HHS Public Access

Author manuscript

Brain Lang. Author manuscript; available in PMC 2021 November 14.

Published in final edited form as:

Brain Lang. 2019 August ; 195: 104642. doi:10.1016/j.bandl.2019.104642.

Inconsistency of Findings due to Low Power: A Structural MRI Study of Bilingualism

Brandin A. Munson[1], **Arturo E. Hernandez**[2]

[1]Department of Health & Human Performance, University of Houston

[2]Department of Psychology, University of Houston

Abstract

Research on structural brain differences between monolinguals and bilinguals remains inconsistent, and this has been proposed by some to be due in part to inadequate sample sizes. The aim of the present study is to reveal the expected degrees of uncertainty among neuroimaging findings by analyzing random samples of varying sizes from a larger-than-average sample. Bilinguals (n = 216) were compared with monolinguals (n = 146) using grey matter volume measures across region-of-interest tests. Variability among findings were compared with the true full-sample findings, and taken in the context of expected differences within the larger bilingualism neuroimaging literature. Results demonstrate excessive variability across the lowest sample sizes (e.g. samples totaling 20 – 80 participants), and this is explored through the trends of subsample outcomes and effect sizes across sample sizes. The results of this study illustrate the influences of power on expected variability among sample findings.

Keywords

power; sample size; reproducibility; neuroimaging; MRI; bilingualism

1. Introduction

Recent work in the neuroimaging literature has found some controversy regarding the reproducibility of differences between monolinguals and bilinguals. In the case of neuroanatomical differences between both groups there is no consensus about where in the brain they consistently appear (Li, Legault, & Litcofsky,2014; García-Pentón, Garcia, Dunabeitia & Carreiras 2016). The current paper seeks to address this issue by using a larger sample size than is typically used in these studies. After the neuroanatomical differences in the comparisons between the larger group comparison is established, we will look at how sample size relates to the detection of true and false positives. Although some work has been done to establish reproducibility or lack thereof with simulated data, the current approach of

corresponding author.

using real data should help by providing a more concrete example. It will also help to inform researchers who are actively investigating differences between bilinguals and monolinguals as well as across different populations with experiences that could be linked to changes in brain anatomy as observed with structural MRI.

Researchers have known and warned about a problem with the reproducibility of findings in psychology for over a decade. In 2005, Ioannidis published an article which estimated the rate of false positives in psychology to be greater than 50% – suggesting that fewer than half of all studies would be able to be reproduced under similar testing conditions. In 2015, Aarts and colleagues published a paper in Science testing these claims empirically. The authors chose 100 influential studies from Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition. In attempting to reproduce the 100 studies, 39% of effects reproduced findings from the original studies. They conclude by observing, “Scientific progress is a cumulative process of uncertainty reduction that can only succeed if science itself remains the greatest skeptic of its explanatory claims” (p. 7).

This question gets to the root of the academic discipline: are experiments and investigations really saying something informative about the population of interest? The inability of a test to accurately portray the characteristics of a population would mean that, in fact, researchers’ conclusions are not as generalizable to the population at large as they would like to think.

1.2. Statistical Concerns

Reproducibility can be thought of as the ability for an independent study to find the same test result while using the same testing methodology (referred to as “results reproducibility” by Goodman, Faneli, & Ioannidis, 2016). In considering reproducibility, achieved statistical power among studies often comes into question (e.g. Button et al., 2013). The tradeoffs between Type I errors where a researcher finds a false positive effect and Type II errors where a researcher misses a true positive effect is well known, and one researchers must often contend with when creating an experimental design (Craiu & Sun, 2008). There is an unavoidable tradeoff between attempts to control for Type I vs Type II errors, where researchers may try to increasingly control for Type I error rate by choosing more stringent alpha values, but thereby also increase the Type II error rate.

In the neuroimaging literature, a high number of comparisons are, by necessity, commonly controlled for with a more restrictive alpha value (Yarkoni, 2009). Such corrections, though necessary to ensure a low rate of false positives, can lead to a lack of sensitivity to detect effects. This is an issue that has been addressed widely in past neuroimaging research (e.g. Genovese, Lazar, & Nichols, 2002; Bennett, Wolford, & Miller, 2009). Though these and other past studies have explored a variety of treatments for alpha corrections for multiple comparisons, this paper focuses primarily on the Bonferroni type- commonly used in neuroimaging region-of-interest analyses.

Few studies have investigated power-related issues in the neuroimaging literature. Of those that did, only a handful (Desmond & Glover, 2002; Mumford & Nichols, 2008; Murphy &

Garavan, 2004; Thirion et al., 2007) have tried to estimate the sample sizes necessary to gain sufficient power, and even then, all four studies were within-subject functional magnetic resonance imaging (fMRI) designs. For between- vs. within-person designs, Yarkoni (2009) noted that correlational effects may involve 5%−10% of the power to detect within-subject effects of similar magnitude. This paper involves a comparison between monolinguals and bilinguals which rely on between-group differences by necessity, and therefore start with less power overall than within-group investigations.

One common method for determining power in the neuroimaging literature is to use estimates based on previous studies. Yarkoni (2009) demonstrated the unacceptability of combining small sample sizes with stringent alpha levels, a design commonly seen in the MRI / fMRI literature. Using a region of interest (ROI) test as an example, the authors show that, for example, for a sample of 20 subjects with 10 comparisons and a $p = 0.005$ (0.05 corrected for 10 comparisons), the power for detecting a true effect is only 13%. Importantly, this also means that the critical value for detecting an effect becomes a Pearson's $r = 0.6$, a large effect within the psychology literature (Cohen, 1988). The authors show how this causes observed effect sizes to become greatly inflated- all significant observations being above a higher threshold cause the mean observed effect size to increase.

The failure to detect a true effect is another potential outcome of underpowered studies. Vadillo, Konstantinidis & Shanks (2016) reveal how, especially in research focusing on lack of an effect (as in unconscious learning), studies that have too little power can fail to find effects which are actually present. Such an inability to find true effects is another power-related factor potentially influencing a lack of wider consistencies in findings.

1.3. Reproducibility of fMRI Research

Cremers, Wager, and Yarkoni (2017) published a thorough investigation into the effects of underpowered samples on researchers' abilities to make accurate inferences in whole-brain fMRI analyses. The authors created simulated brain slices of 10,000 subjects, and drew 2,000 random subsamples at sample sizes ranging from 10–150. First, and expectedly, they found that the vast majority of the smaller random samples did not show effects which were present in the full sample – confirming that the samples were in fact underpowered. Secondly, though the number of significant voxels was found to increase with the increase in sample size, the average degree of significance actually exponentially decreased as sample size increased. Third, the author's suggested hypothesis-driven method to increase power (by decreasing the large number of tests, and thus reducing the necessary alpha correction) is to use ROIs. Whereas this method does allow researchers to use the literature to drive their predictions, it still requires the use of stringent alpha correction for multiple comparisons.

Despite the advances in stastical modelling, studies on statistical power alone are unlikely to change trends in statistical methodology (Sedlmeier & Gigerenzer, 1989). Szucs & Ioannidis (2017) suggest that over 50% of studies in psychology and cognitive neuroscience are false-positives due to low sample sizes that have not improved in the last 50 years. In light of this, it makes sense to take a more concrete approach, where the consequences of a lack of power can be viewed in the context of tangible conclusions (or lack thereof) due to researcher practices. One field which has seen an increase in neuroimaging studies recently is that of

bilingualism. In order to ensure best researcher practices, as well as giving a literature-based perspective of predictions and effect sizes, a model of neuroimaging studies on bilingualism will be used in order to create grouping variables with evidence-based predictions for what differences ought to be observable, and within which regions.

1.4. Bilingual-Monolingual Neuroanatomy Literature

Several studies, including Li, Legault, and Litcofsky (2014), and García-Pentón, Garcia, Dunabeitia and Carreiras (2016), have reviewed the bilingual neuroanatomy literature in order to better grasp which structural differences are most consistently found between monolinguals and bilinguals, as well between bilinguals of varying language backgrounds. A large number of brain regions have been tied to neuroanatomical differences due to language experience, which are covered extensively in the aforementioned meta-analyses. Interestingly, though there is much overlap between the studies included in these meta-analyses, reviewers have come to different conclusions in terms of whether there are consistent findings of differences across bilinguals and monolinguals.

For instance, in a review which included findings from 10 bilingual-monolingual brain comparison studies, Li, Legault & Litcofsky (2014) concluded that “the evidence reviewed so far portrays a picture that is highly consistent with structural neuroplasticity observed for other domains: second language experience-induced brain changes, including increased grey matter density and white matter integrity, can be found in children, young adults, and the elderly” (p. 301). However, in a separate review of 11 studies (6 of which were the same as those covered in the 2014 Li et al. review), García-Pentón, Garcia, Dunabeitia and Carreiras (2016) concluded that, aside from the IFG and certain white matter connections, present research fails to consistently point to specific neurophysiological differences between monolinguals and bilinguals. García-Pentón et al. then propose certain methodological inconsistencies between studies which may cause unexpected variability in findings, including 1) differing corrections used for multiple comparisons, 2) inadequate descriptions of participant backgrounds, especially related to bilingual language experience (see Hernandez et. al, 2015 for further discussion), and 3) small sample sizes.

