (1.24)	3.50 (**2.75**)	**5.01**	**4.97**	**2.59**
I ate more than I should have	2.50 (2.20)	4.28 (2.44)	4.57 (2.23)	2.00 (1.45)	2.14 (1.77)	3.92 (2.81)	**5.49**	**5.52**	**1.59**

We also checked the the ANOVA values (*F* statistics) with `rpsychi`, which should not have been used for factors with more than two levels that are unbalanced. However, `rpsychi` is accurate for the "Effect of price" as that only has two levels, and given the problems with that factor it could be argued that problems with the other factor and the interaction were likely. Indeed, every single *F* statistic was incorrect. It is unclear how the original ANOVAs were calculated.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Correct Table 2

	$4 (Discounted-price)			$8 (Full-price)			*F* statistics		
	One piece (N = 18)	Two pieces (N = 18)	Three pieces (N = 7)	One piece (N = 19)	Two pieces (N = 21)	Three pieces (N = 12)	Effect of price	Effect of pieces	Effect of price×pieces
I ate more pizza than I should have	2.63 (2.06)	4.82 (**2.56**)	6.00 (2.00)	1.76 (1.82)	4.05 (**2.80**)	4.92 (3.23)	2.65	12.08	0.02
N	**16**	**17**	**7**	**17**	**21**	**12**			
I feel guilty about how much I ate	2.39 (1.94)	3.44 (**2.48**)	3.71 (**1.50**)	2.26 (1.79)	2.19 (2.18)	3.33 (2.39)	1.59	1.95	0.72
N	**18**	**18**	**7**	**19**	**21**	**12**			
I am physically uncomfortable	2.17 (**1.89**)	2.94 (**2.13**)	2.43 (1.51)	1.95 (1.68)	1.45 (0.94)	2.25 (1.82)	2.81	0.17	1.60
N	**18**	**18**	**7**	**19**	**20**	**12**			
I overate	2.11 (1.81)	3.89 (2.59)	3.71 (**1.80**)	1.67 (1.28)	1.67 (1.24)	3.50 (2.75)	5.01	4.97	2.59
N	**18**	**18**	**7**	**18**	**21**	**12**			
I ate more than I should have	2.50 (2.20)	4.28 (2.44)	4.57 (2.23)	2.00 (1.45)	2.14 (1.77)	3.92 (2.81)	5.49	5.52	1.59
N	**18**	**18**	**7**	**19**	**21**	**12**			

Article 4, Table 3

The original Table 3 ought to have shown the same summary statistics as the original Table 2, but remarkably it did not. As a result, Table 3 shared some of the same granularity errors as Table 2 while also having its own unique errors. Many of the test statistics were also wrong.

Original Table 3

	1 Piece			2 Pieces			3 Pieces		
	$4 (N = 18)	$8 (N = 19)	*F* test	$4 (N = 18)	$8 (N = 21)	*F* test	$4 (N = 7)	$8 (N = 12)	*F* test
I ate more pizza than I should have	2.63 (2.06)	1.76 (1.82)	1.62	4.82 (2.55)	3.53 (2.39)	2.47	6.00 (2.00)	4.40 (3.24)	1.34
I feel guilty about how much I ate	2.39 (1.94)	2.26 (1.79)	0.04	3.44 (2.48)	1.68 (1.42)	7.13	3.71 (1.50)	2.90 (2.08)	0.78
I am physically uncomfortable	2.17 (1.89)	1.955 (1.68)	0.14	2.94 (2.13)	1.28 (0.46)	8.11	2.43 (1.51)	2.10 (1.91)	0.14
I overate	2.11 (1.81)	1.67 (1.28)	0.72	3.89 (2.59)	1.53 (1.02)	1.63	3.71 (1.79)	3.50 (2.95)	0.03
I ate more than I should have	2.50 (2.20)	2.00 (1.45)	0.67	4.28 (2.44)	2.05 (1.72)	10.36	4.57 (2.23)	4.00 (3.02)	0.18

Amusingly, the Response Table changes several values which did not need to be changed, thus introducing errors instead of fixing them. The Response Table correctly fixes the test statistics, but it is unclear how to reproduce the original test statistics.

