AN EXPLORATION OF VARIABILITY DUE TO LOW POWER IN STRUCTURAL MRI STUDIES OF BILINGUALISM

A Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

By

Brandin A. Munson

December, 2018

AN EXPLORATION OF VARIABILITY DUE TO LOW POWER

IN STRUCTURAL MRI STUDIES OF BILINGUALISM

_____

Brandin A. Munson

APPROVED:

_____

Arturo E. Hernandez, Ph.D.
University of Houston
Committee Chair

_____

David J. Francis, Ph.D.
University of Houston

_____

Benjamin J. Tamber-Rosenau, Ph.D.
University of Houston

_____

Christine Chiarello, Ph.D.
University of California-Riverside

_____

Antonio D. Tillis, Ph.D.
Dean, College of Liberal Arts and Social Sciences
Department of Hispanic Studies

AN ABSTRACT OF A DISSERTATION ON

AN EXPLORATION OF VARIABILITY DUE TO LOW POWER IN STRUCTURAL MRI STUDIES OF BILINGUALISM

_____

A Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

_____

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

_____

By

Brandin A. Munson

December, 2018

ABSTRACT

The adequacy of replicability among psychological findings has previously been questioned, especially for neuroscientific fields of research. Researchers increasingly point towards the negative effects of low power on replicability of findings. Though decreased sensitivity in smaller samples is a well-known consequence of inadequate power, many overlook the increased likelihood of inflated observed effects and weakened positive predictive values. The aim of this study is to reveal the expected degrees of uncertainty among neuroimaging findings by conducting tests in different sample sizes from a larger-than-average sample, in an area of research with wide-ranging findings that have been proposed by some to be due in part to inadequate sample sizes: bilingual-monolingual structural brain differences. Bilinguals (n = 216) were compared with monolinguals (n = 146) using grey matter density in whole-brain analyses and 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 extent to which infrequently utilized methods such as multivariate analyses of covariance (MANCOVAs) and Bayes Factors can improve the accuracy of results at lower sample sizes were also explored. It is our hope that this study helps to demonstrate the influences of power on expected variability among sample findings, especially for bilingual researchers and any researchers interested in exploring group differences using neuroimaging.

Keywords: replicability, sample size, power, neuroimaging, MRI, bilingualism

An Exploration of Variability due to Low Power in Structural MRI Studies of Bilingualism

Researchers have known and warned about a problem with the replicability 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 not be able to be replicated 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 replicate the 100 studies, 39% of effects replicated 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.

Statistical Concerns

In considering replicability, achieved statistical power among studies often comes into question. Statistical power is the likelihood of successfully finding an effect in a sample when it exists within the population (in other words, the likelihood of rejecting a false null

hypothesis). Therefore, inadequate power results in a greater likelihood of a Type II error, where a true population effect is ignored within the sample. Statistical power increases as the sample size increases, as the observed effect size increases, as standard deviation decreases, for larger (less restrictive) alpha values, and is higher for one-tailed than for two-tailed tests. The likelihood of a Type I error, where a significant effect is found in the sample even though it is not present in the population, decreases as the cutoff alpha value decreases.

Therefore, there is often a 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. The importance of this balance may be somewhat overlooked in the neuroimaging literatures, where (by necessity) a high number of comparisons are commonly controlled for with a more restrictive alpha value. With many comparisons and a small alpha, it becomes necessary to consider other variables in order to reach appropriate levels (often 0.8) of power: sample size, effect size, standard deviation, experiment structure, and various other 'researcher degrees of freedom'. In other words, neuroimaging researchers already do what they can to minimize Type I error, but this is sometimes not adequately balanced with higher sample sizes, which is one of the few researcher options to attain both more stringent alpha values with adequate Type II error rates.

It has been noted (Yarkoni, 2009) that 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-

potentially limiting generalizability to structural MRI studies. Referring to the power to detect effects in between- vs. within-person designs, Yarkoni (2009) stated, "The importance of this point is difficult to overstate: Under reasonable assumptions, the power to detect correlational effects may be as little as 5%-10% of the power to detect similar-sized within-subject effects" (p. 295). Some areas of research, such as those investigating differences between monolinguals and bilinguals, rely on between-group differences by necessity, and therefore start with less power overall than within-group investigations.

Accurate estimates of population effect sizes are necessary in order to adequately predict necessary sample sizes for future studies. However, certain researcher practices (beyond consistently reporting statistical effects for all tests) make accurate effect size estimates very difficult to obtain. 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 $r = 0.6$, a large effect within any psychology literature. This causes observed effect sizes to become greatly inflated.

The failure to detect a true effect is another potential outcome of underpowered studies. Vadillo, Konstantinidis and 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.

Replicability of fMRI Research

In November 2017, Cremers, Wager, and Yarkoni 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. Their findings clearly demonstrate the harmful effects of underpowered samples, especially when combined with studies that contain many comparisons.

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. Interestingly, even their largest sample (N = 150), on average, only detected 9% of the true effects in the population / full sample. Thus, at best, only 12.5% of what researchers often consider to be the minimum acceptable power (0.80) was actually attained.

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. This occurs where tests have both low power and strict alpha corrections. In order for an underpowered sample to be seen as significant when the alpha is very low, the differences have to be even more extreme, and this is often more extreme than what the true population differences are- whether or not those population differences are also significant.

The authors note that a useful hypothesis-driven method to increase power (by decreasing the large number of tests, and thus reducing the necessary alpha correction) is to

use ROIs. This allows for tests to be conducted in areas already supported by the literature, sidestepping the need to check every voxel, as is the case for whole-brain analyses. However, they mention it is important to remember that many ROIs, as are often used in neuroimaging, still necessitate a strict (though not as strict as whole-brain analyses) alpha correction for multiple comparisons – meaning that larger sample sizes are still important in order to achieve ideal levels of power.

One might suggest that, since the power and assumptions of past studies in the neuroimaging literature on bilingualism can be calculated post-hoc and compared to what is recommended, the analysis of a novel sample is unnecessary in order to demonstrate a trend for underpowered studies – or even simulations performed viewing power in samples of fabricated data, as in Cremers, Wagner, and Yakoni (2017) above. However, it has been shown (Sedlmeier & Gigerenzer, 1989) that studies on statistical power alone are unlikely to change trends in statistical methodology. Szucs and Ioannidis (2017) also demonstrate that studies in the psychology and cognitive neuroscience literatures have not shown any improvements in power or effect sizes over the past 50 years, likely due to low sample sizes. The authors support the idea that over 50% of studies in psychology are false-positives, and that this is likely even worse in cognitive neuroscience specifically.

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.

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's 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 and 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, and 3) small sample sizes.

Recall Yarkoni (2009), who 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 replicability 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.

The present author conducted a brief review of 14 studies was conducted in order to better glimpse the average effect sizes found for studies reporting anatomical differences between bilinguals and monolinguals. 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. 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.

How might we 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 'full-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. Beyond exploring variance due to certain 'researcher-degrees-of-freedom' such as sample size and alpha restrictiveness, an extension of the same ideas using either different methods, such as multivariate analyses of covariance (MANCOVAs), or different forms of statistical inference, such as using calculated Bayes Factors in place of traditional effect sizes could prove informative.

Whole-brain Analyses

Whole-brain MRI and fMRI analyses are, in certain ways, more ideal than selecting ROIs. ROI tests use predetermined regions to group voxels according to estimated brain structure locations, whereas whole-brain tests include every brain voxel without pre-grouping

voxels according to brain region. As such, whole-brain analyses are less hypothesis-dependent tests. However, to an even greater extent than for ROI tests with a high number of comparisons, there are drawbacks when this is done without adequate sample sizes.

As Cremers, Wager and Yarkoni (2017) demonstrated using simulated whole-brain fMRI data, "with smaller samples, statistically significant effect sizes appear to be much larger than the true effect sizes" (section 4.2). This results from the combination of small sample sizes, a very large number of statistical comparisons, and stringent alpha corrections (to correct for the many comparisons). However, the authors limited their analyses to both uncorrected $p < 0.001$ thresholds, and without using cluster-based selection (where only results with voxel differences in cluster sizes beyond a pre-determined threshold are considered significant). These two factors were ignored for the sake of a demonstration of the effects of insignificant power.

So, to what extent would these results be consistent in a non-simulated sample using structural MRI? Whole-brain analyses, which often include many thousands of statistical comparisons, allow us to test the extreme in terms of number of comparisons- and ROI analyses, as the focus of this paper, could then be thought of as addressing one major concern of whole-brain analyses, by severely reducing the number of comparisons to selected regions based on previous literature. Higher degrees of inflated effects would be expected for whole-brain analyses than for ROI analyses, due to the difference in number of comparisons. Moreover, the combined usage of family-wise error rate (FWE) correction and cluster-based selection of results within whole-brain analyses should help to lessen the extent of inflated effects across sample sizes.

Bayesian Statistics

Bayesian statistics is a form of statistical inference which stems from Bayes' Theorem (Bayes, Price, & Canton, 1763). The most recognizable difference between Bayesian inference and traditional frequentist approaches to inference is that, in Bayesian inference, researchers begin with a 'prior,' where an educated guess based on past studies or theory is made about the conditions relating to the event of interest before the actual test is conducted. The prior affects the likelihood of the outcomes, and therefore the results of the test itself. Following the definition of a prior, models are used to predict the prior and then compared in terms of their explanatory value, resulting in a 'Bayes Factor' which gives the relative weight for one model versus another.

In the context of the present project, it would be informative to determine whether Bayesian modeling shows the same or similar influences of sample and test characteristics on power and effect size as frequentist approaches. Using a Bayesian-defined model both with and without a variable of interest (here, bilingual or monolingual language status) would allow for Bayes Factors to be calculated, which could then be explored in terms of accuracy and potential degree of inflation at lower effect sizes. This should help to clarify the utility or risks of Bayesian inference with analyses involving low sample sizes.

Multivariate Analyses

There are certain alternative statistical analyses which may be used in place of individually testing each ROI in a univariate test, in order to conserve power. Specifically, multivariate analyses of covariance (MANCOVAs) might be used. Multivariate analyses (Anderson, 1958; Morrison, 2005) are statistical tests with more than one dependent variable.

Here, rather than testing every single ROI to determine whether specific regions are significantly related to predictors, the volumes for each ROI are included as dependent variables within the same test.

Multivariate tests reveal whether the predictors, including language status (what we are interested in), explain a significant amount of variance in brain volume across selected brain regions as a whole. Importantly, though dependent upon the inter-relatedness of the dependent variables (Cole, Maxwell, Arvey, & Salas, 1994), the power for conducting a single multivariate test is greater than conducting a single univariate test (such as, in the case of this paper, a univariate multiple regression) for each dependent variable. So, in order to determine the extent to which inferences and accuracy would be improved through the use of multivariate tests, the eta-squared estimates, standardized and structural coefficients, and accuracy of MANCOVAs across samples were also investigated. Interestingly, multivariate methods are very infrequently utilized in fMRI or MRI research, with one exception being Batty et al., 2010. The present authors are unaware of any neuroimaging analyses of bilingual differences which make use of multivariate methods.

Hypotheses

The present study treated the large sample of monolingual and bilingual structural MRIs as the population, in order to determine variability in randomly sampled results across sample sizes and statistical methods.

ROI Univariate Regressions

1. Greater accuracy, here referring to fewer false positives, more true positives, and fewer false negatives, will be observed at higher subsample sizes than at lower subsample sizes.
2. Sensitivity, or power, as well as PPV, will increase with subsample sizes.
3. Stricter alpha corrections will decrease the total number of false positives, but increase the number of false negatives, across subsample sizes.
4. Sensitivity, or power, will be smaller, and PPV will be greater, for stricter alpha corrections.
5. Inflated effects relative to the full-sample effect size will be seen for smaller subsample sizes.
6. More inflated effects will be seen when alpha corrections are stricter.

Whole-Brain T-tests

7. Whole-brain analyses will show a greater degree of inflated effects at smaller subsample sizes relative to ROI analyses.
8. Whole-brain analyses with results selected above 20-voxel cluster thresholds will show a lesser degree of inflated effects at smaller subsample sizes relative to whole-brain analyses without using a cluster-based threshold.

MANCOVAs

9. Sensitivity, or power, as well as PPV, will be greater for MANCOVAs than for univariate multiple regressions where the alpha threshold is $p = 0.025$.
10. Degree of inflated effects at smaller subsample sizes will be smaller for MANCOVAs than for univariate multiple regressions where the alpha threshold is $p = 0.025$.

Bayes Factors

11. Measures of test accuracy will be negatively influenced at smaller subsample sizes to a lesser extent when Bayes Factors are used to determine test outcomes than where p-values are used to determine test outcomes.
12. Degree of inflated effects will occur at smaller subsample sizes to a lesser extent when Bayes Factors are used to determine test outcomes than where p-values are used to determine test outcomes.

Method

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. There were originally 376 participants in all 8 studies, but some scans were not usable due to reconstruction issues (7), technical MRI scanning issues (6), or brain trauma (1). 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. All participants were screened for background factors incompatible with MRI. See *Table 2* in the appendix 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.

Voxel-Based Morphometry (VBM)

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 Magnetization Prepared Rapid Gradient Echo (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 x 0.479 x 1.0 mm, 192 slices.

The T1 scans were preprocessed through modulated normalized segmentation in order to create measures of grey matter volume, and without non-modulated normalized segmentation in order to create measures of grey matter density, 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 VBM, and both intracranial volume (ICV; used to control for overall brain size) and ROI volume data values extracted, using the SPM Computational Anatomy Toolbox (CAT12) SPM package (http://www.neuro.uni-jena.de/cat/).

ROI Analyses

ROI grey matter volume data was extracted from SPM and 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 ICV as predictors were conducted on each of 10 ROIs. 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 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 subsamples, smaller true effects and more restrictive alphas) were graphed and summarized. The calculation of a 'confusion matrix' (*Figure 1*), which groups the percentage of significant or non-significant subsample 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 subsamples. 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 subsample 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, a more stringent Bonferroni-corrected cutoff ($p = 0.005$ for 10 total comparisons), and one less stringent ($p = 0.025$) – in order to explore differences in findings across alpha cutoff stringency.

Whole-Brain t-tests

Whole-brain family-wise error rate (FWE)-corrected t-tests comparing grey matter density between monolinguals and bilinguals were conducted for random subsamples of the same description as in the *ROI Analyses* section. Test significance, voxel cluster sizes, and peak voxel *z* scores were output by SPM for each of the random subsamples. A threshold of 20-voxel clusters was selected a priori as the cutoff for a cluster of meaningful difference.

In order to approximate the degree of overlapping clusters between subsample clusters and full-sample clusters, spheres of varying sizes, centered around peak cluster voxel differences, were manually calculated in R. The spheres varied in size proportionately with the number of voxels within each cluster, such that roughly the same number of voxels were within each sphere. A test was conducted to determine whether spheres from subsamples overlapped with those from the full sample (the 'true effect'), and returned a 'True Positive' results for an overlap. Any subsample sphere overlap with a full-sample sphere would count

as a single 'True Positive' instance, and more overlapping spheres were not counted as additional instances. If a cluster from a subsample was significant but did not overlap with any from the full sample, this was considered a 'False Positive.' The total number of full-sample spheres without any subsample overlap per subsample were counted as 'False Negatives.' Both the PPV and sensitivity were then calculated.

Peak *z* scores within each significant cluster were modeled across subsample sized in order to determine the degree of inflation in estimated effects. As an exploration of the influence of a cluster-based threshold on observed effect sizes, peak *z* scores were also explored across subsample sizes without using a 20-voxel cluster threshold cutoff. This ought to demonstrate a much higher degree of effect size inflation in lower subsamples than when a cutoff is used.

In order to determine the effect of sphere size upon estimations of overlapping clusters, spheres with roughly the same number of voxels as the original cluster around which they are centered were used in one exploration, and spheres with roughly twice the number of voxels were used in another.

Also, due to the time-intensive nature of these analyses, whole-brain tests were only conducted on 250 random subsamples per sample size. The results nonetheless demonstrate relatively smooth transitions in averages and counts between subsample sizes, suggesting that this is not an insufficient size of subsamples to accurately gauge changes in estimates with sample size.

Bayes Factors

Within the R statistical software, the BayesFactor package (Morey, Rouder, & Jamil, 2015) allows for Bayes Factors to be calculated from general linear models which have been created and compared with other general linear models. That is, a model which includes a term of interest, in this case *ROI Volume ~ Language Status + ICV*, compared with a model which does not, in this case *Volume ~ ICV*, allows for the relative contribution of *Language Status* alone towards explaining the variance in Volume to be calculated in terms of a Bayes Factor. Using this package, Bayes Factors were calculated for the Language Status term, while using the default Jeffreys' prior (1961) for each tested model. This was done following the same randomized sample data selection process which is described in the *ROI Analyses* section.

Of interest is 1) the accuracy of Bayes Factors when a predefined cutoff is used, and 2) the extent of inflation, if any, for Bayes Factors at lower subsample sizes. The original cutoff chosen prior to testing was a Bayes Factor equal to 10, as this is described in Jeffreys (1961) as the cutoff between 'strong' and 'very strong' support of the evidence. See *Table 3* for Jeffreys' (1961) scale of strength of evidence when interpreting Bayes Factors. However, in order to ensure consistency in the number of full-sample ROIs that were considered significant as with the *p*-value cutoffs for the original linear models, this was slightly adjusted to a Bayes Factor of 15.

It should be noted that it is not recommended to create a cutoff between 'significant' and 'non-significant' when unnecessary, and the continuous nature with which Bayes Factors are treated is considered an advantage over the sharply separated p-value cutoffs- similar to effect sizes. However, this cutoff is being used in order for a confusion matrix to be created

which can intuitively display the degree of accuracy in subsamples to predict the state of the full sample.

Multivariate Analyses of Covariance

For each subsample across subsample sizes, a single Multivariate Analysis of Covariance (MANCOVA) with the predictors of language status (monolingual or bilingual) and ICV was conducted, using the same 10 ROIs described in the *ROI Analyses* section as outcome variables: bilateral STG, BG, ACC, IFG, and IPL. The true positive rates were explored, along with eta-squared effect size estimates, and both the standardized and structure discriminant function coefficients. The R candisc package (Friendly & Fox, 2017) was used to extract eta-squared values and discriminant function coefficients.

Eta-squared are similar to r-squared values, with the difference being that the proportion of variance being explained by the predictor (language status) includes all of the outcome variables together. The standardized discriminant function coefficients parallel beta weight coefficients in univariate ANOVA, where a larger observed coefficient suggests a stronger relationship between individual outcome variables (each ROI) and the predictor of interest, language status. The structure discriminant function coefficients can be thought of as factor loadings of individual outcome variables (each ROI) upon the discriminant function itself. This can help to give researchers an idea of the characteristics of the discriminant function which was selected by the analysis (Bray & Maxwell, 1985; Smith, 1958).

Network Analyses

As an extended exploration into different treatments of the dependent ROIs, two separate combinations of ROIs were created as approximations of regions associated with

certain cognitive function 'networks': a cognitive control network, which includes the LIPL, bilateral BG, and bilateral ACC; and a language network, which includes the LSTG and LIFG. Both were taken as combinations of regions from the original 10 ROIs, which had been selected based on findings from past bilingual-monolingual neuroimaging studies. Cognitive control network regions were selected based on Abutalebi and Green (2007), while language network regions were selected based on Hickok and Poeppel (2004). A single measure of brain volume was created for each network by summing the ROIs associated with each.

The same linear regression and MANCOVA analyses described above were conducted on these network estimates of overall volume within regions related to each network (cognitive control and language). Because multiple ROIs were summed, this means that a single univariate regression can test differences in the total volume across ROIs within each network, rather than one ROI at a time, which increases the critical alpha in order to correct for multiple comparisons. The MANCOVA analyses were again single tests, this time including only 1) the cognitive control network ROIs, and 2) the language network ROIs. Language status and ICV were again the predictors of both the regressions and MANCOVAs.

Results

Whole-brain Full-sample T-tests

T-tests comparing grey matter density between bilinguals and monolinguals revealed a large number of FWE-corrected significantly different clusters greater than 20 voxels in size, where bilinguals show greater density than monolinguals. Information on the 22 significant clusters can be seen in *Table 4*, including region name, cluster sizes, peak voxel Z statistic, and MNI coordinates of the peak voxel differences. No FWE-corrected full-sample differences were significant, regardless of cluster size, where monolinguals show greater density than bilinguals. Because the subsample significance (FP) rates were so low across subsample sizes for comparisons of monolinguals greater than bilinguals, with a range of about 2 percent significance at the lowest subsample sizes to about 11 percent significance (with mostly single small clusters) at the largest subsample sizes, these were not further explored.

A number of the 5 bilateral regions selected for the ROI analyses were also found to have significantly greater volume in bilinguals than in monolinguals in the whole-brain t-test comparisons. These include the bilateral IPL, right IFG, and right STG. The left dorsolateral prefrontal cortex (DLPFC) was also a significant cluster, which has been previously found in fMRI studies of differences in bilingual and monolingual abilities to switch tasks (Hernandez, Martinez, & Kohnert, 2000). The 5 largest clusters surrounded the bilateral anterior temporal lobes, right posterior cingulate gyrus, right IFG, and left primary visual cortex. These also include 5 of the 6 most significant voxel differences.

Whole-brain T-test Accuracy

The accuracy of subsamples to predict the full-sample results was assessed by simplifying and measuring the number of overlapping subsample and full-sample clusters. Using spheres of voxels around each voxel of peak significance within each cluster (given in *Table 4*; spheres consisted of half the number of voxels as the original cluster), the number of full-sample clusters which overlap with any cluster from a subsample were counted as TPs. Missed full-sample clusters were counted as FNs, and subsample clusters which did not overlap with any full-sample clusters were counted as FPs.

Figure 2 shows the counts of these whole-brain cluster-specific outcomes for each subsample. For subsample sizes below 50 or so per group, there is very little change in either the number of TPs or FPs. Above 50 per group, TPs increase at a faster rate, but FPs are, interestingly, also seen to slightly increase with subsample size.

Figure 3 shows those same counts in the form of PPV and sensitivity. As one would expect from the previous figure, sensitivity of whole-brain tests does not noticeably increase until roughly 60 participants per group. PPV increases at a fast rate, reaching a maximum value of 0.80 at 120 / 130 participants per group. However, it is important to note that this cutoff lies beyond the threshold of the average sample size seen in bilingual neuroimaging studies- which, as calculated from the selected studies earlier, lies at roughly 26 participants per group.

Interestingly, if the sizes of spheres are greatly inflated to include 20 times as many voxels as the original clusters, similar rates of test outcomes are seen (*Figure 4*). This suggests that the test of overlapping peak voxels is not necessarily insensitive to see what

would otherwise be overlapping clusters, but instead that underpowered subsamples are in fact inadequately finding clusters that are truly present in the full sample. Sensitivity in the lenient overlapping tests are the same, while PPV does increase significantly faster (*Figure 5*). This suggests that the clusters are simply not being found at all (with a cutoff of 20 voxels per cluster) in the subsamples- while all full-sample clusters that are present in each subsample are being found (thus the high PPV), sensitivity remaining low means that this is a small proportion of the full-sample clusters that should be found.

Whole-brain Effect Size Inflation

Contrary to what was expected based on Cremers, Wager, and Yarkoni (2017), who showed inflated effects in small sample sizes for simulated fMRI data, the effect sizes of peak voxel differences are not inflated at lower subsample sizes (*Figure 6*). In fact, it appears that higher subsample sizes generally show higher peak *z* score voxel differences, which is what would be expected with increased sample sizes generally- especially for true effects.

One major difference with Cremers, Wager, and Yarkoni (2017) is that they did not perform corrections for multiple comparisons, nor did they create a cluster size cutoff for significance. The authors instead used a general alpha = 0.01 significance threshold for descriptive purposes. Therefore, in order to determine the extent to which the cluster cutoff influenced our estimates of inflated effects, a peak voxel *z* scores across subsamples were also explored where all significant clusters were included (*Figure 7*). Ignoring a cluster cutoff does take away much of the difference in effect size, but there is still no hint of inflated effects in smaller subsample sizes. If there are differences across subsample sizes, it

appears that larger subsamples have more large effects than smaller subsamples- though this doesn't seem to substantially change the median value.

