The Treacherous Path to Generalizable Results

Daniel Wilmott

A Thesis submitted to
The Department of Psychology
Rutgers University

Written under the direction of
Professor Arnold Glass
Of the Department of Psychology
Rutgers University
New Brunswick, New Jersey

Abstract

Psychologists have historically used a within-subjects ANOVA or *t*-test to make conclusions about populations of humans and items such as words, images, and social situations. However, these conclusions are overgeneralizations due to the statistics' vulnerability to sampling bias, which restricts the scope of a justified conclusion. Also, the tests necessarily analyze data at the group level, and hence are blind to the presence of a treatment effect in a single observation. They assume the proportion of observations exhibiting an effect in a treatment condition, or sample pervasiveness, is always 100%. Computer simulations were performed to examine the relationship between effect size, pervasiveness, and the results of several significance tests, standard deviations, and effect size measurements for various experimental designs. Results indicate that effect pervasiveness weakens the generalizable strength of these statistics. Finally, an algorithm is derived based on theoretical considerations to estimate sample effect pervasiveness. Preliminary results showed the estimate did not predict pervasiveness. Further research is required to develop an accurate estimator of pervasiveness, with the ultimate goal of applying such an estimate to real psychological research data and exploring real world implications.

Psychologists have long utilized the ANOVA to test a hypothesis about a treatment factor. The large amount of variability in human subjects' responses requires treating subjects as a random effect in the statistical design. In contrast, when words are used in an experiment as a testing medium, the differences among the responses to different words have historically been treated as a fixed effect (Clark 1973; Santa, Miller & Shaw 1979; Wickens & Keppel 1983). However, this methodological decision assumes rather than tests the assumption that there is less variability in the responses to different items in the sample of items than there is variability in the sample of humans. This issues arises whenever the sample of items is drawn from a large population of items, and has direct implications for the statistical test that should be used. Treatment of an items factor as a fixed effect invariably affects the measurement of a treatment effect in the same way as treating a subject factor as a fixed effect does, increasing the probability of a Type I error. This problem was raised by Clark (1973), who built upon arguments proposed by Coleman (1964) in an attempt to provide a solution to this so-called "fixed-effects fallacy". For example, consider the model for recognition created by Atkinson & Juola (1973) that formed a basis for modern memory models. For their series of experiments, they gathered a sample of subjects and words, and theorized that familiarity with words would lead to a decrease in reaction time. Their design collapsed over items so that there was one observation per treatment-subject combination, and used a subjects-as-sole-random-effect ANOVA for analysis. However, their samples of words are just as random as their sample of subjects was. This is the fixed-effects fallacy, for the random items factor is being treated as if all the words differ equally within subjects.

Clark's (1973) proposed resolution to the fixed-effects fallacy was the use of the quasi-F, an approximation to an F distribution that treats both subjects and items as random effects.

Because of what were a prohibitive number of computations required for the quasi-*F* in 1973, the min-F was suggested; the minimum value the quasi-*F* statistic can take on. The min-*F* could be calculated through the use of mean squares obtained from ANOVAs that modeled either subjects or items as a fixed effect, while modelling the other as a random effect (Clark 1973; Winer 1971). Santa et al. (1979) demonstrated that the min-*F* was robust with respect to actual Type I error probabilities across a range of distributions, while Wickens and Keppel (1983) showed that the min-*F* was successful in reducing Type I error rates when there is a relatively large amount of variability associated with samples of either subjects or items. Even with this evidence, the use of the min-*F* for significance was short-lived; for example, Raajimakers, Schrijnemakers, & Gremmen (1999) showed that, between 1975 and 1985, the proportion of papers published that reported the min-*F* in the *Journal of Verbal Learning and Verbal Behavior/ Journal of Memory and Language* dropped from 1.0 to 0.4; that number fell to 0.0 by 1997. This in part stemmed from a conclusion by Wickens & Keppel (1983) that, if items are balanced by explicit blocking, the use of subjects as the sole random effect in an ANOVA that includes a blocking factor gives the most powerful and least biased test. In our previous example from Atkinson & Juola (1973), they used various word sampling techniques between experiments, from random sampling from two pools of test and distractor words, to complex stratified sampling to control for concreteness, frequency, and number of syllables. Under these conditions, a subjects-as-sole-random-effect *F* test has far less positive bias than if no balancing had taken place. However, researchers did not take the next step advocated by Wickens & Keppel (1983) and explicitly include item differences as a blocking factor. Instead, they mistakenly assumed that whenever they use an equal number of items and subjects in an experiment, they are able to justify their conclusion using these significance tests. Raajimakers et al. (1999) found that in the 1980s, psychologists began

performing both a subjects-as-sole-random-effect ANOVA and items-as-sole-random-effect ANOVA, and claimed significant results if both tests produced p-values < 0.05.

More recently, increases in computing power has made possible the use of linear mixed-effect regression models (LMEM) to examine a fixed treatment effect and the inclusion of two blocking factors as crossed random effects in a wide range of science and engineering fields (Faraway 2006) More recently, these techniques have spread into the fields of memory and language processing (Baayen, Davidson, & Bates 2008). This regression technique measures effects for individual levels of random factors. These effects are the estimated coefficients for the levels of the factors in the regression model An effect can be measured through modification of an intercept, a slope or both; the choice of which depends on the underlying theory of the experiment and statistical significance. Software for generating LMEMs exists in commercial statistical programs, such as SAS and SPSS; as well as in programming languages such as R. A good review of the use of LMEM statistical software packages can be found in West, Welch, & Gallechki (2007).

Significance tests measure the probability that you would obtain a result found in a sample of subjects and items from their respective populations. When the aggregate treatment effect in a sample is large enough, an F-test or t-test will indicate a significant result, and a psychologist would claim that an effect exists for the **entire** populations of subjects and items. These generalizations are what Wickens & Keppel (1983) referred to when distinguishing between two types of arguments used as evidence for a conclusion: an appeal by statistical argument, which uses quantitatively justified results to make appropriate generalizations, and an appeal by extrastatistical argument, which relies on assumptions implicitly accepted by the experimenter and their audience. Extrastatistical arguments lead to generalizations that are not

quantitatively justified; which result in the possibility that a conclusion may be false. Consider the goal of Atkinson & Juola (1973); they wanted to make claims about human recognition for all individual humans and all individual words. These hypotheses imply the belief that a treatment effect exists for **every** individual response of every subject to every item in a treatment condition, hence characterizes the responses of all subjects to all items. However, it is possible that a treatment effect's existence in a sample is less than 100%. The observed treatment effect for a sample of subjects or items has two components; the magnitude of the effect, which we will refer to as effect size, and the proportion of subjects or items in a sample showing it, which will be called effect pervasiveness. In this paper, we perform simulations to examine the relationship between effect size, pervasiveness, and the performance of several hallmark statistics including variants of the *F*-test, *t*-test, and contemporary effect size calculations. We go on to demonstrate a procedure for estimating sample effect pervasiveness, and examine its performance at measuring the true sample effect pervasiveness.