Response Table 3

	1 Piece			2 Pieces			3 Pieces		
	$4 (N = 18)	$8 (N = 19)	*F* test	$4 (N = 18)	$8 (N = 21)	*F* test	$4 (N = 7)	$8 (N = 12)	*F* test
I ate more pizza than I should have	2.63 (2.06)	1.76 (1.82)	1.62	4.82 (2.55)	**4.05** (1.82)	**0.78**	6.00 (2.00)	**4.92** (**3.23**)	**0.63**
I feel guilty about how much I ate	2.39 (1.94)	2.26 (1.79)	0.04	3.44 (2.47)	**2.19** (**2.18**)	**2.82**	3.71 (1.49)	**3.33** (**2.39**)	**0.14**
I am physically uncomfortable	2.17 (1.88)	1.95 (1.68)	0.14	2.94 (2.12)	**1.45** (**0.94**)	**8.11**	2.43 (1.51)	**2.25** (**1.82**)	**0.05**
I overate	2.11 (1.81)	1.67 (1.28)	0.72	3.89 (2.59)	**1.67** (**1.24**)	**12.26**	3.71 (1.79)	**3.50** (**2.75**)	**0.03**
I ate more than I should have	2.50 (2.20)	2.00 (1.45)	0.67	4.28 (2.44)	**2.14** (**1.77**)	**9.96**	4.57 (2.23)	**3.92** (**2.81**)	**0.28**

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Correct Table 3

	1 Piece			2 Pieces			3 Pieces		
	$4 (N = 18)	$8 (N = 19)	*F* test	$4 (N = 18)	$8 (N = 21)	*F* test	$4 (N = 7)	$8 (N = 12)	*F* test
I ate more pizza than I should have	2.63 (2.06)	1.76 (1.82)	1.62	4.82 (2.56)	4.05 (2.80)	0.78	6.00 (2.00)	4.92 (3.23)	0.63
N	16	17		17	21		7	12	
I feel guilty about how much I ate	2.39 (1.94)	2.26 (1.79)	0.04	3.44 (2.48)	2.19 (2.18)	2.82	3.71 (1.50)	3.33 (2.39)	0.14
N	18	19		18	21		7	12	
I am physically uncomfortable	2.17 (1.89)	1.95 (1.68)	0.14	2.94 (2.13)	1.45 (0.94)	8.11	2.43 (1.51)	2.25 (1.82)	0.05
N	18	19		18	20		7	12	
I overate	2.11 (1.81)	1.67 (1.28)	0.72	3.89 (2.59)	1.67 (1.24)	12.26	3.71 (1.80)	3.50 (2.75)	0.03
N	18	18		18	21		7	12	
I ate more than I should have	2.50 (2.20)	2.00 (1.45)	0.67	4.28 (2.44)	2.14 (1.77)	9.96	4.57 (2.23)	3.92 (2.81)	0.28
N	18	19		18	21		7	12	

List of Inconsistencies

In addition to providing Response Tables, the authors provided a point by point response of the inconsistencies we detected. While we appreciate the effort of the authors to thoroughly respond to our findings, we do not find the explanations completely accurate or transparent. For example, while it is true that most of the granularity inconsistencies are due to cell sizes that differ from those stated in the articles, and it is true that there are missing responses, the actual cell sizes are often **larger** than what was stated in the articles (thus excluding "missing responses" as the sole cause). Furthermore, the authors' responses neglect to point out that the original statistics included data which they shouldn't have. Of course we did not point this out in our original critique as we had no way of knowing about these problems, but it seems disingenuous to simply label all of the problems as due to nonresponses when in fact the problems go much deeper.

The list also states whether each inconsistency is "unique and valid". Each inconsistency is obviously unique—we wouldn't list an inconsistency twice—but there is some value in considering whether the *source* of the inconsistency is unique. For example, if there is a typo for the sample size, that would cause numerous downstream problems, but would only need one correction to fix all of the problems. However, we did not find that to be the case for this data set.

It is also unclear what "valid" means. Of course, if we made a mistake (such as the BMI calculations), then the inconsistency is not valid, but if the reported number is in fact mathematically impossible it is hard to understand how it is not "valid".