Whole-brain t-tests were also conducted where only a general alpha = 0.01 was used, and this demonstrates even an even stronger difference in peak Z-score effect sizes across subsample sizes, in the opposite direction than expected (*Figure 8*). It is possible that this relates to the very strong whole-sample effects, where bilinguals show much greater grey matter volume than monolinguals, whereas Cremers, Wager, and Yarkoni (2017) found the strongest amount of inflation to be in a whole-brain effect which was small but dispersed throughout many brain regions.

Full-sample ROI Univariate Multiple Regressions

Table 5 contains the R-squared effect sizes and *p* values for the linear multiple regressions which included all 362 participants. The full-sample results, which represent the 'true' outcomes to which subsamples are compared, reveal volume in the RACC to be predicted by language status at $p = 0.028$ ($t = 2.21$), and in the RIPL at $p = 0.0008$ ($t = 3.37$), when controlling for ICV. Each of the other eight ROIs were non-significant with p's > 0.20. Because alpha cutoff corrections for multiple comparisons were chosen a priori as $p = 0.005$ (Bonferroni correction; more stringent) and $p = 0.025$ (less stringent), the full-sample RIPL is the only ROI which survives correction. This is therefore the only region of the 10 which serves as a 'correct' detection of a significant full-sample effect.

ROI Replication Rates Using a Cutoff Value

Figure 9 shows changes in the accuracy of findings across subsample sizes. For each of the 10 ROIs, the amount of variance contributed by language status (bilingual or

monolingual) is either significant ($p < 0.005$) or non-significant ($p > 0.005$). 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 subsample findings are called 'Positive', and non-significant subsample findings are called 'Negative'. If the subsample 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 subsample. 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 subsample. 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 subsample. 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 subsample. See *Figure 1* for a simple visualization. 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 subsample 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 subsample 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 subsample sizes. Figure 9 is cut off at 1,000, but the true total number of tests for each subsample size is 10,000.

ROI Multiple Regression Accuracy

Figure 9 demonstrates that, at a Bonferroni-corrected critical cutoff of $p = 0.005$, as subsample size increases, the likelihood of finding a true positive effect also increases, the likelihood of finding a false negative decreases, and the likelihood of finding a false positive is stable, with a slight decrease. As we would expect, increasing the subsample size of a statistical test has positive effects on the accuracy of that 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 subsample sizes. At the lowest subsample 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 subsamples with 70 or more participants in each group are attained.

What is often thought of as a minimum level of power, 0.80, is not even achieved with the largest subsamples 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 ROIs, 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 10*, which shows the same accuracy metrics for subsample vs. full samples as *Figure 9*, with the only difference being that the threshold of significance was changed from $p = 0.005$ in *Figure 9* to $p = 0.025$ in *Figure 10*.

Figure 10 demonstrates an increase in the rates of True Positives across all subsample sizes, especially as subsample sizes increase- since the threshold to significance is lower, it is more likely to find a truly significant difference in the random subsamples. 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 subsample sizes increase. This means that at the highest subsample 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 subsamples 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 subsample sizes, they would be *much* more likely to reach such a misleading conclusion.

Positive Predictive Value (PPV), or the number of True Positive findings out of the total number of positive subsample 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 11*, with a critical alpha cutoff of $p = 0.005$,

shows that although the PPV increases with subsample size, it is very unlikely (about a 25% chance) in many of the smaller subsample 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 subsample will return a positive (significant) result when it should. In the context of this test and these ROIs, the only positive full-sample outcome is the RIPL. So, here, sensitivity refers to the likelihood of a subsample finding a significant different in the RIPL. *Figure 12*, again with a critical alpha cutoff of $p = 0.005$, shows a dismal sensitivity across subsample sizes for a test to find a significant difference in the RIPL, where it should be found.

ROI Effect Size Inflation

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 subsample 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 subsample 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 subsamples which are small than from large subsamples.

The present data reflected the phenomenon described in the above paragraph. Looking at the variability in effect sizes across subsamples, *Figure 13* shows that as subsample size increases, the average *significant* observed R-squared effect size (where the Bonferroni-corrected alpha = 0.005) decreases in size, especially for subsamples 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-squared 0.015 for the only statistically significant difference in the RIPL comparison (see *Table 5* for all full-sample test effect sizes). So, smaller subsamples are more likely to see inflated effects when significant, and observed effect sizes asymptotically approach the true full-sample effect size as subsample 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-squared effect size. *Figure 14*, shows that with a less stringent alpha cutoff, smaller subsample sizes differ less in average observed R-squared values relative to larger subsamples. So, increased power that results from less stringent alpha cutoffs does lead subsamples 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 subsample sizes, and more consistent and powerful statistical methods across studies.

Full-Sample ROI Bayes Factors

Table 6 contains the Bayes Factors for the model comparisons which included all 362 participants. These full-sample results, which again refer to the 'true' outcomes to which the subsequently explored subsamples are compared, show that the relative evidence *substantially-to-strongly* favors a model with RACC as the outcome (Bayes Factor = 10.40), while evidence *decisively* favors a model with RIPL as the outcome (Bayes Factor = 243.60). All other ROIs resulted in Bayes factors between 1.0 and 2.0, which is considered evidence *barely worth mentioning*. This mirrors the linear multiple regression results, which also showed the RACC and RIPL to be significant at 0.05, and only the model predicting volume in the RIPL to be significant at 0.005. Therefore, this is again the only region of the 10 which serves as a 'correct' detection of a 'significant' full-sample effect.

ROI Bayes Factor Accuracy

The use of Bayes Factors was shown to be very similar to the original regressions in terms of degree of accuracy towards predicting outcomes in the full sample, when results for a Bayes Factor cutoff of 15 is compared with those for alpha value cutoffs of 0.025. *Figure 15* demonstrates a very comparable plot of the rate of true positives, false positives, and false negatives across subsample sizes using Bayes Factors relative to the standard linear multiple regressions at a 0.025 alpha cutoff (*Figure 10*). TP rates are nearly identical, but the average FP rate is overall lower- from an average of 230 FPs per 10,000 to 125. Likewise, *Figure 16* shows extremely similar PPV and sensitivity curves across subsample sizes relative to *Figure 12*. In terms of accuracy of effects detected (or not) at a cutoff, the primary difference in accuracy for using a Bayes Factor cutoff of 15 vs. an alpha cutoff of 0.025 appears to be similar to a simple difference in cutoff stringency. That is, a slightly more stringent alpha

cutoff, such as between 0.025 and 0.005, would have a similar effect where both FPs and TPs are slightly decreased.

ROI Bayes Factor Effect Size Inflation

On the other hand, *Figure 17* demonstrates substantially different influences of subsample sizes for Bayes Factors. While effect sizes were clearly inflated beyond the largest full sample effect sizes in linear multiple regressions (Figures 6 and 7), Bayes Factors were not inflated in smaller subsample sizes (*Figure 17*). In fact, sizes of Bayes Factors only reach the size of the truly 'significant' full-sample Bayes Factor of 243.60 (RIPL) a handful of times in the smallest of subsample sizes (group subsample sizes = 10 or 20), and overestimations of the size of Bayes Factors occur more frequently from there. This suggests that the use of Bayes analyses actually somewhat underestimates the sizes of the Bayes Factors in smaller subsamples.

ROI Full-sample MANCOVAs

MANCOVAs, which in this context treat multiple ROIs as dependent variables in a single test, were also explored. Within the full sample, there was a statistically significant difference between bilinguals and monolinguals on the combined ROI dependent variables after controlling for ICV, $p(10, 350) = 0.0002$, $F = 3.47$, Pillai's trace = 0.09. The obtained eta-squared for ICV is 0.69, and the obtained eta-squared for language status 0.079. *Table 7* shows both the standardized and structural coefficients for language status in the full sample MANCOVA test. Consistent with the univariate multiple regression results, the RIPL has the largest of both standardized and structural coefficients, and the RACC is estimates among the second largest structural and third largest standardized coefficient (very slightly below

LIPL), demonstrating the strength of relationship between each specific ROI and language status.

MANCOVA Accuracy

Because only one outcome is true in a single MANCOVA test, each subsample outcome was compared to the significant full-sample result. *Figure 18* therefore shows only the rate of TPs and FNs across subsample sizes, both of which are a difference of the opposite measure from the total number of subsamples per group (1000).

It is clear that the rate of true outcomes is much more favorable where the MANCOVA was used. At about 60 participants per language group, an equal number of true (TP) and false (FN) outcomes are seen, and beyond this a consistently linear approach to 100% TP outcomes is seen. However, samples of about 20 or 30 per language group are only 25% likely to show the significant finding. This means that the most commonly used sample sizes in structural neuroimaging studies of bilingualism, while having improved overall accuracy relative to many univariate regressions, are still not consistently finding a strong positive outcome in the full sample.

MANCOVA Effect Size Inflation

Eta-squared, a similar measure to r-squared in univariate regression tests, was explored here to examine the consistency of effect sizes attributed per subsample size to language status in the MANCOVA test. *Figure 19* shows a substantial amount of increased inflation in smaller subsample sizes for MANCOVA tests than was observed in even the more stringently-corrected univariate regression tests (*Figure 13*). While r-squared estimates in univariate tests are very inflated for 10 participants per language group, they stabilize

(decrease to 0.10 and below) for subsamples with 20 and more participants per group. Meanwhile, the eta-squared estimates in MANCOVAs are even more inflated at the lower subsample sizes, with nearly 75% of subsample eta-squared estimates resting above 0.40 in size where subsamples include 20 participants per language group, and estimates being about double the true effect size on average even at roughly 50 participants per group.

This could be due to the fact that the MANCOVA, as a single test with a high eta-squared estimate (0.079), is more likely to show inflated effects than the many univariate tests, of which only two (RIPL and RACC) r-squared slightly exceeded 0.01 in size. This appears to present a divergence in terms of the overall likelihood of an accurate test vs. the degree of inflated effect sizes. While power is preserved in multivariate tests and this allows for a higher chance of an accurate test outcome, the inflation of multivariate eta-squared effect sizes appears to be greater than the inflation of univariate r-squared effect sizes.

MANCOVA Coefficient Consistencies

MANCOVA coefficients allow for interpretations of the degree of estimated influence of each ROI towards a calculated discriminant function. The consistencies of these coefficients were explored across subsample sizes, in the form of visualizing both moderately large and very small ROI values for both standardized (*Figure 20*) and structural (*Figure 21*) coefficients. *Table 7* includes each of the estimated coefficients for all 10 of the ROIs.

Figure 20 demonstrates that, for a very small coefficient estimate (LACC), while the interquartile range is consistently around the true value (-0.02), the amount of variability decreases as subsample sizes increase. Meanwhile, for a moderately large coefficient estimate (LIPL), estimates in the smallest sample sizes begin by underestimating the true size

of the coefficient (0.55); for samples with 60 or fewer participants per group, about 75% of the subsamples underestimated the coefficient size. Estimates normalize around the true value as subsample sizes increase.

Figure 21 shows similar trends, this time for structural coefficients. Again, a very small coefficient estimate (LBG) is centered near its true value (0.02) across subsample sizes, and decreases in the amount of variability around this value as subsample sizes increase. Meanwhile, a moderately large coefficient estimate (RACC) underestimates the size of the coefficient in smaller subsamples, and approaches the true value (-0.40) as subsample sizes increase. Again, more than 75% of subsamples underestimate the size of the true effect, and these appear to require a larger number of participants per group to normalize around the true size than was seen for the LIPL in *Figure 20*.

Brain Network Full-sample Analyses

While ROI analyses are commonly approached as separate univariate analyses in the bilingual neuroimaging literature, one potential way to treat the many ROIs may be to combine them into meaningful groups based on overlapping cognitive functions. The chosen networks were a cognitive control network, consisting of the LIPL, bilateral ACC, and bilateral BG, and a language network, consisting of the LIFG and LSTG. In order to create aggregated measures of volume in these regions, the selected ROIs were summed into single 'network scores' which were investigated to indicate differences in total volume in regions considered important to these cognitive abilities. Each of these network variables were investigated in univariate regressions, as well as in MANCOVAs where the ROIs were included as individual dependent variables.

The full-sample univariate multiple regression results reveal neither an aggregated cognitive control network measure (p = 0.20, R-squared = 0.002) nor an aggregated language network measure (p = 0.67, R-squared = 0.0002) to be significantly predicted by language status when controlling for ICV. *Table 8* contains the estimated mean differences, test statistics, p values, and R-squared effect sizes for both network full-sample univariate regressions.

The MANCOVA results, meanwhile, found a significant amount of variability to be explained in only one of the chosen networks. Within the full sample, there was a statistically significant difference between bilinguals and monolinguals on the cognitive control-related ROI dependent variables after controlling for ICV, $p(5, 355) = 0.033$, F = 2.47, Pillai's trace = 0.034. The obtained eta-squared for ICV is 0.59, and the obtained eta-squared for language status 0.019. *Table 9* shows both the standardized and structural coefficients for language status in the full-sample cognitive control network MANCOVA test. RACC, the second-most influential dependent variable towards the discriminant function of all 10 ROIs, becomes most influential using both standardized and structural coefficient measures among the cognitive control network ROIs (where RIPL is removed). There was *not* a statistically significant difference between bilinguals and monolinguals on the language-related ROI dependent variables after controlling for ICV, $p(2, 358) = 0.093$, F = 2.39, Pillai's trace = 0.013. The obtained eta-squared for ICV is 0.59, and the obtained eta-squared for language status 0.007. *Table 10* shows both the standardized and structural coefficients for language status in the full-sample language network MANCOVA test.

Brain Network Regression Accuracy and Effect Size Inflation

Figure 22 shows the accuracy of subsample outcomes for univariate regression tests of the aggregated cognitive control network ROI measure. Because the true full-sample effect is non-significant, only FPs and TNs are possible outcomes. It is clear that there is little change across subsample size, with slightly less than 10% of all outcomes being untrue FPs until the highest subsample sizes, when a slight decline can be seen. *Figure 23* shows the R-squared effect size distributions across subsample sizes for univariate regression tests of the cognitive control network measure. A similar trend of inflated effects in the smallest subsample sizes can be seen as with *Figure 14*, which shows R-squared effect sizes for all 10 univariate ROI regression tests across subsamples.

With similar full-sample outcomes, the exploration of subsample accuracy and effect sizes do not greatly differ for aggregated language network ROI outcomes. *Figure 24* shows the accuracy of subsample outcomes across subsample sizes for language network univariate regression tests, and here we again see little change across subsample sizes. The total number of FPs does seem to be smaller relative to cognitive control network tests, which may be due to an extremely small effect size (R-squared = 0.0002), even relative to the already-small effect size observed in the full-sample cognitive control network test (R-squared = 0.002). *Figure 25* also shows an extremely similar trend for inflated R-squared effects in the smallest subsample sizes as was seen in *Figure 23*. While the use of a single outcome variable for each of the networks of interest allows for alpha corrections for multiple comparisons to be avoided, the lack of full-sample significance make deeper interpretation of the influences of subsample size difficult.

Brain Network MANCOVA Accuracy, Effect Size Inflation, and Coefficients

The same ROI networks were explored using MANCOVAs, this time without first combining the ROIs into a single measure. The MANCOVA tests overall relatedness of the predictors, here language status and ICV, with each of the outcome variables, here whichever of the ROIs are entered in as dependent variables. This is a more holistic test of whether the predictors influence the outcome variables, and is truer to considering the variance between each of the outcome ROIs than when they are simply summed into a single aggregate measure, as was done in the univariate multiple regressions.

Figure 26 shows the accuracy of subsample outcomes for MANCOVA tests of cognitive control-related ROIs. Because this was the only significant full-sample brain network test, outcomes in subsamples are either TPs or FNs. Though the number of TPs is lower than is ideal for many of the subsample sizes (below 25% TP until 80 participants per group and above, and never reaches 50% TP), the accuracy of outcomes (number of observed TPs) does increase linearly across subsample sizes. Consistent with the MANCOVA tests of all ROIs (*Figure 19*), *Figure 27* shows that the Eta-squared effect sizes are inflated in the lower subsample sizes, though to a lesser degree. Also consistent are the changes in both standardized (*Figure 28*) and structural (*Figure 29*) MANCOVA coefficients across subsample sizes. Again, the more influential ROI (here, RACC in both standardized and structural coefficients) is seen to be underestimated in smaller subsample sizes.

Figure 30 shows the accuracy of subsample outcomes for MANCOVA tests of language-related ROIs. A non-significant full-sample finding means that subsamples are either FP or TN, and a relatively high rate of FPs across subsample sizes, at just below 25% of outcomes, can be seen. The number of FPs do seem to actually, very slightly, increase

across subsample sizes. Eta-squared effect sizes, shown in *Figure 31*, appear to only be slightly inflated in lower subsample sizes, and again approach the true full-sample effect size estimate as subsample sizes increase.

Figure 32 shows changes in estimates of standardized coefficients, and *Figure 33* changes in structural coefficients, across subsample sizes for LIFG and LSTG, the only two ROIs in the language network. Standardized coefficients are consistent with earlier displays of changes in moderately large coefficient accuracy across subsample sizes, but structural coefficients appear to differ. For both LIFG and LSTG, there is a very large amount of variability in structural coefficient estimates across subsample sizes, and this decreases somewhat with increases in subsample sizes. However, structural coefficient estimates of LIFG are uniquely overestimated in the smallest subsample sizes, and actually somewhat decrease to approach closer estimates as subsample sizes increase. This is likely related to the extreme variance for both estimates, with the entire range of coefficients being included within 1.5 standard deviations of the median value, which itself lies above the true coefficient value for the LIFG in the majority of subsample sizes.

Discussion

The present study illustrates the inaccuracies which might be expected from underpowered samples in neuroimaging, specifically when investigating 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 for any study which uses the frequentist statistical approach to experimental testing. Also explored are a number of infrequently utilized options for exploring group differences in more powerful ways, which has revealed certain benefits as well as potential drawbacks.

Whole-brain T-tests

The whole-brain t-tests showed that monolinguals did not have greater grey matter volume in any brain regions relative to bilinguals, while bilinguals had greater grey matter in 22 separate clusters. This supports previous findings that have found bilingual experience to be related to greater grey matter volume.

However, the finding was not present in smaller subsamples whatsoever, regardless of the leniency of the test- no clusters greater than 20 voxels in size survived the FWE correction. TP clusters were more consistently found as subsample sizes increased, with the overall FP rate only slightly increasing. This is a demonstration of the importance of using adequate sample size when strenuous whole-brain FWE correction is used. Until roughly 70 participants per group were included in the whole-brain t-tests, no overlapping clusters were found between subsamples and full-samples, regardless of the leniency of overlap testing. The main influence of increased test leniency was an improved PPV at moderately large

subsample sizes (30-70 participants per group), though this is simply indicating improved accuracy for the few clusters which were found to be significant at these subsample sizes.

The inflation of peak cluster effect sizes is unexpectedly absent from whole-brain findings, regardless of whether clusters were cut off at greater than 20 voxels per cluster or not. This is likely due to FWE correction being used, which was not used (for descriptive purposes) in Cremers, Wager and Yarkoni (2017). This suggests that, for whole-brain t-tests of brain density comparing groups, such effect size inflation may not be as much a problem as with uncorrected fMRI testing.

ROI Univariate Multiple Linear Regressions

Full-sample univariate multiple linear regressions which separately tested language status for each of the 10 ROIs showed that the RIPL was the only region which survived Bonferroni corrections ($p = 0.0008$), with the RACC approaching significance ($p = 0.028$). The RIPL replicates a finding by Grogan et al. (2012), though Grogan et al. had found this region when using grey matter density as an outcome, and not grey matter volume. The RACC is not quite significant, but is consistent with grey matter volume findings by Abutalebi et al. (2015).

Several regression tests of ROIs resulted in the expected lack of accuracy due to insufficient sample sizes. 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- not because of increased FPs, but because of an increased proportion of FPs relative to the number of TPs. Because bilingual neuroimaging studies lie

on the smaller end of the explored subsample sizes at an average N = 26, this paints a picture of inadequate sample sizes clearly

Beyond this, smaller samples which result in significant findings are more likely to present inflated effect sizes. This is true to an even greater extent where more stringent alpha corrections are used. ROI analyses do not appear to increase power enough to improve inflated effect sizes in the smallest sample sizes- which are also the range of sample sizes most commonly achieved in bilingual neuroimaging studies. 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.

ROI Bayes Factors

The exploration of Bayes Factors across sample sizes serves as an initial foray into expanding replicability investigations into relevant areas of statistics, but is not meant to represent an in-depth investigation of the factors influencing replicability using Bayesian inference. Bayes Factors are here simply calculated and compared with output from traditional frequentist analyses without manipulating prior values beyond the identical frequentist null-model likelihoods.

The use of Bayes Factors calculated from univariate multiple regressions, with a Bayes Factor cutoff of 15 in order to determine significance, revealed an extremely similar distribution of accuracy of outcomes as what was found for the multiple regressions with an alpha cutoff of 0.025. This is consistent with expectations as long as a similar cutoff is used, as the Bayes Factors were calculated from those same multiple regression analyses which were earlier explored.

However, the degree of effect size inflation is greatly reduced relative to the original regressions. Whereas significant R-squared effects were inflated for the smallest subsample sizes (10-30 per group), significant Bayes Factors are not at all inflated at the smaller subsample sizes. Effects actually appear to increase across subsample sizes. This suggests that the use of Bayes Factors may pose benefits in terms of estimating effect sizes at lower sample sizes. This is likely due to the prior assumptions made in the BayesFactor R statistical package to calculate Bayes Factors, which assumes large effects are less likely than smaller effects (Morey, Rouder, & Jamil, 2015; Ioannidis, 2008a; Ioannidis, 2008b).

Though it would make for an interesting extension, an investigation into the influences of informative vs. uninformative priors towards the necessity for a large sample size is beyond the scope of this project. In a comparison of Bayesian and classical frequentist approaches, Sadia and Hossain (2014) demonstrate how the Bayesian method requires smaller sample sizes when more informative prior information is used. An applied exploration of the extent to which samples can be optimized with informative prior use would help to clarify the utility of Bayesian analyses towards small-sample research questions.

ROI MANCOVAs

The full-sample MANCOVA including all 10 ROIs was very significant, suggesting that language status, while controlling for ICV, is significantly related to the included ROIs together. Post-hoc discriminant analyses demonstrated consistent results with the univariate multiple regressions, where the sizes of both RIPL and RACC coefficients showed them to be *most* related to the calculated discriminating function. The use of these coefficients may pose a more holistic method to compare ROIs on predictors than to use a univariate test of

each, as this allows for overlapping ROI variances to be considered within both the entire test and the post-hoc ROI-specific discriminant analyses.

Subsamples revealed the number of observed TPs to increase linearly with subsample sizes, with a relatively high value at lower subsample sizes. The high number of overall TPs is likely related to the significance of the effect, though the sensitivity of a MANCOVA analysis to detect differences in the many ROIs is here demonstrated relative to univariate tests.

Interestingly, the degree of Eta-squared effect size inflation is much greater at the smaller subsample sizes relative to the univariate tests. Smaller-sampled MANCOVA tests, while more likely to return a TP result in this scenario, were also more likely to return inflated effect sizes. Conversely, discriminant function coefficients were shown to be underestimated at smaller subsample sizes if they were moderately large.

Although multivariate analyses do not test the significance of each individual dependent variable, their use in determining whether predictors are related to multiple dependent variables was shown here. Because it is only a single test, stringent correction for multiple tests is also able to be avoided. Moreover, post-hoc discriminant analyses allowed for each ROI to be examined in terms of their influence towards the predictors. Two potential drawbacks that researchers should be mindful of, however, are 1) a greater tendency for returning inflated Eta-squared effect sizes, and 2) a tendency for moderately large discriminant function coefficients (both standardized and structural) to be underestimated at smaller subsample sizes.

