Testing for a Lack of Relationship Among Categorical Variables

Tanja Shishkina (orcid.org/0000-0002-1500-5520)

Linda Farmus (orcid.org/0000-0002-5303-6408)

Robert A. Cribbie (orcid.org/0000-0002-9247-497X)

Quantitative Methods Program

Department of Psychology

York University

Abstract

Determining a lack of association among two or more categorical variables is frequently necessary in psychological designs such as comparative outcome analyses, assessments of group equivalence at a baseline level, and therapy outcome evaluations. Despite this, the literature rarely offers information about, or technical recommendations concerning, the appropriate statistical methodology to be used to accomplish this task. This paper explores two equivalence tests for categorical variables, one introduced by Rogers, Howard, and Vessey (1993) and another by Wellek (2010), as well as a proposed strategy based on Cramer's *V*. A simulation study was conducted to examine and compare the Type I error and power rates associated with these tests. The results indicate that an equivalence-based Cramer's *V* procedure is the most appropriate method for determining a lack of relationship among categorical variables in two-way designs.

Keywords: equivalence testing, categorical variables, frequency tables

Testing for a Lack of Relationship

Among Categorical Variables

In psychological research, investigators are often interested in confirming a lack of association among two or more categorical variables. This is frequently the case for designs such as comparative outcome analyses and the assessment of group equivalence on categorical variables at a baseline level. For instance, researchers may wish to demonstrate that two or more treatments are equal based on their therapeutic outcomes, such as return to work, reduction of debt, medication adherence, or abstinence from gambling (Chen, Tsong, & Kang, 2000). In addition, experimental groups are frequently assessed at the start of a study in an effort to establish the equivalence of demographic and/or clinical variables (Rogers, Howard, & Vessey, 1993). In these circumstances, it is crucial for researchers to employ the appropriate methodology to conclude that categorical variables are equivalent across the groups.

Buoli, Cumerlato Melter, Caldiroli, and Altamura (2015) provide a practical example of statistical efforts to establish equivalency. The authors examined the efficacy of differing pharmacological classes of antidepressants on the long-term treatment of major depressive disorder. At baseline, the investigators needed to demonstrate the equivalence of the experimental groups on demographic characteristics (e.g., work and marital status) and clinical characteristics (e.g., family history of mental problems, number of suicide attempts, and substance abuse involvement). The authors applied chi-square tests of independence to determine whether the experimental groups were equivalent on these variables. The goal was to find a non-significant result so as to retain the null hypothesis that the groups showed equal outcomes.

Another example derives from a study by Bailine et al. (2010) who sought to assess whether bipolar and unipolar depression patients responded equally to electroconvulsive therapy. It was necessary at baseline for the examiners to prove the equality of the unipolar and bipolar subjects in terms of demographic traits (such as gender, race, education, and marital status) and clinical variables (such as the presence or absence of psychosis). Moreover, to evaluate the hypothesis that electroconvulsive therapy would affect these groups equally, the authors compared the two groups in terms of their responses to treatment, identifying a 50% reduction from baseline on the Hamilton Rating Scale for Depression as a positive outcome. Correspondingly, chi-square tests of independence were conducted for both the baseline comparisons and the response to treatment hypotheses, with the goal of demonstrating the absence of statistical significance (i.e., the traditional null hypothesis is not rejected).

These two examples highlight investigations in which researchers explore a lack of relationship among categorical variables; they also illustrate the misuse of traditional null hypothesis tests for identifying such an absence. It is important to state outright that we are not interested in criticizing the decisions made by the authors in the above-discussed examples; indeed, these two studies were selected from among countless others that followed similar procedures. Rather, we wish to emphasize that suitable tests for assessing lack of relationship among categorical variables are not widely known at present nor are they regularly available in statistical software packages. Furthermore, explicit discussions of the limitations of traditional methods for assessing equivalence among categorical variables are scarce. Also, while these examples provide illustrations of situations that explore homogeneity (i.e., equivalence of the proportions across groups), the same issues apply when investigating independence. This paper addresses the relative dearth of information regarding the insufficiencies of conventional methods for assessing the absence of a relationship among categorical variables and suggests more robust methodological tools better suited to this task.

Introduction to Equivalence Testing

As the previous examples demonstrate, researchers commonly try to infer the equivalence of groups or establish the lack of a relationship among variables based on the absence of significant differences or associations. However, this method is not appropriate for several reasons. First, as Quertemont (2011) notes, non-significant results are often due to insufficient statistical power. Thus, previously statistically insignificant differences between groups may actually become significant once the sample size is increased sufficiently. Second, the failure to reject a null hypothesis does not mean that the null hypothesis is true; it simply means that there is inadequate evidence at present to conclude that it is incorrect (Walker & Nowacki, 2010). The theoretical statement of the null hypothesis for equivalence tests is exactly opposite to the assertion of the null hypothesis for traditional difference-based tests (Cribbie, Arpin-Cribbie, & Gruman, 2009). These dynamics suggest the need for statistical procedures dedicated specifically to testing for a lack of association among variables, which is precisely the aim of equivalence testing.

To be able to test for a lack of a relationship among variables, investigators begin by choosing the smallest degree of association that their study will recognize as practically significant. In practice, sampling error makes nil associations (e.g., identical means, zero correlations) impossible (Counsell & Cribbie, 2015). The purpose of equivalence testing is not to test for a total lack of association among variables, but rather to examine whether the differences discovered are relevant (Cribbie, Gruman, & Arpin-Cribbie, 2004). To accomplish this task, researchers must quantify their conception of irrelevant difference by deciding upon a specific range of values called an equivalence interval, often denoted symmetrically using (-δ, δ); δ may represent any effect of interest, such as a lack of correlation or an irrelevant difference in proportions. The equivalence interval generally has both an upper and a lower limit, with that particular range representing the smallest association (e.g., a difference in population proportions) that the framework of the study would consider meaningful.

The null hypothesis of equivalence testing asserts that the relationship among the variables is at least as large as the effect specified by the investigator through the equivalence interval. Conversely, the alternative hypothesis contends that the relationship among the variables is smaller than the one specified through the equivalence interval. Equivalence or lack of association is established when the data provide enough evidence to conclude that the magnitude of the relationship falls within the equivalence interval (Schuirmann, 1987; Walker & Nowacki, 2011). There are no fixed rules for establishing equivalence margins; their justification depends heavily on the nature of the research, the outcome variable of interest, previous findings in specific research areas, and the risk/benefit judgments of relevant experts (Committee for Medicinal Products for Human Use, 2006). For example, O'Reilly et al. (2007) tested the equivalence of telepsychiatry and face-to-face psychiatric consultation. One of the outcome measures was the proportion of participants with psychiatric admissions during the twelve months after the initial assessment. The investigators, in consultation with psychiatrists, decided that a difference in proportions of 10% between groups would be the smallest clinically significant difference, resulting in a nondirectional equivalence margin of (-δ, δ) = (-.10, .10).