Simulation 1: A Two Factor Design, One Fixed and One Random Effect

Consider an experiment with two factors: a fixed treatment factor, and a random subject factor. For example, a sample of subjects is gathered and the effect of group size on group rule adherence is examined. If there are more than two levels (there are more than two group sizes being examined) then a within-subjects *F*-test is an appropriate statistic to detect if a treatment effect is likely to exist in the human population (or if group size has an effect on rule adherence). If there are only two levels of a treatment, then a paired-differences *t*-test can also be used appropriately for the experimental design. In Atkinson & Juola (1973), their design consisted of three factors: a fixed treatment, random subjects, and random items. However, by collapsing over items, their design essentially simulates the two-factor design explained above. In this case, if the test rejected the null hypothesis at some nominal alpha level α, an experimenter would claim a treatment effect is significant for the entire population of subjects (or, that an effect group size has on rule adherence in a sample can be generalized to all individual humans). However, this test does not provide evidence for such a claim, as the possibility remains that an effect was not present for all observations in the sample; hence, the effect might not be present for all members of the population. In this case, experimenters would commit an error of overgeneralization when drawing their conclusions.

Procedures

Simulations in this paper were performed using the *R* programming language. Simulation 1 examines the relationship between effect size, pervasiveness, and the likelihood that a paired-differences *t*-test will result in a *p*-value less than 0.05. There were three independent variables: effect size *c*; pervasiveness *p*, and the variance of the normal distribution sampled from. The range of *c* was from 0 to 1.5, representing effects that range from 0 to 1.5 times the S.D. of a sampled normal distribution; *c* took on values incremented by 0.5. The range of *p* was from 0.0 to 1.0, representing the proportion of observations in a treatment condition that exhibit an effect, and took on values incremented by 0.02. Finally, population variance was held constant at 1.0 to represent a mesokurtic normal distribution. There were 100 trials performed for each combination of *c* and *p*, and the dependent variables were averaged across trials to obtain stable measurements. Each trial consisted of two phases: a data generation phase, and an analysis phase.

In the data generation phase, 50 values (R.E.) are randomly sampled from the standard normal distribution with mean = 0, variance = 1; these values represent means for individual subjects' performance, implying the variability found in a random effect. Next, two observations were randomly sampled for each R.E. from a normal distribution with mean = R.E., variance = 1. The two values represent the performance of an individual subject in a control and a treatment condition where a treatment alters performance positively with respect to a control (Initial analyses included both a positive and negative treatment condition, yet results were the same for both; hence, we only report results for the positive condition). The resultant 100 observations were separated into 2 groups of 50: the control condition (C) and the positive treatment condition (T+). In order to simulate a treatment effect, a proportion of observations, equal to 50 * *p*, were randomly selected from T+ and modified by adding *c*. Selection of observations for modification

was randomized, so that different observations in the treatment were chosen for each trial. Finally, we took the difference of the observations in the control and treatment conditions, hence generating a group of 50 observations. Upon completion of the data generation phase, we recorded the standard deviations of both experimental conditions and of the difference condition, as well as effect size measurements and the *p*-values of a paired-differences *t*-test in response to effect size- pervasiveness combinations.

Results

Figure 1 depicts the relationship between effect size (ES), pervasiveness (P), and the proportion of trials that the *t*-test rejected the null hypothesis with an alpha level of 0.05. For an ES of 0.5 S.D. units, a significant result is only likely when the treatment has a high degree of pervasiveness. However, when ES > 0.5 S.D. units, a significant result is more likely than not when the pervasiveness of the effect exceeds 50%.

Figure 2 shows the relationship between ES, P, and contemporary effect size measurements for a sample treatment effect (ESM). Sample effect size estimates are drastically affected by pervasiveness. There is a direct linear relationship between ESM and pervasiveness. The formula used to calculate ESM is given by Equation 1:

$$(1) \quad ESM = (\bar{X}_T - \bar{X}_C) / SD_C$$

When an effect is not present for 100% of a treatment sample, the presence of observations for which no effect is expressed lowers the estimate of the effect size. This suggests that the formula for a sample effect size can be quantified by Equation 2:

$$(2) \quad SES = P * ES$$

Where ES is the treatment effect exhibited by a single observation, and P is the proportion of observations in a sample that express it. The SES is the difference between the means of the control and treatment condition:

$$SES = (\bar{X}_{treatment} - \bar{X}_{control})$$

$$P * ES = (\bar{X}_{treatment} - \bar{X}_{control})$$

$$ES = (\bar{X}_{treatment} - \bar{X}_{control}) / P$$

Simple algebra shows us that the individual effect size can be estimated from the difference between condition means by dividing by P. In order to convert this measurement into the formula used in Equation 1, simply divide the quantity by the SD of the control condition:

$$\text{ESM} = (\bar{X}_T - \bar{X}_C) / SD_C$$

$$\text{(3)} \quad \text{IESM} = (\bar{X}_T - \bar{X}_C) / (P * SD_C)$$

Where IESM is the individual effect size measurement. It becomes clear that ESM neglects the possibility that an effect is not present for every observation in the sample.

Figure 3 shows the relationship between ES, P, and the standard deviation of observations in a treatment condition (SD_T) and control condition (SD_C). When P is 0% or 100%, SD_T was an accurate estimate of the population standard deviation of 1.0. However, as P approached 50% from either direction, SD_T increased. Note that SD_C is not different from 1.0 across all levels of P, as well as all three ES levels.

Simulation 2: A Three Factor Design, One Fixed and Two Random Effects

Consider an experiment with three factors: a fixed treatment factor, a random subject factor, and a random item factor. For example, consider the model for automaticity of social behavior created by Bargh, Chen, & Burrows (1996). They sought to test whether social behavior was capable of automatic activation by the presence of features of the current environment. The design of their first two experiments consisted of a random sample of subjects given one of three sets of 30 items, designed to prime either a rude, polite, or neutral construct. An item was group of five words in random order, and for each item the subject was asked to create a meaningful sentence from the words. In each item set, 15 of the 30 items had a priming word; that is, a word thought to prime the given social construct, and all item sets consisted of the entire respective sample of words (there was no stratified sampling). In their third experiment, they gave subjects a tedious computer task while subliminally flashing either an African American face or a non-African American face before each trial. On the 130th trial, a system failure was reported and behavioral indications of agitation were recorded to test whether the activation of the stereotype influenced hostile behavior. For all three experiments, they used a subjects-as-random-effect ANOVA to test their hypotheses about different priming materials. However, Bargh et al. committed an error of generalization because they attempted to generalize to "the perceptual and actional representations of...behavior" (1996). Implicitly, they are making claims about **all** individual humans and **all** individual types of perceptual information, yet they have not provided sufficient evidence for such a claim.

The reason their analyses are inadequate is because of their inability to exclude the possibility that their significant results are only true of a subset of the population. As we showed in Simulation 1, the ANOVA can reject the null hypothesis when a treatment effect is not present

for the entire sample. If an effect is not present for an entire sample, by definition it cannot be true of the entire population.

Procedures

Simulation 2 examines the relationship between effect size, pervasiveness, and the likelihood that a subjects-as-sole-random-effect and items-as-sole-random-effect *F*-test will result in a *p*-value less than 0.05. In order to accurately model this design, we chose a balanced, fully-crossed three-factor design, so that the sum of variances for a single observation mirrors Equation 1; that is, there is an interaction term for all combinations of factors included in the model. Therefore, there were four independent variables: effect size *c*; pervasiveness *p*, the size of the interaction effects *s*, and the variance of the normal distribution sampled from. The range of *c* was from 0 to 1.5, representing effects that range from 0 to 1.5 times the S.D. of a sampled normal distribution; *c* took on values incremented by 0.5. The range of *p* was from 0.0 to 1.0, representing the proportion of observations in a treatment condition that exhibit an effect, and took on values incremented by 0.02. There were two levels of *s*: 0.5 and 1.0 S.D. units, representing a small or large amount of variability. Finally, population variance was again held constant at 1.0 to represent a mesokurtic normal distribution. There were 100 trials performed for each combination of *c*, *p*, and *s*, and the dependent variables were averaged across trials to obtain stable measurements. Each trial consisted of two phases: a data generation phase, and an analysis phase.