ROI Network Analyses

Neither the univariate nor multivariate full-sample network analyses appeared to be more sensitive overall tests, nor did subsample sizes appear to be uniquely related with either network relative to analyses which included all 10 ROIs. Both non-significant univariate multiple regression explorations showed a consistent number of FPs across subsample sizes, and similar amounts of effect size inflation in the smallest subsamples. While MANCOVA using cognitive control brain network ROIs did result in a significant test, the observed trends across subsamples was consistent with earlier findings. Only for the non-significant MANCOVA using language brain network ROIs was a difference found. For the larger structural coefficient of the two ROIs, LIFG, the trend in lower subsample sizes was for slightly inflated coefficient estimates. This is the opposite of what was seen for other moderately large MANCOVA coefficients, and may relate to the very high amount of variance in this coefficient across subsample sizes.

Conclusion and Future Directions

The negative influences of inadequate sample sizes in testing the influence of language experience on measures of brain structure were explored in this paper. Significant drawbacks in wholebrain t-tests and ROI analyses were found, while multivariate analyses and Bayes Factors offered certain alternatives to mitigate some of these drawbacks. While MANCOVAs were seen to have better power and more sensitive tests of differences, they also showed even more inflated Eta-squared effect sizes than were seen for R-squared effects in the univariate regressions. Also, while Bayes Factors showed nearly identical test accuracy where a cutoff is defined, their utility in measuring the effect sizes across subsample sizes, without having a trend for inflated effects in smaller subsample sizes, was shown. It is important for these tools to be more fully explored in the context of underpowered studies. A multivariate Bayesian analysis may prove to be the best of both worlds, but could pose its own unique risks as well. In the field of bilingual neuroimaging, such an analysis could be particularly useful, but remains to be investigated.

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.'

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 influencer 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 replicability 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. Also, 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 is an exploration of the influences of inadequately powered studies in the hopes of having a more direct impact on researcher practices. 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 amounts of observed variability in samples of 10 to 30 participants per group suggest that researchers ought to strongly consider some of the aforementioned options for addressing power in studies.

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 neuroimaging studies on bilingualism. It is our hope that this study helps to open the eyes of bilingual researchers who use neuroimaging, as well as researchers in other areas, to the negative inferential effects that coincide with inadequate statistical power.

ACKNOWLEDGEMENTS

My sincerest gratitude to my loving wife, Stella Tamesis Munson; parents, Mark and Deborah Munson; and brother, Matthew Munson.

Thank you to current and former members of the Laboratory for the Neural Bases of Bilingualism (LNBB) who originally collected and cleaned the neuroimaging scans, including Dr. Kailyn Bradley, Dr. Pilar Archila-Suerte, Dr. Aurora Ramos, Dr. Maya Greene, Kelly Vaughn, and Hannah Claussenius-Kalman. Thank you also to Juliana Ronderos and Tres Bodet of the LNBB, who gave helpful feedback in the final drafts of this written dissertation. Thanks to Dr. Peggy Lindner who assisted with the writing of the MATLAB code for the study. Finally, thanks to my committee, including Dr. Arturo Hernandez, Dr. David Francis, Dr. Benjamin Tamber-Rosenau, and Dr. Christine Chiarello, who gave much helpful feedback and support as I completed each stage of this dissertation.

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, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., & Fedor, A. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), 1-8.

Abutalebi, J., Canini, M., Della Rosa, P. A., Green, D. W., & Weekes, B. S. (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, P. A., Gonzaga, A. K. C., Keim, R., Costa, A., & Perani, D. (2013). The role of the left putamen in multilingual language production. Brain and language, 125(3), 307-315.

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.

Batty, M. J., Liddle, E. B., Pitiot, A., Toro, R., Groom, M. J., Scerif, G., ... & Hollis, C. (2010). Cortical gray matter in attention-deficit/hyperactivity disorder: a structural magnetic resonance imaging study. Journal of the American Academy of Child & Adolescent Psychiatry, 49(3), 229-238.

Bray, J. H., Maxwell, S. E., & Maxwell, S. E. (1985). Multivariate analysis of variance (No. 54). Sage.

Cohen, J. (1992). A power primer. Psychological bulletin, 112(1), 155.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. Psychological Bulletin, 115(3), 465.

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

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage, 9(2), 179-194.

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

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. Journal of neuroscience methods, 118(2), 115-128.

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.

Fischl, B., & Dale, A. M. (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, D. H., van der Kouwe, A. J., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. Neuroimage, 23, S69-S84.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., ... & Caviness, V. (2004). Automatically parcellating the human cerebral cortex. Cerebral cortex, 14(1), 11-22.

Friendly, M., & Fox, J. (2017). candisc: Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.8-0. https://CRAN.R-project.org/package=candisc

García-Pentón, L., Fernández García, Y., Costello, B., Duñabeitia, J. A., & 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.

Grogan, A., Jones, Õ. P., Ali, N., Crinion, J., Orabona, S., Mechias, M. L., ... & Price, C. J. (2012). Structural correlates for lexical efficiency and number of languages in non-native speakers of English. Neuropsychologia, 50(7), 1347-1352.

Hernandez, A. E., Martinez, A., & Kohnert, K. (2000). In search of the language switch: An fMRI study of picture naming in Spanish-English bilinguals. Brain and language, 73(3), 421-431.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition, 92(1-2), 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5), 393.

Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. Epidemiology, 640-648.

Ioannidis, J. P. (2008). Effect of formal statistical significance on the credibility of observational associations. American journal of epidemiology, 168(4), 374-383.

Jeffreys, H. (1961). Theory of probability (3rd ed.). Oxford: Oxford University Press, Clarendon Press.

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, J. K., & Watkins, K. E. (2014). Age of language learning shapes brain structure: a cortical thickness study of bilingual and monolingual individuals. *Brain and Language*, *131*, 20-24.

Li, P., Legault, J., & Litcofsky, K. A. (2014). Neuroplasticity as a function of second language learning: anatomical changes in the human brain. Cortex, 58, 301-324.

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

Mechelli, A., Crinion, J. T., Noppeney, U., O'doherty, J., Ashburner, J., Frackowiak, R. S., & Price, C. J. (2004). Neurolinguistics: structural plasticity in the bilingual brain. Nature, 431(7010).

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9, 9, 2014.

Morrison, D. F. (2005). Multivariate analysis of variance. Encyclopedia of biostatistics, 5.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage, 39(1), 261-268.

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.

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.

Sadia, F., & Hossain, S. S. (2014). Contrast of bayesian and classical sample size determination. Journal of Modern Applied Statistical Methods, 13(2), 23.

Sawilowsky, S. S. (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.

Smith, H. F. (1958). A multivariate analysis of covariance. Biometrics, 14(1), 107-127.

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

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

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

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., and Müller, K. (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

Woodcock, R. W. (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.

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.

APPENDIX

Table 1

Year, sample sizes (Total 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.

	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

Harold Jeffreys' (1961) scale of strength of evidence for Bayes Factors.

Strength of Evidence	Bayes Factor
Negative (supports other model)	$BF < 1.0$
Barely Worth Mentioning	$1.0 < BF < 3.16$
Substantial	$3.16 < BF < 10$
Strong	$10 < BF < 31.6$
Very Strong	$31.6 < BF < 100$
Decisive	$100 < BF$

Table 4

Full-sample (216 bilinguals; 146 monolinguals), FEW-corrected, whole-brain density t-test results, where bilinguals > monolinguals, including cluster size, peak z score voxel values, coordinates (x, y, z) for each peak z score statistic, and corresponding hemisphere and brain region. No regions were significant in a comparison of monolinguals > bilinguals.

Region	Hem	Cluster Size	Peak Z	x	y	z {mm}
Posterior Cingulate Gyrus	R	1606	7.36	26	-50	8
Anterior Temporal Lobe	R	3743	6.92	50	15	-36
Primary Visual Cortex	L	885	6.58	-18	-68	4
Anterior Temporal Lobe	L	4366	6.52	-50	20	-22
Inferior Parietal Lobule	R	155	6.36	18	-58	68
Inferior Frontal Gyrus	R	4474	6.31	56	32	16
Medial Parietal Lobe	R	171	6.18	58	-33	27
Inferior Temporal Gyrus	L	280	6.04	-40	-27	-28
Inferior Temporal Gyrus	R	194	5.95	51	-22	-28
Inferior Parietal Lobule	L	120	5.59	-18	-66	64
Superior Temporal Gyrus	R	218	5.58	58	2	-4
Angular Gyrus	L	75	5.51	-48	-62	39
Primary Somatosensory Cortex	R	82	5.5	32	-46	68
Dorsal Posterior Cingulate Gyrus	R	86	5.48	2	-70	24
Superior Parietal Lobule	L	34	5.27	-18	-62	48
Primary Somatosensory Cortex	R	34	5.24	6	-38	74
Angular Gyrus	R	37	5.24	56	-56	22
Sensory Association Area	L	25	5.16	-6	-40	74
Premotor Cortex	R	64	5.16	62	2	26
Sensory Association Area	L	23	5.09	-10	-38	66
Thalamus	L	61	5.08	-16	-34	8
Inferior Temporal Gyrus	R	84	5.06	34	9	-46
Dorsolateral Prefrontal Cortex	L	65	4.95	-40	30	44

Table 5

Full-sample (216 bilinguals; 146 monolinguals) R-squared effect sizes, estimates, t statistics, 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-squared	Estimate	Statistic	p value
LACC	0.002	0.06	1.04	0.30
RACC	0.011	0.14	2.21	0.028
LIFG	0.002	-0.10	-1.11	0.27
RIFG	0.00005	0.02	0.17	0.87
LIPL	0.001	0.15	0.95	0.34
RIPL	0.015	0.56	3.37	0.0008
LSTG	0.0005	0.04	0.65	0.51
RSTG	0.0000007	-0.01	-0.07	0.94
LBG	0.0003	-0.01	-0.39	0.70
RBG	0.0002	-0.01	-0.33	0.74

Table 6

Full-sample (216 bilinguals; 146 monolinguals) Bayes Factors for each tested grey matter volume ROI from each regression. These values demonstrate the unique explanatory value of 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	Bayes Factor
LACC	1.70
RACC	10.40
LIFG	1.84
RIFG	1.01
LIPL	1.56
RIPL	243.60
LSTG	1.23
RSTG	1.00
LBG	1.08
RBG	1.06

Table 7

MANCOVA standardized and structural coefficients where all 10 ROIs are included as dependent variables, with both language status and ICV as predictors.

ROI	Standardized Coefficients	Structural Coefficients
LACC	-0.02	-0.20
RACC	-0.50	-0.40
LIFG	0.43	0.11
RIFG	-0.06	-0.07
LIPL	0.55	-0.17
RIPL	-1.25	-0.49
LSTG	-0.16	-0.13
RSTG	0.38	-0.04
LBG	0.14	0.02
RBG	0.22	0.003

Table 8

Full-sample (216 bilinguals; 146 monolinguals) cognitive control and language network R-squared effect sizes, estimates, t statistics, and p values from each regression. Specifically, these values are for the bilingual-monolingual comparisons within each regression, while controlling for intra-cranial volume. Cognitive Control Network = LIPL + LBG + RBG + LACC + RACC, and Language Network = LSTG + LIFG.

Network	R-squared	Estimate	Statistic	p value
Cognitive Control	0.002	0.33	1.28	0.20
Language	0.0002	-0.05	-0.43	0.67

Table 9

MANCOVA standardized and structural coefficients where cognitive control network-related ROIs are included as dependent variables, with both language status and ICV as predictors. A test of Pillai's trace on language status is significant, p (5, 355) = 0.03, F = 2.47, Pillai's trace = 0.03. Eta-squared for ICV is 0.59, and eta-squared for language status 0.019.

ROI	Standardized Coefficients	Structural Coefficients
LACC	0.01	0.53
RACC	0.92	0.87
LIPL	0.36	0.52
LBG	-0.78	0.13
RBG	0.23	0.17

Table 10

MANCOVA standardized and structural coefficients where language network-related ROIs are included as dependent variables, with both language status and ICV as predictors. A test of Pillai's trace on language status is non-significant, p (2, 358) = 0.09, F = 2.39, Pillai's trace = 0.013. Eta-squared for ICV is 0.59, and eta-squared for language status 0.007.

ROI	Standardized Coefficients	Structural Coefficients
LIFG	0.98	0.56
LSTG	-0.77	-0.23

		Population Outcome		
		Positive	Negative	
Sample Outcome	Positive	True Positive (TP)	False Positive (FP; Type I Error)	Positive Predictive Value = (TP / (TP + FP))
	Negative	False Negative (FN; Type II Error)	True Negative (TN)	
		Sensitivity = (TP / (TP + FN))		

Figure 1

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.

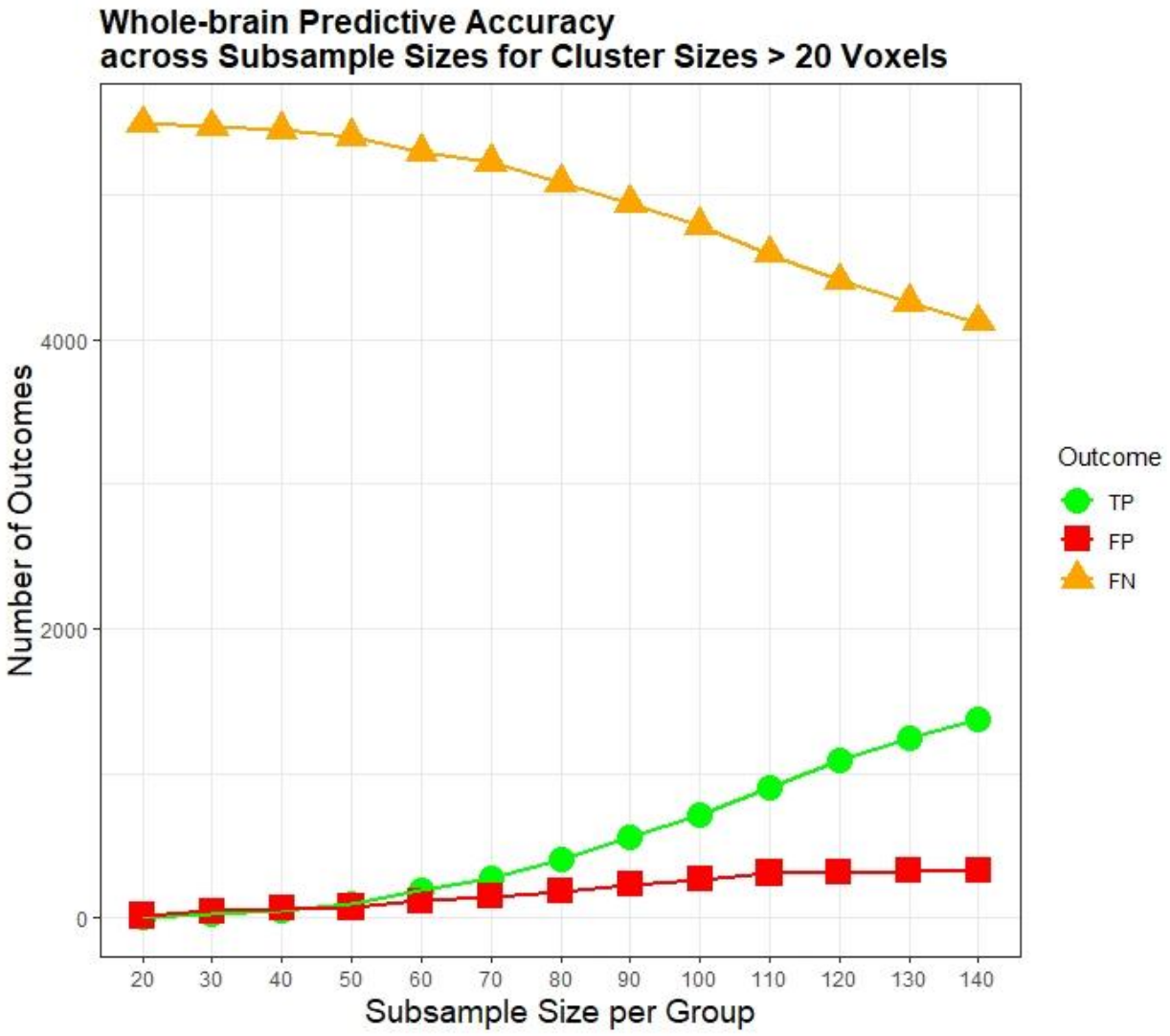

Figure 2

Whole-brain accuracy of subsample test outcomes relative to the full sample across subsample sizes per group, FWE corrected, for clusters > 20 voxels in size. Accuracy is here determined by counting the number of overlapping clusters, simplified as spheres with roughly the same number of voxels as the original clusters, to full-sample clusters.

False Negatives (FN; the yellow triangles) are the most common outcome, and slightly decrease as the subsample size increases. False Positives (FP; the red squares) are least common, and very slightly increases in the higher subsample sizes. True Positives (TP; the green circles) increase as subsample size increases. There is little change in accuracy until subsample sizes reach about 70 per group.

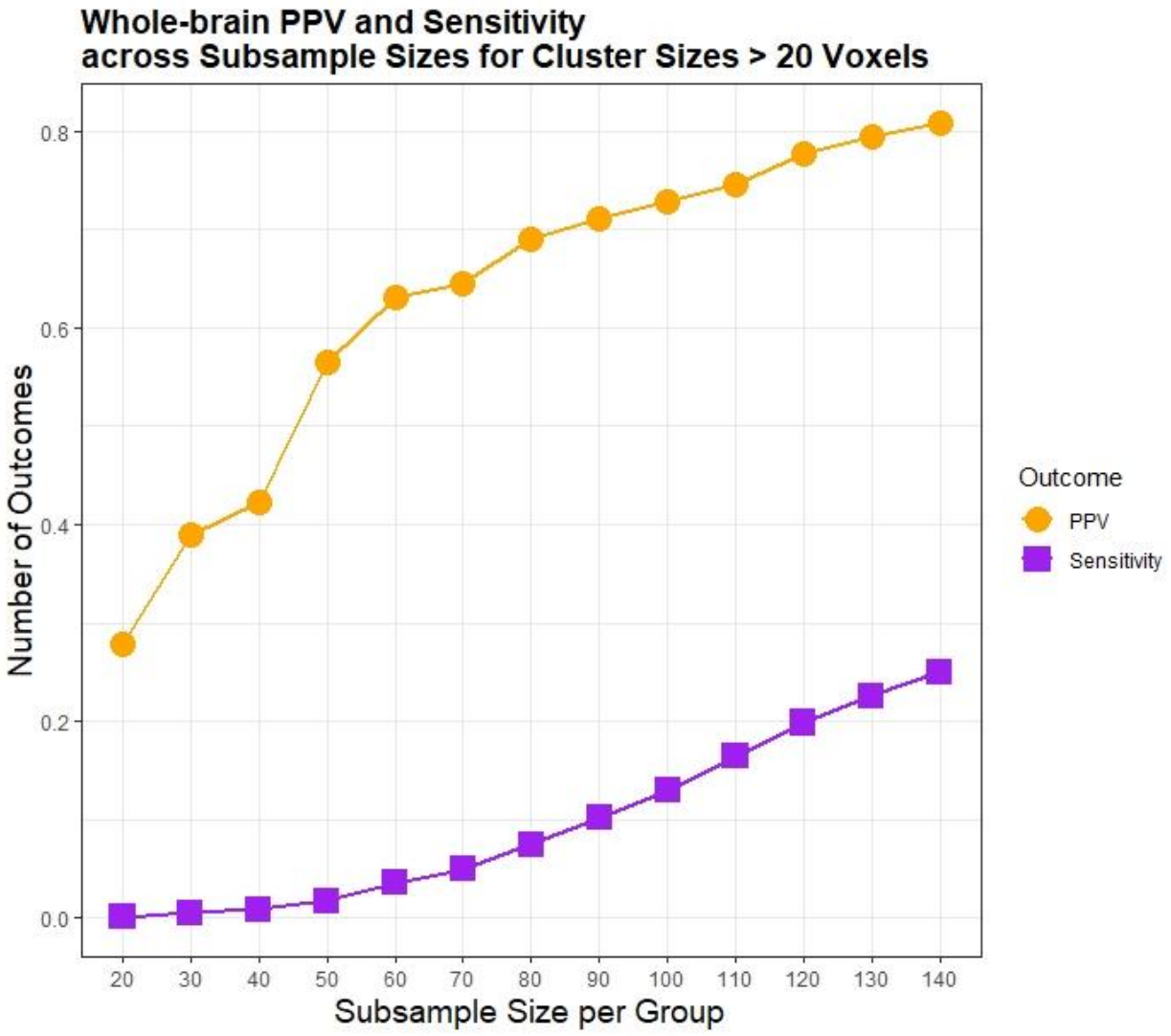

Figure 3

Average sensitivity (the yellow circles), also known as power, and Positive Predictive Value (PPV; the purple squares) across subsample sizes per group, for FWE-corrected whole-brain analyses, in terms of the detection of overlapping (simplified, spherical) subsample clusters and full-sample clusters. Sensitivity increases more quickly at lower subsample sizes, and eventually reaches 0.80 where 140 participants are included per group. PPV, however, remains below 0.10 until 90 participants are included in each group, and below 0.25 in all subsamples.

Sensitivity, or power, is defined as the proportion of TPs to the sum of TPs and FNs (TP / (TP + FN)), and is therefore, for these whole-brain analyses, a measure of the likelihood of observing an overlapping subsample cluster of voxels with a full-sample cluster of voxels, given the total number of significant full-sample clusters. 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 significant subsample cluster of voxels is also a cluster which overlaps with a full-sample cluster.