As noted earlier, Yarkoni (2009) demonstrated that small sample sizes are associated with inflated significant effect sizes relative to the true population effect size. This might have significant ramifications on the overall reproducibility of a group of findings. The decreased likelihood of finding significant effects which are true in the population is a clear mistake to be avoided by researchers, but a more overlooked outcome might be the inability to accurately design future studies with enough power.

To address this issue with regard to bilingual / monolingual neuroanatomical differences, a brief review of 14 studies in order to better glimpse the average effect sizes found for studies reporting anatomical differences between bilinguals and monolinguals was conducted. Studies were selected through 1) the Li et al. (2014) review, 2) the García-Pentón et al. (2015) review, and 3) a Google Scholar search of “bilingual monolingual structural MRI.” For these 14 studies, effect sizes were calculated wherever possible; 4 studies did not present sufficient information for Cohen’s d effect sizes to be calculated, and 5 others did not include comparable results of bilingual and monolingual neuroanatomy; some investigated

only differences in effects of factors such as ages of acquisition in bilinguals (e.g. Berken et al., 2015), while others investigated interhemispheric differences (e.g. Felton et al., 2017). See Table 1 for study-specific details, including sample sizes, mean within-study effect sizes and other details. This left 5 studies which were used to estimate effect sizes of bilingual-monolingual differences in the literature.

The mean total sample size for all 14 studies was 52 participants; on average, 24 monolinguals were compared with 28 bilinguals. Of the 5 studies which reported adequate information for bilingual-monolingual comparison effect sizes to be calculated, the average Cohen's d effect size for significant findings (1 mean value per study) was 1.21. Cohen's d represents the standardized differences between the means of two groups; so, if it is equal to 0.5, then the mean of one group is half of a standard deviation greater than or less than the other. For all 10 studies with adequate information to calculate effect sizes (which is more of a measure of general within- and between-group neurophysiological differences due to language experience), an average Cohen's d effect size of 1.16 was found. Putting this into perspective, Cohen (1992) suggested a Cohen's d of 0.2 could be described as 'small,' 0.5 as 'medium,' and 0.8 as 'large.' Seventeen years later, Sawilowski (2009) suggested an effect size of 1.2 might be described as 'very large.' Such a description would suggest that researchers are generally finding very large neuroanatomical differences between monolinguals and bilinguals. However, this does not quite fit with some of the noted inconsistencies within the literature (e.g. García-Pentón et al., 2015). If these are truly large differences within the populations of interest, they should then be more consistently observed between studies. This inconsistency may be explained by the lack of power to 1) detect true population effects that are present in the population, and 2) accurately estimate the sizes of true effects in the population, rather than overestimating effect sizes with underpowered samples.

A power analysis using G*Power (Erdfeler, Faul, & Buchner; 1996) revealed that, for a very large effect size, a two-tailed, independent samples t-test (alpha = 0.05) with a very large effect size (Cohen's d = 1.2, as with the average published effect size for those reviewed in the bilingual neuroimaging literature) would require a total sample size of 12 per group, 24 total, in order to achieve 0.80 power. This is a very achievable number of participants, and is in fact slightly below the average observed in this area of research. However, the assumption of finding such a large effect size is considerably risky. If the effect size were decreased to a medium Cohen's d = 0.5, the necessary sample size then becomes 64 per group, 128 total for a reliable detection rate (0.80 power). The issue is also exacerbated where the alpha is corrected for multiple comparisons. If alpha = 0.005, 21 participants per group are then necessary to reliably detect a very large effect size, and 109 participants per group for a medium effect size. Hence, there is clearly a wide range of potential scenarios where very small or very large sample sizes are necessary.

How might one then pin down the extent to which small sample sizes, and other researcher degrees of freedom, are affecting variability in study findings? One possibility is to take a larger-than-normal sample of bilingual and monolingual scans, and conduct simulated studies of smaller sample sizes ('subsamples') within this group. The 'population,' or 'whole-sample,' findings being known, this would reveal the extent to which variability

of 1) effect size inflation, 2) sensitivity (likelihood that a true finding in the population will be observed as a true positive in a sample), and 3) positive predictive values (likelihood that an observed true finding in a sample is actually true in the population) are due to factors such as sample size. If variability in this literature actually is due to inadequate sample sizes as has been suggested, then very inconsistent findings among smaller subsamples relative to whole-sample differences would support this theory, and display concretely to researchers that further steps need to be taken in future studies in order to more consistently find differences that actually are present in the population. If, however, these small-sample findings are able to adequately represent findings in the population, it would suggest that other reasons for the variability may be the primary cause of inconsistent results in the literature, such as inconsistent definitions of bilinguals and bilingual language experiences. The present study uses a sample of monolinguals and bilinguals much larger than average in bilingual neuroimaging studies (356 total brain scans) in order to determine achieved decreases in expected variability at varying researcher degrees of freedom.

2. Methods

2.1. Participants

A total of 362 participants were scanned at the Center for Advanced Magnetic Resonance Imaging (CAMRI) in Houston, TX (234 females; 216 Spanish-English bilinguals) across 8 separate studies. Analyses of prior collected data was approved through University of Houston Internal Review Board (IRB) study ID ‘STUDY00000015.’ Participants were primarily University of Houston students, as well as members of the greater Houston community. Compensation was given in the form of either 1) Starbucks or Target gift cards, or 2) course extra credit, with participants choosing which. All participants were screened for background factors incompatible with MRI. See Table 2 for means and standard deviations, split between bilinguals and monolinguals, of background variables of interest, including age, language proficiencies, and age of second language acquisition.

Monolinguals, who reported limited knowledge of any language other than English, were asked to complete the Boston Naming Test (Kaplan et al., 1983) and/or the following subtests of the Woodcock-Muñoz Language Survey – Revised: picture vocabulary, followed by either passage comprehension or English listening comprehension (for detailed explanation of each subtest see Woodcock, Muñoz-Sandova, Ruef, & Alvarado, 2005). Spanish-English bilinguals were asked to complete the above measures both in English and Spanish to ensure qualification as a bilingual participant.

T1-weighted high-resolution images were obtained from a Siemens Magnetom Trio 3-T MRI scanner at the Center for Advanced Magnetic Resonance Imaging (CAMRI) at Baylor School of Medicine in Houston, Texas. The T1-weighted MPRAGE scans were collected using the following parameters for the eight studies: repetition time (TR), 1200 ms; echo time (TE), 2.66 ms; flip angle (FA), 12°; voxel size, 0.479 × 0.479 × 1.0 mm; 192 slices.

2.2. Analyses

The T1 scans were preprocessed through modulated normalized segmentation in order to create measures of grey matter volume, using the Statistical Parametric Mapping (SPM) software (Ashburner et al., 2014; http://www.fil.ion.ucl.ac.uk/spm/). All images were checked to confirm consistent orientation. Region of interest (ROI) grey matter volume values were estimated using voxel-based morphometry, and were generated from the automated anatomical labeling (AAL) brain atlas (Tzourio-Mazoyer et al., 2002). Both intracranial volume (ICV; used to control for overall brain size) and ROI volume data values were extracted using the SPM Computational Anatomy Toolbox (CAT12) SPM package (http://www.neuro.uni-jena.de/cat/).

ROI grey matter volume data extracted from SPM was analyzed using the R statistical software (R Team, 2000). Participants were randomly selected from the full sample at total sample sizes ranging from 20 to 280 in increasing increments of 20 (20, 40, … 280). Every sample was controlled such that the proportion of bilinguals to monolinguals was 50:50. For instance, in a single sample size of 20 participants, 10 would be randomly selected bilinguals, and 10 would be randomly selected monolinguals. For each sample size, 1,000 randomized subsamples were created without replacement.

Regressions that included language status (bilingual or monolingual) and intra-cranial volume (ICV; used to control for overall brain size) as predictors were conducted on each of 10 ROI's. These included bilateral superior temporal gyrus (STG), basal ganglia (BG) anterior cingulate cortex (ACC), inferior frontal gyrus (IFG), and inferior parietal lobule (IPL). ROIs were selected prior to data analysis based on 1) published findings showing differences in either volume (left BG: Zou et al., 2012; left STG: Ressel et al., 2012; bilateral ACC: Abutalebi et al., 2015), density (left IFG: Mechelli et al., 2004; right IPL: Grogan et al., 2012) or cortical thickness (bilateral IFG: Klein et al., 2013), and 2) the 2007 Abutalebi and Green model for regions associated with control during bilingual language processing, which includes the basal ganglia, ACC, IPL, and prefrontal cortex (including the IFG). Significant differences in grey matter volume, grey matter density, and cortical thickness, as some of the most commonly used phenotypes in bilingual neuroimaging, served as determiners of potential regional brain differences. The 'true' full-sample (N = 362) effects were calculated for each ROI, and compared with findings between subsamples.

