Two Methods of Estimating a Study's Replicability

Isadore Newman
University of Akron
Keith McNeil
New Mexico State University
John Fraas
Ashland University

Abstract

This article presents two methods of estimating a study's replicability that researchers should consider reporting along with their statistical significant and effect size findings. One method of estimating the replicability of the findings deals with replication in the exact same system. The second method, which may contain subjective probability values, is used to estimate the replicability of a study's findings in a system that may differ from the initial system with respect to salient variables. The incorporation of the replicability estimates delineated in this paper would provide critical information to decision makers about the likelihood that the implementation of a particular method or treatment would produce similar results in their systems.

Most researchers would agree that the statistical significance levels and effects sizes reported in a study are important pieces of information (Fraas and Newman, 2000; Levin and Robinson, 2000; Robinson and Levin, 1997; Thompson, 1996, 1997, 1999a, 1999b). We take the position that what may be the most relevant piece of information practitioners and policy makers need to glean from a study is the ability of the study's findings to replicate (assuming the intent of the study is inferential not descriptive). If a study's findings are unlikely to replicate, the study's significance levels and effect sizes are virtually meaningless to interested practitioners and policy makers. Thus we as applied statisticians have the responsibility not only to provide estimates of a study's replicability but also delineate the assumptions on which these estimates are based.

In this article we define two types of replications and present methods by which researchers can provide estimates of each type. The definition of replicability that we are developing is the extent to which a curriculum, treatment, etc., can be successfully implemented in two types of systems. One system is assumed to be an exact replication of the one used in the initial study. The replicability estimate that deals with this exact same system criterion is based on the same underlying assumptions of calculating the p value on a random sample of a known population. The other replicability estimate assumes the system of interest is not an exact replication of the system used in the original study. This replicability estimate is based not only on random sampling assumptions but also on probability estimates, which will be somewhat subjective, that certain key variables will differ between the system used in the initial study and the system of interest. It is important to note that this second type of replication estimate can be calculated before as well as after the initial study is implemented. If it is calculated before the implementation of the study, it could assist in the re-design of the study before it is actually implemented. If calculated after the study is implemented, it will have implications for practitioners and decision makers.

Statistically Significant Exact Replications of a Study

One value we believe should be contained in research reports is the likelihood that the study's findings are replicable in the same system. Such a value should not take the place of statistical significance tests but rather should be reported along with them. We agree with Robinson and Levin (1997) who expressed the position that the probability value (p value) produced by a statistical test is an important piece of information to report in a quantitative study. Robinson and Levin stated that "authors should *first* indicate whether the observed effect is a statistically improbable one (e.g., is the difference greater than what would be expected by chance?)" (p. 22).

It is important, however, not to misinterpret a p value with respect to the likelihood of the replication of results (Nickerson, 2000). This point was addressed by Posavac (2002) who stated that "some believe [incorrectly] that rejecting a null hypothesis means that at least 95% of replications would be statistically significant" (p. 102). Posavac does take the position, however, that rejecting a null hypothesis should increase the researcher's expectation that replications of the research would yield similar results.

Using p Values to Estimate Statistically Significant Exact Probabilities

Greenwald, Gonzalez, Harris, and Guthrie (1996) presented an analytic method by which a p value can be converted into a probability estimate that an exact replication of the research would produce a statistically significant result. Posavac (2002), who elaborated on the method proposed by Greenwald, et al., noted that an "*exact replication*

means that the initial experiment is repeated using the same independent and dependent variables with the same number of participants selected in the same way from the same population" (p. 102). In this type of replication the difference between the replication and the original study is due to random variation. We believe that this is one type of replication that should be addressed by researchers.

Green et al. (1996) and Posavac (2002) suggest that the probability of a statistically significant exact replication (SSER) can be estimated from the probability of the statistical test. As a means of demonstrating how a researcher could convert a *p* value from a statistical test to an estimate of the SSER probability, we present a brief discuss of the procedure. It is beyond the scope of this paper to present the rationale on which this procedure is based. We encourage interested readers to review the works published by Greenwald et al. and Posavac for a more in-depth discussion of this concept.