Equivalence Tests for the Relationship among Categorical Variables

In this project we examined three approaches for testing for a lack of association among categorical variables. The first, described by Rogers et al. (1993), is a modified version of the two one-sided tests (TOST; Schuirmann, 1987) procedure, which aims to examine the equivalence of two proportions, denoted as p_1 and p_2. The test is based on the normal approximation of the difference between two proportions. We will refer to this test of the equivalence of two proportions as the "EP" test. Although framed as a test of homogeneity, it can also be used to assess the independence of two categorical variables. For example, researchers could utilize this test to show that the proportion of males and females in the control and experimental groups are similar, or that sex (male/female) is minimally related to choice of pain medication (Drug A vs Drug B).

The first null hypothesis, H_{01}: $p_1 - p_2 \leq -\delta$, is rejected if $z_1 \geq z_{1-\alpha}$, and the second, H_{02}: $p_1 - p_2 \geq \delta$, is rejected if $z_2 \leq z_\alpha$, where:

$$z_1 = \frac{(\hat{p}_1 - \hat{p}_2) - (-\delta)}{s_{\hat{p}_1 - \hat{p}_2}},\ z_2 = \frac{(\hat{p}_1 - \hat{p}_2) - \delta}{s_{\hat{p}_1 - \hat{p}_2}},$$

$\hat{p}$ is the sample proportion, and $z_{1-\alpha}$ and z_α are values from a standard normal distribution that cut off the lower 1-α and α proportions of the distribution, respectively. Note that $(-1)z_\alpha = z_{1-\alpha}$. The standard error of the difference between two proportions can be calculated using:

$$s_{\hat{p}_1 - \hat{p}_2} = \{[\hat{p}_1\,(1 - \hat{p}_1)/n_1] + [\hat{p}_2\,(1 - \hat{p}_2)/n_2]\}^{1/2},$$

where n_1 and n_2 are the sample sizes for groups one and two, respectively. When both null hypotheses of the EP method are rejected, investigators can reject the null hypothesis that the difference in the proportions is greater than δ ($-\delta < p_1 - p_2 < \delta$; the difference in the proportions falls within the equivalence interval) or, in other words, that there is no relationship among the two dichotomous variables. The EP procedure is operationally comparable with the simple asymptotic interval (SAI) approach (Barker, Rolka, Rolka, & Brown, 2001). According to the SAI method, if the $\hat{p}_{1} - \hat{p}_2 \pm z_\alpha s_{\hat{p}_1 - \hat{p}_2}$ confidence interval (CI) for p_1-p_2 is within the equivalence interval (-δ, δ), then both H_{01} and H_{02} are rejected at a predetermined α level. An important limitation of the EP test is that it is only applicable to 2 x 2 designs.

Another equivalence testing procedure, described by Wellek (2010), is based on Euclidean distance (i.e., the distance between two points in Euclidean space). We will refer to it as the "ED" procedure. The null hypothesis for this test states that the sum of the squared distances (D^{*2}) between the observed cell probabilities, denoted as π, and the expected cell probabilities (the product of marginal totals), denoted as g(π), in the population, is at least as large as the critical distance. Wellek suggested ε = .15 as the largest acceptable distance between π and g(π) (i.e., between the vector of observed probabilities and the vector of expected probabilities), however researchers should consider what value of ε is most appropriate given the nature of their researcher (and more research is required on understanding the magnitude of ε in order to assist researchers in setting appropriate values for ε). Again, this test could be used for investigating either independence or the homogeneity of group proportions. Thus, H_0: $D^{*2} \geq \varepsilon^2$ is rejected if:

$$D^2 + z_{1-\alpha} v_n / \sqrt{n} < \varepsilon^2,$$

where $v_n / \sqrt{n}$ is the standard error and:

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\pi_{ij} - g(\pi_{ij})\right)^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\pi_{ij} - \pi_{i+}\pi_{+j}\right)^2,$$

where *i*,*j* specifies the row and column, respectively, *r* is the number of rows, *c* is the number of columns, π_{i+} are the sum of the observed probabilities for row *i*, and π_{+j} are the sum of the observed probabilities for column *j*. Another way to frame the ED test is that the null hypothesis of an important relationship among the two categorical variables can be rejected if the upper limit of the CI for D^2 is less than ε^2. The variance, v_n^2, can be expressed as:

$$v_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\hat{d}_{ij}^2\pi_{ij} - \sum_{i_1=1}^{r}\sum_{j_1=1}^{c}\sum_{i_2=1}^{r}\sum_{j_{2=1}}^{c}\hat{d}_{i1j1}\hat{d}_{i2j2}\pi_{i1j1}\pi_{i2j2},$$

where:

$$\hat{d}_{ij} = 2\{(\pi_{ij} - \pi_{i+}\pi_{+j}) - \sum_{a=1}^{r}[(\pi_{aj} - \pi_{a+}\pi_{+j})\pi_{a+}] - \sum_{b=1}^{c}[(\pi_{ib} - \pi_{i+}\pi_{+b})\pi_{+b}]\}.$$

Finally, we propose an approach based on Cramer's *V* (Cramer, 1946), referred to in this paper as CV. Cramer's *V*, an effect size measure for the association among categorical variables, takes into account the dimensions of the frequency table, implying that *V* for tables of different dimensions can be meaningfully compared (Smithson, 2003). Thus, Cramer's *V* can be used to determine a lack of association among categorical variables in general two-way (or higher) tables, and since *V* ranges from 0 to 1 the task of finding an appropriate equivalence interval is made easier.

To reject the null hypothesis of nonequivalence for the CV approach (H_0: $V \geq \delta$), the upper limit of the Cramer's *V* CI should lie below the prespecified equivalence bound δ. Cramer's *V* is computed as:

$$V = \sqrt{\chi^2/[n(k-1)]},$$

where *k* is the smaller of the number of rows *r* or columns *c*. Following Smithson (2003), the CI for Cramer's V is calculated as:

$$V_L = \sqrt{(\Delta_L + m)/[n(k-1)]}\text{, and}$$

$$V_U = \sqrt{(\Delta_U + m)/[n(k-1)]}$$

where *m* = (*c* – 1)(*r* – 1) and Δ_L and Δ_U represent the lower and upper confidence limits of the noncentrality parameter for noncentral χ^2 (which are determined through iteration, see Signorell, 2017). We have included an applied example for all three methods (see Appendix A).