Several aspects of test accuracy were explored. Achieved power per test, as well as the degree of effect size inflation (expected to be greater for significant effects within smaller samples, smaller true effects and more restrictive alphas) were graphed and summarized. The calculation of a 'confusion matrix' (Figure 2), which groups the percentage of significant or nonsignificant sample tests vs. true or false full-sample tests, allowed for 1) sensitivity and 2) positive predictive values, both positively associated with levels of achieved power, to be graphed across samples. These are measures of both the consistency and accuracy of tests relative to the actual differences within the population. Sensitivity, or power, is calculated as the number of true positive findings (those which are both significant within a tested sample and true in the population) divided by the total number of true findings (true positive findings combined with false negative findings), and can be interpreted as the likelihood that a study is going to find a significant effect when there

is a true effect present in the population. Positive predictive value (PPV) is calculated as the number of true positive findings divided by the total number of positive findings (true positive findings combined with false positive findings), and can be interpreted as the likelihood that a significant effect found within a study is a true effect present in the population. Because each of these are calculated by creating a cutoff value, two separate p-value cutoffs were explored - one stringent Bonferroni-corrected ($p = 0.005$ for 10 total comparisons), and one less stringent ($p = 0.025$) - in order to view differences in findings across alpha cutoff stringency.

3. Results

3.1. Replication Rates Using a Cutoff Value

Figure 1 shows changes in the accuracy of sample findings across sample sizes. For each of the 10 ROI's, the amount of variance contributed by language status (bilingual or monolingual) is either significant ($p < 0.005$) or nonsignificant ($p > 0.005$). This was compared to the full-sample results (all 362 subjects) for each ROI, where findings were also tested at a Bonferroni-corrected p value of 0.005. Significant sample findings are called 'Positive', and nonsignificant sample findings are called 'Negative'. If the sample finding matches the full sample finding, it is 'True'; if it does not match, it is 'False'.

Thus, in a 'True Positive' finding for a single ROI, a significant amount of variability in the region (measured with volume) is explained by language status in the full sample, and this is also found in the smaller random sample. In a 'False Positive' finding, a significant amount of variability in the region is *not* explained by language status in the full sample, but language status is still found to be significant in the smaller random sample. In a 'True Negative' finding, a significant amount of variability in the region is *not* explained by language status in the full sample, and language status is also *not* found to be significant in the smaller random sample. In a 'False Negative' finding, a significant amount of variability in the region is explained by language status in the full sample, *but* language status is *not* found to be significant in the smaller random sample. See Figure 2 for a simple table. For instance, if a significant amount of variability in right ACC volume is explained by language status in the regression for the full sample, but is not found to be significant in a random sample of 30 monolinguals and 30 bilinguals, this would count as one instance of a 'False Negative' for N = 30.

True Negatives were not included, as they 1) did not change significantly across sample sizes, and 2) were much more numerous than the other three outcomes, making it more difficult to compare the other outcomes. True Negatives occurred in roughly 5 out of 6 tests across sample sizes. Figure 1 is cut off at 1,000 (10%), but the true total number of tests for each sample size is 10,000.

3.2 . True / False Positives

In the full sample, only the RIPL region was found to significantly differ between monolinguals and bilinguals, $p = 0.0008$; for all 9 of the other ROIs, $p > 0.01$. Table 3 lists the p values and R^2 effect sizes for each full-sample ROI test. Figure 1 demonstrates

that, at a Bonferroni-corrected critical cutoff of $p = 0.005$, as sample size increases, the likelihood of finding a true positive effect (that is, a difference in the RIPL region) also increases, the likelihood of finding a false negative decreases, and the likelihood of finding a false positive is stable. As we would expect, increasing the sample size of a statistical test has positive effects on the accuracy of that test to guess at the 'true population-level' group differences.

However, what researchers view as 'acceptable' rates of true vs. false findings (often a power of 0.80, or false negative rates limited to 20% where findings in the full sample are actually positive) is not even approached at the highest sample sizes. At the lowest sample size, 10 monolinguals vs. 10 bilinguals, a very small proportion of tests (less than 4%) are detecting the only truly significant full-sample effect of RIPL. At this rate of true positive findings, tests are actually more likely to be falsely detecting a difference which is actually not significant within the full sample (roughly 5%). The rate of true positive findings only becomes greater than false positives where tests are conducted with 30 monolinguals and 30 bilinguals in each group- the difference between a true positive and a false positive is roughly a coin flip, which lasts until samples of 70 or more in each group are attained.

What is often thought of as a minimum level of power, 0.80, isn't even achieved with the largest samples consisting of 140 participants per group (280 total). As covered by Yarkoni (2009), the factors of 1) small effects (which are often an issue in neuroimaging studies), 2) multiple comparisons, such as the case here of using many ROI's, and 3) a stringent alpha restriction (Bonferonni-corrected $p = 0.005$, used here, is somewhat stringent, though not so when compared to whole-brain analyses) all combine to reduce achieved power. With the purpose of clarifying the effects of alpha stringency on test accuracy, the outcome of using a relatively less stringent alpha of $p = 0.025$ was explored. This is detailed in Figure 3, which shows the same accuracy metrics for subsample vs. full samples as Figure 1, with the only difference being that the threshold of significance was changed from $p = 0.005$ to $p = 0.025$.

Figure 3 demonstrates an increase in the rates of True Positives across all sample sizes, especially as sample sizes increase- since the threshold to significance is lower, it is more likely to find a truly significant difference in the random samples. It is also clear that False Negative rates, nearly 100% for a more stringent alpha correction, start off lower (roughly 85%) and decrease more rapidly as sample sizes increase. This means that at the highest sample size of 140 per group, a power of 80% is nearly reached – but still not quite. However, this is a tradeoff with increased overall False Positive rates. For samples below 60 per group, researchers would be more likely to falsely conclude that a test was significant than to accurately do so – and at the lowest sample sizes, they would be *much* more likely to reach such a misleading conclusion.

3.3. Positive Predictive Values (PPV) and Sensitivity

Positive Predictive Value (PPV), or the number of True Positive findings out of the total number of positive sample findings, is (again) a metric used to measure the likelihood that an observed positive (significant) finding is reflective of a finding that is actually positive within the full sample. Figure 4, with a critical alpha cutoff of $p = 0.005$, shows that although the PPV increases with sample size, it is very unlikely (about a 25% chance) in

many of the smaller sample sizes that a positive result actually reflects a true population finding.

Sensitivity, or the number of True Positive findings out of the total number of true full-sample findings, is a metric used to measure the likelihood that a sample will return a positive (significant) result when it should. In the context of this test and these ROI's, the only positive full-sample outcome is the RIPL. So, here, sensitivity refers to the likelihood of a sample finding a significant different in the RIPL. Figure 4, again with a critical alpha cutoff of $p = 0.005$, shows a dismal sensitivity across sample sizes for a test to find a significant difference in the RIPL, where it should be found.

Figure 5 shows the same PPV and sensitivity metrics for subsample vs. full samples as Figure 4, with the only difference being that the threshold of significance was changed from $p = 0.005$ to $p = 0.025$. This demonstrates the tradeoff of improved sensitivity, but decreased PPV, that would come with a less stringent alpha correction.

3.4. Inflated Effect Sizes

Yarkoni (2009) has shown with simulated fMRI data that underpowered tests combined with strict alpha corrections are more likely to have inflated significant outcomes. This is at first counterintuitive, in that lower sample sizes often mean smaller observed effect sizes. This is true when we think of an individual statistical test, without regard for whether it is significant. But, as discussed by Yarkoni, when studies are restricted to findings with very restrictive alpha cutoffs, this creates a scenario where smaller sample sizes need to have larger effects in order to become significant, on average. So, with a higher critical cutoff and many potential comparisons being looked at, researchers would be more likely to find higher-than-actual effect sizes from samples which are small than from large samples.

The present data reflected the phenomenon described in the above paragraph. Looking at the variability in effect sizes across samples, Figure 6 shows that as subsample size increases, the average *significant* observed R^2 effect size (where the Bonferroni-corrected alpha = 0.005) decreases in size, especially for samples less than 40 per group in size. This variability is seen to 'stretch' the interval of observed effects away from the true average, which is closer to the observed R^2 0.015 for the only statistically significant difference in the RIPL comparison (see Table 3 for all full-sample test effect sizes). So, smaller samples are more likely to see inflated effects when significant, and observed effect sizes asymptotically approach the true full-sample effect size as sample size increases.

Also consistent with Yarkoni (2009), less stringent alpha cutoffs (p = 0.025) show a smaller amount of average inflation away from the true full-sample R^2 effect size. Figure 7 shows that with a less stringent alpha cutoff, smaller sample sizes differ less in average observed R^2 values relative to larger samples. So, increased power that results from less stringent alpha cutoffs does lead samples to more accurately estimate the true effect sizes. However, it should be noted that this is just demonstrated for illustrative purposes; it is *not* recommended to trade increases in Type I errors, which are potentially more damaging false conclusions for researchers to make, for decreases in Type II errors. This is likely a part of the reason why stringent alpha cutoffs are often prioritized over adequate power for statistical tests. The

primary ways to address these issues, addressed in further detail below, would be to strive for increased sample sizes, and more consistent and powerful statistical methods across studies.