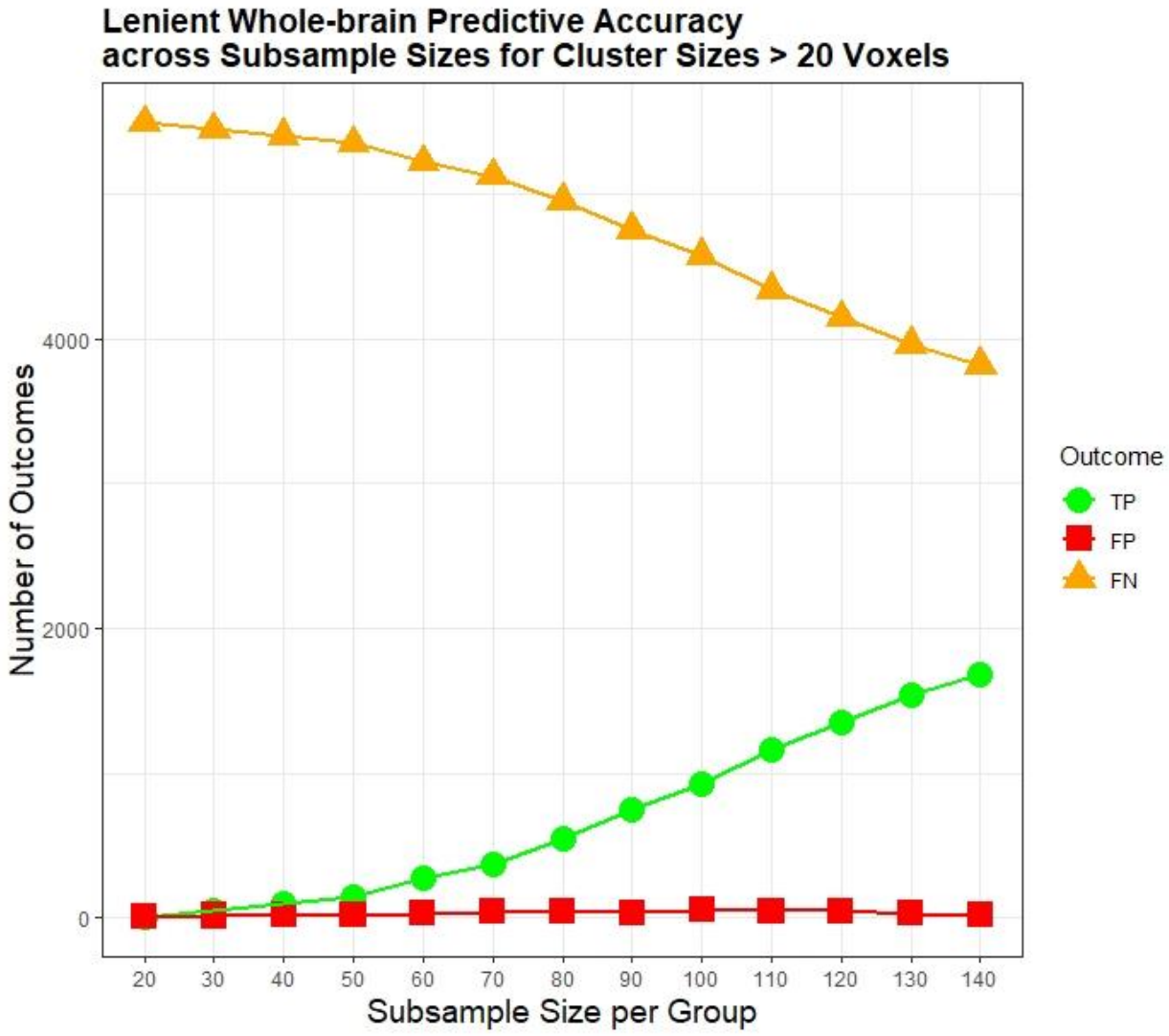

Figure 4

Lenient whole-brain accuracy of subsample test outcomes relative to the full sample across subsample sizes per group, FWE corrected, for clusters > 20 voxels in size. Accuracy is here determined by counting the number of overlapping clusters, simplified as spheres, to full-sample clusters. Spheres here are significantly increased in size to include roughly 20 times as many voxels as the original clusters, and yet little difference is seen in the resulting number of outcomes vs. the normally-sized spheres in Figure 2.

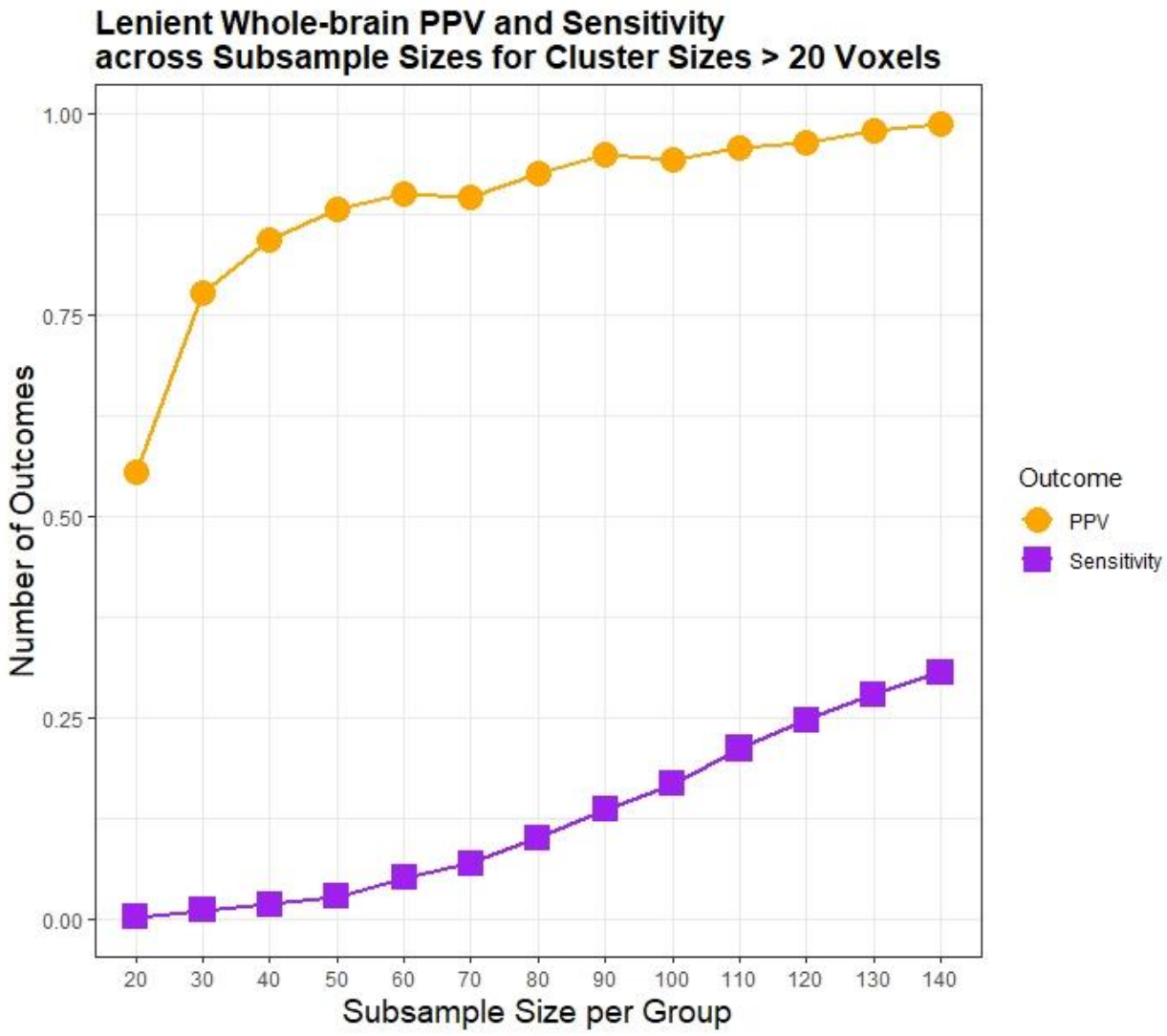

Figure 5

Average lenient sensitivity (the yellow circles), also known as power, and PPV (the purple squares) across subsample sizes per group, for FWE-corrected whole-brain analyses, in terms of the detection of overlapping (simplified, spherical) subsample clusters and full-sample clusters. Here, the overlapping spheres are 20 times as large as the original clusters, and yet the trends do not seem to meaningfully differ in sensitivity from the original visualization using the same number of voxels in overlapping spheres (Figure 3), though PPV increases much more quickly.

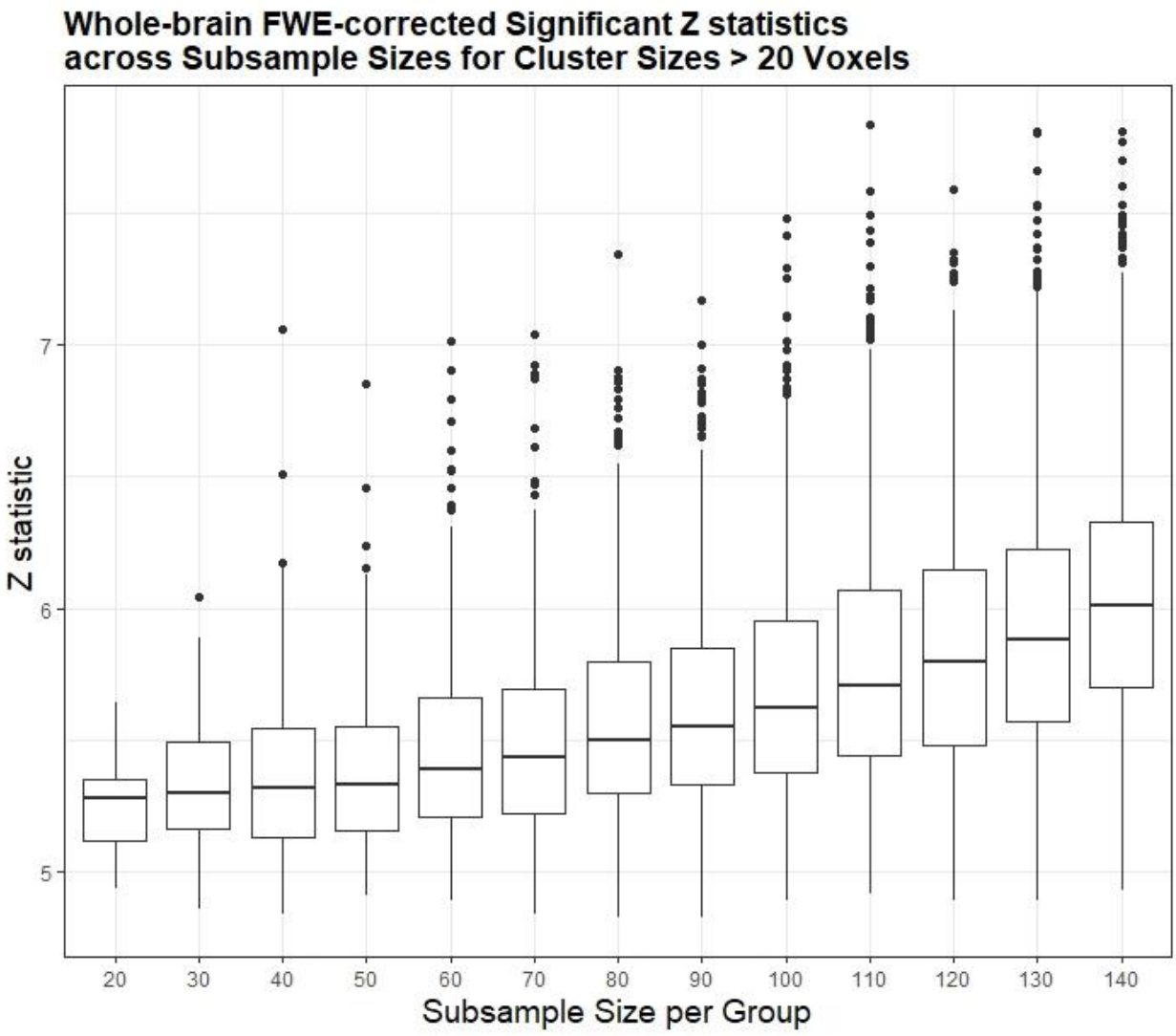

Figure 6

Boxplot of FWE-corrected peak whole-brain cluster Z statistics across subsample sizes where clusters are greater than 20 voxels in size. Effects appear to grow larger as subsample sizes increase- reflecting the expected relationship between subsample size and effect size.

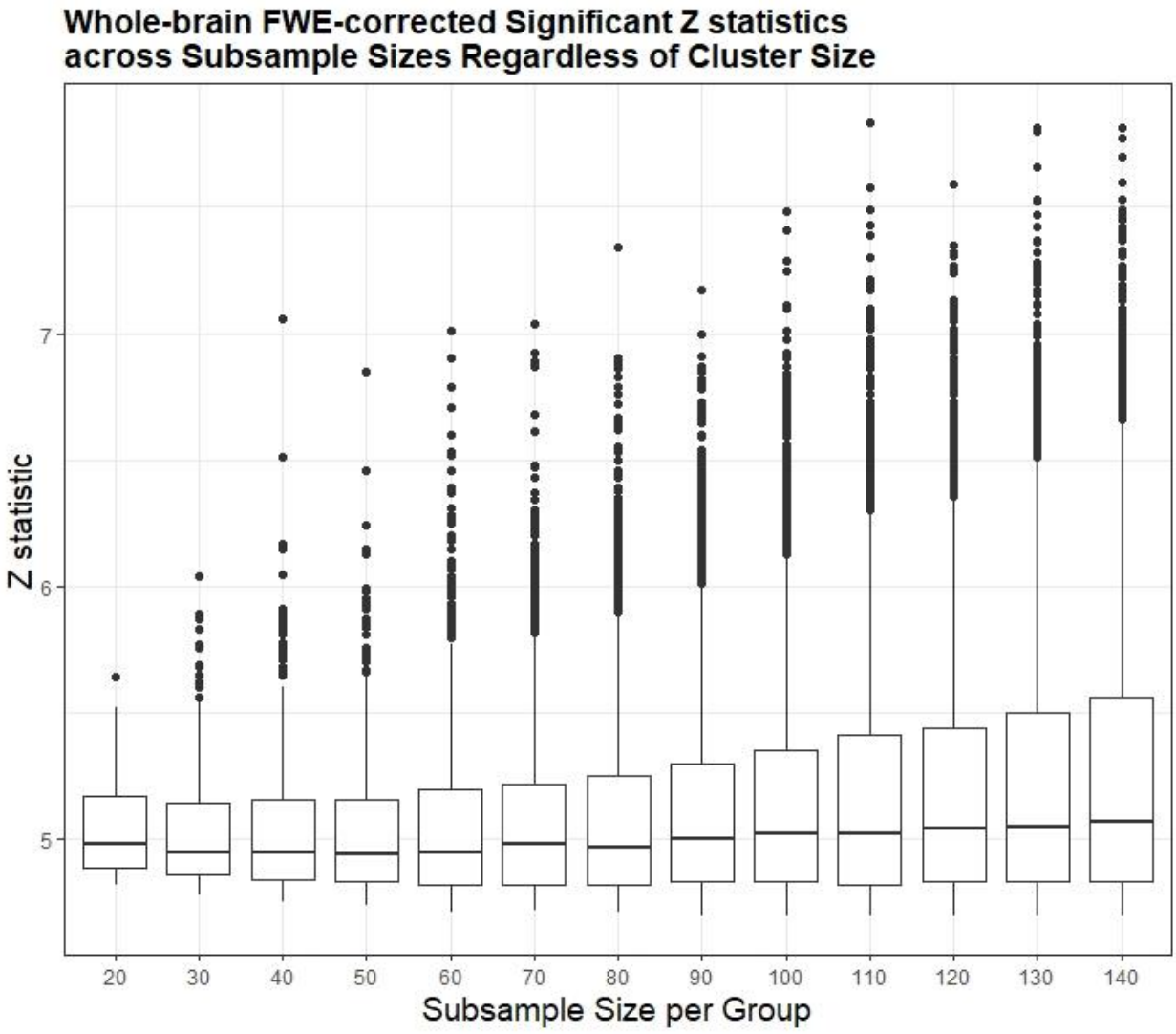

Figure 7

Boxplot of FWE-corrected peak whole-brain cluster Z statistics across subsample sizes regardless of cluster size. Peak effects do not appear to differ greatly across subsamples, though more extreme effects are seen in larger subsample sizes.

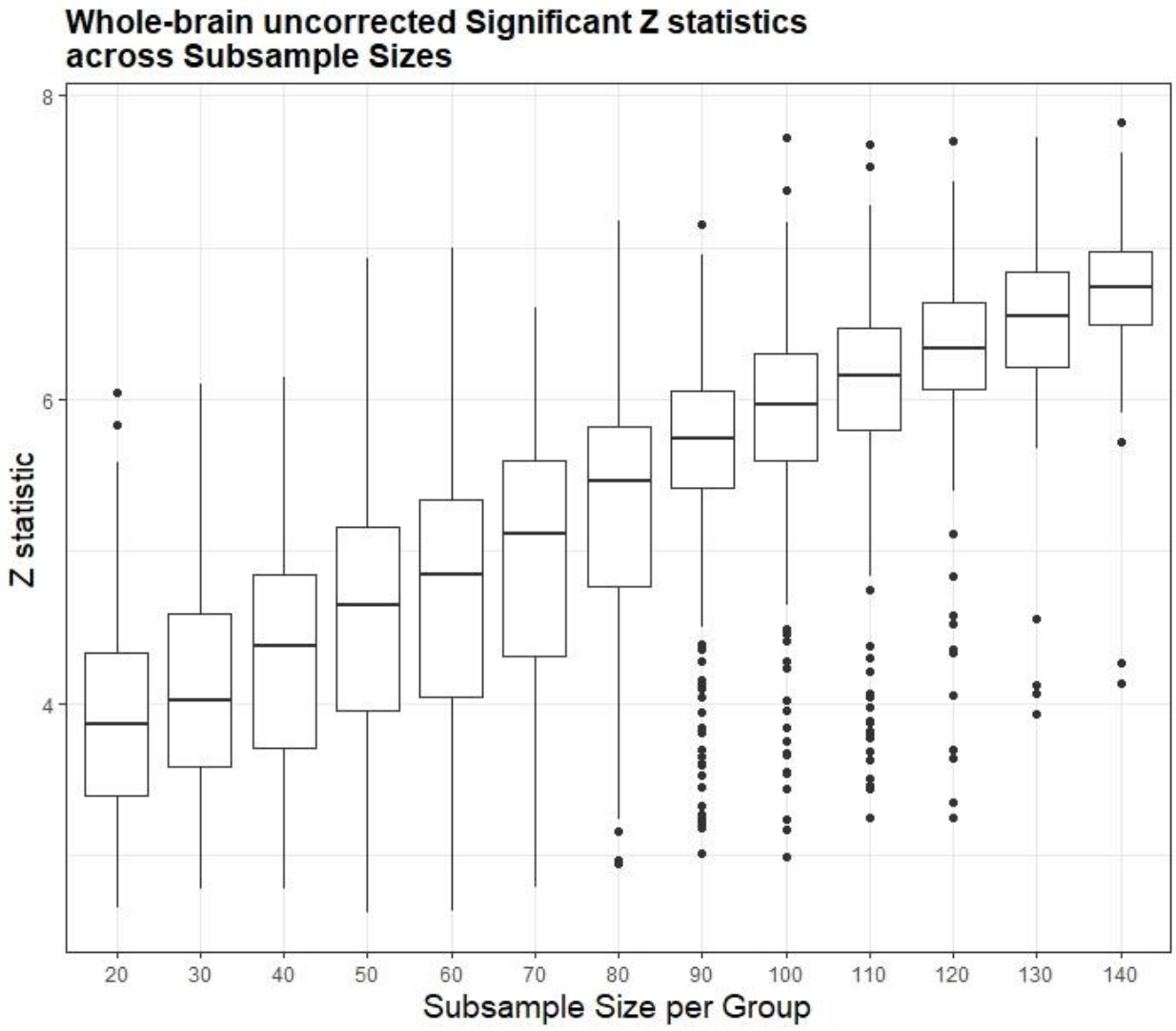

Figure 8

Boxplot of uncorrected ($p < 0.01$) peak whole-brain cluster Z statistics across subsample sizes regardless of cluster size. Contrary to expectations, uncorrected peak effects are not inflated in the smaller subsamples, and in fact increase greatly as subsample sizes increase. Clusters which survived at $p < 0.01$ were all greater than 20 voxels in size.

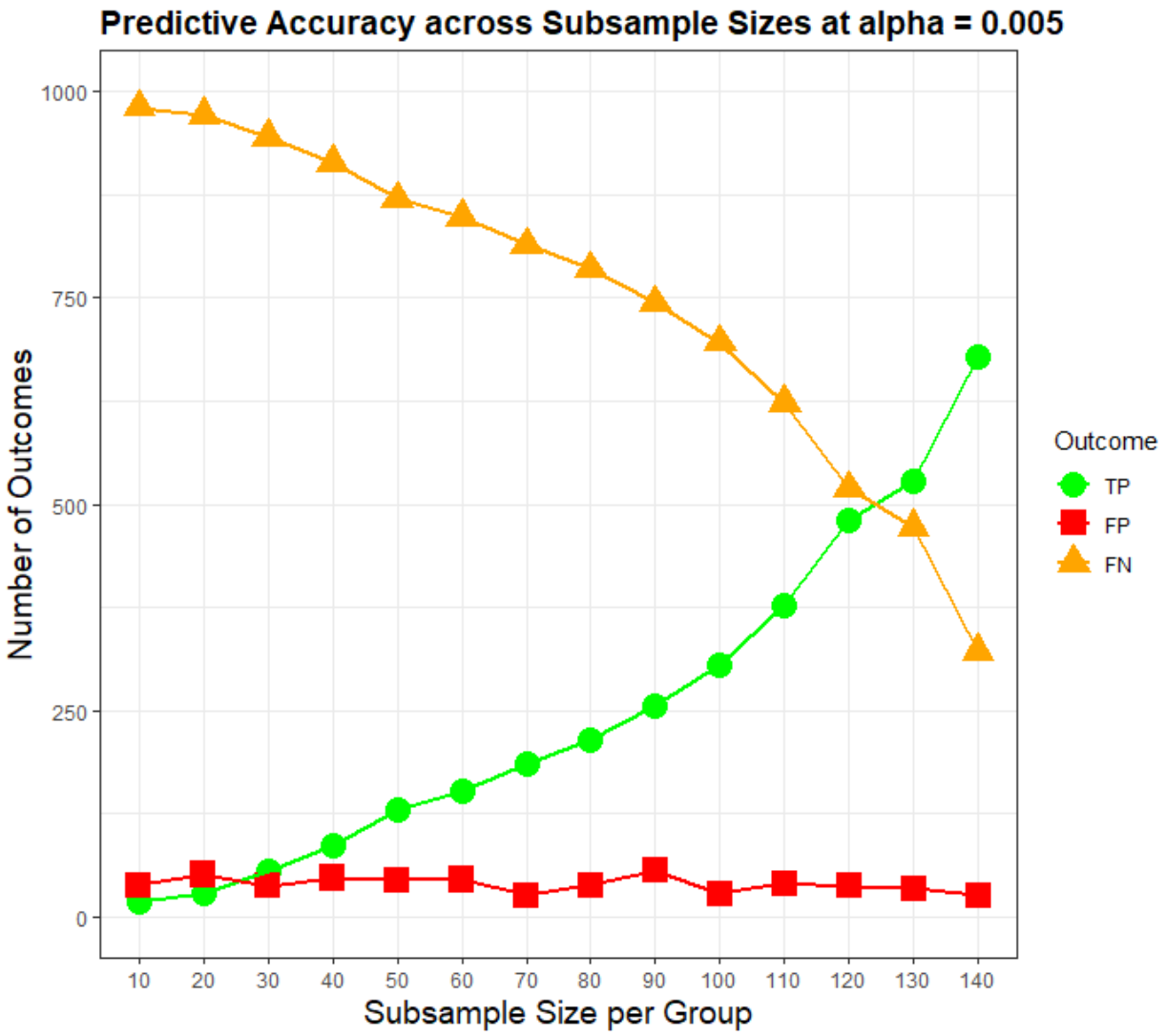

Figure 9

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 ROIs 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. 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.