In the data generation phase, two sets of 20 values, subject random effects (SRE) and item random effects (IRE), are randomly sampled from the standard normal distribution with mean = 0, variance = 1; these values represent means for individual subjects' and individual items' performance respectively, implying the variability found in the random effects. Next, we randomly sampled values to represent the means for the effects of the interaction terms for the subject-treatment (SxT), item-treatment (IxT), and subject-item (SxI) from the standard normal

distribution. The three-way interaction SxIxT was confounded into our error term. For example, for our 20 subject, 20 item simulation, we generated 40 distinct levels for SxT; an effect for all 20 subjects in one of two treatment conditions. Similarly, we sampled 40 values for IxT, and 400 values for SxI (because there are 20^2 combinations of subjects and items).

Next, two observations were randomly sampled for each combination of SRE and IRE from a normal distribution with mean = (SRE + IRE + ST + IT + SI), variance = 1. There were 400 combinations total. The two values represent the performance of a single subject and single item in both a control and a treatment condition, where a treatment alters performance positively with respect to a control. The resultant 800 observations were separated into 2 groups of 400: the control condition (C) and the positive treatment condition (T+). In order to simulate a treatment effect, a proportion of observations, equal to 400 * *p*, were randomly selected from T+ and modified by adding *c*. Selection of observations for modification was randomized, so that different observations in the treatment were chosen for each trial. Finally, we took the difference of the matching observations in the control and treatment conditions, hence generating a group of 400 residual observations. Upon completion of the data generation phase, we recorded the standard deviations of both experimental conditions and of the difference condition, as well as effect size measurements and the *p*-values of the two *F*-test variants mentioned above in response to effect size- pervasiveness combinations.

Results

The likelihood the subjects-as-random-effect ANOVA and items-as-random-effect ANOVA reject the null hypothesis was affected by effect size (ES) and pervasiveness (P), as well as the size of the interaction effects present in the data, as shown in Figure 4. The subjects-as-random-effect ANOVA and items-as-random-effect ANOVA performed nearly identically under each condition. Interestingly, for a small ES, the tests failed to reject more than 20% of the time even when pervasiveness was 100%, across both small and large interaction effects. This lack of power is in stark contrast to the results from Simulation 1, where under the small ES condition, the *t*-test detected a treatment effect 60% of the time when P = 1.0. The presence of variability from interaction terms appears to reduce the power of the significance tests. As in Simulation 1, with larger ES, the tests rejected the null hypothesis when P was less than 1.0; for ES of 1 SD they reject 50% of the time when P is only 75%. Also, we witnessed the same relationship in Figure 5 between ES, P, and the sample ES estimate (ESM) as in Simulation 1. However, the presence of an interaction effect negatively affected the performance of the ESM for P larger than 20%. With large amounts of interaction variability in the model, ESM will underestimate ES even when P is 100%. This implies that ESM is inadequate for most experimental designs in cognitive and social psychology.

Finally, the average standard deviations of the control and treatment conditions were plotted in Figure 6. Again, SD_T increases when P is different from 0 or 100%. However, we also witness a positive correlation between the size of the interaction effects and the standard deviation. Intuitively, this reflects the addition of extra variability via interaction effects.

Discussion

The results from Simulations 1 & 2 show how the proportion of observations in a sample that exhibit a treatment effect should be considered when using an ANOVA or *t*-test. A significant result does not indicate that P is 100%, because if it P was less than 1.0, it would be likely that it would reject the null hypothesis. Also, as the size of ES increases, these tests are more likely to reject the null hypothesis when P is less than 100%. When the purpose of the test is to generalize to every member of a theoretically infinite population, a significant result does not prove that an effect is completely expressed in a treatment condition, which is necessary to generalize to 100% of a population.

Also, ESM measures the product of P and ES, as derived in Equations 2 and 3. This leads to the statistic underestimating the true sample ES whenever P is less than 100%. The use of averages to compute ESM can dilute the measurement of the true effect size, which is either expressed or not expressed by individual observations, if pervasiveness in a treatment condition is not absolute. Interestingly, variability from interaction effects present in an experiment can decrease the estimate of ESM; even when P = 100%, SES will exhibit this negative bias. This is due to an increase in overall model variability and hence, an increase in the value of SD_C.

Sample standard deviation calculations are affected by effect pervasiveness. Consider our experimental design from Simulation 2. If a treatment effect is completely absent or totally present in the sample of treatment observations, then SD_T will equal SD_C. Otherwise, the treatment condition consists of two subsets of observations: those that express an effect, and those that do not. When pervasiveness is 50%, these subsets are of equal size, and hence the average variability will be at a maximum. This implies the standard deviation of a treatment condition should always be equal to or greater than the control condition. Moreover, in the

presence of interaction effect variability, the statistic will not be an accurate measure of the true population standard deviation. The added variability in the model causally increases the sample standard deviation. In experimental designs with a large amount of significant interaction effects, the standard deviation will likely be much larger than the true population standard deviation. These findings about the effect of pervasiveness and interaction effect variability on SD invokes questions that go beyond the scope of this paper, and will be addressed in future research.

Simulation 3: Estimating Sample Effect Pervasiveness

The results of these simulations demonstrate the need for an estimate of sample effect pervasiveness. Consider an experiment with three factors: a Treatment, with *a* fixed levels; Subjects, with *n* random levels; and Items, with *m* random levels. Then Table 1 represents the design of the data structure for a single level of the Treatment. Before the experiment is performed, these factors are random variables. Each cell's value is equal to the sum of the effects of the three factors above, as well as interaction effects, and experimental noise, as shown in Equation 4:

$$(4)\quad Y_{ijk} = \mu + S_i + I_j + T_k + SxI_{ij} + SxT_{ik} + IxT_{jk} + SxIxT_{ijk} + \varepsilon_{ijk}$$

Where *i* indicates the individual subject, *j* the individual item, and *k* the treatment level. Simple algebra leads us to an equation that solves for the treatment effect of an individual observation:

$$(5)\quad T_k = Y_{ijk} - \mu - S_i - I_j - SxI_{ij} - SxT_{ik} - IxT_{jk} - SxIxT_{ijk} - \varepsilon_{ijk}$$

Once an experiment is performed, we will obtain estimates of the random parameters on the right side of Equation 5 except for SxIxT and ε, and subtract them from the observed value in a given cell. These estimates are the coefficients for individual levels of the random effects derived from the use of an LMEM. For example, for the observation corresponding to Subject *i*, Item *j*, and Treatment 1, we would subtract the mean for Treatment 1, as well as the effects for S_i, I_j, SxI_{ij}, SxT_{i1} and IxT_{j1}:

$$(6)\quad T_{ij1} = y_{ij1} - (\bar{X}_1 + \hat{S}_i + \hat{I}_j + \widehat{SxI}_{ij} + \widehat{SxT}_{i1} + \widehat{IxT}_{j1})$$

If the absolute value of the residual is large enough, we would say that a treatment effect exists for that observation. We believe this decision criterion is related to the MS_{error} of the current model. We believe that if the residual value is larger than our estimate of the confounded factors SxIxT and ε, it is likely that a treatment effect has been expressed; or that $T_{ij1} > 0$ in Equation 6.

In the following simulation, we test the use of different multiples of MS_{error} as a critical value in detecting P.