Article 1, Table 1, Granularity inconsistencies

Response	Unique and Valid
"This is due to non-response by some participants"	"No"

There are missing responses, but the correct cell sizes are actually larger than what was reported in the article, which makes it mathematically impossible for missing responses to be the sole explanation. A careful inspection actually reveals that data for every single diner is included in this table, which directly contradicts the article which states that 11 diners were removed because of nonresponses, and 6 diners were removed because of age. It is difficult to understand how an inaccuracy as large as this, with an accompanying false statement in the text, is not "unique and valid".

Article 1, Table 1, Test statistics

Response	Unique and Valid
"This statistic [1.34] is revised in our reanalysis due to an error found in the classification of one group. The reported value reflects the statistic produced when this error is uncorrected. The inconsistency is due to misreporting the number of respondents in the table (n=133 versus n = 122 reported in the paper)."	"Yes"

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

This is an example of a more transparent response to one of our findings. Why this type of response is not included for the granularity inconsistencies in this same table is unclear. However, it is also unclear what the original error was, as it is hard to see what "one group" will cause the sample size to go from 133 to 122; indeed, it is also unclear how the authors determined what the original problem was. We speculate that they might have had access to the original analysis scripts.

Article 1, Table 2, Granularity inconsistencies

Response	Unique and Valid
"This is due to non-response by some participants"	"No"

Most of the granularity inconsistencies are due to cell sizes being different from what was reported, which is a result of missing responses in combination with more participants than originally reported. However, the statistics for "The middle piece of pizza I ate tasted really great" cannot be reproduced from the data. We speculate that the statistics for the row below were accidentally duplicated. How these typos are not unique and valid is unclear.

Article 2, Issues with the regression models

Issue	Response	Unique and Valid
"In the regression models in Article 2, the dependent variable (Overall evaluation of all slice consumed) seem to be conceptually indisguisable[sic] from the predictors (individual slices)"	"This is not and[sic] inconsistency. Moreover, this is a common issue in the study of how the evaluation of components of an experience translate into overall evaluations."	"No"
"Acute problems with multicollinearity conflicting to repeated-measures ANOVAs"	"This is not an inconsistency"	"No"

In Article 2 the authors constructed various linear models to predict how diners rated the taste of the pizza using ratings such as how the pizza tasted. The authors are correct that the multicollinearity issue is not an "inconsistency" as the authors did not make a mistake and fully intended to perform the regressions in the article, but that is the problem. One of the assumptions of linear regression is that the predictors are independent, which is clearly not the case when the predictors are taste of the first, middle, and last slice. The problem is easily seen in the regression coefficients. For the full price group, the "taste of last slice" has a coefficient of .97 when it is the sole predictor, meaning that for every 1 increase in "taste of last slice" there will be a .97 increase in the rating for overall taste. However, in the total model, which includes ratings for the first and middle slice, the coefficient for "taste of last slice" is $-.02$. The correlation went from a near perfect one to one relationship to a **negative** relationship.

Article 3, Table 2, Granularity inconsistencies/Test statistics

Response	Unique and Valid
"This is due to non-response by some participants"	"No"

This statement does not appear to be accurate. While there are nonresponses in this data set (as there are in any real-world data set), we could not reproduce the original table with the released data. It is a complete mystery how the original values were calculated.

Article 3, Table 3, Granularity inconsistencies/Test statistics

Response	Unique and Valid
"This is due to non-response by some participants"	"No"

Like the previous table it is unclear how to reproduce the original values given the data, and it is unclear why the authors believe the problems to be due to nonresponses.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

Article 3, Metric conversions

Response	Unique and Valid
"We converted the raw reports and then took a mean. The critique converts the rounded mean."	"No"

As described above, our conversions are not accurate for BMI, but we believe that they are accurate for Weight and Height. Given the problems with the Weight and Height conversions we have concerns about the BMI calculations, but are unable to check them.

Article 4, Tables 2/3, Granularity inconsistencies/Test statistics

Response	Unique and Valid
"This was due to an improper handling of outliers in the original analysis."	"Yes"

It is unclear how the authors know how the original statistics were calculated, but presumably they have access to the original scripts. It would be useful to make that information public so that we could see exactly how these outliers were improperly handled.