4. Discussion

The present study provides an example of the type of inaccuracies which might be expected from underpowered samples in neuroimaging, specifically when using ROIs to investigate bilingual-monolingual differences in brain volume. Though this is the framework through which the results are being viewed, as shown in Cremers, Wager, and Yarkoni (2017), these effects are generalizable to MRI / fMRI studies, and likely to any study which uses the frequentist statistical approach to experimental testing.

As expected, inadequate power is related to an inability to find true sample differences, as well as a higher likelihood of showing significant effects that are not truly significant within the population. Beyond this, inflated effect sizes are more common among small, significant samples. Such inflated effects could hinder meta-analyses and calculations for necessary power analyses in future studies by giving researchers inaccurate measures for expected effect sizes. To be clear, effect size inflation may be increased where more stringent alpha values are seen, but that does not make alpha corrections for multiple comparison 'bad practice' whatsoever- it does mean, though, that larger samples are then necessary in order to achieve higher accuracy and reproducibility.

A power analysis revealed that a very large effect size as determined from the bilingual neuroimaging literature average, Cohen's d = 1.2, would require roughly 21 participants per group to consistently find an effect. Because the only full-sample effect size that was found to be significant was medium-small in size (d = 0.3; r = 0.015), this calculation differs greatly from the actual necessary sample size, which would be closer to 298 per group. Such a number is often unrealistic (though not without precedent for fMRI explorations of individual differences; see Dubois & Adolphs, 2016), and likely relates to our inability to find a large effect. This also, however, reveals the dissonance between the observed effect size in our sampled data, and those achieved in significant findings in published research in bilingual neuroimaging. Caution ought to be used when utilizing such prior studies for future power analyses, to the extent it is difficult to parse relevance to the sample to be collected. A better solution may be to conduct a small pilot study (where possible) to collect preliminary data to determine an estimate of effect sizes, and to use that information alongside the literature to make estimates of necessary sample sizes. It also ought to be noted that, although the observed trends for inaccuracies would remain consistent across sample sizes, their degree would differ relative to what was observed if full-sample findings were varied (e.g. more clusters were found to be significant, or with a greater degree of significance). This was not explored in this paper, though aspects of inflation varying with full-sample characteristics are covered in Yarkoni, 2009.

While the influences of power within bilingual neuroimaging analyses were explored here, other potentially influential factors also ought to be considered in planning and conducting future studies. García-Pentón et al. (2015) suggest that these include ensuring

randomized (less region- and population-specific) sample selection, as well as clearer and more consistent operationalization of variables between studies, such as the definition for a “bilingual” versus a “monolingual.” Without a clear and consistent definition of what constitutes a ‘bilingual,’ it is very difficult, if not impossible, to study ‘bilingual-monolingual differences.’ Other methods more specific to MRI analyses may also allow larger effects to be extracted from the same sample sizes in studies, such as Vaden et al.’s (2012) demonstration of the utility of multiple imputation in fMRI.

Although this study focused on structural MRI data, issues with reproducibility as it relates to achieved power extend to many research areas, especially where there are many variables of interest. For instance, reproducibility in biomedicine has been shown to be very low (Begley & Ellis, 2012), with potential causes ranging from low power (explored here) to more general publication bias, where novel, significant findings are often published in preference to non-significant or replication findings (Prinz, Schlange, & Asadullah, 2011).

One of the first responses to requests for larger samples in research is, understandably, “Okay, then. Give us the money and we’ll collect more participants!” Larger samples are the most direct way to achieve higher power, and ought to be aimed for whenever certain effects / analyses require it, but sample size is not the only influence on achieved power. Button and colleagues (2013), after criticizing the trend for inadequate power in the neurosciences, list several methods to help improve researcher practices, which would have a positive impact on reproducibility of findings in the long term.

First, if an a-priori power calculation is conducted, researchers will have a good idea as to how many participants would need to be collected in order to run certain statistical tests. This relates to study pre-registration, which holds researchers accountable to their original hypotheses, and (in certain journals) allows for studies to be published based upon their designs and investigations alone, rather than on significant findings- thus also decreasing the “file-drawer” problem of unpublished null results. Finally, considering that larger grants are not always available for optimal sample sizes, Button recommends collaboration between labs with similar data. This would not only make larger sample sizes available, but would also somewhat alleviate the problem of lab- and region-specific findings.

The present study sheds light on the ways in which inadequately powered studies may influence results with the intent of informing research practices more directly. This is aimed towards revealing how accurately studies in the bilingualism literature are approximating population-level brain structure differences between bilinguals and monolinguals given current researcher practices. The high amount of observed variability in small samples suggest that researchers ought to strongly consider some of the aforementioned options for addressing power in studies. Future studies using either real data that captures real-world complexities (as done here) or simulated data which allows for many statistical variables to be controlled for and therefore more clearly explored (e.g. Cremers, Wager, & Yarkoni, 2017) would be beneficial to more fully communicating methodological issues in neuroimaging reproducibility.

This study does not definitively demonstrate that factors such as inadequate power and multiple comparisons are a causal influence behind the observed variability within the bilingual neuroimaging literature. However, it does reinforce the possibility that these factors have negatively affected the accuracy and consistency of bilingual studies. It is our hope that this study helps to open the eyes of bilingual researchers who use neuroimaging, as well as behavioral differences, to the negative inferential effects that coincide with inadequate statistical power.

ACKNOWLEDGEMENTS

Thank you to current and former members of the Laboratory for the Neural Bases of Bilingualism who originally collected and cleaned the majority of the neuroimaging scans used in this study, including Dr. Kailyn Bradley, Dr. Pilar Archila-Suerte, Dr. Aurora Ramos, Dr. Maya Greene, Dr. Kelly Vaughn, and Hannah Claussenius-Kalman. Thank you also to Dr. David Francis and Dr. Benjamin Tamber-Rosenau, whose valuable feedback helped to shape the focus and direction of these methodological explorations. Finally, thank you to Dr. Peggy Lindner, who assisted with the writing of the MATLAB code for this study.

This research was supported by Award Numbers R03 HD050313, R21 HD059103, R03 HD079873, and P50 HD052117, from the Eunice Kennedy Shriver National Institute of Child Health and Human Development to the University of Houston. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health.

REFERENCES

Aarts AA, Anderson JE, Anderson CJ, Attridge PR, Attwood A, & Fedor A. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), 1–8.

Abutalebi J, Canini M, Della Rosa PA, Green DW, & Weekes BS (2015). The neuroprotective effects of bilingualism upon the inferior parietal lobule: a structural neuroimaging study in aging Chinese bilinguals. Journal of Neurolinguistics, 33, 3–13.

Abutalebi J, Della Rosa PA, Gonzaga AKC, Keim R, Costa A, & Perani D. (2013). The role of the left putamen in multilingual language production. Brain and language, 125(3), 307–315. [PubMed: 22538086]

Abutalebi J, & Green D. (2007). Bilingual language production: The neurocognition of language representation and control. Journal of neurolinguistics, 20(3), 242–275.

Ashburner J, Barnes G, Chen C, Daunizeau J, Flandin G, Friston K, & Penny W. (2014). SPM12 manual. Wellcome Trust Centre for Neuroimaging, London, UK.

Bennett CM, Wolford GL, & Miller MB (2009). The principled control of false positives in neuroimaging. Social cognitive and affective neuroscience, 4(4), 417–422. [PubMed: 20042432]

Cohen J. (1992). A power primer. Psychological bulletin, 112(1), 155. [PubMed: 19565683]

Cohen J. (1988). Statistical power analysis for the behavioural sciences.

Craiu RV, & Sun L. (2008). Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. Statistica Sinica, 18, 861–879.

Cremers HR, Wager TD, & Yarkoni T. (2017). The relation between statistical power and inference in fMRI. PloS one, 12(11), e0184923.

Dale AM, Fischl B, & Sereno MI (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage, 9(2), 179–194. [PubMed: 9931268]

De Bruin A, Treccani B, & Della Sala S. (2015). Cognitive advantage in bilingualism: An example of publication bias?. Psychological science, 26(1), 99–107. [PubMed: 25475825]

Desmond JE, & Glover GH (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. Journal of neuroscience methods, 118(2), 115–128. [PubMed: 12204303]

Destrieux C, Fischl B, Dale A, & Halgren E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage, 53(1), 1–15. [PubMed: 20547229]

Dubois J, & Adolphs R. (2016). Building a science of individual differences from fMRI. Trends in cognitive sciences, 20(6), 425–443. [PubMed: 27138646]

Erdfelder E, Faul F, & Buchner A. (1996). GPOWER: A general power analysis program. Behavior research methods, instruments, & computers, 28(1), 1–11.