In order to estimate effect pervasiveness for the treatment condition, we encode the results of Equation 3 into a binary statistic, where

$$(7)\quad d_i \begin{cases} 1 & \text{if } |Ti| \gg K * \text{MSerror} \\ 0 & \text{otherwise} \end{cases}$$

If there are g observations in the condition, then Equation 5 will produce an estimate of effect pervasiveness:

$$(8)\quad \gamma = \left(\sum_{i=1}^{g} d_i \right) / g$$

Note this statistic is distributed according to a binomial distribution with $n = g$ and $p = \gamma$, which allows us to compute a confidence interval around γ with a $100(1 - \alpha)$ % confidence level.

Procedures

Simulation 3 consisted of a data generation phase and analysis phase. In order to generate data, we used the same generation techniques for the three-factor design modelled in Simulation 2, with two levels of a fixed treatment effect and two random effects with 20 levels each; all interaction effects were modelled as well. Therefore, the sum of variances for a single observation is quantified by Equation 9:

$$(9) \quad Y_{ijk} = \mu + S_i + I_j + T_k + SxI_{ij} + SxT_{ik} + IxT_{jk} + \varepsilon_{ijk}$$

Where the only difference beween Equations 4 and 9 is the confounding of the SxIxT term with the error. While generating the random interaction effects, we used a standard normal distribution with mean = 0, variance = SD = 1 to simulate a large amount of interaction variability. As before, our independent variables were ES (0.5, 1.0, and 1.5 SD units) and P (0 to 1.0 by increments of 0.02, 51 levels). For the analysis, we performed an LMEM using the `lmer` package recommended by Baayen et al. (2008) with by-slope random effects for S, I, SxI, SxT, and IxT; rather, we sought to measure the difference in the intercept each level of each random effect caused.

Next, we performed the procedure explained above. For each of the 400 observations in the treatment matrix, we subtracted the estimated coefficients for the corresponding levels of S, I, SxI, Sxt, and IxT effects from each observed value. The resultant residual matrix was then encoded into binary values four times, one for each decision criterion (DC) of interest: 1, 1/2, 1/3, or 1/4 the value of the model $MS_{error.}$ The sample effect pervasiveness measure (SEP) was then recorded for each DC by taking the sum of the binary values and dividing by 400. Each trial was replicated 10 times for each level of P and ES, and the average SEP and a 95% confidence interval was recorded.

Results

Figure 7 depicts the performance of the algorithm in detecting the true sample effect pervasiveness for the four DCs at each combination of the levels of P and ES. All four DC, represented by the colored series, failed to accurately detect P, the diagonal line, across all three sizes of ES. Furthermore, there was a large amount of variability in SEP across replications; the average range of all the confidence intervals was 38%. There does not appear to be a relationship between P and SEP.

Discussion

While our algorithm failed to accurately detect P, Simulation 3 has provided several intuitions about attempting to estimate effect pervasiveness in a sample. Would the use of a different decision criterion, perhaps using another measure of variability or different constants, improve detection? How does the test perform in situations where interaction variability is moderate, small, or nonexistent? Clearly, more research must be performed. But our work is a start in attempting to find a statistic that will accurately detect sample effect pervasiveness. Such a statistic could be used to correct ES measurements using the formula in Equation 3. It could also explain some differences in an experimental treatment condition standard deviation from a control condition. Finally, the statistic could indicate the scope of a generalization that can be made from an experiment when there is a significant result from an ANOVA or *t*-test. If it reliably provided evidence that an effect is not present for all observations in a sample, then the generalization of an effect to all individual members of a population becomes illicit.

Summary

A psychologist inherently attempts to generalize to every individual member of several extremely large populations, such as humans, words, images, and social situations. However, the size of such populations and the significant differences in experience with items between subjects requires the modelling of such factors as random. The ANOVA and *t*-tests used to analyze such designs necessarily provide information at a group level, hence they are blind to individual expressions of a treatment effect. In turn, they are insensitive to the proportion of observations in a treatment condition that exhibit that effect. Thus, the tests can reject the null hypothesis when pervasiveness is less than 100%. The *t*-test and *F*-test are only able to justify conclusions about whether an effect is likely or not to exist in the population as a whole, not to all members of a population.

The sample effect size measurement in Equation 1 is also insensitive to sample effect pervasiveness. When an effect is not absolutely pervasive, the statistic underestimates the true effect size. Furthermore, as variability in a model caused by interaction effects increases, the measurement decreases for combinations of P and ES. This implies that under such experimental conditions, ESM will rarely accurately detect ES. The sample standard deviation is influenced by both of these factors as well. If a treatment effect is either completely absent or totally pervasive and there is no variability due to interaction effects, then SD_T and SD_C provide good estimates of the population SD. However, as P approaches 50%, SD_T increases as well. And when there is significant interaction effect variability, both SD_T and SD_C exhibit positive bias as estimators of the population SD.

Since these statistics are functions of effect pervasiveness, obtaining an estimate would allow for mathematical consideration of P during computation, as well as shed light on the

generalizability of an experiment's results. Further research must be performed in order to obtain an accurate and powerful measure of P. Once this has been established, the application of the statistic to actual data from psychological research would explore the significance of our findings in the real world. If nothing else, our paper has forced the exposure of the ANOVA and *t*-test as inadequate tests when used as evidence for research in the fields of cognitive and social psychology, for they do not provide quantitative evidence to generalize to individual humans, words, images, or social situations.

References

Atkinson, R.C., & Juola, J.F. (1973). Factors influencing speed and accuracy of word recognition. *Fourth International Symposium on Attention and Performance* (583-611).

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Bargh, J.A., Chen, M., & Burrows, L. (1996). Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action. *Journal of Personality and Social Psychology, 71*(2), 230-244.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports, 14*, 219 – 226.

Faraway, J.J. (2006). *Extending the linear model with R*. Boca Raton, FL: Chapman & Hall/CRC.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with 'The language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416-426.

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin, 86*, 37 – 46.

West, B.T., Welch, K.B., & Gallechki, A.T. (2007). *Linear mixed models. A practical guide using statistical software*. Boca Raton: Chapman & Hall/CRC.

Wickens T. D. & Keppel, G. (1983). On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior, 22*, 296 – 309.

Winer, B.J. (1971). *Statistical principles in experimental design*. New York, NY: McGraw-Hill.

	Treatment k			
	Item 1	Item 2	...	Item m
Subj. 1	y_{11k}	y_{12k}	...	y_{1mk}
Subj. 2	y_{21k}	y_{22k}	...	y_{2mk}
⋮	⋮	⋮	⋮	⋮
Subj. n	y_{n1k}	y_{n2k}	...	y_{nmk}

Table 1. Example data structure for a single treatment level of a three-factor fully-crossed design.