An Illustration

To illustrate the calculation of the SSER probability value, assume researchers are testing the difference between sample means of two independent groups consisting of 20 subjects each. Further assume that the *t* value produced by the difference between the two means recorded for their study was 2.150. Since this observed *t* value (t_{obs}) is greater than the two-tailed critical *t* value (t_{crit}) of 2.024 for an alpha level of .05, the researchers would declare the difference between the two group means to be statistically significant. The question we believe is important for these researchers to address is: What is the chance that the difference between the two group means recorded for an exact replication of the study would be declared statistically significant?

Calculation of the SSER probability value. As noted by Posavac (2002), the probability of obtaining a SSER can be obtained by executing three steps. First, the replication *t* value (t_{rep}) is calculated by subtracting the critical *t* value used in the initial study from the study's observed *t* value. Thus the t_{rep} value is calculated as follows for our hypothetical example:

$$t_{rep} = t_{obs} - t_{crit}$$
$$t_{rep} = 2.150 - 2.025$$
$$t_{rep} = .125$$

Second, the researchers would obtain the one-tailed probability for this t_{rep} value of .125 with 38 degrees of freedom. With respect to the procedure used in this step, Posavac (2002) stated that "a one-tailed test is used because one would want a replication to produce means of the same relative magnitudes as found in the first study (p. 108)." The one-tailed probability for the t_{rep} value of .125 with 38 degrees of freedom is .45.

Third, the researchers subtract the .45 probability value from 1.00, which produces a value of .55. This value indicates that the chance that an exact replication will be statistically significant is .55.

Points to note regarding the SSER probability value. Three points should be noted regarding this SSER probability value of .55. First, the SSER probability value is a function of the *p* value. However, practitioners need to be careful not to directly interpret the *p* value as a replicability value. Second, Greenwald et al. (1996) and Posavac (2002) recommended that SSER probability values should be considered upper limits: The reason for this recommendation is based on the fact that "even in a careful replication the participants would be a different sample from the population, the calendar date would be different, the weather would be different and so forth" (Posavac, p. 111). Third, researchers may be surprised that for a study, such as the one used in our example, which had 38 degrees of freedom and an observed *t* value of 2.150 ($p = .038$), the chance that an exact replication will be statistically significant (SSER probability level = .55) is only slightly above the 50-50 level. In fact an observed *t* value for this hypothetical study would need to be 2.874, which produces a *p* value of .01, in order for the a SSER probability level to reach the .80 level.

To further emphasize this third point, a review of values produced by Posavac (2002) reveals that when degrees of freedom value is at least eight and the *p* value is .05, the SSER probability value will be .50. That is, there is a 50-50 chance of replicating significant findings. If the degrees of freedom value is at least eight and the *p* value is .01 for a two-tailed test, the SSER probability value will not be less than .73 or greater than .84. And if the degrees of freedom value is at least eight and the *p* value is .005, the SSER probability value will not be less than .80 and not greater than .92. (It is interesting to note that these replicability values are less for corresponding p values for one-tailed tests.) Thus researchers need to be careful not to assume that statistically significant findings automatically mean that the chance of obtaining statistically significant exact replications for the study will be high. For this reason we believe that researchers should report the SSER probability value along with the probability of the observed t test.

Replication in a Different System

We believe that a second type of replication of findings is important for researchers to address. That is, the type of replication that deals with the question: Would the study's findings replicate in a system different from the one used in the initial study? It should be noted that we consider this type of replication of findings important even if an individual is interested in the same system in which the study was conducted, assuming the system is a dynamic one. That is, the system experiences considerable change with respect to the variables that may influence the replicability of the findings. Since most people attempt to relate research findings to systems that are different from their own or, at least, relate findings to systems that are similar but dynamic, we believe obtaining a likelihood estimate for this type of replication would be most valuable for them. The remaining

portion of this section of the article presents our *preliminary* attempt to develop a procedure for calculating such a likelihood estimate.

Estimate Procedure

The procedure we are proposing for the estimate of the likelihood of replication of findings for a system different from the one in which the study was conducted can best be presented through an example modeled on a study conducted by Benson, Aronson, Desmett, Shaheen, and Showalter (2002), which presented an evaluation of a multiage classroom educational program. In our example we have children in grades 1-3 who were grouped in the same classroom and their teacher stayed with them for the three years. The evaluation indicated that the teachers volunteered for the project and were enthusiastic about the concept of multiage education. The project was supported fully by the principals and was enthusiastically supported by the parents. Achievement scores indicated moderate success of the project as compared to national norms and comparison students in the same school.