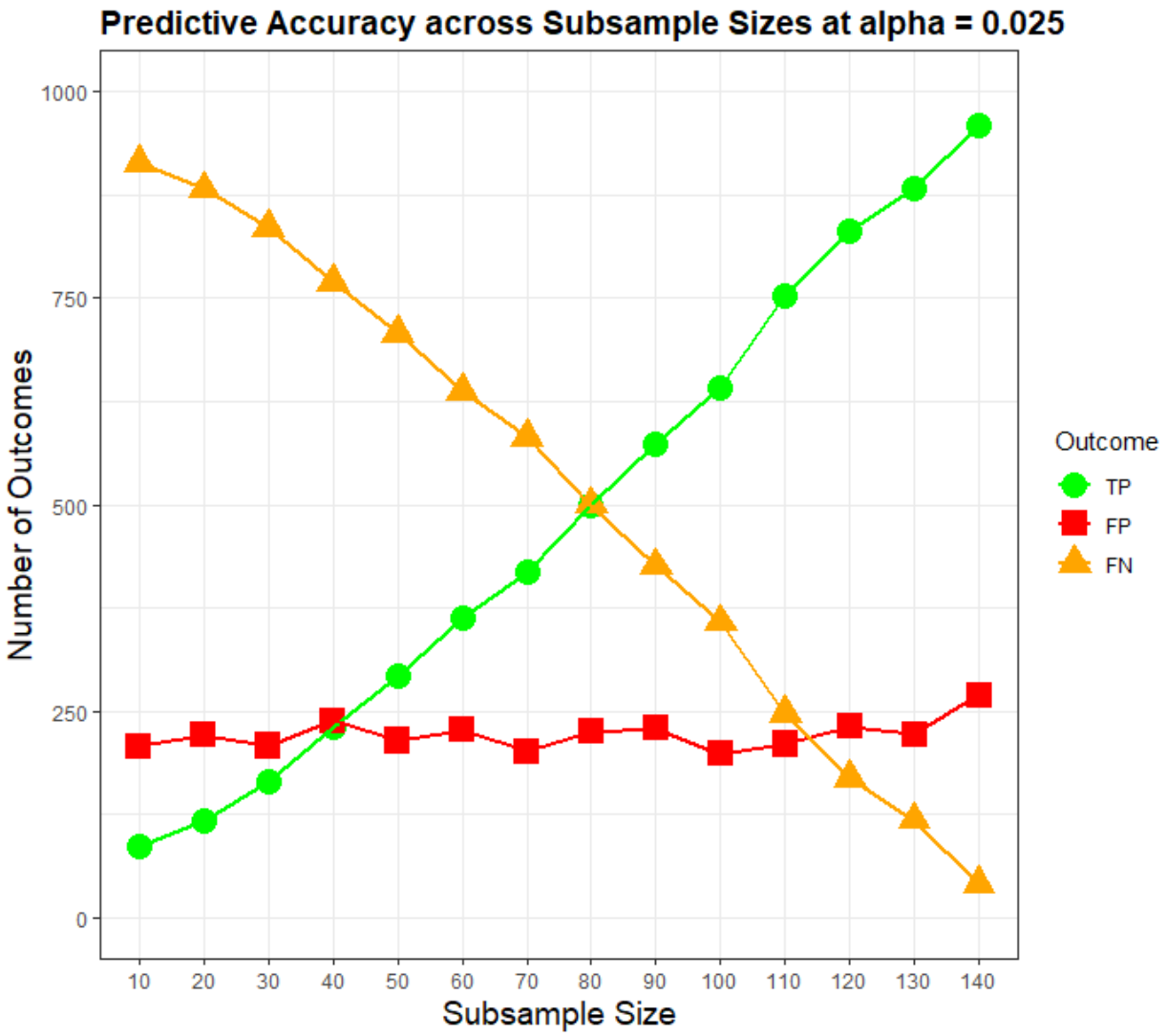

Figure 10

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 subsample 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.

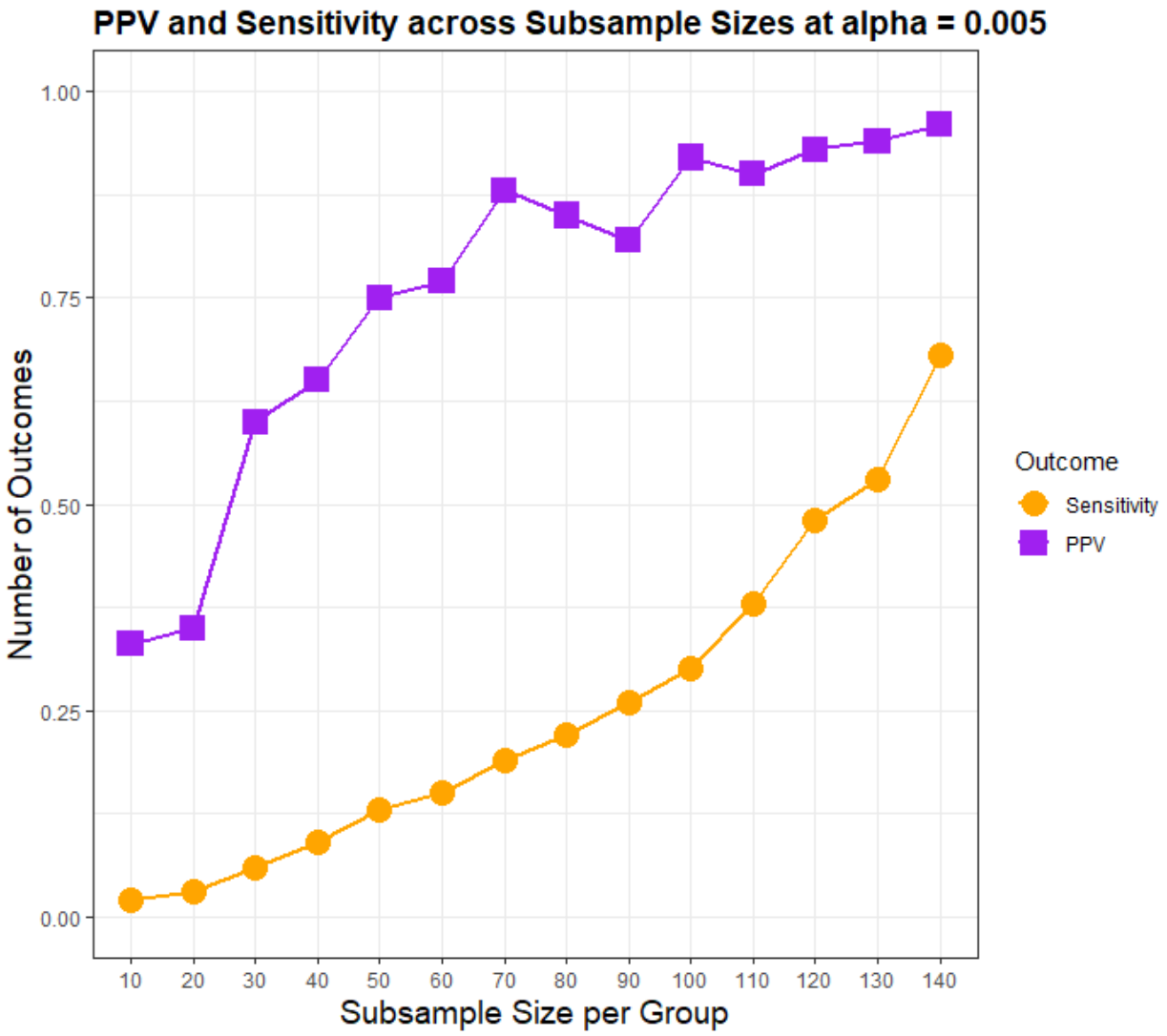

Figure 11

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 mirror 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.

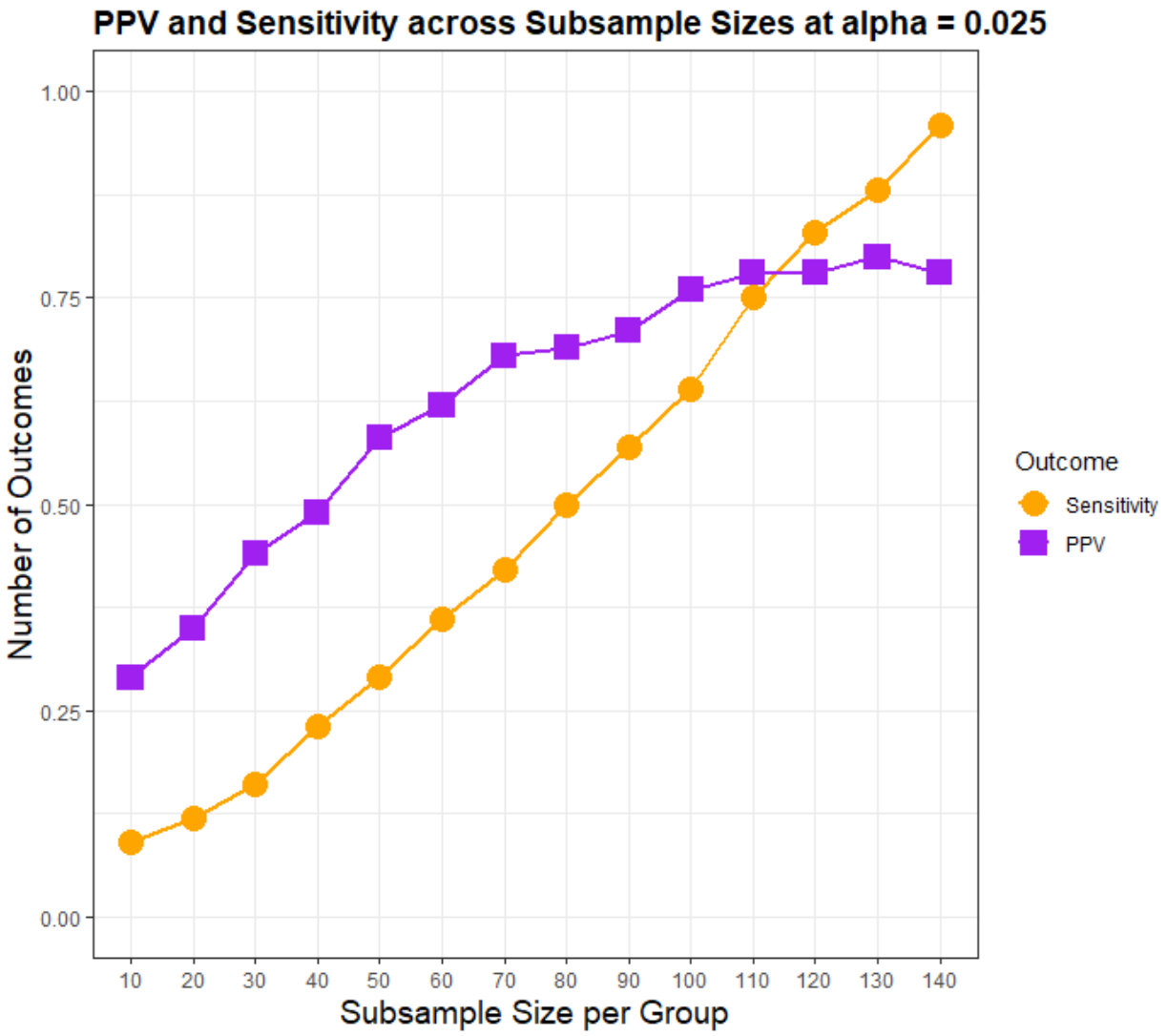

Figure 12

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.

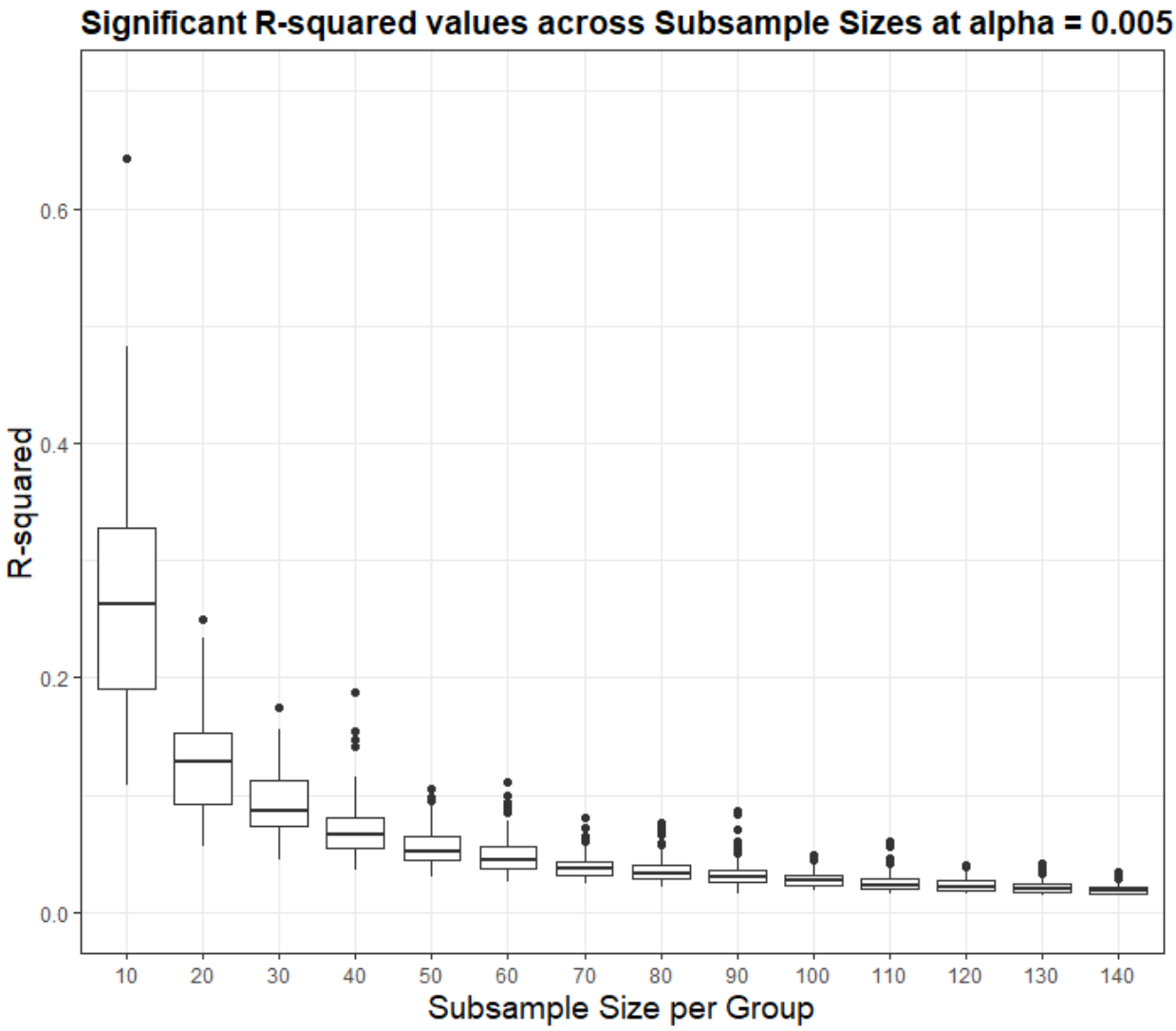

Figure 13

Boxplot of significant R-squared effect sizes across subsample sizes where the stringent critical alpha = 0.005. The average R-squared for subsamples of 10 per group is clearly inflated relative to higher subsample sizes which approach the true full-sample significant R-squared value of 0.015.

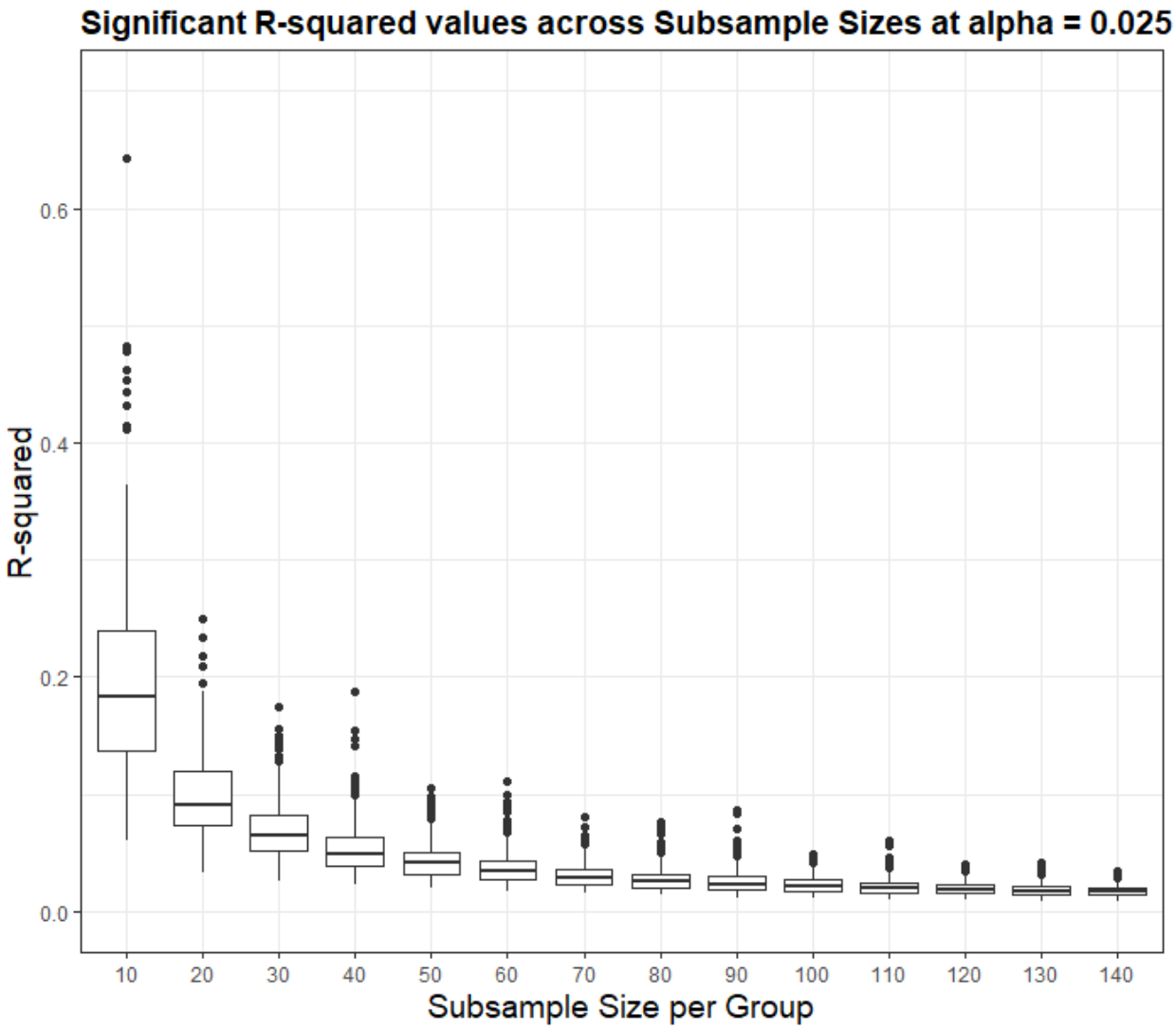

Figure 14

Boxplot of significant R-squared effect sizes across subsample sizes where the lenient critical alpha = 0.025. Again, the average R-squared value for subsamples 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 more accurate estimates of the true effect sizes.

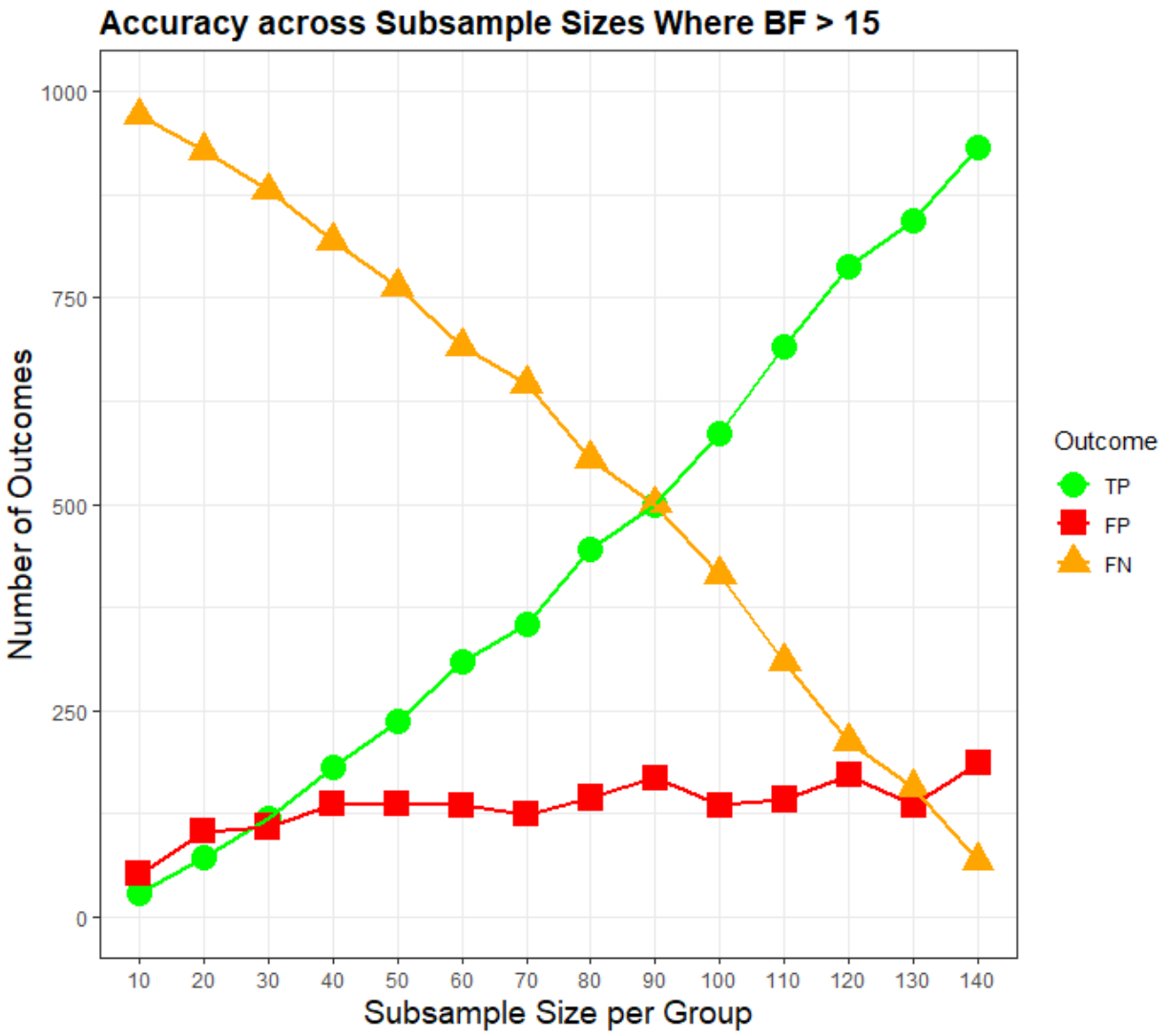

Figure 15

Accuracy of subsample test outcomes relative to the full sample across subsample sizes per group, where the 'critical bayes factor' cutoff = 15. Results are nearly identical to the accuracy of linear multiple regressions to predict differences at p = 0.25 (Figure 10). 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.

Here, where the 'critical Bayes Factor' = 15, True Negatives were seen in about 8,900 of the 10,000 tests across each subsample size.