Method

A Monte Carlo study was conducted to evaluate the Type I error and power rates of the EP, ED and CV equivalence testing procedures for categorical variables in 2 x 2 (EP, ED, CV) and 2 x 4 (ED, CV) designs. We used $\alpha = .05$ and performed 5000 simulations for each condition using the open-source statistical software R (R Development Core Team, 2016). The manipulated variables were sample size, degree of association (for the power condition) and study design (the specific conditions can be found in Table 1). Sample sizes of 50, 100, 200, and 1000 were investigated because these are commonly used in psychological research. We focused only on the case of equal row sums for the 2 x 2 study design, while in the 2 x 4 study design row sums were not necessary equal. Note that Wellek (2010) recognizes the liberal nature of the ED procedure (i.e., the empirical Type I error rates can exceed α) and therefore recommends using an adjusted nominal level $\alpha^{*} < \alpha$ “whenever strict maintenance of the prespecified level is felt to be an indispensible requirement” (p. 277). Thus, a nominal α level of .05 was employed for all tests except the ED, where we used both $\alpha = .05$ and $\alpha^* = .025$.

When the degree of association between variables exactly matches the equivalence interval, the empirical Type I error rate is expected to equal the nominal Type I error rate (α). We used Bradley’s (1978) liberal bounds, $\alpha \pm .5\alpha$, as the criteria for having satisfactory Type I error control. Thus, with $\alpha = .05$, Type I error rates are acceptable if they fall between .025 and .075. Since the EP, ED and CV tests use different scales, the procedures for determining the equivalence intervals were different as well. Following Wellek’s (2010) recommendations for the ED test, the equivalence interval was set as $\varepsilon = (-.15, .15)$. Thus, for the Type I error conditions, we chose values for the population cell proportion that produced the Euclidean distance $d = .15$ (the specific 2 x 4 cases were actually sampled from those derived by Wellek, 2010, p. 276). These cell proportions were then used to calculate comparable equivalence intervals for the EP and CV procedures (i.e., we computed the difference in proportions and Cramer's *V* for the population cell proportions that produce *d* = .15). As a result, in the Type I error conditions, the equivalence bound for the ED test was always the same, ε = .15, while for the EP and CV the interval changed for different population matrices (see Table 1).

For the power conditions, the strength of the association between two dichotomous variables, measured by the Euclidean distance, was set up to be equal to 0, .02 and .10. For the power conditions, the upper limit of the equivalence bound was at ε = .15 for the ED test, .30 for the EP test, and .40 for the CV test. As stated earlier, Wellek (2010) recommends ε = .15 for use with the ED test, and the values for EP and CV were comparable in strength to that for ε = .15.

We also conducted a chi-square test of independence to compare the performance of the traditional approach with that of the equivalence tests. It is important to highlight that since the goal is to demonstrate a lack of association (independence/homogeneity), this test would not be appropriate since the goal would be to not reject the null hypothesis; it is included though since this is often the method used by researchers to demonstrate a lack of association. In order to have a comparable outcome for each test, the outcome variable was the proportion of simulations in which the conclusion related to "no association". For the equivalence tests this means rejecting H_0, but for the traditional chi-square test of independence this means not rejecting H_0. Note, therefore, that in the Type I error condition and the nonzero effect power condition for the equivalence tests that the reported rates for the traditional chi-square test are Type II errors, and for the null effect power condition for the equivalence tests this translates into rates of correct nonrejections for the chi-square test (with an expected proportion of 1- α).

Results

2 x 2 Design

Type I Error Rates. The proportion of cases in which a lack of association between two dichotomous variables is falsely concluded for the EP, ED, and CV tests (Type I error rates), as well as the probability of a Type II error for the chi-square test of independence, are presented in Figure 1.

Both the EP and CV procedures had Type I error rates that fell within Bradley's limits (.025 - .075). Contrary to both the EP and the CV, sample size significantly influenced the Type I error rates of the ED test; with a small sample size (*N*= 50) the Type I rates were twice as large as compared to *N* = 1000 when α = .05, and on average three times larger for *N* = 50 as compared to *N* = 1000 when α = .025. Only with *N* = 1000 does the ED's empirical Type I error rates fall within Bradley's limits.

The proportion of non-rejections of the null hypothesis for the chi-square test, when in fact it is false (Type II error rates), is presented herein for comparison with the equivalence tests. As expected, Type II error rates for the chi-square test are strongly related to sample size. Thus, for small sample sizes, such as *N* = 50, the chi-square test has the greatest probability of declaring equivalence in comparison to all the equivalence tests examined by this paper, whereas it has the lowest probability of declaring equivalence when the sample size increases to *n* = 1000. Both of these results are expected given that what is recorded are Type II errors.

Power Rates. The probabilities of correctly concluding equivalence for the ED, EP, and CV procedures (power rates) are presented in Figure 2. All the equivalence tests examined in this paper produce similar patterns in terms of power rates, i.e., as expected they increase as sample size increases. The ED with α = .025 and CV tests have similar power rates across different degrees of association. The EP procedure displays less power than other equivalence tests when the sample size is *N* = 50. However, this difference in power rates disappears when the sample size increases to *N* = 200. As anticipated, the degree of association among categorical variables has a substantial influence on power, with *d* = 0 producing markedly higher probabilities than *d* = .1. Population cell proportion patterns moderately impact the power for the EP procedure (e.g., .250, .250, .250, .250 versus .100, .400, .100, .400), but not for the ED or CV tests.

In the power condition, when *d* = .02 or *d* = .10, the reported rates for the traditional chi-square test of independence are the probabilities of not rejecting H_0 (Type II error). As Figures 1 and 2 reveal, as expected, a bigger (rather than a smaller) Euclidean distance – which reflects the degree of association between categorical variables – results in smaller Type II error rates.

2 x 4 Design

Type I Error Rates. Figure 3 displays the empirical Type I error rates for the ED and CV tests and the probability of a Type II error for the chi-square test of independence in the 2 x 4 study design. The ED test with α = .05 reveals inappropriate empirical Type I error rates (predominantly for Condition 1) for sample sizes less than *N* = 200. When α = .025, the ED test's Type I error rates remain within Bradley's limits for all but one condition. Although slightly conservative, the Type I error rates for the CV procedure always fell within Bradley's limits for robustness. Both the ED and CV approaches were sensitive to the population cell proportions with Type I error rates varying slightly across the conditions. All the equivalence tests considered for the 2 x 4 study design are influenced by sample size, with smaller Type I error rates occurring with larger samples. As in the 2 x 2 condition, the proportion of cases in which the traditional chi-square test of independence falsely concluded that the association was nil was moderate for the *N* = 50 condition but equal to or near zero for all *N*s > 50.