ADDITIONAL ANALYSES

In our initial critique we could not check every statistic and analysis without access to the data. With the data release we are now able to check these, and will go over a few examples below. Unsurprisingly, we found even analyses which we could not check directly with granularity testing have problems, highlighting the ability of granularity testing to raise concerns about the overall health of a paper.

Article 1, Table 3

When we wrote our critique we were unsure exactly what values were in this table. Cell sizes are not listed, and the terms "1st", "2nd", and "3rd" are used despite not being defined in the article. With access to the STATA code, it is clear that '1st", "2nd", "3rd" in this article are meant to be synonymous with "first", "middle", "last". Table 3 is a within-subjects analysis for diners who ate 3 or more pieces.

Original Table 3

	Effect of price paid			Effect of pieces consumed				Effect of price×pieces						
								$4			$8			
	$4	$8	*F* test (*p* value)	1st piece	2nd piece	3rd piece	*F* test (*p* value)	1st piece	2nd piece	3rd piece	1st piece	2nd piece	3rd piece	*F* test (*p* value)
Pizza taste evaluations	6.84	8.00	7.15 (0.00)	7.77	7.36	7.13	11.09 (0.00)	7.43	6.71	6.38	8.13	8.00	7.88	4.38 (0.02)
Pizza satisfaction ratings	6.89	7.79	3.41 (0.07)	7.63	7.33	7.07	7.06 (0.00)	7.50	6.79	6.39	7.75	7.88	7.76	7.70 (0.00)
Pizza enjoyment ratings	6.67	7.78	6.82 (0.01)	7.55	7.21	6.91	7.10 (0.00)	7.28	6.61	6.11	7.82	7.82	7.71	4.85 (0.02)

The STATA code for this table correctly ignores diners who did not indicate how many pieces they ate. The STATA output for the most part agrees with the original statistics for the $8 piece subgroups, but not with the original statistics for the $4 piece subgroups. The code we provide agrees with the means of the STATA output (we did not attempt to write functions for a repeated-measures ANOVA). It is unclear how to reproduce the values for the $4 piece subgroups.

It retrospect, we probably should have flagged this table as inconsistent given that the averages reported for the "Effect of pieces consumed" columns are simply the averages of the corresponding subgroups for each treatment. They should instead be weighted averages, i.e., the value for "1st piece" should just be the average of all 1st pieces for that question, not the average of the two treatments. These values could only be correct if the two treatments in this table had the exact same sizes, which is highly unlikely.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

STATA output for Table 3

	Effect of price paid			Effect of pieces consumed				Effect of price×pieces $4			Effect of price×pieces $8			
	$4	$8	*F* test (*p* value)	1st piece	2nd piece	3rd piece	*F* test (*p* value)	1st piece	2nd piece	3rd piece	1st piece	2nd piece	3rd piece	*F* test (*p* value)
Pizza taste evaluations	6.90	8.00	6.21 (0.02)	7.73	7.32	7.08	8.66 (0.00)	7.43	6.81	6.48	8.13	8.00	7.88	3.06 (0.05)
N	63	48		37	37	37		21	21	21	16	16	16	
Pizza satisfaction ratings	6.93	7.79	3.20 (0.08)	7.63	7.31	7.03	6.90 (0.00)	7.53	6.84	6.39	7.75	7.88	7.75	7.47 (0.00)
N	56	48		35	35	34		19	19	18	16	16	16	
Pizza enjoyment ratings	6.82	7.78	4.41 (0.04)	7.69	7.25	6.89	14.17 (0.00)	7.58	6.74	6.11	7.82	7.82	7.71	10.41 (0.02)
N	56	51		36	36	35		19	19	18	17	17	17	

Article 2, Table 2

Article 2 primarily focuses on the results of various linear regressions. It was originally impossible for us to check the accuracy of these values without access to the data. With the data release it is clear that like other original analyses, nonsensical responses were included. To see this we reproduced the sample sizes for one of the main tables below.