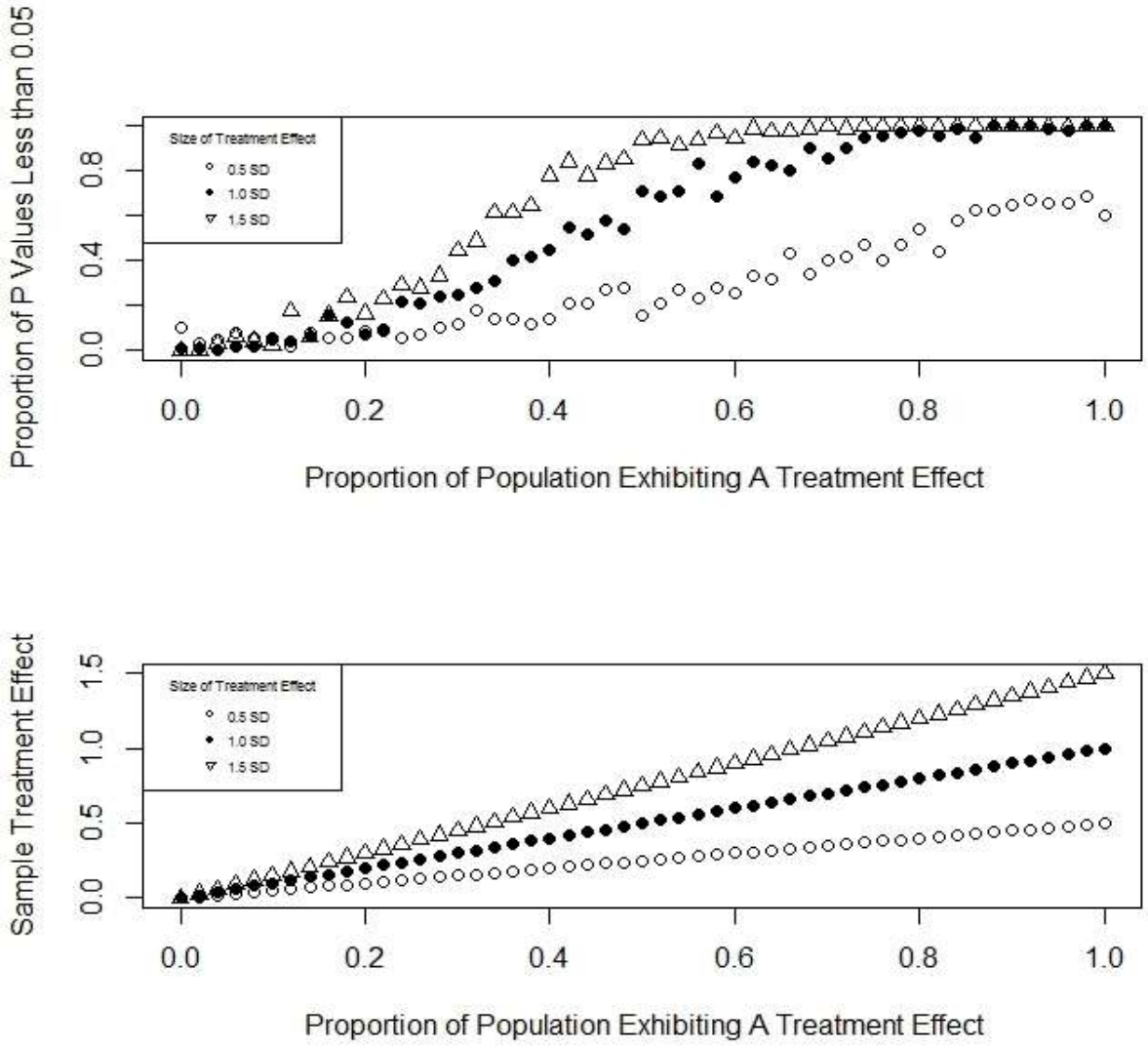

Figure 1. Figure 1A shows the relationship between effect size (ES), pervasiveness (P), and the probability that a paired-differences *t*-test rejects the null hypothesis given a nominal alpha level of 0.05 (A). Clearly, P has a significant effect on the test; for ES of 0.5 S.D. units, A rises above 20% when P is greater than 60%. For ES larger than 0.5 SD units, the rate at which A grows, increases. For ES of 1.0 S.D. units, the *t*-test is almost guaranteed to reject the null hypothesis when pervasiveness is greater than 75%, and is even more likely to reject for larger ES. The actual sample's aggregate treatment effect for each level of P is plotted in Figure 1B.

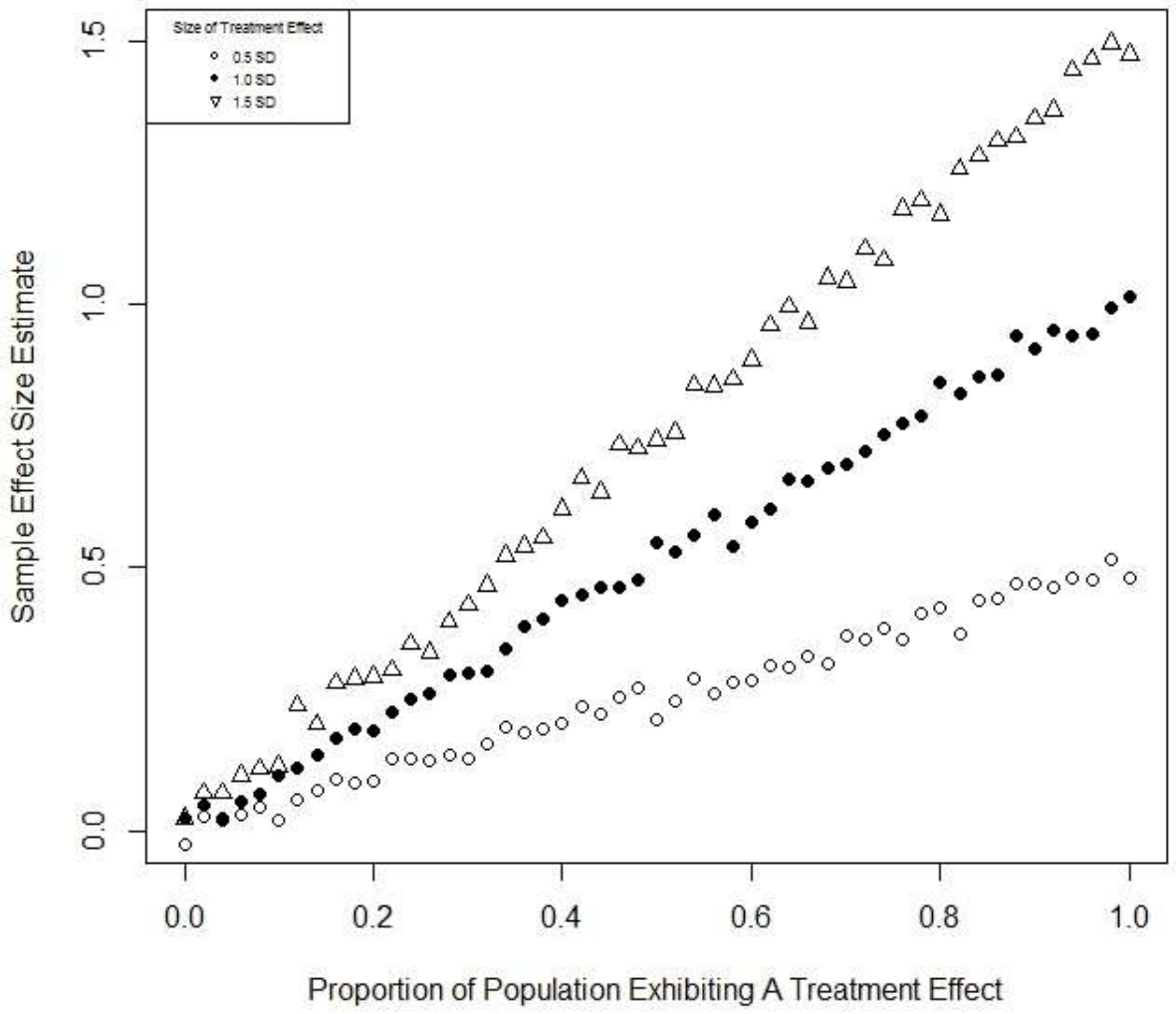

Figure 2. Relationship between true effect size (ES), pervasiveness (P), and sample effect size measurements (ESM). These results indicate that SES has a linear relationship with P; it is directly affected by how many observations exhibit an effect. Whenever P is less than 100%, ESM will underestimate the true effect size. This relationship is quantified by the formula ESM = P * ES.