Internal validity issues are apparent, since parents voluntarily allowed their children into the project (see Campbell and Stanley, 1963, for a discussion of internal validity issues). Enthusiastic teachers might generate better results, no matter what the curriculum. In addition, a supportive principal might be partly (or entirely) responsible for the achievement results. Other internal validity concerns could also be raised.

External validity issues are also of concern with this study. Would the same effects be observed with less enthusiastic teachers? Will the same effects occur after the novelty of the multiage grouping wears off? Other external validity concerns could be raised (see Campbell and Stanley, 1963, for a discussion of external validity issues). The concept of replicability, though, is different from internal validity and external validity. It is based on the realization that any implementation is accomplished in a system and the realization that that system is likely to be dynamic. We believe that the likelihood of replicating a study's findings in a different system or even the same dynamic system is crucial to estimate.

Important variables. The first step in the estimation process is to identify key variables that influenced the findings but may be different in the new system. Let us assume that for our multiage project example four such variables were identified:

1. Twelve volunteer teachers were used.
2. The study involved supportive principals.
3. The study used 240 volunteer (supportive) parents.
4. A total of 5 days of in-service training was given to the teachers on the multiage project.

As an illustration of how these variables could influence the replicability of the findings of the original study, consider the principals variable. If a principal leaves, the project will, in all likelihood, be supported less by the new principal. The new principal may even kill the project, not because the project is ineffective, not because the concept of multiage education is bad, but because the crucial component of the system (the principal) does not believe in or want the project.

The likelihood of each crucial component changing should be taken into account when the project is envisioned. If a particular component is likely to change, then the project should be devised so it is immune to that change—in the case of principal change—the project should be made "principal proof."

Once the variables are identified, the second step is to estimate the proportion of the R^2 value accounted for by each variable, the probability of that variable changing, and the probability of the changed variable being negatively influential on the original findings. Table 1 contains such hypothetical values of these estimates for our example.

Table 1

Proportion of the R^2 Accounted for in the Dependent Variable and the Probability Values for Each Variable

Variable	Proportion of R^2	Estimate of the Probability of Change	Estimate of the Probability of Negative Impact
Teacher	.50	.30	.09
Principals	.20	.70	.60
Parents	.10	.33	.02
Staff Development	.20	.02	.02

The proportion of the R^2 value accounted for by each component is determined. This could be accomplished with GLM if enough implementation sites were available (similar to meta analysis), or conceptualized either before the study started or afterwards. In the example, here we provide "educational guesses." For instance, it is likely that some teachers will leave the project. Some may become disillusioned with the project or with education in general. Others may find a more lucrative job in another district or another profession. Nevertheless, other enthusiastic teachers are likely available, so the systemic effect on the project of teacher change would be minimal.

On-the-other hand, the likelihood of a principal leaving the system is high (estimated to be .70 in a three-year period) and the likelihood of the replacement being equally enthusiastic is low (.40). Indeed, most replacement principals may gut the project, leading to absolutely no replicability from the component of the principal. Therefore, because of the high probability of principal change, and high probability of a different (lower) level of support, the overall replicability is lowered.

Parent turnover will be at least 33% every year, with third graders moving to fourth grade. But we suspect that the parents of the incoming first graders will be just as enthusiastic (maybe even more so if the project is a success).

Thus, the high turnover rate (large system change) of parents will have little effect on replicability—the project is "parent proof." If the staff development is "packaged" then it could easily stay the same from year to year. This part of the system would likely be stable.

Actual probabilities may be quite difficult to determine. To deal with this problem, one might rate the stability of each component on a 1 to 5 scale, with 5 being the most stable. Such estimates and the calculation of the reliability value for the multiage example are listed in Table 2. It should be noted that a replicability value calculated in this manner would produce higher values the more stable key variables are from the system used in the initial study and the system of interest especially for the variables that account for the higher proportion of the R^2 value.