Fischl B, & Dale AM (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences, 97(20), 11050–11055.

Fischl B, Salat DH, van der Kouwe AJ, Makris N, Ségonne F, Quinn BT, & Dale AM (2004). Sequence-independent segmentation of magnetic resonance images. Neuroimage, 23, S69–S84. [PubMed: 15501102]

Fischl B, Van Der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, & Caviness V. (2004). Automatically parcellating the human cerebral cortex. Cerebral cortex, 14(1), 11–22. [PubMed: 14654453]

García-Pentón L, Fernández García Y, Costello B, Duñabeitia JA, & Carreiras M. (2016). The neuroanatomy of bilingualism: how to turn a hazy view into the full picture. Language, Cognition and Neuroscience, 31(3), 303–327.

Genovese CR, Lazar NA, & Nichols T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage, 15(4), 870–878. [PubMed: 11906227]

Goodman SN, Fanelli D, & Ioannidis JP (2016). What does research reproducibility mean?. Science translational medicine, 8(341), 341ps12–341ps12.

Grogan A, Jones ŌP, Ali N, Crinion J, Orabona S, Mechias ML, … & Price CJ (2012). Structural correlates for lexical efficiency and number of languages in non-native speakers of English. Neuropsychologia, 50(7), 1347–1352. [PubMed: 22401989]

Hernandez AE, Greene M, Vaughn K, Francis DA & Grigorenko E. (2015). Beyond the bilingual advantage: The potential role of genes and environment on the development of cognitive control. Journal of Neurolinguistics. 35, 109–119. [PubMed: 30270989]

Ioannidis JP (2005). Why most published research findings are false. PLoS medicine, 2(8), e124. [PubMed: 16060722]

Jeffreys H. (1961). The Theory of Probability. Oxford.

Kaplan E, Goodglass H, & Weintraub S. (1983). The Boston naming test. 2nd. Philadelphia: Lea & Febiger.

Kass R, & Raftery A. (1995). Bayes Factors. Journal of the American Statistical Association, 90(430).

Klein D, Mok K, Chen JK, & Watkins KE (2014). Age of language learning shapes brain structure: a cortical thickness study of bilingual and monolingual individuals. Brain and Language, 131, 20–24. [PubMed: 23819901]

Li P, Legault J, & Litcofsky KA (2014). Neuroplasticity as a function of second language learning: anatomical changes in the human brain. Cortex, 58, 301–324. [PubMed: 24996640]

MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

Mechelli A, Crinion JT, Noppeney U, O'doherty J, Ashburner J, Frackowiak RS, & Price CJ (2004). Neurolinguistics: structural plasticity in the bilingual brain. Nature, 431(7010).

Mumford JA, & Nichols TE (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage, 39(1), 261–268. [PubMed: 17919925]

Murphy K, & Garavan H. (2004). An empirical investigation into the number of subjects required for an event-related fMRI study. Neuroimage, 22(2), 879–885. [PubMed: 15193618]

Prinz F, Schlange T, & Asadullah K. (2011). Believe it or not: how much can we rely on published data on potential drug targets?. Nature reviews Drug discovery, 10(9), 712.

Ressel V, Pallier C, Ventura-Campos N, Díaz B, Roessler A, Ávila C, & Sebastián-Gallés N. (2012). An effect of bilingualism on the auditory cortex. Journal of Neuroscience, 32(47), 16597–16601. [PubMed: 23175815]

Sawilowsky SS (2009). New effect size rules of thumb.

Sedlmeier P, & Gigerenzer G. (1989). Do studies of statistical power have an effect on the power of studies?. Psychological bulletin, 105(2), 309.

Szucs D, & Ioannidis JP (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS biology, 15(3), e2000797.

Team RC (2000). R language definition. Vienna, Austria: R foundation for statistical computing.

Thirion B, Pinel P, Mériaux S, Roche A, Dehaene S, & Poline JB (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. Neuroimage, 35(1), 105–120. [PubMed: 17239619]

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, & Joliot M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage, 15(1), 273–289. [PubMed: 11771995]

Woodcock RW (2005). Woodcock-Muñoz language survey-revised. Itasca, IL: Riverside.

Yarkoni T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al..(2009). Perspectives on Psychological Science, 4(3), 294–298. [PubMed: 26158966]

Zou L, Ding G, Abutalebi J, Shu H, & Peng D. (2012). Structural plasticity of the left caudate in bimodal bilinguals. Cortex, 48(9), 1197–1206. [PubMed: 21741636]

Highlights

- Test accuracy is much lower for smaller and more commonly collected sample sizes.
- Inflated effects are seen in regressions with samples less than 30 participants per group.
- Power calculated a-priori from the bilingualism literature is inadequate for even a very large sample.

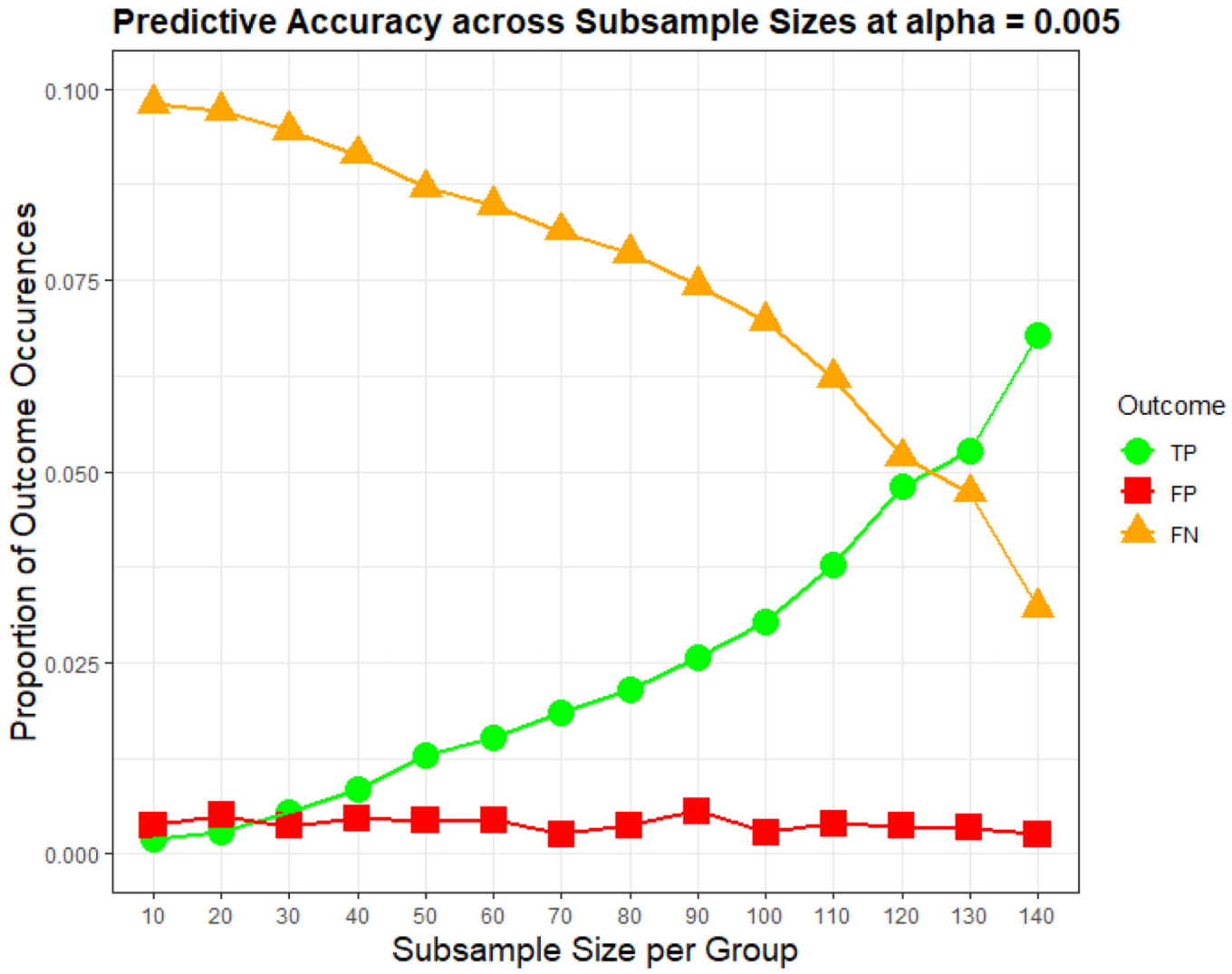

Figure 1.

Accuracy of subsample test outcomes relative to the full sample across subsample sizes per group, where the stringent critical alpha = 0.005. False Negatives (FN; the yellow triangles) are the most common outcome, and decrease as the subsample size increases. False Positives (FP; the red squares) are least common, and remain constant as the subsample size increases. True Positives (TP; the green circles) are seen to increase as subsample size increases. The result of each individual ROI test within each subsample size is included here. Thus, 5 bilateral ROI's multiplied by the number of random samples (1,000) tested at each subsample size makes the total 10,000, though the y-axis is cut off at 1,000 (10%). This is because True Negatives are not included, as they 1) change a very small amount across subsample sizes, and 2) make up a large majority of the test outcomes. Here, where the critical alpha = 0.005, True Negatives were seen in about 8,960 of the 10,000 tests across each subsample size.