Power Rates. An examination of the values in Figure 4 indicates that the ED and CV tests show similar power rates when *d* = 0 (Conditions 1 and 2) and *d* = .02 (Conditions 3 and 4), although the ED test was generally more powerful. When *d* = .1 (Conditions 5 and 6), power rates decrease, particularly for the CV test, because the extent of the association comes very close to the equivalence interval. As expected, as the sample size increases, the power rates for both equivalence tests grow. For the traditional test, when the association is nil, the recorded rates were, as expected, approximately equal to 1 - α. When the association was greater than zero but within the equivalence interval, the Type II error rates decreased for the traditional chi-square test of independence as sample sizes increased. However, these rates highlight the problem with using a difference-based test to evaluate equivalence; with a small sample size the CS test often incorrectly concluded equivalence, whereas with a large sample equivalence was rarely or never concluded. In other words, the probability of inappropriately declaring a lack of association among the variables decreased, rather than increased, as sample sizes increased.

Discussion

Many psychological studies explicitly aim to show that there is no association among the categorical variables under investigation. Often, a researcher wishes to show that, before beginning a study, certain key characteristics (such as ethnicity, job status, health condition, and educational standing) are equal among the different groups. Another typical example is that of an investigator who wishes to demonstrate that different treatment approaches produce similar frequencies for different groups (e.g., males vs. females, or urban vs. rural residents).

Despite the commonality of these circumstances, efforts to prove the equivalence of groups or treatments routinely suffer from two core problems. First, testing for a lack of association is often riddled with complications relating to the selection of an appropriate statistical method. Second, difficulties arise in adequately defining “equivalence”, which are related to the concept of an equivalence bound.

With regard to the issue of selecting an appropriate statistical test, this study evaluated the statistical properties of three different approaches for testing a lack of association among categorical variables in 2 x 2 and 2 x 4 designs. Additionally, we investigated the relationship of these equivalence tests to the traditional chi-square test of independence with the objective of presenting recommendations for behavioural researchers concerning their suitability and practicality. Several key differences distinguish the various equivalence tests examined in this paper. The ED procedure is based on the difference between observed and expected frequencies, and its logic is close to that of the traditional chi-square test of independence. The EP test is rooted in differences between proportions. The CV approach adopts a correlation metric, and thus potential values range from 0 to 1. To compare the statistical properties of these three equivalence tests, it was necessary to determine a way in which their equivalence intervals could be equated (keeping in mind that each is measured according to a different scale). Thus, the equivalence interval for the ED test follows the recommended Euclidean distance, *d* = .15 (Wellek, 2010), which yields a set of two-way frequency tables that allowed us to examine Type I error rates and power (see Table 1). Given these derived population frequency tables, equivalence intervals for the EP and CV tests were then determined in order to match the population association (Type I error conditions) or to be proportional to the Euclidian distance measure used with the ED test. Our results indicate that the proposed test based on Cramer's *V* provided the best balance between Type I error control and power and is available for both 2 x 2 and larger two-way designs.

With regard to the issue of selecting an appropriate equivalence interval, some important issues arose in this this study. As noted in the results tables, a Euclidian distance of .15 relates to differences in proportions and correlations that are theoretically controversial. Although it was necessary for the purposes of this study to equate the intervals for the EP and CV tests with those of the ED test so that fair comparisons could be made among the procedures, many would argue that differences in proportions and correlations greater than .3 are too large to signify a lack of association among variables. For example, consider the equivalence intervals used by Rogers et al. (1993) with the EP test. The authors employed the EP procedure to compare twenty-seven baseline characteristics between two groups of women, one that carried their pregnancies to term and another that aborted. The researchers indicated differences in proportions and equivalence intervals for every characteristic measured. Importantly, the equivalence intervals their investigation utilized (20% of the control groups value) ranged from .001 to .199. Thus, compared to the work by Rogers et al. (1993), the equivalence interval used for the EP test in our study (i.e., .3) appears markedly liberal. For this reason, researchers must decide, based on the metric chosen for investigating a lack of association, the smallest value they would consider meaningful.

Recommendations

Barker et al. (2001) suggest that recommendations regarding which kind of equivalence test ought to be used should be based on the relationship between the empirical and nominal Type I error rates as well as the power of the test under consideration. In the 2 x 2 study design, the ED test’s empirical Type I error rates are substantially higher than the nominal rate. This outcome excludes the ED procedure from further consideration, despite the fact that it shows better power rates than the other procedures in many conditions. Both the CV and EP tests produce viable Type I error rates for the full range of examined conditions, with the CV approach being slightly more conservative than the EP. Thus, we recommend using the CV or the EP tests for the 2 x 2 and the CV approach for 2 x 4 study designs.

To summarize, tests of equivalence allow researchers to assess the research hypothesis that two categorical variables are negligibly related. Establishing a ‘minimally important relationship’ (or equivalence interval) is a difficult and subjective aspect of testing for a lack of association. It is hoped that future discussions will highlight the issues involved in determining an appropriate interval and make it a less daunting task for researchers. However, even permitting a slight amount of subjectivity in establishing an equivalence region is better than inappropriately using the nonsignificance of a traditional chi-square test of independence to explore a lack of association among categorical variables.

Appendix A

Applied Example Using an Equivalence-based Version of Cramer's *V*, the EP method, and the ED method.

A researcher is interested in demonstrating that there is no association between the sex of a child and whether he or she scores high on a measure of attention-deficit disorder (ADD) in grade 8. The data (cell percentages of the total in parentheses) were as follows (note that the data were generated for the purposes of this demonstration):

Sex	Score High on ADD	Do Not Score High on ADD	Total
Boy	25 (7%)	143 (38%)	168
Girl	32 (9%)	172 (46%)	204
Total	57	315	372

Cramer's *V*

$$E_{ij} = \frac{R_i C_i}{N}$$

$$E_{11} = \frac{168(57)}{372} = 25.74$$

$$E_{12} = \frac{168(315)}{372} = 142.25$$

$$E_{21} = \frac{204(57)}{372} = 31.25$$

$$E_{22} = \frac{204(315)}{372} = 172.74$$

$$E = \begin{bmatrix} 25.74 & 142.25 \\ 31.25 & 172.74 \end{bmatrix}$$

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(25-25.74)^2}{25.74} + \frac{(143-142.25)^2}{142.25} + \frac{(32-31.25)^2}{31.25} + \frac{(172-172.74)^2}{172.74} = 0.046$$