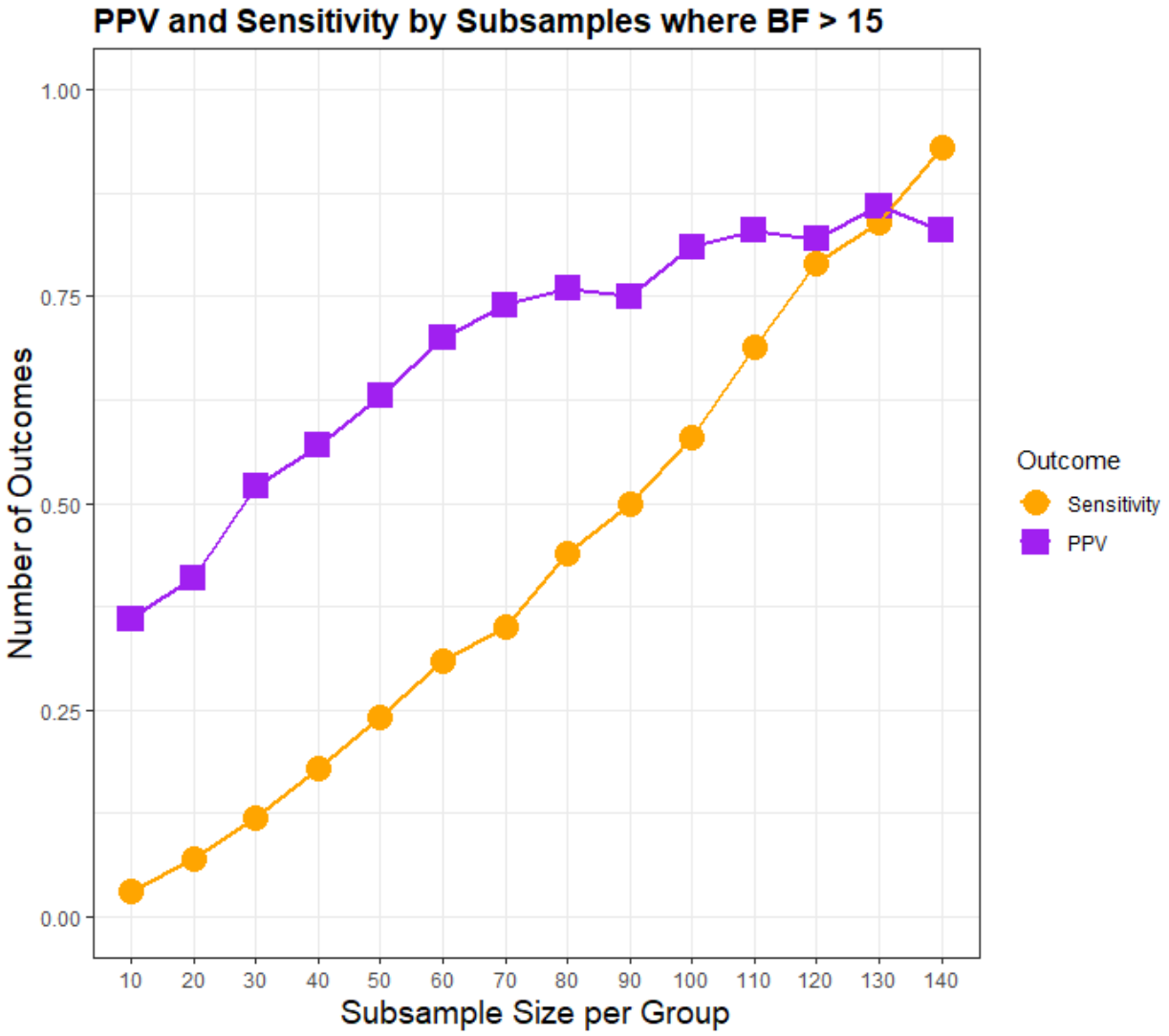

Figure 16

Average sensitivity (the yellow circles), also known as power, and Positive Predictive Value (PPV; the purple squares) across subsample sizes per group, for tests where Bayes Factors are greater 15. As with Figure 15, accuracy for these Bayesian analyses are nearly identical to the accuracy of linear multiple regressions to predict differences at p = 0.25 (Figure 12).

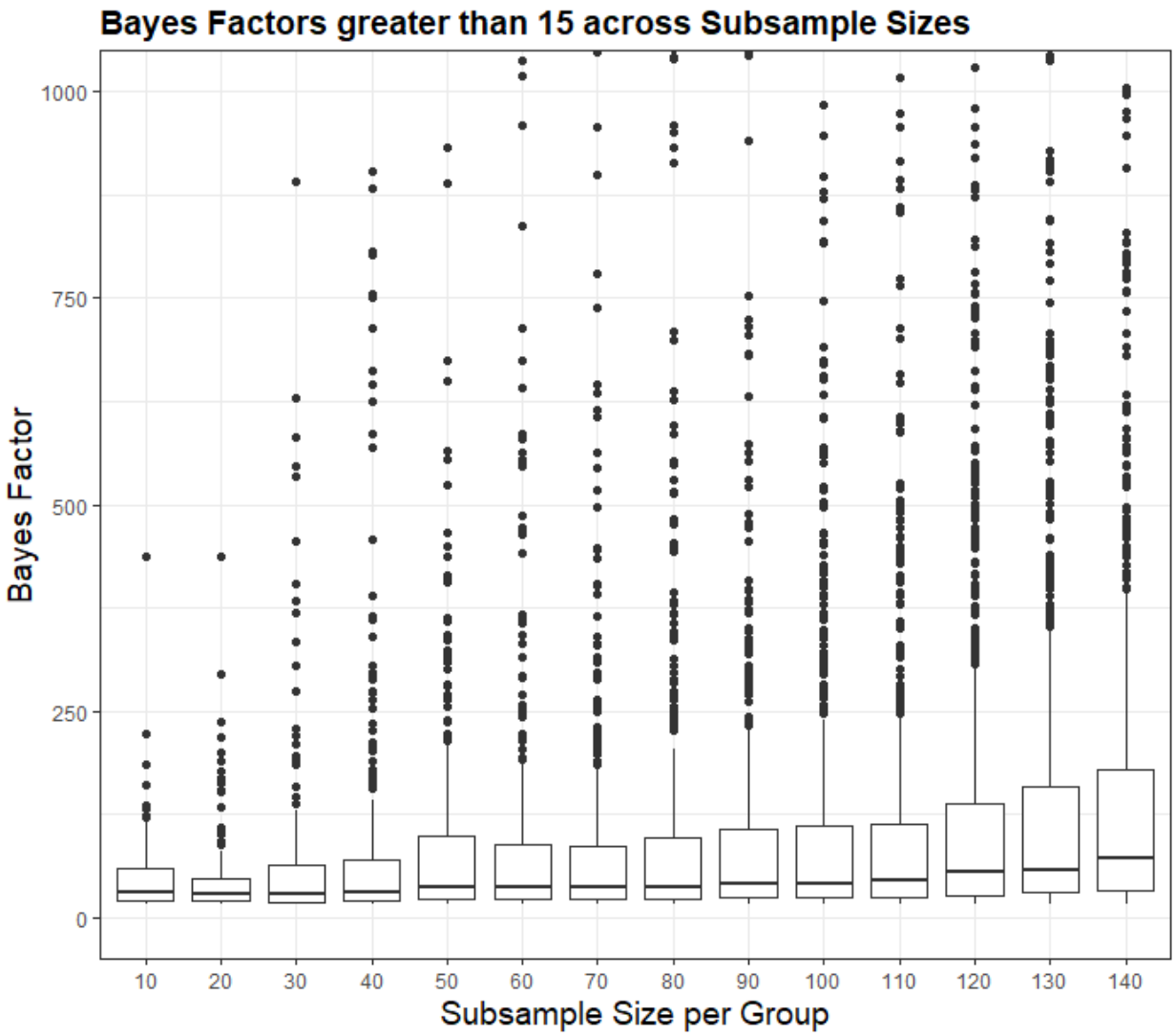

Figure 17

Boxplots of Bayes Factors greater than 15 across subsample sizes. The average Bayes Factor for lower subsample sizes is not inflated relative to larger subsample sizes, which become larger as they approach the true full-sample Bayes Factor of 243.6.

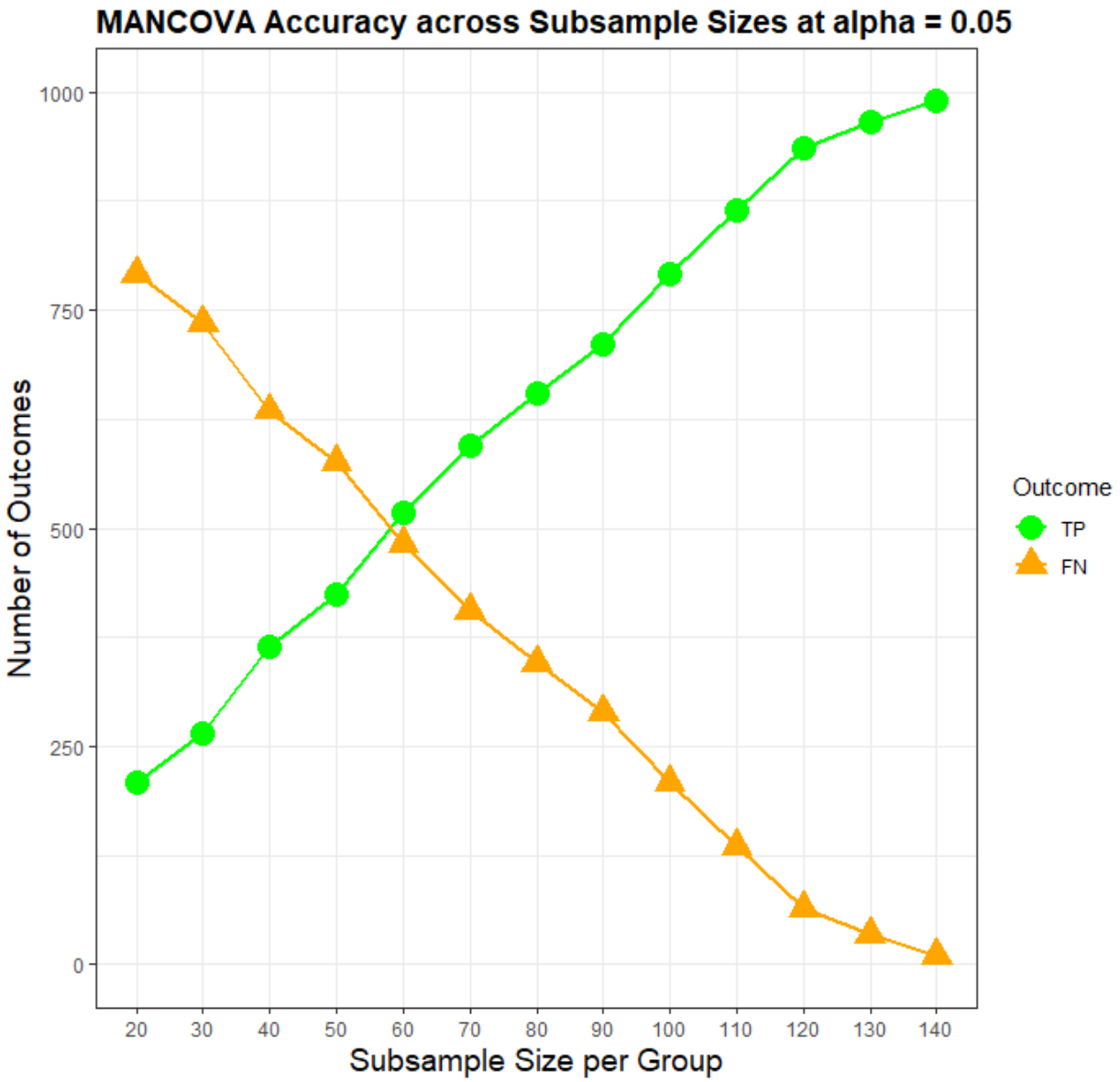

Figure 18

Accuracy of Multivariate Analysis of Covariance (MANCOVA) outcomes using significance of Pillai's trace across subsamples. Because the MANCOVA tested all ROIs in a single test, no FWE correction was necessary, and an alpha cutoff of 0.05 was used. TP's increase linearly as subsample sizes increase, reaching 50% at roughly 60 participants per group, and approaches 100% accuracy at about 120+ participants per group.

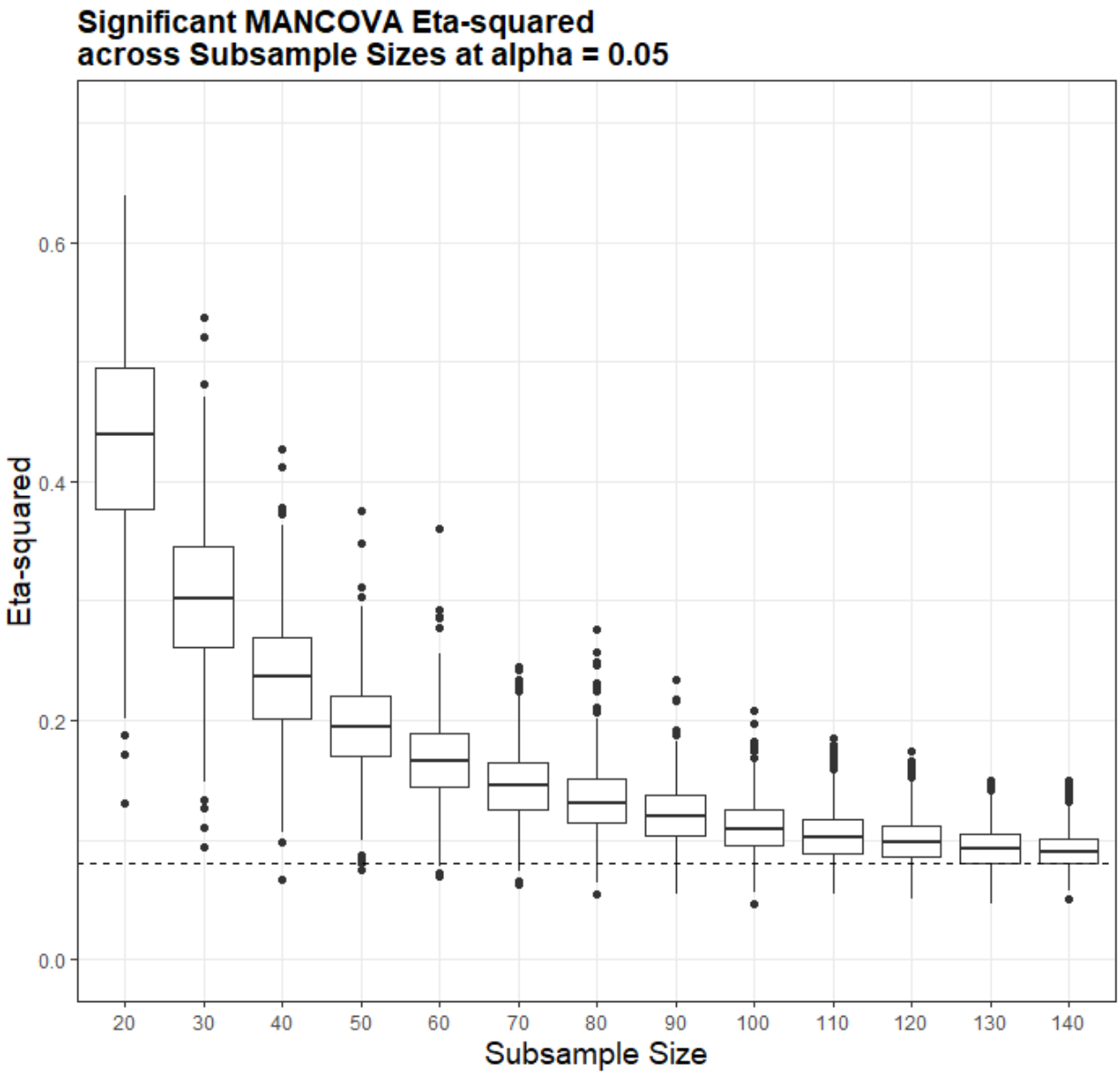

Figure 19

Boxplots of Eta-squared for significant MANCOVAs across subsample sizes. Observed Eta-squared values are very inflated when significant for smaller subsample sizes, and appear to require more participants per group to approach the true effect size than a linear regression (Figure 13).

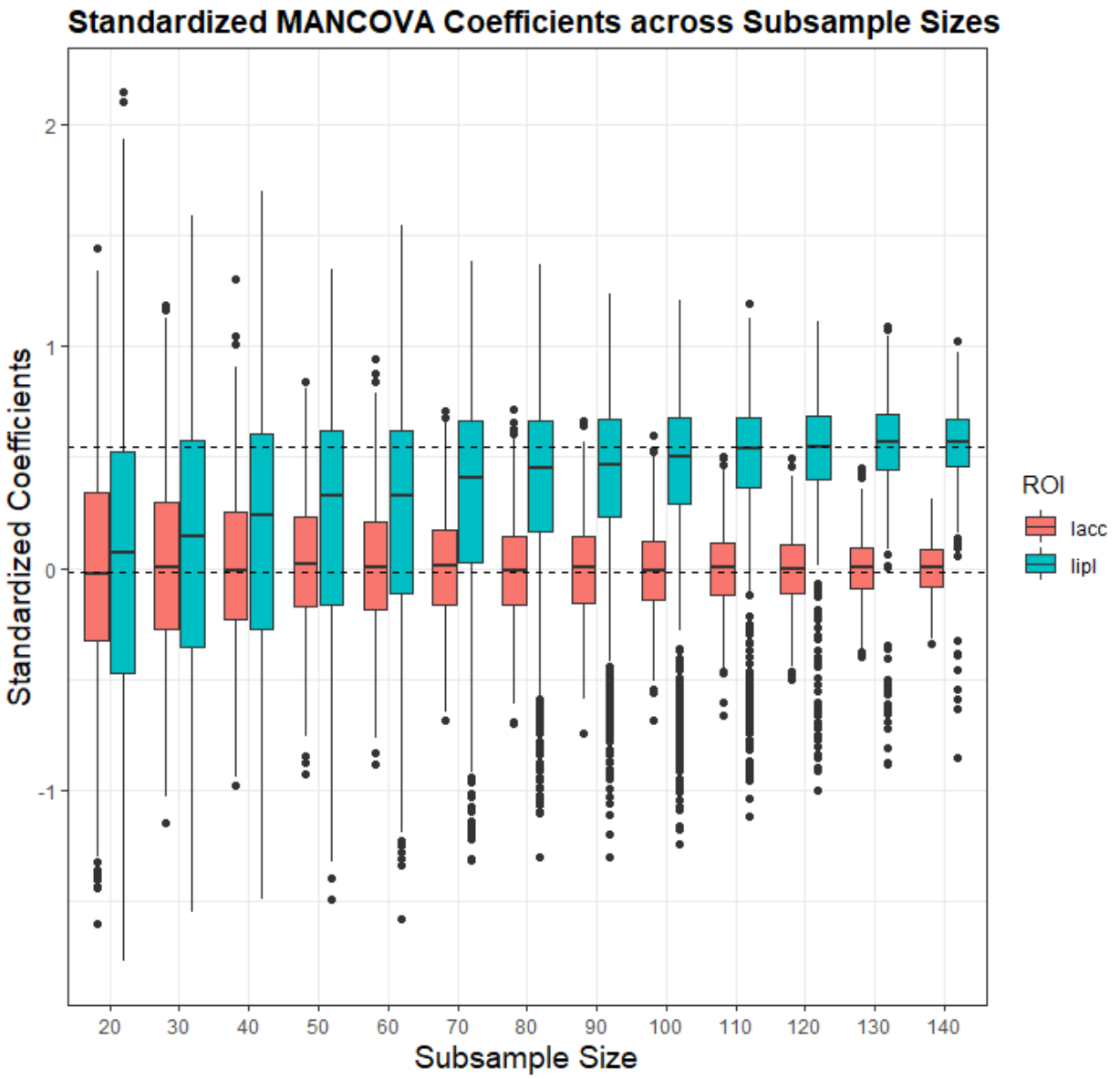

Figure 20

Variability in standardized MANCOVA coefficients across subsample sizes for the LIPL (largest standardized coefficient) and the LACC (smallest standardized coefficient). While the LACC hovers around its estimated coefficient of roughly 0, and slightly decreases in the amount of variability around this median as subsample sizes increase, the LIPL underestimates the size of its standardized coefficient at lower subsample sizes, and approaches it as subsample sizes increase. At around 80-100 participants per group, estimates of the LIPL standardized coefficients begin to normalize around its true value.

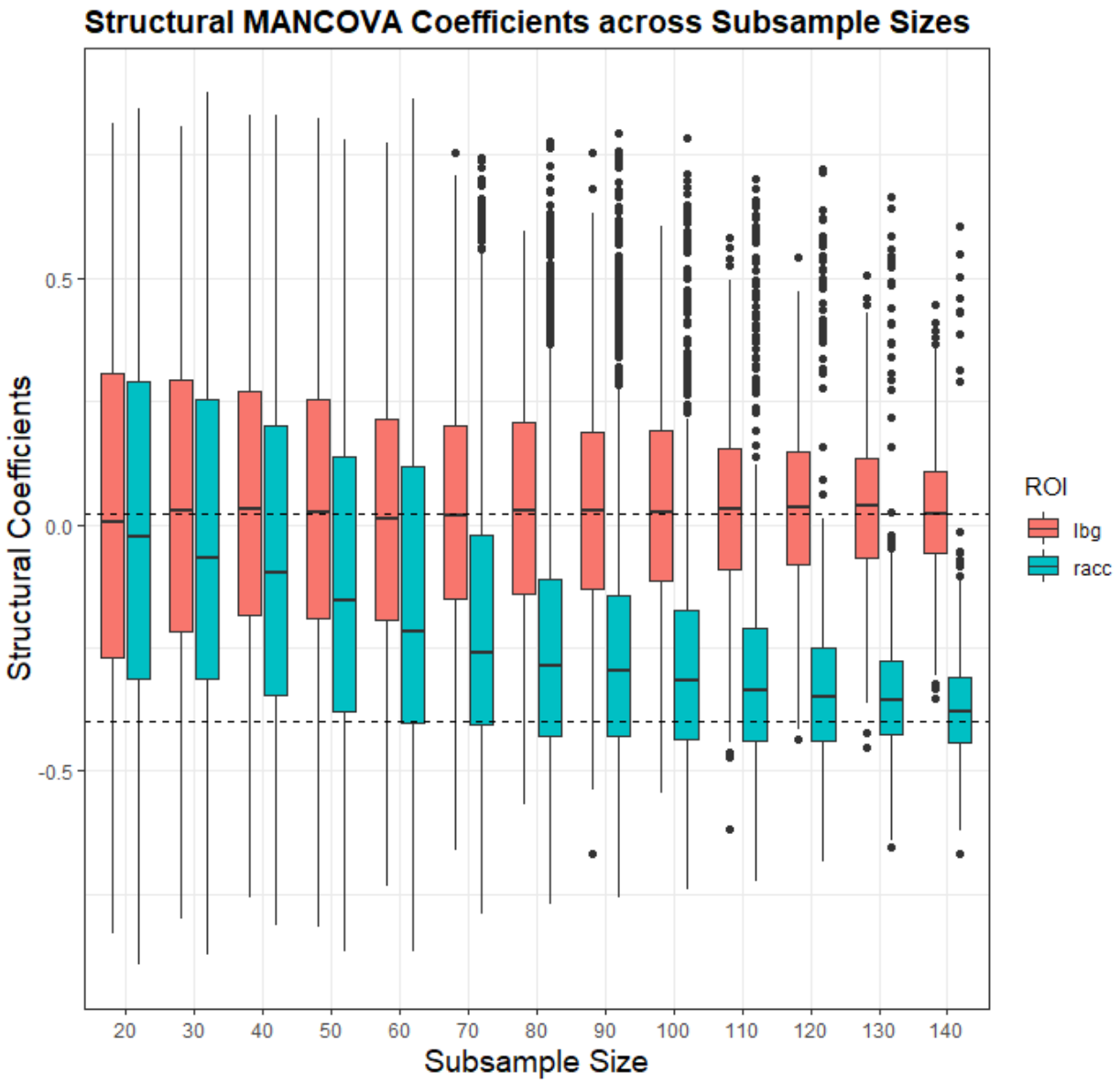

Figure 21

Variability in structural MANCOVA coefficients across subsample sizes for the RACC (largest structural coefficient) and the LBG (smallest structural coefficient). While the LBG hovers around its estimated coefficient of roughly 0, and slightly decreases in the amount of variability around this median as subsample sizes increase, the RACC underestimates the size of its structural coefficient at lower subsample sizes, and approaches it as subsample sizes increase. Only for the highest subsample sizes do estimates of the RACC standardized coefficients begin to normalize around its true value.