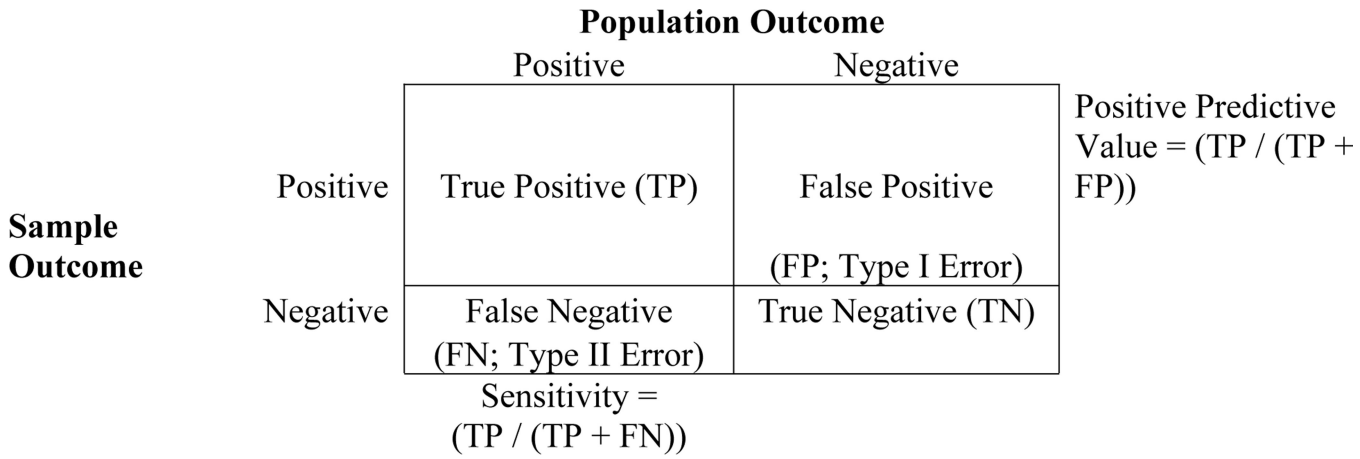

Figure 2.

A simplified confusion matrix. Population (in this paper, full-sample) outcomes are separated by columns, whereas sample (in this paper, sub-sample) outcomes are separated by rows. The calculations used to create measures of both Sensitivity and Positive Predictive Values are given.

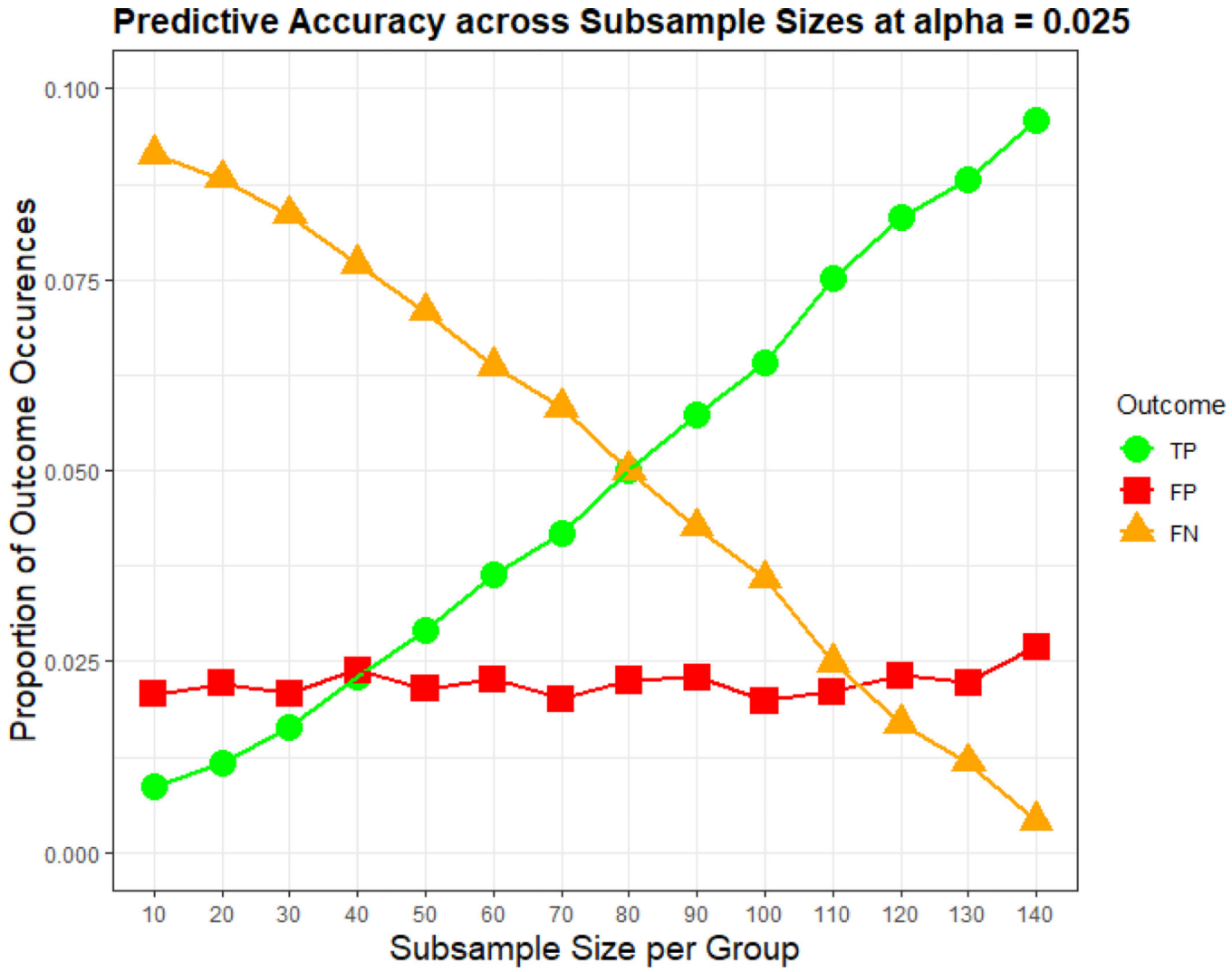

Figure 3.

Accuracy of subsample test outcomes relative to the full sample, where the more lenient critical alpha = 0.025. False Negatives (FN; the yellow triangles) again decrease as the subsample size increases, here at a greater rate- and even become less frequent than True Positives (TP; the green circles) where the subsample size >= 100 per group. False Positives (FP; the red squares) are now seen to be more common than TP in lower sample sizes and overall more frequent. As expected, a less stringent alpha is a trade-off between resulting in both more TP and FP. Here, where the critical alpha = 0.025, True Negatives were seen in about 8,800 of the 10,000 tests across each subsample size.