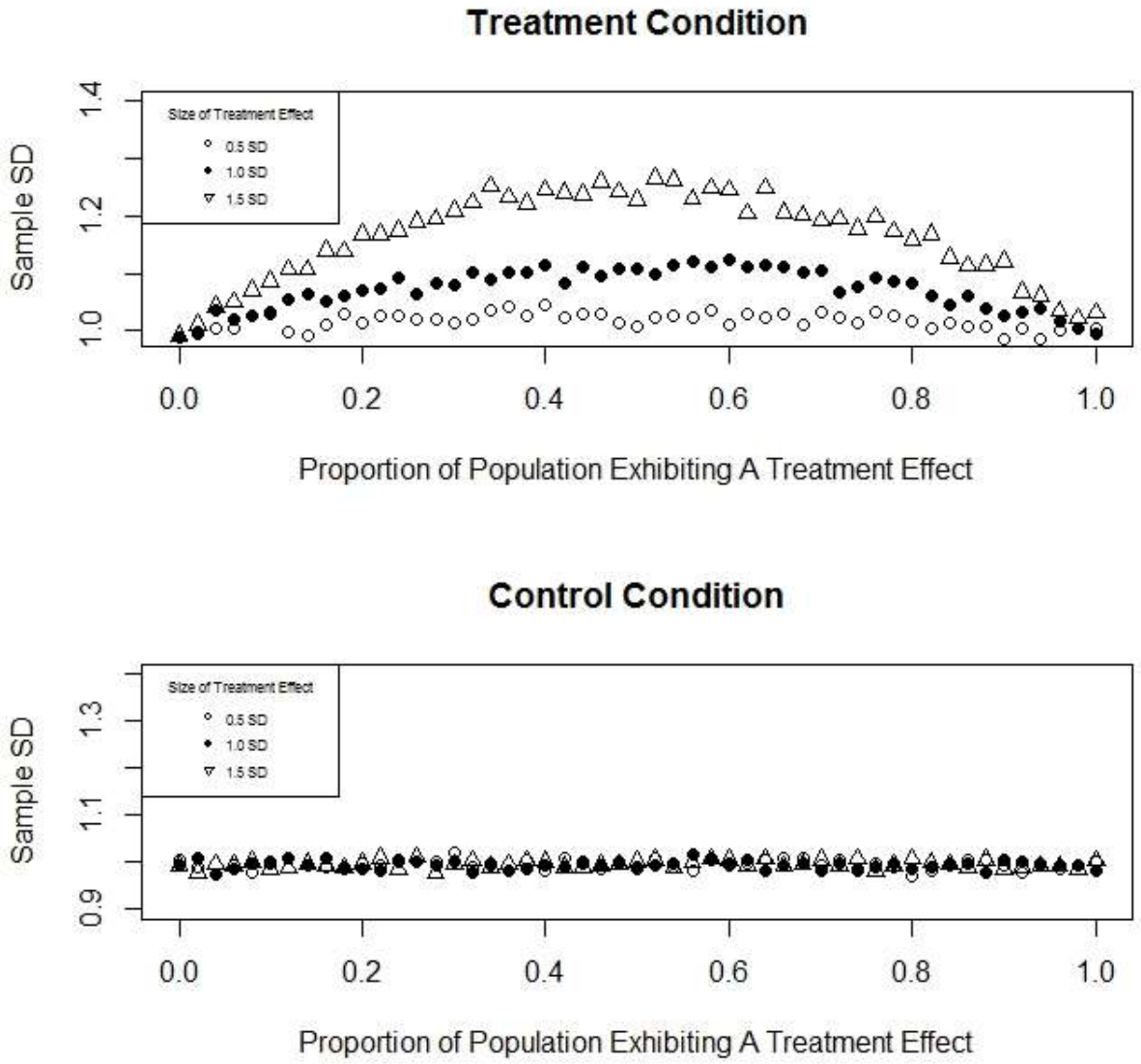

Figure 3. Figure 3A shows the relationship between effect size (ES), pervasiveness (P), and the sample standard deviation (SD) of a treatment condition. For a given ES, the treatment condition SD increases as P approaches 50% from either side. SD_T is only equal to SD_C when P is 0% or 100%, and larger than the control SD otherwise, which is shown in Figure 3B.