$$V = \sqrt{\chi^2/[n(k-1)]}$$

$$V = \sqrt{0.046/[372(2-1)]} = 0.0111262$$

$$V_L = \sqrt{(\Delta_L + m)/[n(k-1)]} = \sqrt{(\Delta_L + 1)/[372(2-1)]}$$

$$V_U = \sqrt{(\Delta_U + m)/[n(k-1)]} = \sqrt{(\Delta_U + 1)/[372(2-1)]}$$

The researcher evaluates the null hypothesis H$_0$: $V \geq \delta$ against the alternate hypothesis that H$_1$: $V < \delta$. Given the lack of theoretical background, the equivalence bound was set at δ = .3, which was found to be the approximate value at which correlations between variables become meaningful (Beribisky, 2018). Recall that Δ_L and Δ_U represent the lower and upper CIs for the noncentrality parameter for the χ^2 distribution. There is no direct method for computing Δ_L and Δ_U, and therefore we utilize a function in R called *CramerV* from the **DescTools** package (Signorelli et al., 2017). This function uses an iterative approach to determine values for Δ_L and Δ_U (see Smithson, 2003). For example, the code *DescTools::CramerV(mat)*, where 'mat' is the matrix of observed frequencies returns the 90% CI (.000, .082). Since the upper bound on the 90% CI for Cramer's *V* (.082) falls below the equivalence bound of δ = .3, we reject H$_0$: $V \geq \delta$ and conclude that there is a negligible relationship between the sex of the child and scoring high on ADD. The value of Cramer's *V* (.011) can be used as an effect size and, since it can be interpreted along the lines of a positive correlation, we could say that the effect is very small.

Equivalence of Two Proportions (EP)

H_{01}: p_1 - $p_2 \leq -\delta$ and H_{02}: p_1 -$p_2 \geq \delta$ are rejected if the 90% CI for $\hat{p}_1 - \hat{p}_2$ falls completely within the equivalence interval (in this case set at (-δ, δ) = (-.10, .10)

$$\hat{p}_1 = \frac{25}{168} = .149$$

$$\hat{p}_2 = \frac{32}{204} = .157$$

$$\hat{p}_1 - \hat{p}_2 = .149 - .157 = -.008$$

The standard error of the difference between two proportions is calculate by:

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\{[\hat{p}_1\ (1 - \hat{p}_1)/n_1\,] + \ [\hat{p}_2\,(1 - \ \hat{p}_2)/n_2\,]\}}$$

$$= \sqrt{\{[.1488\,(1 - .1488)/168] + \ [.1569(1 - \ .1569)/204]\}} = 0.037$$

$$90\%\text{ CI} = \hat{p}_1 - \hat{p}_2 \pm z_{1\text{-}\alpha}\, s_{\hat{p}_1 - \hat{p}_2} = -.008 \pm 1.65(0.037) = (\text{-}0.070,\ 0.054)$$

Since the CI for $\hat{p}_{1-}\hat{p}_2$ falls completely within the equivalence interval (-δ, δ = -.10, .10) both H_{01} and H_{02} are rejected and we conclude that the proportions for boys and girls are equivalent.

Euclidean Distance (ED)

Following Wellek (2010), H_0: $D^{*2} \geq \varepsilon^2$ is rejected if the upper limit of the 1-α CI for D^2 falls below ε^2, where the upper limit for D^2 is calculated as:

$$D^2 + \ z_{1-\alpha} se_{D^2}\ ,$$

where $se_{D^2} = \ v_n/\sqrt{n}$ and:

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\pi_{ij} - g(\pi_{ij})\right)^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\pi_{ij} - \gamma_i(\pi)\eta_j(\pi)\right)^2,$$

Where $\pi_{ij} = \frac{n_{ij}}{N}$ represents the observed probabilities and $g(\pi_{ij})$ represents the expected probabilities.

$$\pi_{11} = \frac{25}{372} = .067, \pi_{12} = \frac{143}{372} = .384, \pi_{21} = \frac{32}{372} = .086, \pi_{22} = \frac{172}{372} = .462$$

$$\pi = \begin{bmatrix} .067 & .384 \\ .086 & .462 \end{bmatrix}$$

$$g(\pi_{ij}) = \pi_{i+}\pi_{+j}$$

$$g(\pi_{11}) = (.067 + .384)(.067 + .086) = .069$$

$$g(\pi_{12}) = (.067 + .384)(.384 + .462) = .382$$

$$g(\pi_{21}) = (.086 + .462)(.067 + .086) = .084$$

$$g(\pi_{22}) = (.086 + .462)(.384 + .462) = .464$$

$$g(\pi) = \begin{bmatrix} .069 & .382 \\ .084 & .464 \end{bmatrix}$$

$$d_{ij} = \pi_{ij} - g(\pi)$$

$$d_{11} = .067 - .069 = -.002, d_{12} = .384 - .382 = .002$$

$$d_{21} = .086 - .084 = .002, d_{22} = .462 - .464 = -0.002$$

$$d = \begin{bmatrix} -.002 & .002 \\ .002 & -.002 \end{bmatrix}$$

$$\pi = \begin{bmatrix} .067 & .384 \\ .086 & .462 \end{bmatrix}$$

$$D^2(\pi, g(\pi)) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\pi_{ij} - \gamma_i(\pi)\eta_j(\pi)\right)^2 = \sum {d_{ij}}^2 =$$

$$D^2 = (-.002^2) + (.002^2) + (.002^2) + (-.002^2) = .00002$$

The variance, v_n^2, can be expressed as:

$$v_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\hat{d}_{ij}^2\pi_{ij} - \sum_{i_1=1}^{r}\sum_{j_1=1}^{c}\sum_{i_2=1}^{r}\sum_{j_{2=1}}^{c}\hat{d}_{i1j1}\hat{d}_{i2j2}\pi_{i1j1}\pi_{i2j2},$$

where:

$$\hat{d}_{ij} = 2\left\{\left(\pi_{ij} - \pi_{i+}\pi_{+j}\right) - \sum_{a=1}^{r}\left[\left(\pi_{aj} - \pi_{a+}\pi_{+j}\right)\pi_{a+}\right] - \sum_{b=1}^{c}\left[\left(\pi_{ib} - \pi_{i+}\pi_{+b}\right)\pi_{+b}\right]\right\}$$

The computation of v_n^2 is extremely cumbersome, even for this simple 2 x 2 matrix. Thus, using the function *gofind_t* from the **EQUIVNONINF** package, υ_n = .003. Finally, to determine if the upper limit of the confidence interval for $D^2 < \varepsilon^2$,

$$CI_{1-\alpha}(D^2) = D^2 + z_{1-\alpha}\frac{v_n}{\sqrt{n}} = .00002 + 1.65\frac{.003}{\sqrt{372}} = .0003$$

Since $CI_{1-\alpha}(D^2)$ (.003) < ε^2 (.15^2 = .0225) the null H$_0$: $D^{*2} \geq \varepsilon^2$ is rejected and a lack of association is concluded.