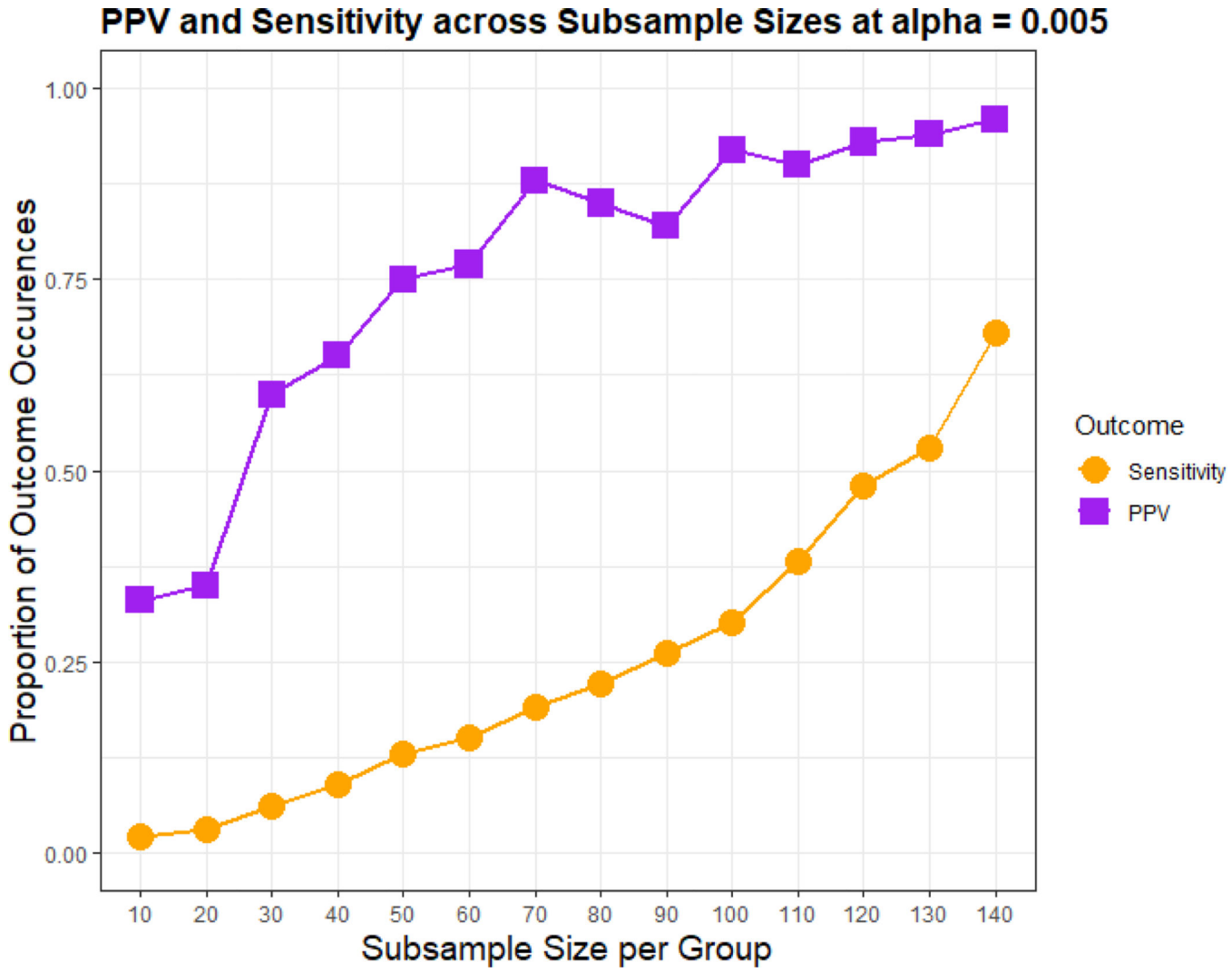

Figure 4.

Average Sensitivity (the yellow circles), also known as Power, and Positive Predictive Value (PPV; the purple squares) across subsample sizes per group, where the stringent critical alpha = 0.005. Both Sensitivity and PPV can be seen to steadily increase with subsample size, though sensitivity remains below 0.25 for the majority of the subsample sizes. Sensitivity, or Power, is defined as the proportion of TPs to the sum of TPs and FNs (TP / (TP + FN)), and is therefore a measure of the likelihood that a positive outcome in a binary statistical test will be able to detect a significant (positive) difference in the full sample. PPV is defined as the proportion of TPs to the sum of TPs and FPs (TP / (TP + FP)), and is therefore a measure of the likelihood that a positive outcome in a binary statistical test accurately reflects a significant (positive) difference in the full sample.

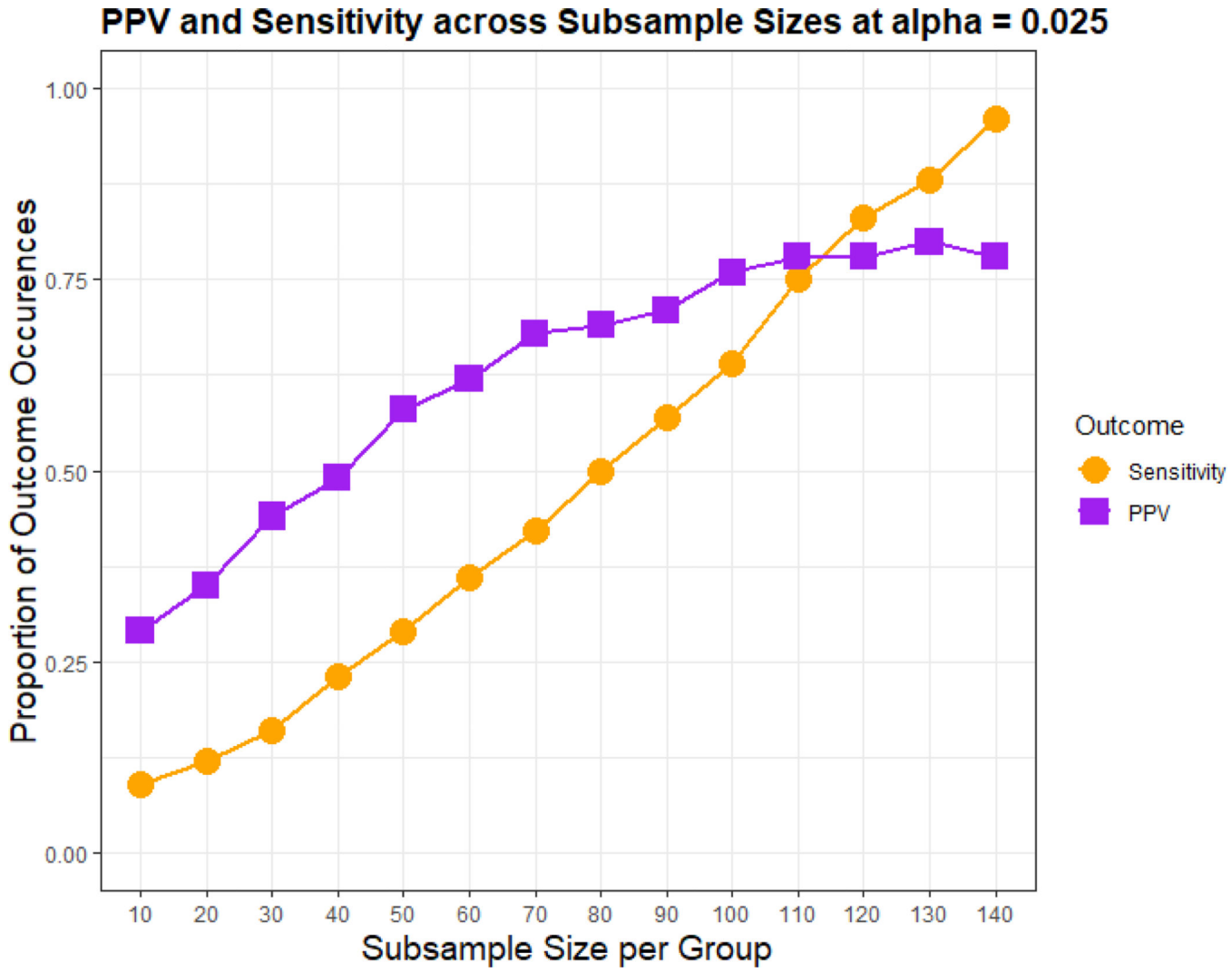

Figure 5.

Average Sensitivity (the yellow circles), also known as Power, and Positive Predictive Value (PPV; the purple squares) across subsample sizes per group, where the lenient critical alpha = 0.025. Both Sensitivity and PPV can still be seen to steadily increase with subsample size. However, the lenient alpha cutoff results in overall increased Sensitivity / Power, with the tradeoff of a decreased PPV.