Sample sizes of Original Table 2

Half price ($4) Models					Full price ($8) Models				
Beginning	Total	End	Peak	Peak-end	Beginning	Total	End	Peak	Peak-end
$N = 62$	$N = 41$	$N = 47$	$N = 62$	$N = 47$	$N = 60$	$N = 26$	$N = 38$	$N = 60$	$N = 38$

The STATA code correctly eliminates nonsensical responses. However, the code includes diners who did not report how many pieces they ate, which is consistent with their STATA code for Article 1, Table 2, but not consistent with their STATA code for Article 1, Table 3. We believe these values should also be removed, and the provided Python code shows an example of how this can be done (linear regressions are performed with `sklearn`). Although we did not attempt to reproduce every value from the original article, it appears that the original statistics are likely consistent with the data if the nonsensical responses are retained.

STATA sample sizes for Table 2

Half price ($4) Models					Full price ($8) Models				
Beginning	Total	End	Peak	Peak-end	Beginning	Total	End	Peak	Peak-end
$N = 62$	$N = 24$	$N = 39$	$N = 62$	$N = 39$	$N = 58$	$N = 17$	$N = 35$	$N = 58$	$N = 35$

CONCLUSIONS

The problems with these four articles extend far beyond minor typos or missing responses. Articles 1 and 2 deal with responses to questions involving the first, middle, and last piece of pizza. However, many of these responses are logically impossible given the number of pieces that the diners ate. Despite this, the original analyses included these data. While Articles 3 and 4 did not deal with these questions and thus were unaffected by this problem, we are unable to work out how the original statistics in these articles were calculated given the data.

This data set contained enough questionable and impossible responses to render it meaningless, yet it was used as the basis of four publications. Even if the data set were perfect, the statistics in the 4 papers would still be uninterpretable given the salami-slicing and *p*-hacking. And even if the data were reliable and the articles were not the result of *p*-hacking, there would still be the issue of the misleading reporting and impossible statistics.

The senior author was more than happy to feature these four publications in a now infamous blog post (Wansink, 2016), suggesting this study is work that the lab is proud of, and an example of what to expect from them. Our investigation into

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

other work from this lab is consistent with this, and we believe the scientific community should exercise caution when interpreting any results from this lab.

ACKNOWLEDGMENTS

We would like to thank Cody DeHaan for providing the STATA output.

COMPETING INTERESTS

JA operates omnesres.com, oncolnc.org, and prepubmed.org. TvdZ has a blog entitled "The Skeptical Scientist" at timvanderzee.com. NJLB has a blog at sTeamTraen.blogspot.com that hosts advertising; his earnings in 2016 were €3.88.

REFERENCES

Anaya, J. (2016). The GRIMMER test: A method for testing the validity of reported measures of variability. *PeerJ Preprints*, 4:e2400v1.

Bartlett, T. (2017). Spoiled science. *The Chronicle of Higher Education*, http://www.chronicle.com/article/Spoiled-Science/239529.

Brown, N. J. L., Anaya, J., van der Zee, T., Heathers, J. A. J., and Chambers, C. (2017). An open letter to dr. todd shackelford. *Nick Brown's blog*, http://steamtraen.blogspot.com/2017/04/an-open-letter-to-dr-todd-shackelford.html.

Brown, N. J. L. and Heathers, J. A. J. (2016). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, Advance online publication:1–7.

Carberry, J. (2017). Cornell university statement regarding questions about professor brian wansink's research. *Food and Brand Lab - Cornell University*, http://mediarelations.cornell.edu/2017/04/05/cornell-university-statement-regarding-questions-about-professor-brian-wansinks-research/.

Cohen, B. H. (2002). Calculating a factorial anova from means and standard deviations. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 1(3):191–203.

Enserink, M. (2017). Paper about how microplastics harm fish should be retracted, report says. *Science*, http://www.sciencemag.org/news/2017/04/paper-about-how-microplastics-harm-fish-should-be-retracted-report-says.

Epskamp, S. and Nuijten, M. B. (2015). Statcheck: Extract statistics from articles and recompute p values (R package version 1.0.1).