References

Bailine, S., Fink, M., Knapp, R., Petrides, G., Husain, M. M., Rasmussen, K., ... & Kellner, C. H. (2010). Electroconvulsive therapy is equally effective in unipolar and bipolar depression. *Acta Psychiatrica Scandinavica*, *121*(6), 431-436. doi:10.1111/j.1600-0447.2009.01493.x

Barker, L., Rolka, H., Rolka, D., & Brown, C. (2001). Equivalence testing for binomial random variables: Which test to use? *The American Statistician*, *55*(4), 279-287. doi: 10.1198/000313001753272213

Beribisky, N., Davidson, H., & Cribbie, R. (submitted). What is the smallest meaningful relationship among two variables? *Current Psychology.*

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144-152. doi:10.1111/j.2044-8317.1978.tb00581.x

Buoli, M., Cumerlato Melter, C., Caldiroli, A., & Altamura, A. (2015). Are antidepressants equally effective in the long-term treatment of major depressive disorder? *Human Psychopharmacology: Clinical and Experimental*, *30*(1), 21-27. doi:10.1002/hup.2447

Chen, J. J., Tsong, Y., & Kang, S. H. (2000). Tests for equivalence or noninferiority between two proportions. *Drug Information Journal*, *34*(2), 569-578. doi:10.1177/009286150003400225

Committee for Medicinal Products for Human Use. (2006). Committee for medicinal products for human use (CHMP) guideline on the choice of the non-inferiority margin. *Statistics in Medicine*, *25*(10), 1628-1638. doi:10.1002/sim.2584

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 292-309. doi:10.1111/bmsp.12045

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press. doi:10.1515/9781400883868

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, *60*(1), 1-10. doi:10.1002/jclp.10217

Cribbie, R. A., Arpin-Cribbie, C. A., & Gruman, J. A. (2009). Tests of equivalence for one-way independent groups designs. *The Journal of Experimental Education*, *78*(1), 1-13. doi:10.1080/00220970903224552

O'Reilly, R., Bishop, J., Maddox, K., Hutchinson, L., Fisman, M., & Takhar, J. (2007). Is telepsychiatry equivalent to face-to-face psychiatry? Results from a randomized controlled equivalence trial. *Psychiatric Services*, *58*(6), 836-843. doi:10.1176/appi.ps.58.6.836

Quertemont, E. (2011). How to statistically show the absence of an effect. *Psychologica Belgica*, *51*(2), 109-127. doi:10.5334/pb-51-2-109

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553-565. doi:10.1037//0033-2909.113.3.553

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Bio-pharmaceutics*, *15*, 657–680. doi:10.1007/bf01068419

Signorell, A. et al. (2017). DescTools: Tools for descriptive statistics. R package version 0.99.23.

Smithson, M. (2003). *Confidence intervals: No. 140*. *Quantitative Applications in the Social Sciences Series*. Belmont, CA: Sage. doi:10.4135/9781412983761

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*(2), 192-196. doi:10.1007/s11606-010-1513-8

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Boca Raton, FL: CRC Press. doi:10.1201/ebk1439808184

Table 1.

Conditions for the Monte Carlo Study.

2 x 2 Design									
Condition	a_{11}	a_{12}	a_{21}	a_{22}	*d*	EB$_{ED}$	PD	EB$_{EP}$	EB$_{CV}$
Type I Error Conditions									
1	.175	.325	.325	.175	.150	.150	.300	.300	.298
2	.250	.250	.100	.400	.150	.150	.300	.300	.312
3	.050	.450	.200	.300	.150	.150	.300	.300	.344
Power Conditions									
1	.250	.250	.250	.250	0	.150	0	.300	.400
2	.100	.400	.100	.400	0	.150	0	.300	.400
3	.270	.230	.250	.250	.020	.150	.040	.300	.400
4	.260	.240	.240	.260	.020	.150	.040	.300	.400
5	.050	.450	.150	.350	.100	.150	.200	.300	.400
6	.200	.300	.300	.200	.100	.150	.200	.300	.400

2 x 4 Design											
Condition	a_{11}	a_{12}	a_{13}	a_{14}	a_{21}	a_{22}	a_{23}	a_{24}	*d*	EB$_{ED}$	EB$_{CV}$
Type I Error Conditions											
1	.050	.050	.050	.050	.659	.050	.050	.041	.150	.150	.507
2	.100	.050	.250	.150	.181	.100	.050	.119	.150	.150	.410
3	.200	.050	.050	.050	.166	.300	.150	.034	.150	.150	.399
4	.150	.150	.150	.150	.253	.050	.050	.047	.150	.150	.382
Power Conditions											
1	.125	.125	.125	.125	.125	.125	.125	.125	0	.15	.4
2	.120	.120	.120	.120	.130	.130	.130	.130	0	.15	.4
3	.145	.105	.125	.125	.125	.125	.125	.125	.02	.15	.4
4	.149	.120	.120	.120	.130	.130	.130	.101	.02	.15	.4
5	.200	.070	.190	.220	.080	.090	.125	.025	.1	.15	.4
6	.100	.150	.134	.100	.150	.120	.046	.200	.1	.15	.4

Note. a_{ij} = cell proportions in the 2 x 2 and 2 x 4 tables; *d* = Euclidean distance; EB$_{ED}$ = equivalence bound for the ED test; PD = p_1 - p_2; EB$_{EP}$ = upper bound on the equivalence interval for the EP test; EB$_{CV}$ = equivalence bound for Cramer’s *V* test.