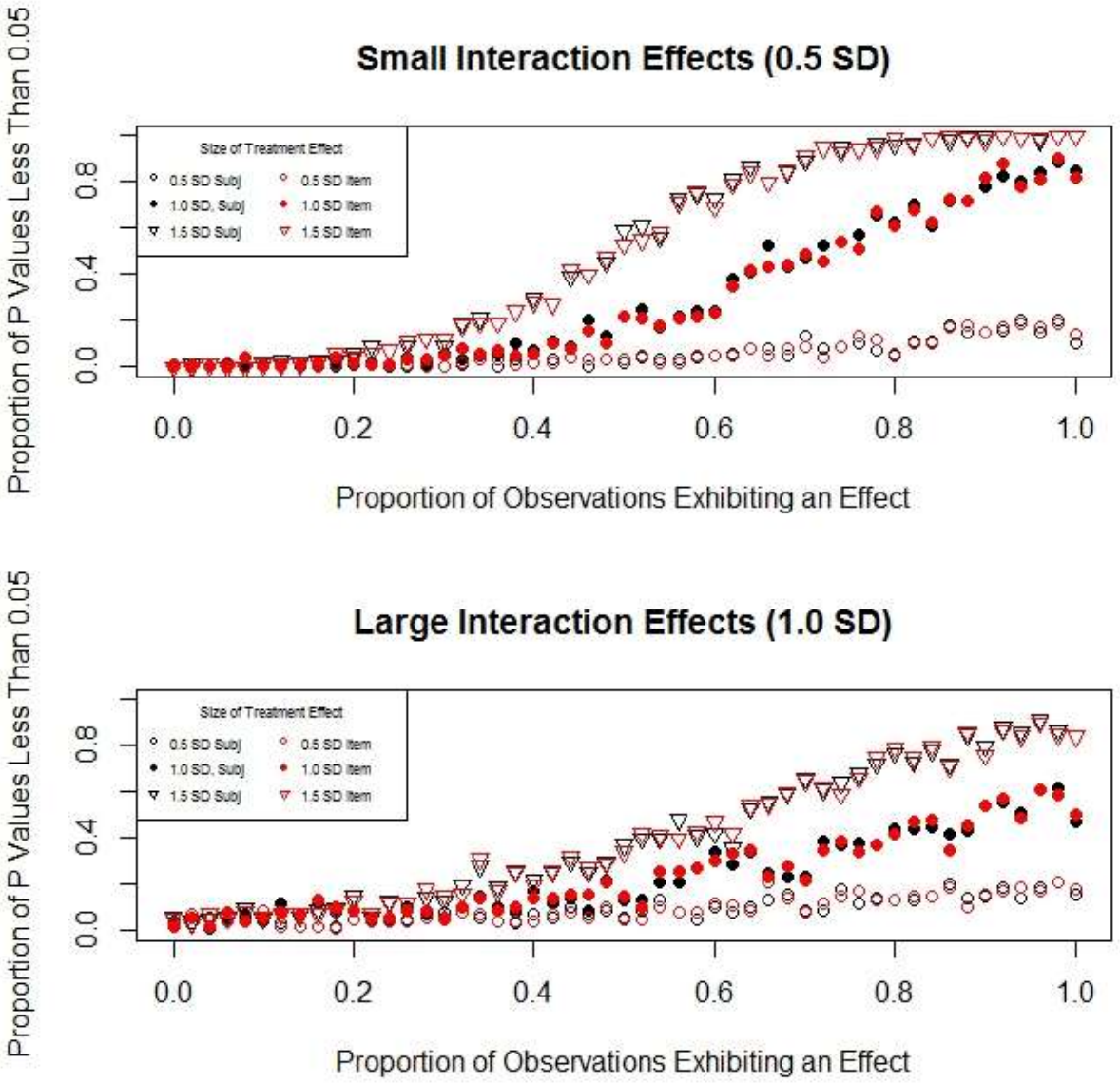

Figure 4. Figure 4A shows the relationship between effect size (ES), pervasiveness (P) and the likelihood that a subjects-as-sole-random-effect *F*-test and items-as-sole-random-effect *F*-test rejects the null hypothesis with $\alpha = 0.05$ for relatively small interaction effects. Figure 4B shows the same, for relatively large interaction effects. These results show P exhibits a significant effect. When ES is 0.5 S.D. units, the tests are not likely reject the null hypothesis irregardless of the value of P, and across both interaction effect conditions. For larger sizes of ES, these tests reject the null hypothesis an unacceptable amount of the time for P as low as 40%; also, more interaction effect variability decreased the probability of a significant result

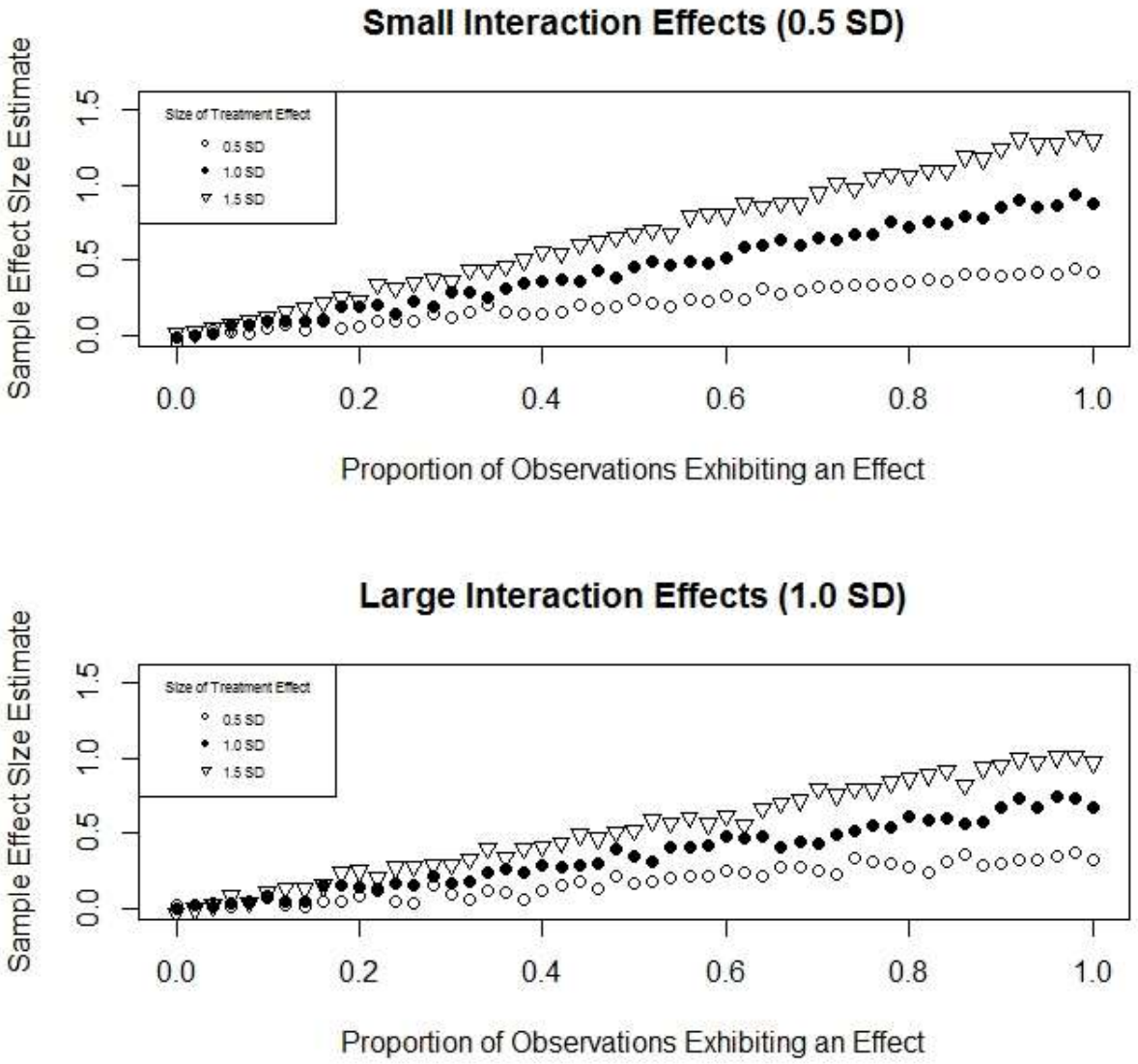

Figure 5. Figure 5A shows the relationship between effect size (ES), pervasiveness (P), and the accuracy of a sample effect size measurement (ESM) using the formula in Equation 6 for relatively small interaction effects. Figure 5B shows the same, but for relatively large interaction effects. Clearly, ESM fails to estimate ES when P is less than 100%, and exhibits the same direct linear relationship with P as in Simulation 1. Also, as interaction variability increases, ESM decreases as well. This is likely due to an increase in SD_C causing negative bias.