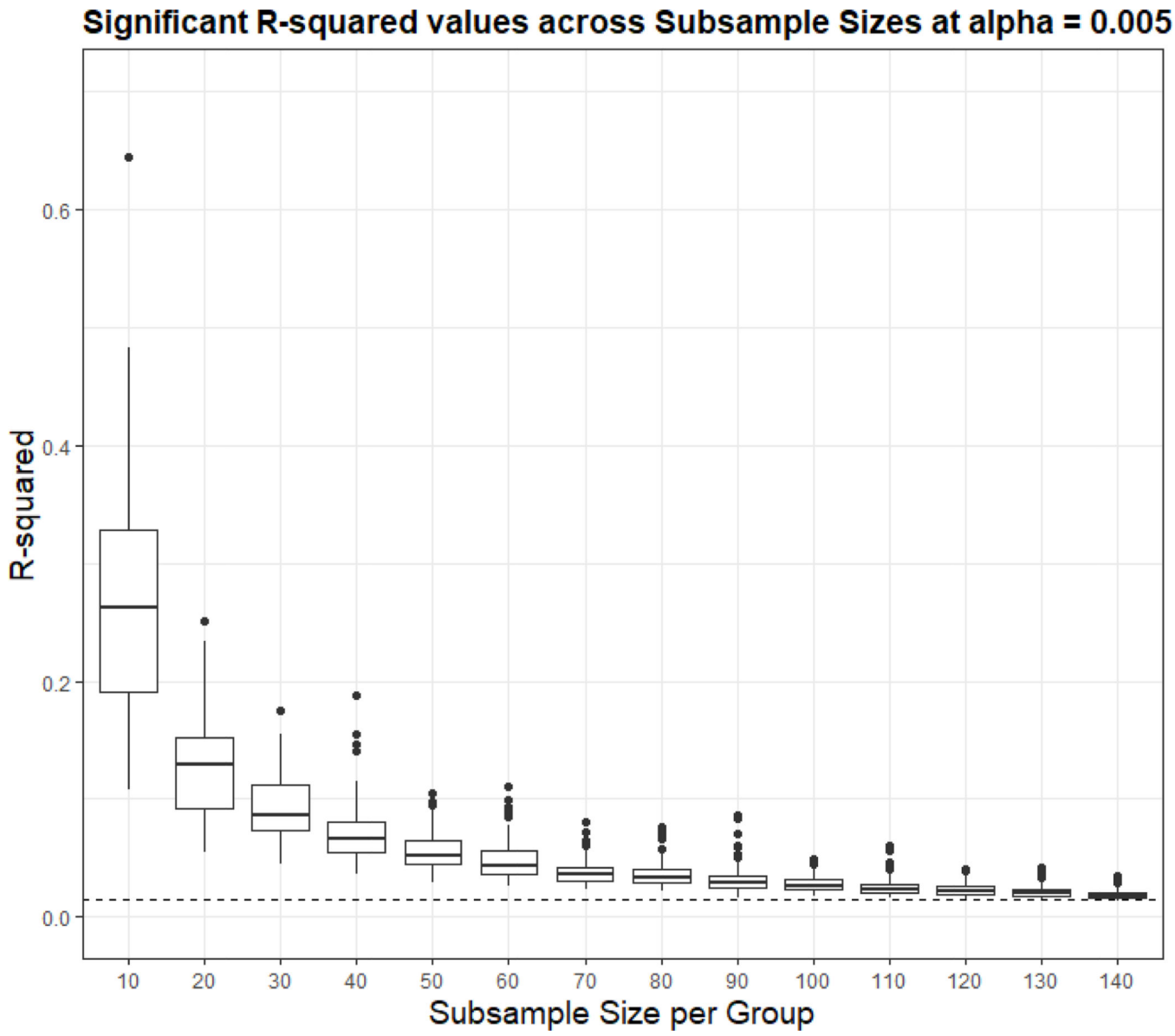

Figure 6.

Boxplots of significant R^2 effect sizes for tests of the RIPL across subsample sizes per group where the stringent critical alpha = 0.005. The average R^2 for samples of 10 per group is clearly inflated relative to higher sample sizes which approach the true full-sample significant R^2 value of 0.015 (dotted line).

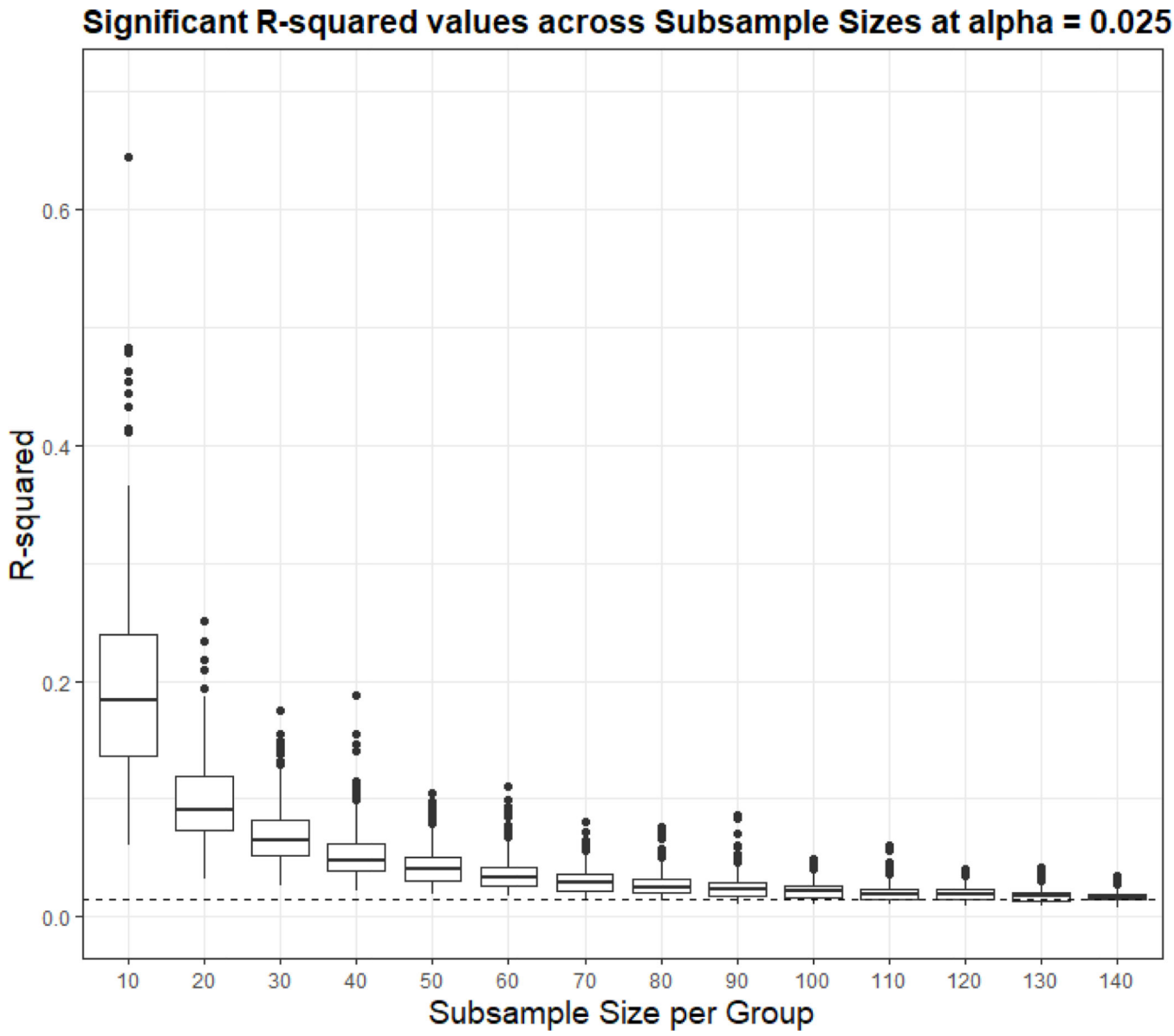

Figure 7.

Boxplots of significant R^2 effect sizes for tests of the RIPL across subsample sizes per group where the lenient critical alpha = 0.025. Again, the average R^2 value for samples of 10 per group is inflated relative to others, though the overall degree of inflation among lower-N groups is somewhat decreased. Higher power due to the more lenient critical alpha relates to slightly more accurate estimates of the true effect sizes.

Table 1.

Year, sample sizes (N), comparison of interest, and mean effect sizes for 15 bilingual-monolingual structural comparisons conducted between 2012 and 2017 (with one conducted in 2004). Where insufficient information was available in a manuscript to calculate a statistic, N/A is given.

	Year	Total N	N Monolingual	N Bilingual	Comparison	Mean Cohen's D significant Effect Size
Abutalebi et al.	2013	28	14	14	B > M	N/A
Abutalebi et al.	2014	46	23	23	B > M	N/A
Abutalebi et al.	2015	38	19	19	B > M	N/A
Abutalebi et al.	2015	60	30	30	Age and AoA	1.67
Berken et al.	2015	34	N/A	34	AoA	1.58
Burgaleta et al.	2016	88	46	42	B > M	0.9
Felton et al.	2017	78	39	39	Asym.*	0.689
Gold et al.	2013	40	20	20	B > M	N/A
Grogan et al.	2012	61	31	30	Mult > B	0.754
Klein et al.	2014	88	22	66	Various*	0.953
Mechelli et al.	2004	83	25	58	B > M	1.57
Olsen et al.	2015	28	14	14	B > M	1
Pliatsikas et al.	2014	39	22	17	M > B	N/A
Ressel et al.	2012	44	22	22	B > M	0.724
Zou et al.	2012	27	13	14	B > M	1.73

Table 2.

Group means and standard deviations for participants averaged within each language group. Standard deviations are given in parentheses. Variables include age (in years), English and Spanish proficiencies (mean accuracy, on a scale of 0–1), and age of acquisition (in years).

	Age	English Proficiency	Spanish Proficiency	Age of Acquisition
Bilingual	23.53 (4.8)	0.74 (0.1)	0.67 (0.14)	8.13 (5.78)
Monolingual	22.72 (4.39)	0.79 (0.07)	NA	NA

Table 3.

Full-sample (216 bilinguals; 146 monolinguals) R^2 effect sizes and p values for each tested grey matter volume ROI from each regression. Specifically, these values are for the bilingual-monolingual comparisons within each regression, while controlling for intra-cranial volume. ACC = Anterior Cingulate Cortex. IFG = Inferior Frontal Gyrus. IPL = Inferior Parietal Lobule. STG = Superior Temporal Gyrus. BG = Basal Ganglia.

ROI	R^2	p value
LACC	0.002	0.30
RACC	0.011	0.028
LIFG	0.002	0.27
RIFG	0.00005	0.87
LIPL	0.001	0.34
RIPL	0.015	0.0008
LSTG	0.0005	0.51
RSTG	0.0000007	0.94
LBG	0.0003	0.70
RB	0.0002	0.74