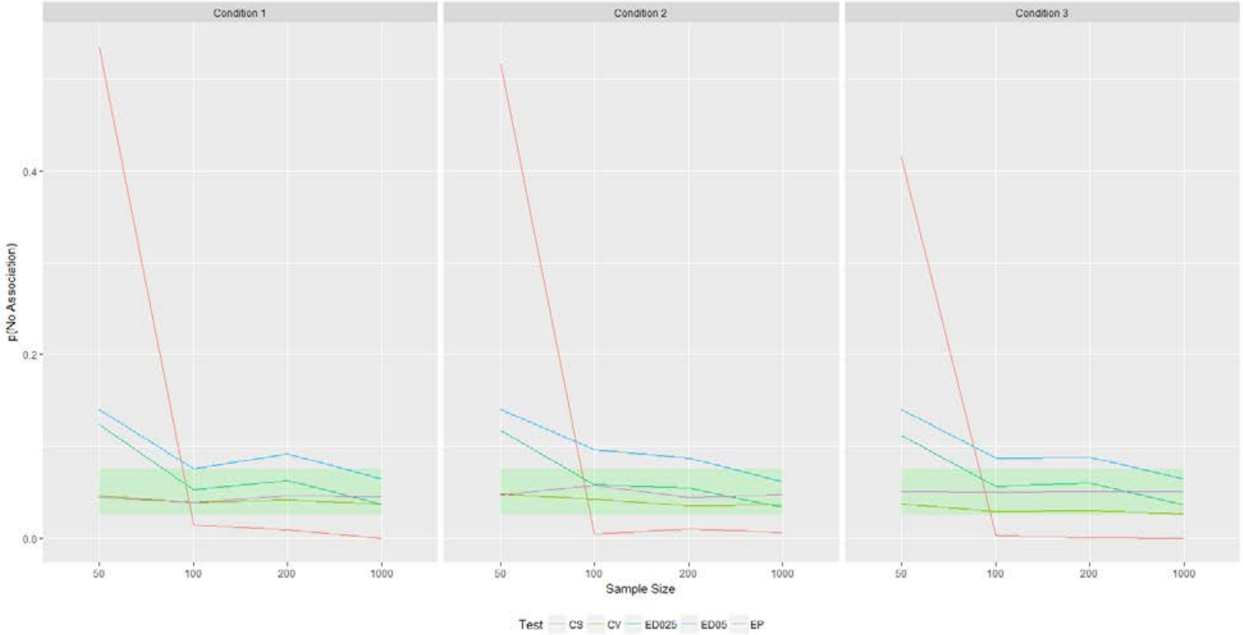

Figure 1. Type I error rates for the equivalence-based tests, Type II error rates for the traditional χ^2 test of independence in the 2 x 2 design. CS = traditional χ^2 test of independence, CV = Cramer's *V*, ED025 = Euclidian distance test (α = .025), ED05 = Euclidian distance test (α = .05), EP = equivalence of proportions test; green highlighted area = Bradley's liberal limits (.025-.075). See Table 1 for condition information.

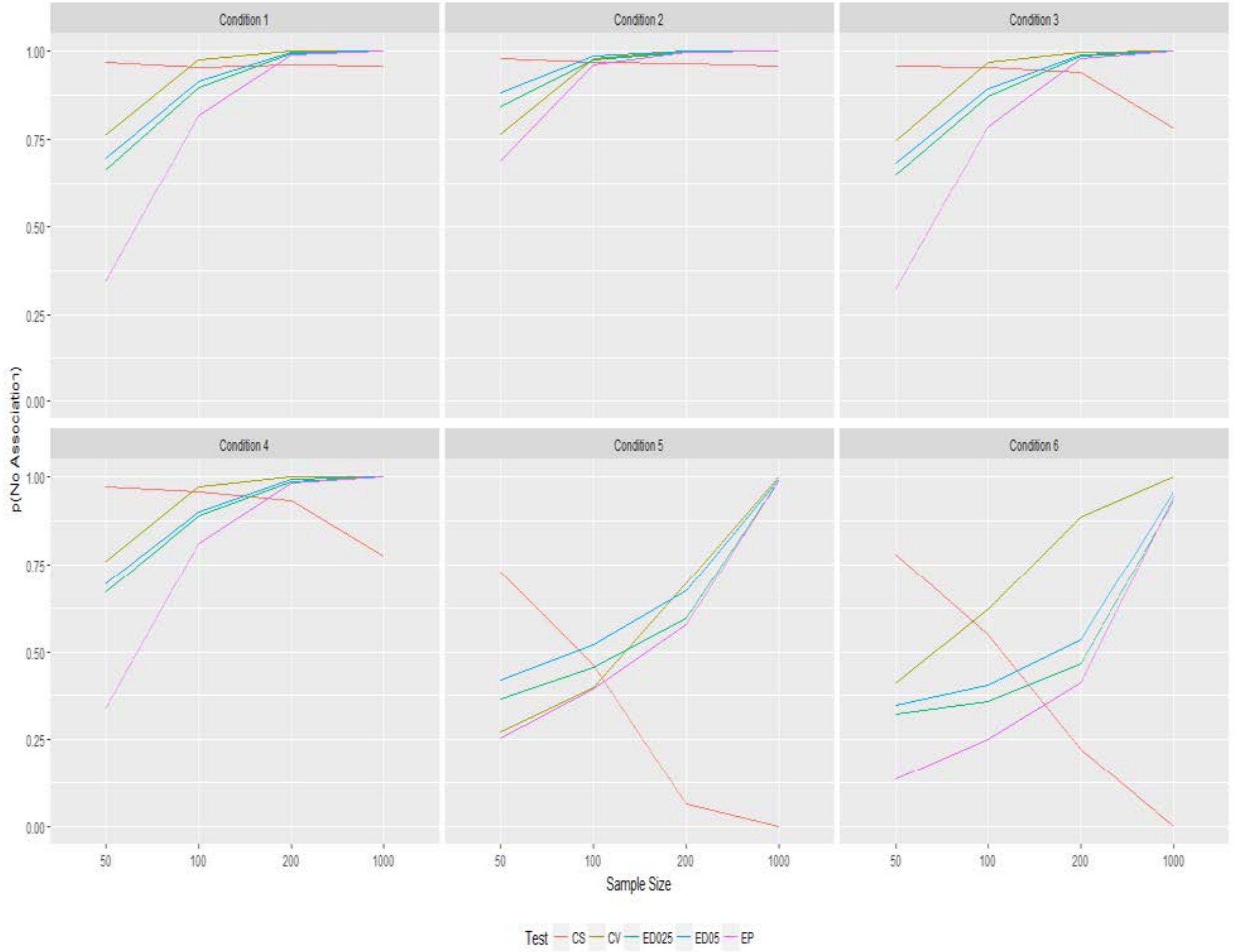

Figure 2. Power rates for the equivalence-based tests, correct decision (Conditions 1, 2)/Type II error rates (Conditions 3-6) for the traditional χ^2 test of independence in the 2 x 2 design. CS = traditional χ^2 test of independence, CV = Cramer's *V*, ED025 = Euclidian distance test (α = .025), ED05 = Euclidian distance test (α = .05), EP = equivalence of proportions test. See Table 1 for condition information.