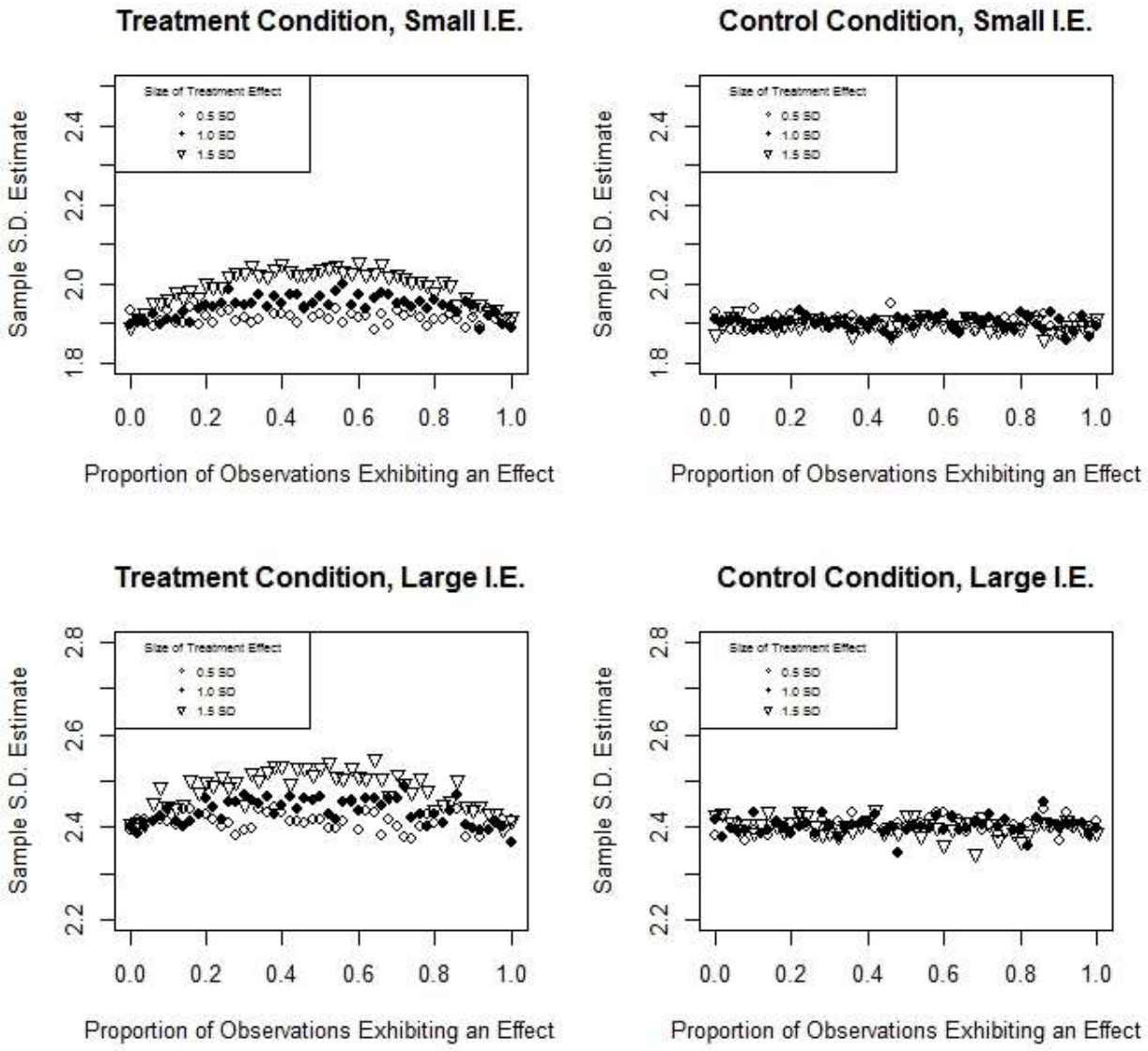

Figure 6. Figure 6 depicts the relationship between effect size (ES), pervasiveness (P) and the sample standard deviations of both a treatment and control condition. The leftmost graphs show the results for treatment SDs, while the rightmost shows the control SDs. Also, the top and bottom graphs represent relatively small and large interaction effects, respectively. SD_T is equal to SD_C when P is 0 or 100%, and is greater as P approaches 50%. Furthermore, as interaction variability increased, the value of both SD_T and SD_C increases significantly from the true population SD.

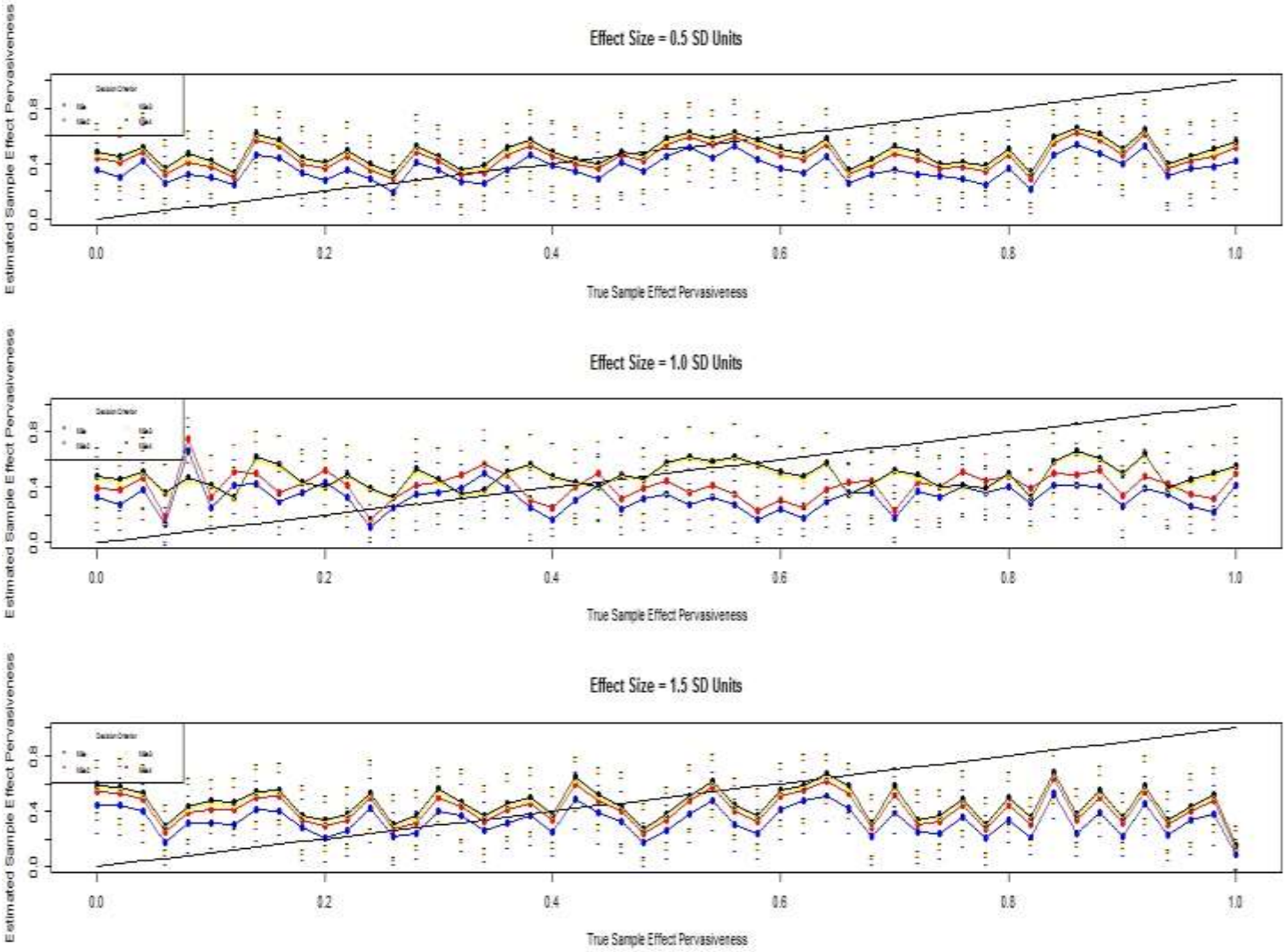

Figure 7. The performance of the algorithm using different decision criterion for detecting sample effect pervasiveness under varying values of ES presented in Simulation 3.The population SD for the randomly generate interaction effects was held constant to 1. Figures 7A, 7B, and 7C show results for ES of 0.5, 1.0, and 1.5 SD units respectively. Individual series for each figure represent the different decision criterion used; blue, red, green, and grey correspond with 1, 1/2, 1/3, and 1/4 the value of the model MS_{error} respectively. The true P is plotted using the black diagonal line. The average estimated sample effect pervasiveness (SEP) for each level of P is plotted, along with 95% confidence intervals for each sample. There appears to be no relationship between EP and P.