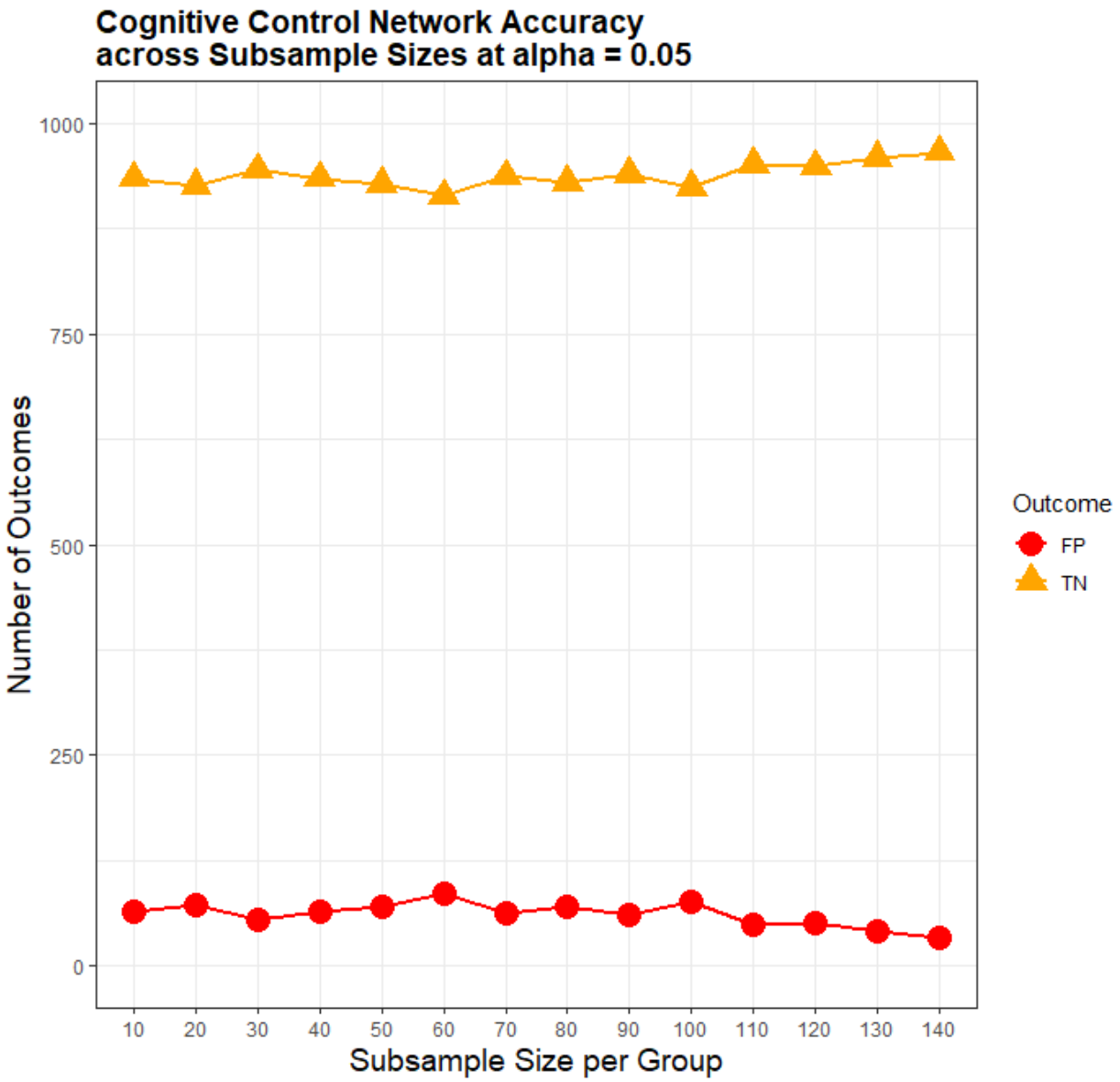

Figure 22

Number of FP and TN outcomes using combined (summed) Cognitive Control Network ROIs as the dependent variable in a linear multiple regression analysis. Because the full-sample test was non-significant at $p < 0.05$, subsample outcomes are only either FP or TN. Rates do not appear to change across subsample sizes.

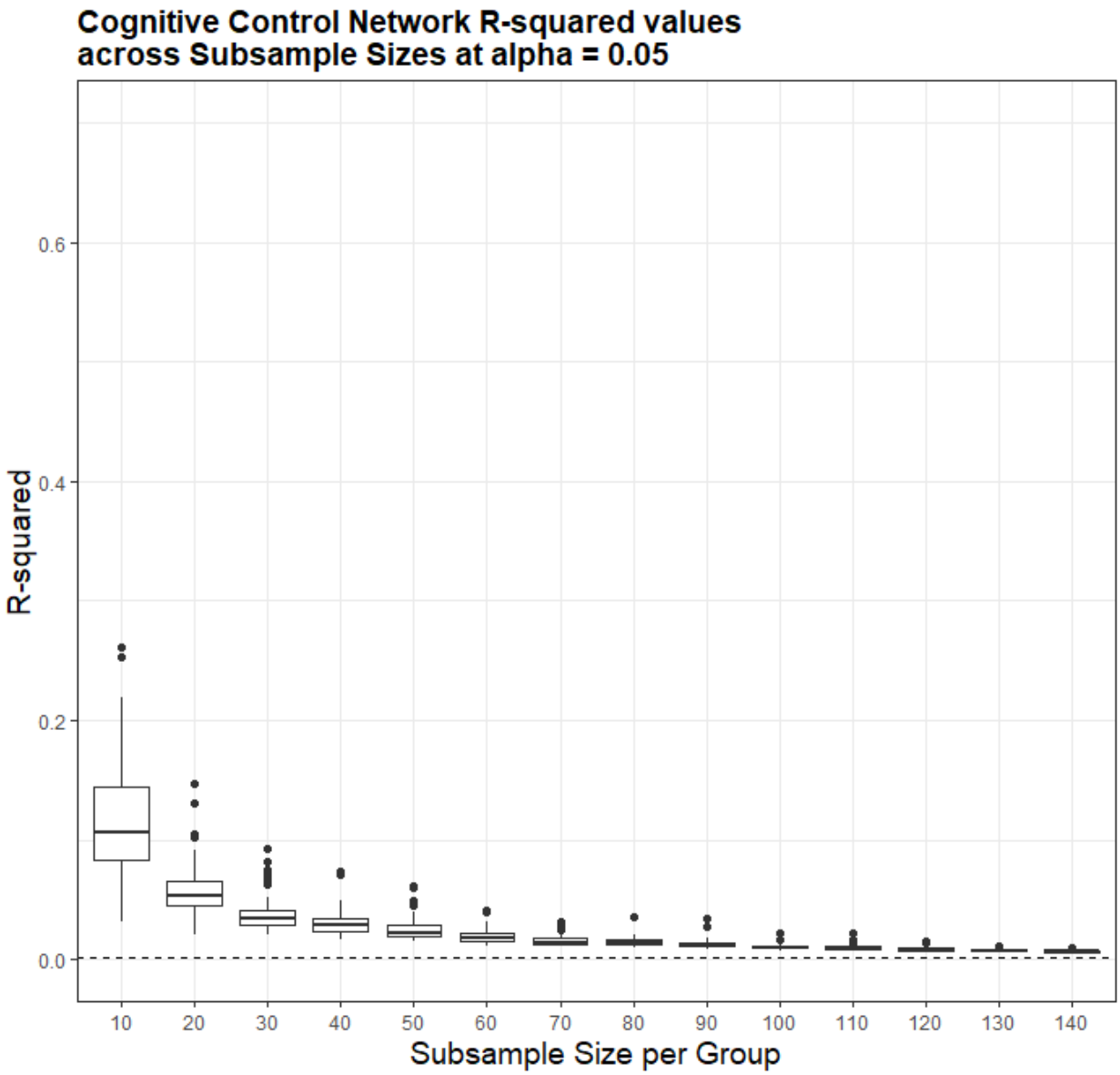

Figure 23

Boxplot of significant R-squared effect sizes across subsample sizes where $p < 0.05$, using the summed Cognitive Control Network ROIs as the outcome variable. Again, the average R-squared for subsamples of 10 per group is inflated relative to higher subsample sizes which approach the very small, true full-sample significant R-squared value of 0.002.

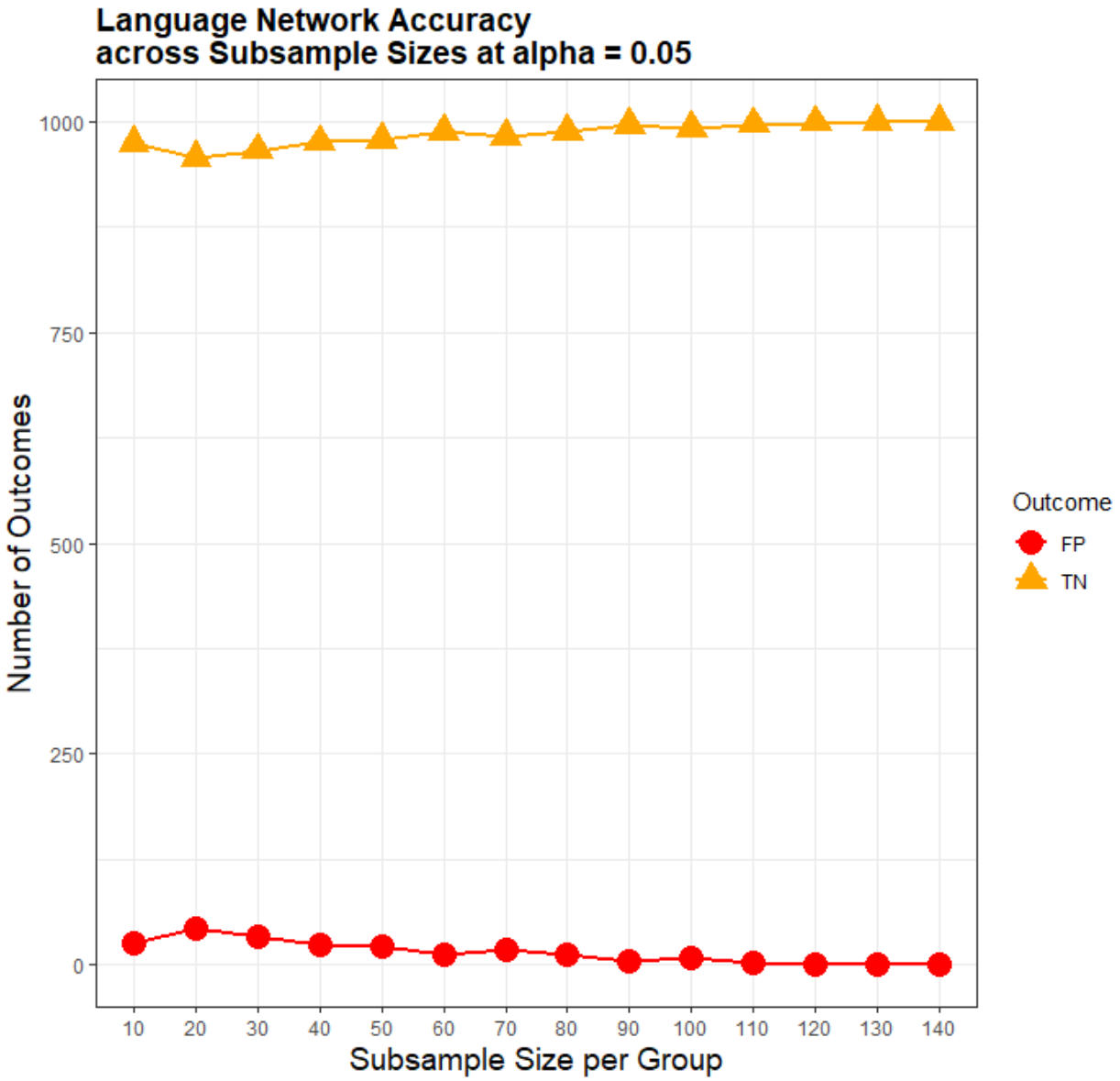

Figure 24

Number of FP and TN outcomes using combined (summed) Language Network ROIs as the dependent variable in a linear multiple regression analysis. Because the full-sample test was non-significant at $p < 0.05$, subsample outcomes are only either FP or TN. Rates do not appear to change across subsample sizes.

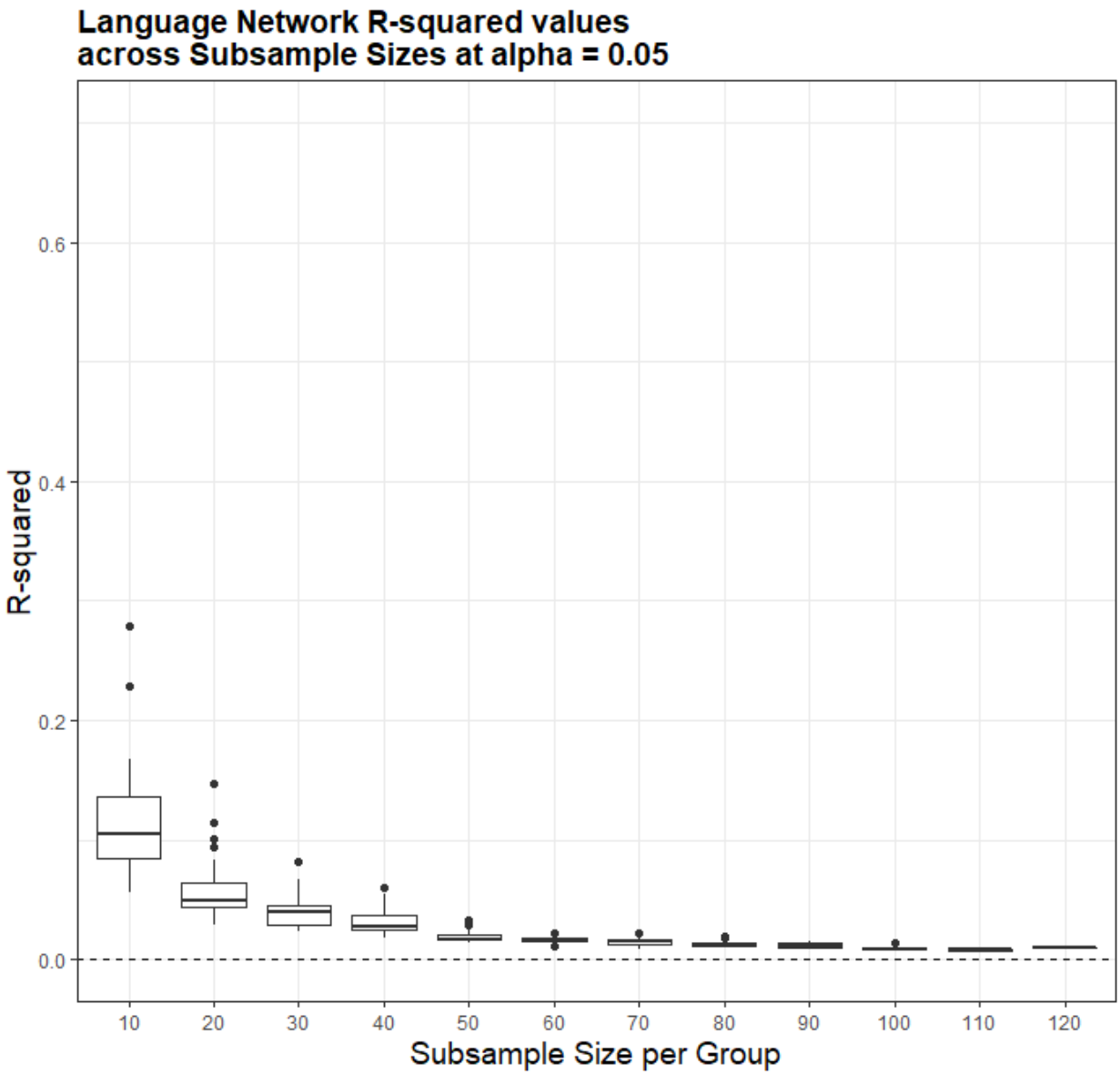

Figure 25

Boxplot of significant R-squared effect sizes across subsample sizes where $p < 0.05$, using the summed Language Network ROIs as the outcome variable. Again, the average R-squared for subsamples of 10 per group is inflated relative to higher subsample sizes which approach the very small, true full-sample significant R-squared value of 0.0002.

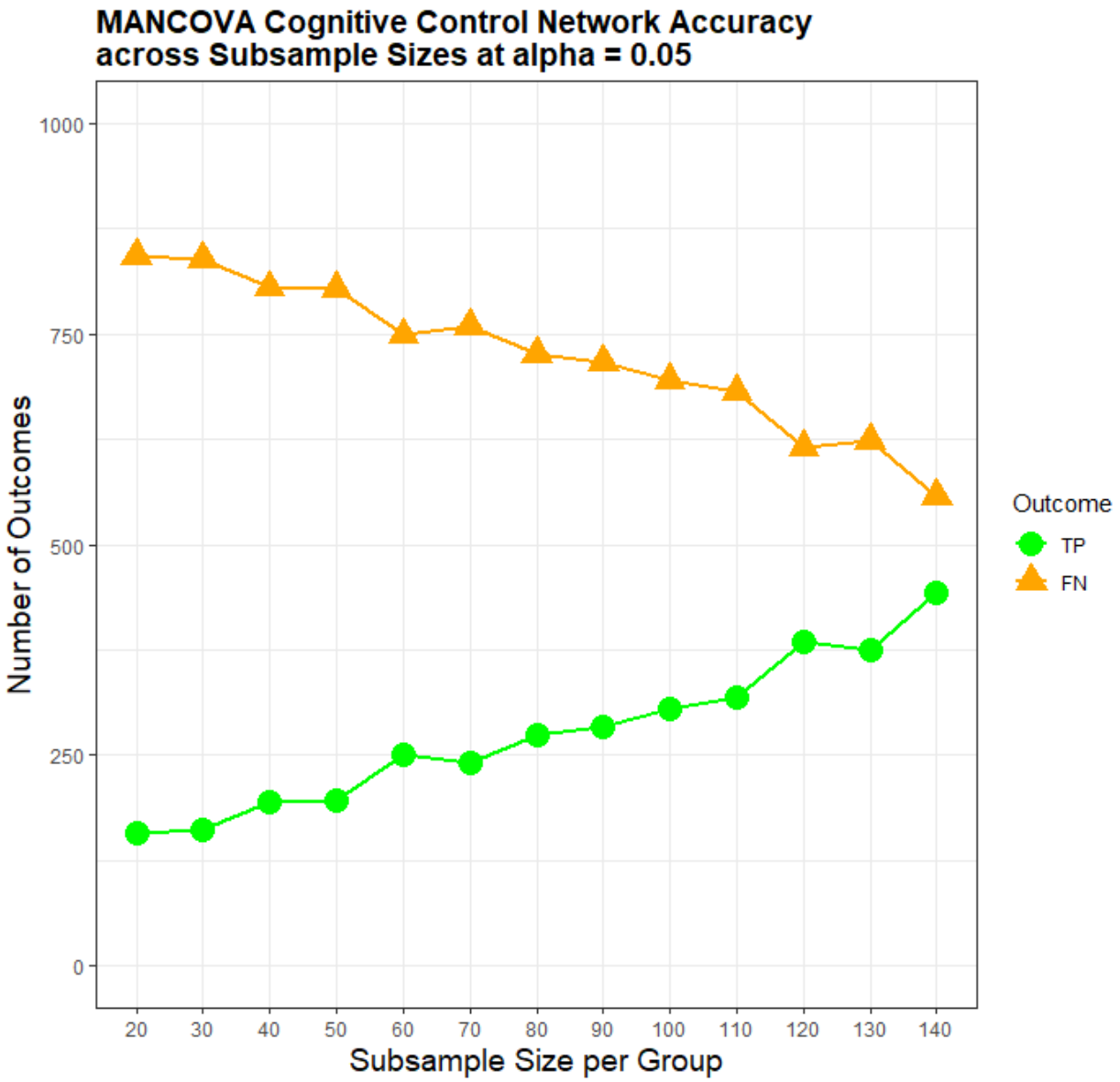

Figure 26

Number of TP and FN outcomes using combined (summed) Cognitive Control Network ROIs as the dependent variable in a linear multiple regression analysis. Because the full-sample test was significant at $p < 0.05$, subsample outcomes are only either TP or FN. TP rates increase across subsample sizes, showing an improvement in accuracy, though it doesn't quite reach 50% of TP outcomes even at 140 participants per group- likely due to the very small Eta-squared effect size, even though it is a significant test.

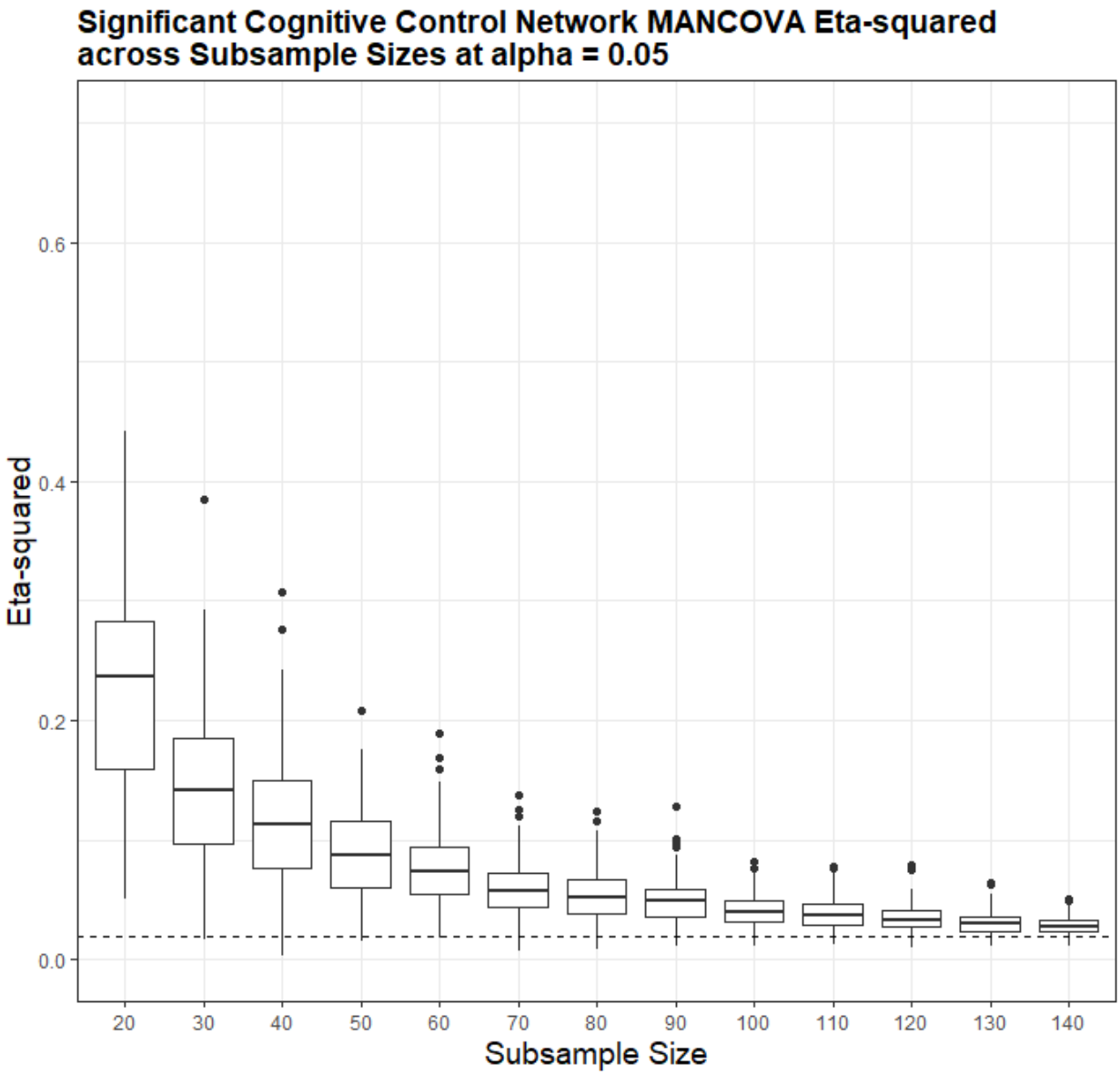

Figure 27

Boxplots of Eta-squared for significant Cognitive Control Network MANCOVAs across subsample sizes.