Table 2
Calculation of the Replicability Value

Variable	Proportion of R^2	Stability	(Proportion of R^2) * (Stability)
Teacher	.50	4	.50*4 = 2.00
Principals	.20	1	.20*1 = 0.20
Parents	.10	5	.10*5 = 0.50
Staff development	.20	5	.20*5 = 1.00
			Replicability = 3.70 / 5 = .74

Estimating replicability before implementing. If a researcher calculated replicability before first implementing a new project, and obtained a low replicability value, the researcher might try to re-conceptualize the project by either doing something to minimize system change or to minimize the effects of the change within any one component. One could minimize system change in the multiage project by getting the school board to mandate multiage in all elementary schools or find another district where all the principals are supportive of multiage programs. It should be noted that a replicability value calculated in this manner would produce higher values the more stable key variables are from the system used in the initial study and the system of interest, especially for the variables that account for the higher proportion of the R^2 value.

Minimizing the effects of change in teachers could be accomplished by each project teacher identifying a non project teacher who would like to be in the project and then keeping that teacher informed about multiage grouping during the year. This "information partnership" actually becomes a new component of the project (or at least modifies the teacher component.) Curricula that purport to be "teacher proof," such as highly prescriptive direct instructional methods, are another example of minimizing the effects of teacher change.

If the replicability index is low, and the researcher cannot identify changes or strategies that would make it higher, then the project should not be implemented. The time of teachers, principals, parents, staff developers, and especially students should not be wasted. If there is very little hope for replication of a particular project, then we have no business investigating the effectiveness of that project.

How about changes in system components not relevant to the project? Changes in components that are not relevant to the project will not affect the replicability, by definition. Nor will these changes affect the index, as the percent of variance accounted for is 0 and the contribution of that component would be 0. Unfortunately, in most educational systems, many components can influence the success of a project.

Implications

The implications of this article relate our position that statistical significance and effect size are important concepts, but they must be examined in light of replicability. Replicability is, in and of itself, not a one-dimensional concept but a multi-dimensional one. In this paper we identified two types of replication estimates. The first type is the SSER probability estimate, which is based on traditional statistical assumptions and probability concepts.

The second type is related to design and subjective probability issues. This approach provides a number of advantages. First, it can assist in the teaching of research design. That is, teaching this replication estimate emphasizes the need for researchers to attempt to identify the relevant variables in a study. Second, it can improve communication among researchers regarding relevant variables in a study in order to improve the design of such studies. Third, it encourages the use of meta-analysis to identify relevant variables. Fourth, it provides a method of simulating the effects of the relevant variables on replicability of findings. One can simulate small changes or large changes on relevant variables and the impact of these changes on replicability. As one can see, this second estimate is not a static approach but a dynamic one and may only be limited by the investigators' creativity and insight.

An emphasis on replications has implications for researchers regarding the research methodology they use. That is, researchers should consider conducting partial replications. Partial replication can be conducted by one of two approaches. First, half of the study could be an exact replication, and the other half could be an extension (into another grade level, using different in-service materials, or checking on efficacy in another bureaucratic situation). Second, the researcher could put a slight twist on the implementation, by reducing or eliminating a component, shortening the period, streamlining in-service, or monitoring more closely the actual implementation.

We believe that an emphasis on replicability estimates are as important to analyzing the data contained in a study as are the statistical test results and effect size estimates. The value of a study's results can be better assessed by researchers and practitioners when all three types of information (i.e., replicability estimates, statistical test results, and effect size estimates) are reported.

References

Benson, S. N. K., Aronson, E., Desmett, P., Shaheen, M., and Showalter, J. (2002, October). *Multiage education: A process and product evaluation.* Paper presented at the annual meeting of the Midwestern Educational Research Association, Columbus, OH.

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally College Publishing Co.

Fraas, J. W., and Newman, I. (2000, October). *Testing for statistical and practical significance: A suggested technique using a randomization test.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

Greenwald, A. G., Gonzalez, R., Harris, R.J., and Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175-183.

Levin, J. R., and Robinson, D. H. (2000). Rejoinder: Statistical hypothesis testing, effect-size estimate, and the conclusion coherence of primary research studies. *Educational Researcher, 29*(1), 34-36.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

Posavac, E. J. (2002). Using *p* values to estimate the probability of a statistically significant Replications. *Understanding Statistics, 1*(2), 101-112.

Robinson, D. H., and Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29-32.

Thompson, B. (1999a). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory and Psychology, 9*(2), 191-196.

Thompson, B. (1999b). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Educational Psychology Review, 11*(2), 157-169.