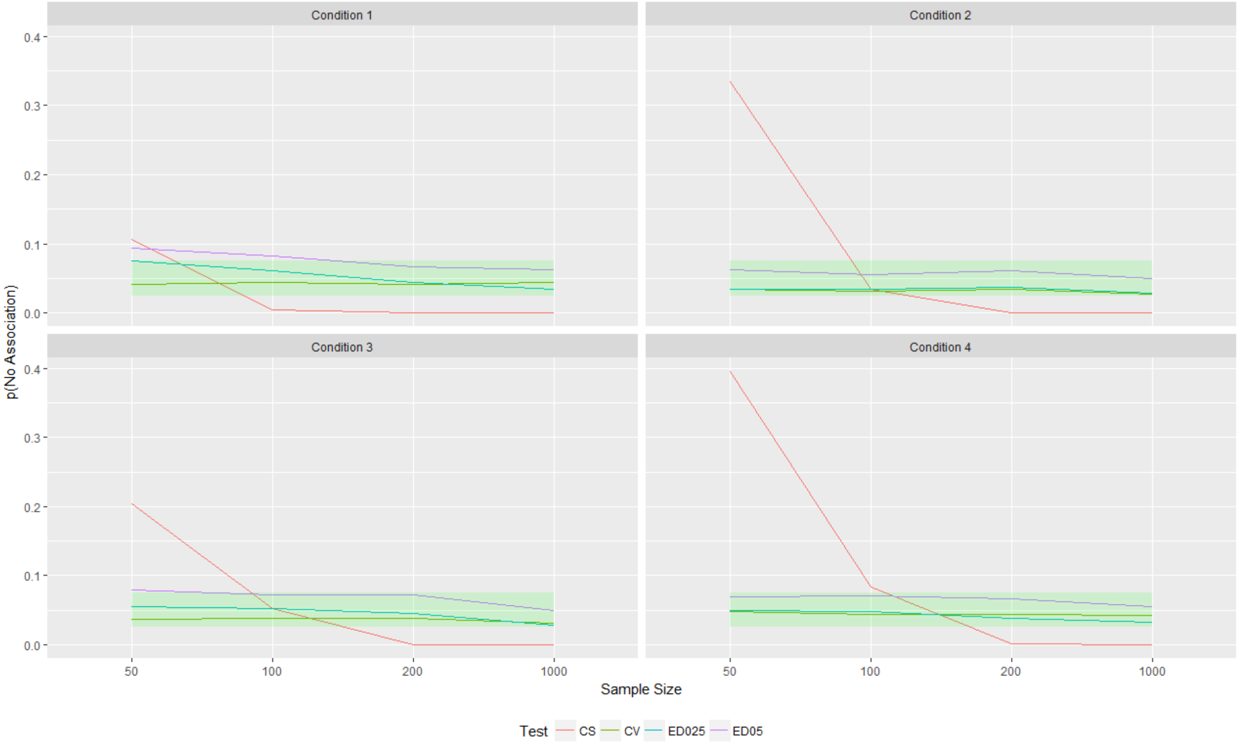

Figure 3. Type I error rates for the equivalence-based tests, Type II error rates for the traditional χ^2 test of independence in the 2 x 4 design. CS = traditional χ^2 test of independence, CV = Cramer's *V*, ED025 = Euclidian distance test (α = .025), ED05 = Euclidian distance test (α = .05), green highlighted area = Bradley's liberal limits (.025-.075). See Table 1 for condition information.

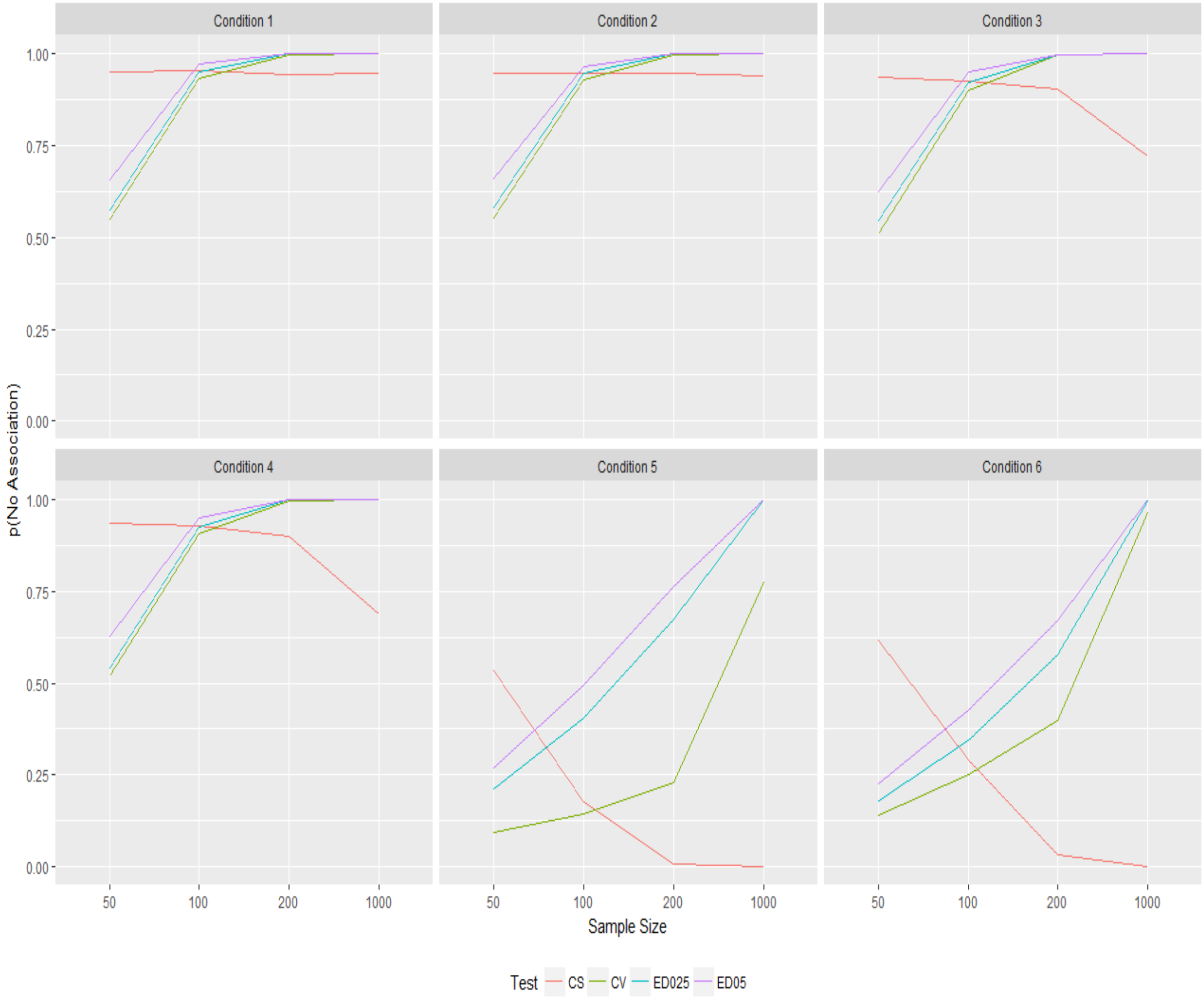

Figure 4. Power rates for the equivalence-based tests, correct decision (Conditions 1, 2)/Type II error rates (Conditions 3-6) for the traditional χ^2 test of independence in the 2 x 4 design. CS = traditional χ^2 test of independence, CV = Cramer's *V*, ED025 = Euclidian distance test (α = .025), ED05 = Euclidian distance test (α = .05). See Table 1 for condition information.