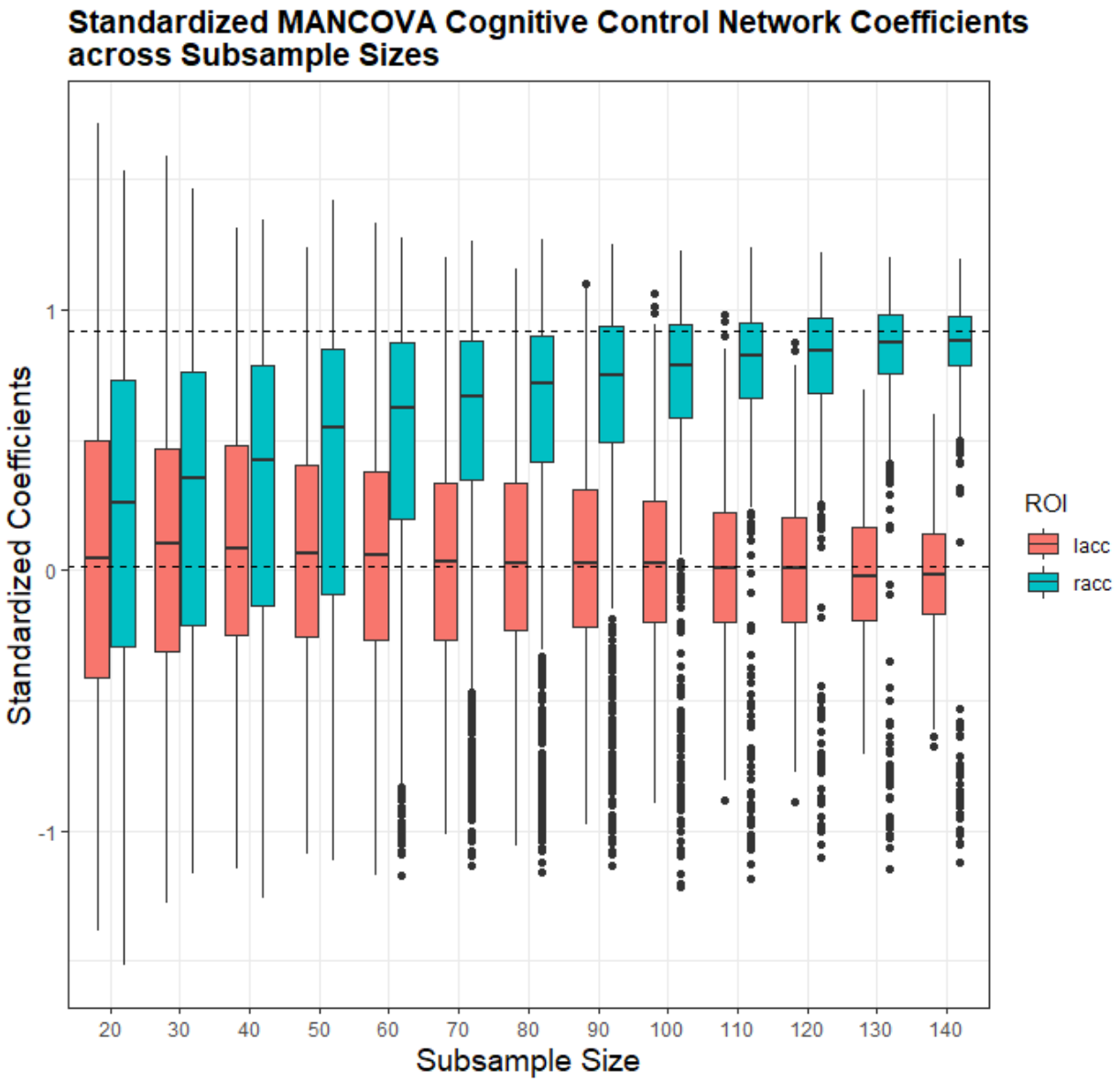

Figure 28

Variability in standardized MANCOVA coefficients across subsample sizes for the LACC (largest standardized coefficient) and the RACC (smallest standardized coefficient) of Cognitive Control Network dependent ROIs.

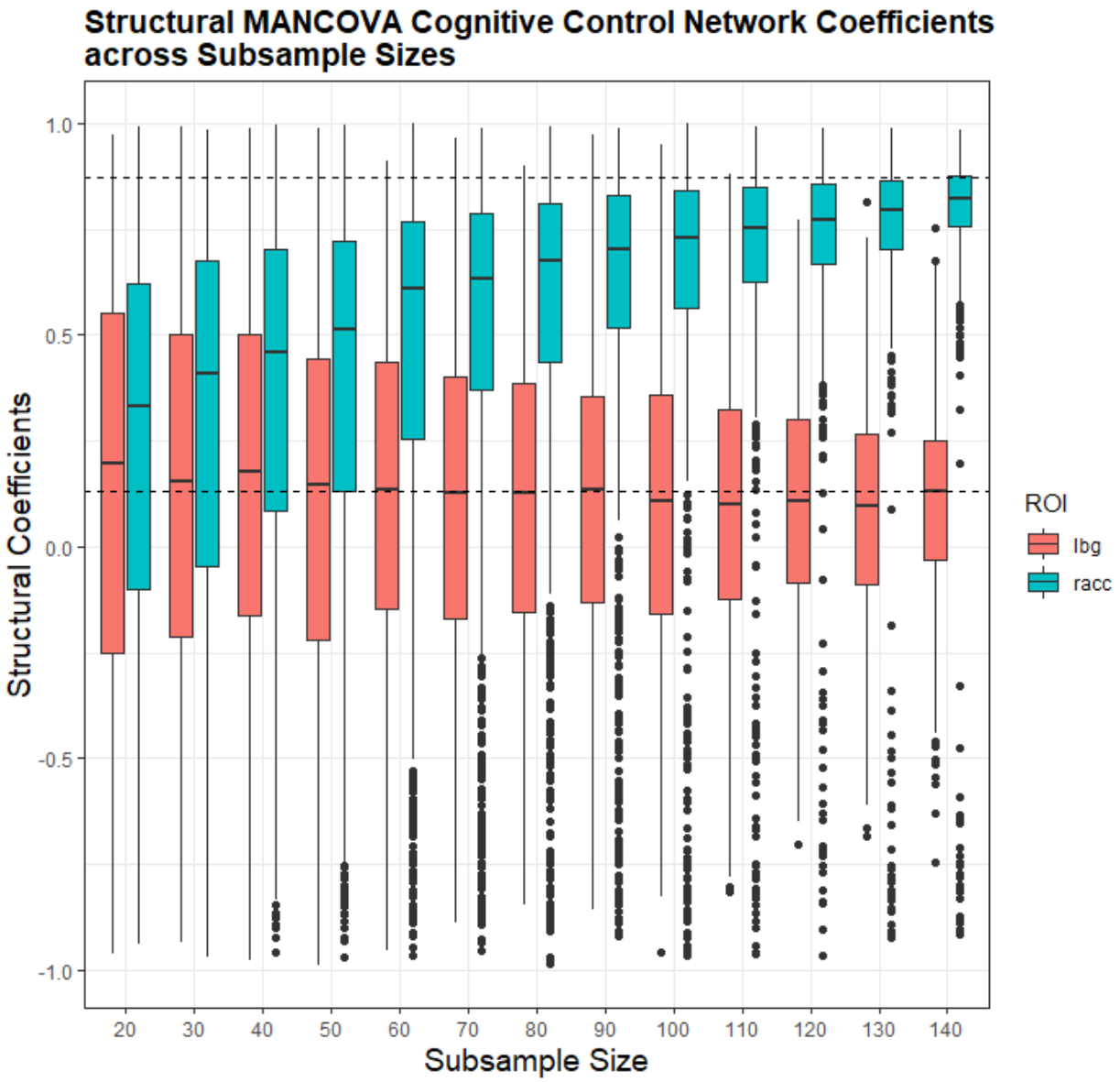

Figure 29

Variability in standardized MANCOVA coefficients across subsample sizes for the RACC (largest structural coefficient) and the LBG (smallest structural coefficient) of Language Network dependent ROIs.

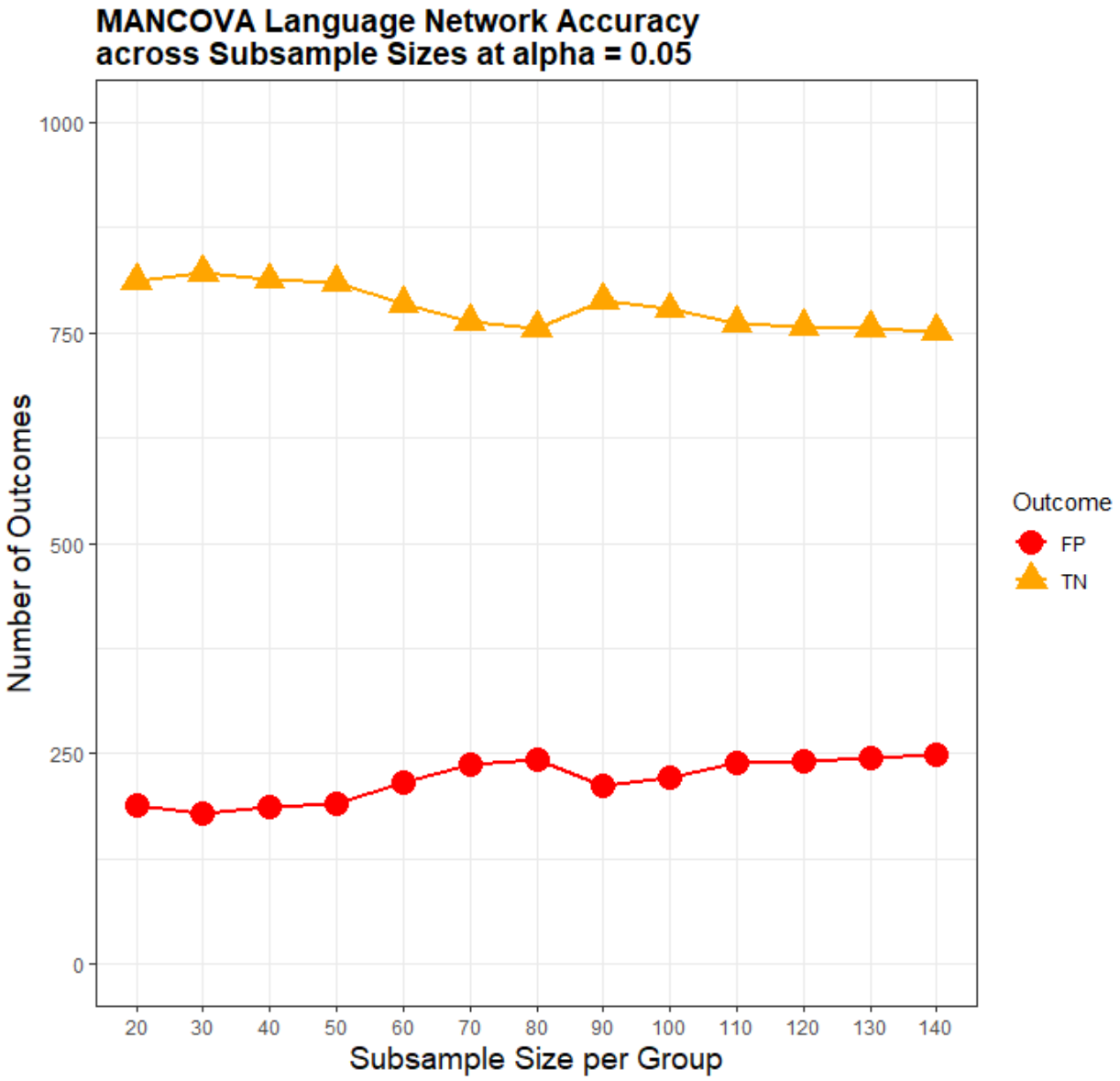

Figure 30

Number of FP and TN outcomes using combined (summed) Language Network ROIs as the dependent variable in a linear multiple regression analysis. Because the full-sample test was significant at $p < 0.05$, subsample outcomes are only either FP or TN. FP rates are relatively high throughout (just below 1 in 4 outcomes), and very slightly increase as subsample sizes increase.

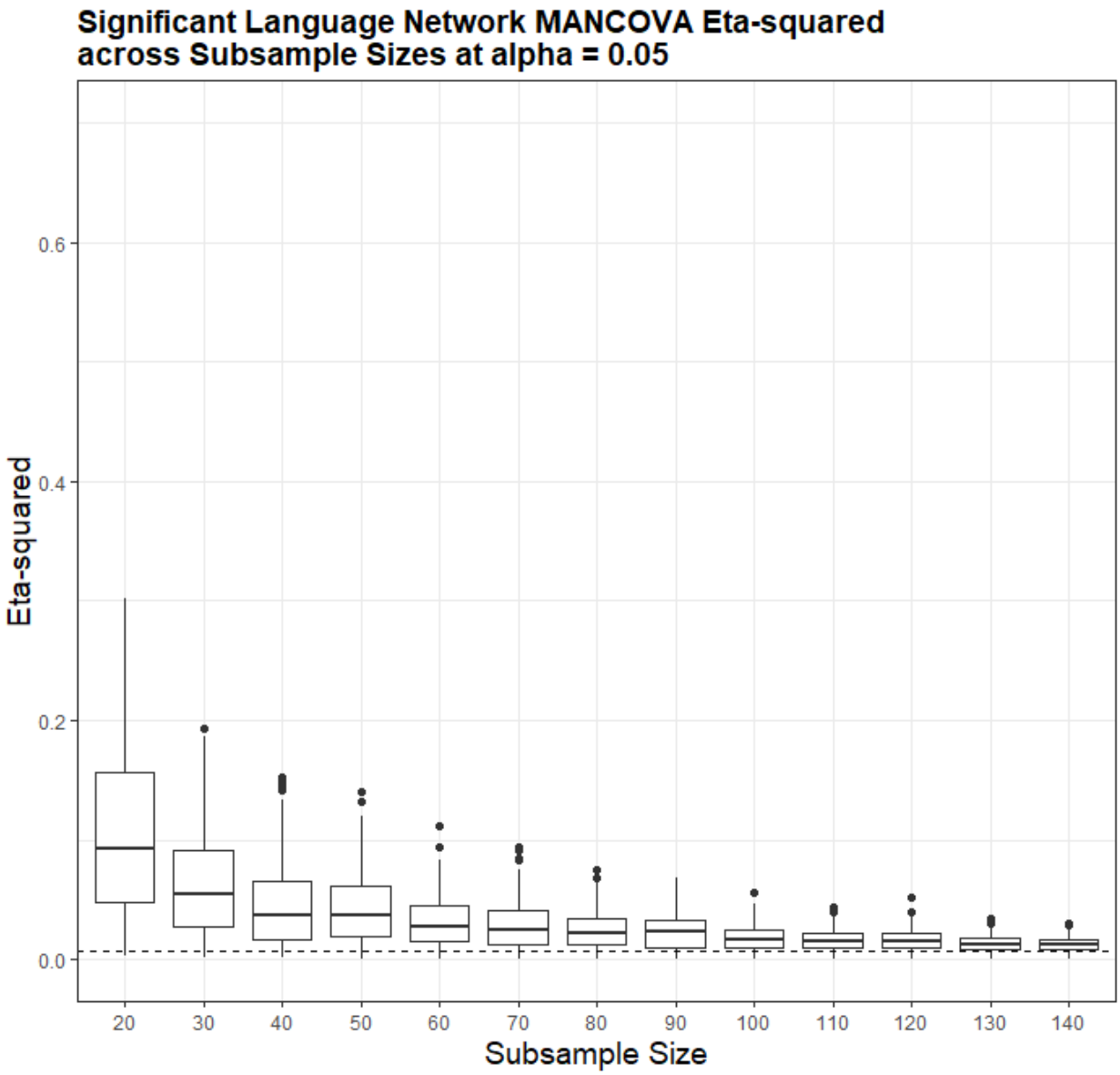

Figure 31

Boxplots of Eta-squared for significant Language Network MANCOVAs across subsample sizes.

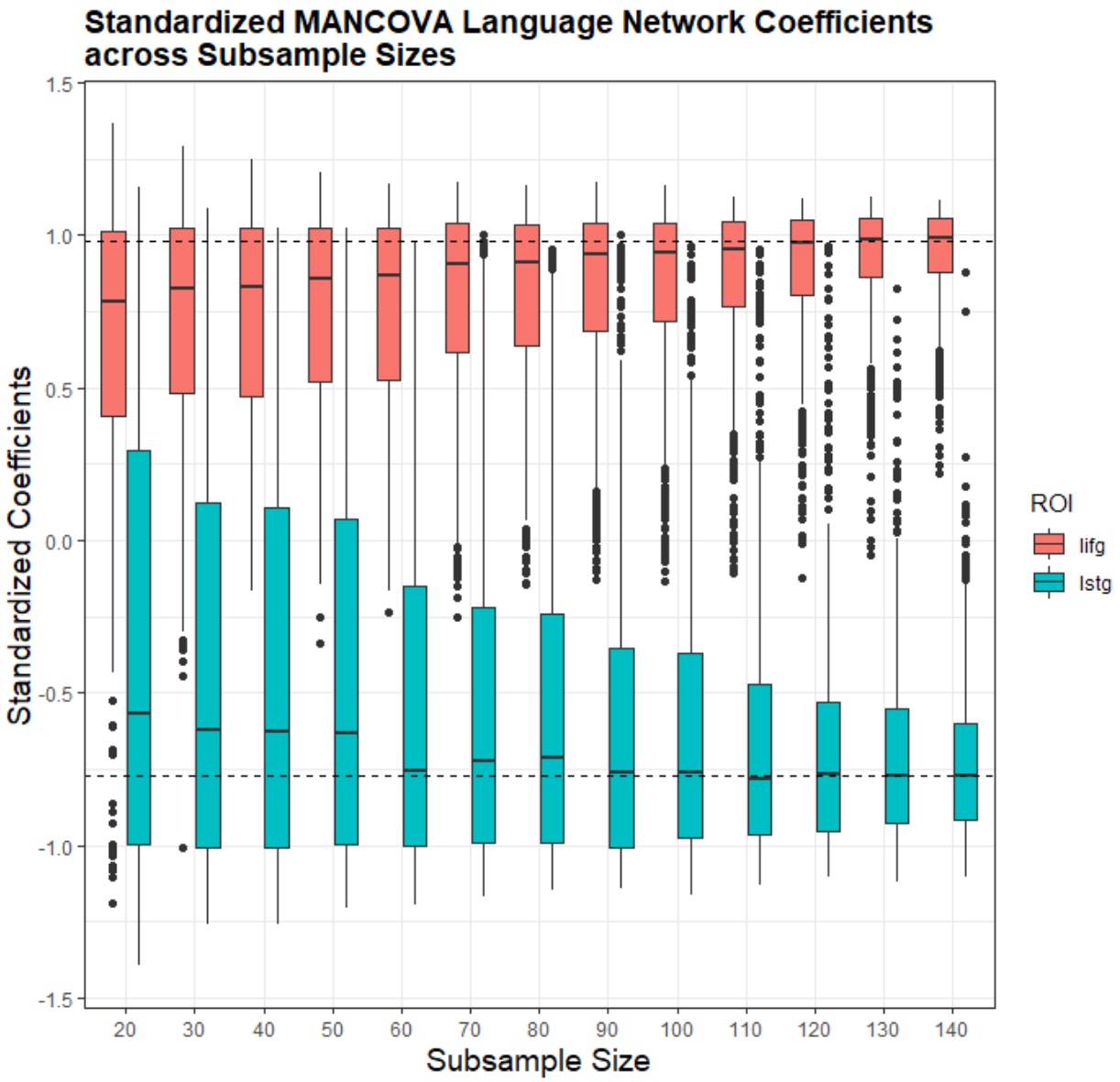

Figure 32

Variability in standardized MANCOVA coefficients across subsample sizes for the LIFG and the LSTG as the only Language Network dependent ROIs.

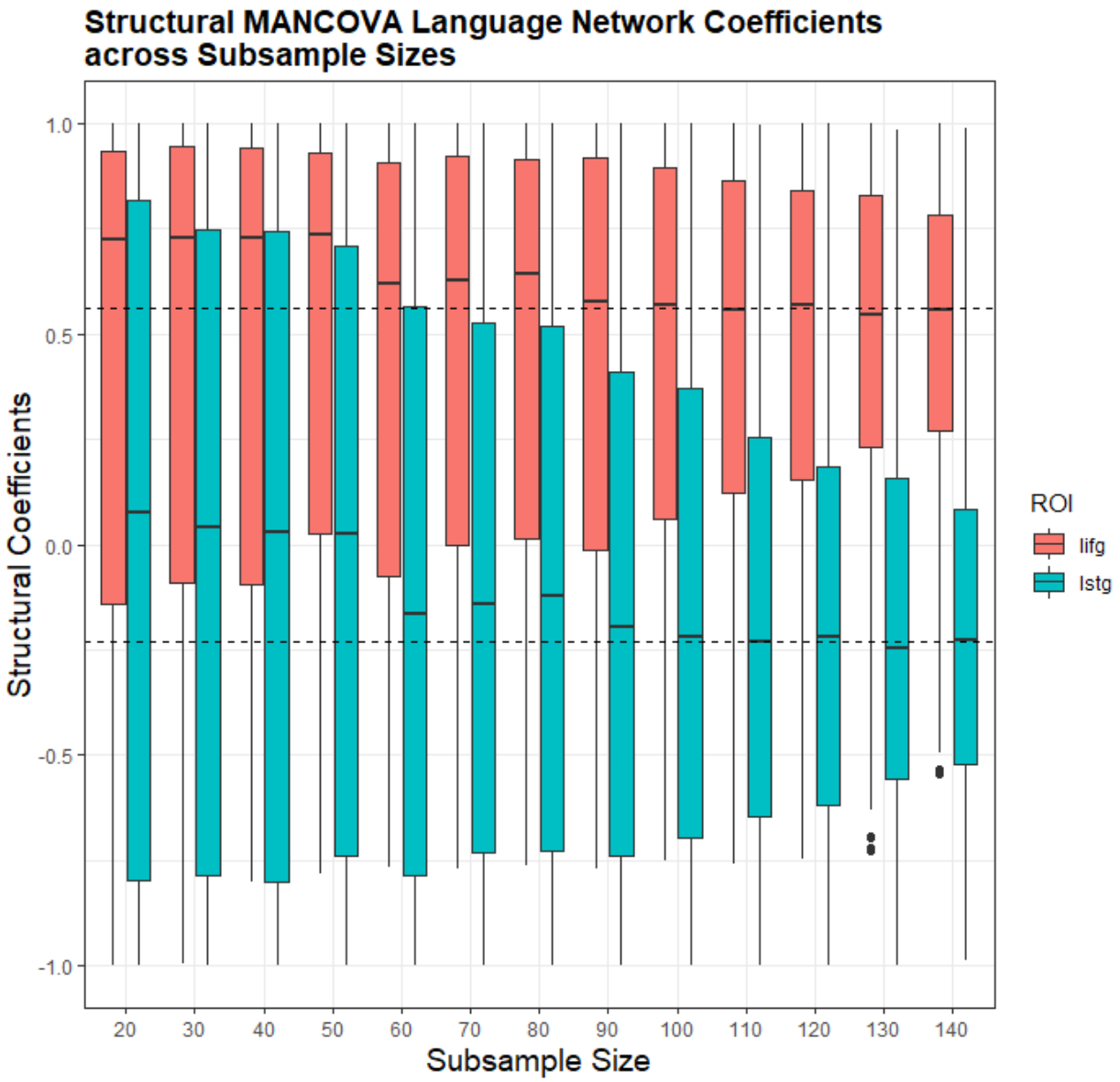

Figure 33

Variability in structural MANCOVA coefficients across subsample sizes for the LIFG and the LSTG as the only Language Network dependent ROIs.

Interestingly, estimates of the structural coefficient for LIFG are slightly inflated in the lower subsample sizes, and stabilize around the true value as subsample sizes increase. All other MANCOVA structural coefficient figures show trends for underestimations of MANCOVA structural coefficients at lower subsample sizes.