Gelman, A. (2017). Pizzagate, or the curious incident of the researcher in response to people pointing out 150 errors in four of his papers. *Statistical Modeling, Causal Inference, and Social Science*, http://andrewgelman.com/2017/02/03/pizzagate-curious-incident-researcher-response-people-pointing-150-errors-four-papers-2/.

Heathers, J. A. J. (2017). Introducing sprite (and the case of the carthorse child). *Hackernoon*, https://hackernoon.com/introducing-sprite-and-the-case-of-the-carthorse-child-58683c2bfeb.

Just, D. R., Sığırcı, Ö., and Wansink, B. (2014). Lower buffet prices lead to less taste satisfaction. *Journal of Sensory Studies*, 29(5):362–370.

Just, D. R., Sığırcı, Ö., and Wansink, B. (2015). Peak-end pizza: prices delay evaluations of quality. *Journal of Product & Brand Management*, 24(7):770–778.

Just, D. R. and Wansink, B. (2011). The flat-rate pricing paradox: conflicting effects of "all-you-can-eat" buffet pricing. *The Review of Economics and Statistics*, 93(1):193–200.

Kniffin, K. M., Sığırcı, Ö., and Wansink, B. (2016). Eating heavily: Men eat more in the company of women. *Evolutionary Psychological Science*, 2(1):38–46.

Lazarz, A. (2007). The psychology of food. *Spectrum News*, http://www.twcnews.com/archives/nys/central-ny/2007/11/17/the-psychology-of-food-NY_37384.old.html.

McCook, A. (2017a). Backlash prompts prominent nutrition researcher to reanalyze multiple papers. *Retraction Watch*, http://retractionwatch.com/2017/02/02/backlash-prompts-prominent-nutrition-researcher-reanalyze-multiple-papers/.

McCook, A. (2017b). "social science isn't definitive like chemistry:" embattled food researcher defends his work. *Retraction Watch*, http://retractionwatch.com/2017/02/16/social-science-isnt-definitive-like-chemistry-embattled-food-researcher-defends-work/.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017

O'Grady, C. (2017). "mindless eating," or how to send an entire life of research into question. *Ars Technica*, https://arstechnica.com/science/2017/04/the-peer-reviewed-saga-of-mindless-eating-mindless-research-is-bad-too/.

Oransky, I. (2015). Data "were destroyed due to privacy/confidentiality requirements," says co-author of retracted gay canvassing study. *Retraction Watch*, http://retractionwatch.com/2015/05/29/data-were-destroyed-due-to-privacyconfidentiality-requirements-says-co-author-of-retracted-gay-canvassing-study/.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shackelford, T. K. (2017). Editorial note. *Evolutionary Psychological Science*, pages 1–1.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Sığırcı, Ö. and Wansink, B. (2015). Low prices and high regret: How pricing influences regret at all-you-can-eat buffets. *BMC Nutrition*, 1(1):36.

van der Zee, T. (2017). The wansink dossier: An overview. *THE SKEPTICAL SCIENTIST*, http://www.timvanderzee.com/the-wansink-dossier-an-overview/.

van der Zee, T., Anaya, J., and Brown, N. J. L. (2017). Statistical heartburn: An attempt to digest four pizza publications from the cornell food and brand lab. *PeerJ Preprints*, 5:e2748v1.

Wansink, B. (2016). The grad student who never said "no". *Healthier & Happier*, https://web-beta.archive.org/web/20170312041524/http://www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no.

Wansink, B. (2017a). A note from brian wansink on research. *Food and Brand Lab - Cornell University*, https://web.archive.org/web/20170220053655/http://foodpsychology.cornell.edu/note-brian-wansink-research.

Wansink, B. (2017b). A note from brian wansink on research. *Food and Brand Lab - Cornell University*, https://web.archive.org/web/20170407054445/http://foodpsychology.cornell.edu/research-statement-april-2017.

Wansink, B. and Payne, C. (2007). All you can eat pizza buffet field study. *CISER Data Archive: Online Catalog*, http://ciser.cornell.edu/ASPs/search_athena.asp?IDTITLE=2783.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3025v1 | CC BY 4.0 Open Access | rec: 14 Jun 2017, publ: 14 Jun 2017