ISSA Proceedings 2002 – Assessing The Problem Validity Of Argumentation Templates: Statistical Rules Of Thumb

Burden of proof, a central concept in argumentation theory, situates the requirements for good argument within bodies of substantive knowledge and practical action (Gaskins, 1982). To respond to the burden of proof associated with any claim means providing grounds for acceptance that are adapted to a constellation of related beliefs and prior experience. Burden of proof should not be assumed to be a set of logical requirements, but instead should be understood as an outline of what is known so far that might constitute grounds for challenging claims of some particular sort within some particular substantive domain. The burden of proof that structures scientific argument in any field should be expected to change over time, as disagreement over particular claims reveals general grounds for disagreement with whole classes of claims. For this reason, scientific arguments contain myriad allusions to argumentative failures of the past, answering objections no one may actually have, simply because someone *could* have that objection or has had that objection to some other claim in the past.

Within expert fields of all kinds, and especially scientific fields, the burden of proof to be discharged may evolve over time as new issues emerge from research and theorizing. Among the discoveries of scientific fields are discoveries of things that can go wrong in drawing conclusions about the subject matter. Such discoveries are likely to stimulate the invention of new methods for guarding against the things that can go wrong, including routinized safeguards applied in research procedures (like “double-blind” administration of experiments or use of drug placebos). These routinized safeguards and boilerplate arguments associated with them often come to be understood by scientists themselves as their methods (McCloskey, 1985).

Disciplinary research practices may be seen as a kind of technology of reasoning and argumentation, embodied in new devices (such as statistics) that have been

designed to serve an argumentative purpose and that may become interactionally stabilized in scientific discourse. As distinct from natural, commonsense reasoning, disciplined argumentation has a "designed" quality that comes from the tuning of argumentation to the requirements of the subject matter. As pointed out by Walton (1997), the more specialized these become, the more impenetrable they become for anyone other than a specialist. In this paper we illustrate how relatively impenetrable expert practices such as statistical testing can be opened to theoretical analysis, blending concepts and methods from pragma-dialectics with systematic computer simulation of certain designs for arguing.

1. Pragma-dialectics

Pragma-dialectics is a theoretical, critical, and empirical research program built on a view of argumentative discourse as an exchange of speech acts directed to the resolution of doubt and disagreement. Dialogue, the interaction between a protagonist of a viewpoint and an antagonist who questions or disputes this viewpoint, is a central theoretical construct, applied not only to discussion and debate, but also to individual texts occurring within broad controversies. Argumentation is assumed to be a set of methods for isolation and repair of disagreements emerging from virtually any form of practical action, shaped by norms of reasonableness embodied in an ideal model for critical discussion (van Eemeren, Grootendorst, Jackson & Jacobs, 1993).

The underlying critical ideal applies to argumentation occurring in all fields of endeavor, from ordinary conversation (where it serves to regulate misalignment among interactants in belief and action) to technical and scientific discourses (where it serves to regulate change in disciplinary understandings of phenomena). Pragma-dialectical theory asserts a fundamental set of field-independent rules for the conduct of argumentation, and it also acknowledges the existence of specialized rules within individual fields such as law and policy. In particular, any field may have its own associated procedures for evaluating new assertions as they are introduced into a discussion. These are known as intersubjective testing procedures (van Eemeren & Grootendorst, 1984, p. 167).

Intersubjective testing procedures are methods agreed to by discussants in advance of any particular local disagreement, and in canonical pragma-dialectical theory the ITP is part of the bundle of mutually accepted starting points identified in the opening stage of an argument. Both protagonist and antagonist must agree on the sufficiency of the ITP, though if this agreement is not already established, the discussants may make the ITP itself a matter of meta-discussion. When the

meta-discussion over an ITP must be conducted by experts external to the primary-level discussion, the ITP ends up having the same strengths and weaknesses as other forms of authority-dependent argumentation.

For the most part, ITPs in expert fields must operate as Walton (1996) describes for other forms of "presumptive argument." The ITP, once established within the field's practice, can be applied wherever relevant to produce conclusions that enjoy a very strong presumption. An assertion that might be doubted or contradicted within a discourse, once passed to the ITP, acquires a presumptive status, either as verified or as falsified by the ITP. The acceptability of the ITP does not have to be defended in each occasion of use; what has to be defended is a refusal to accept the results of the ITP as an adequate defense of the tested assertion.

Much depends on the reliability of the ITP, since in many ways it functions as an argumentative 'black box" that generates presumptions for or against particular assertions. In pragma-dialectics, the reliability of an ITP or any other argumentative move is known as its problem validity (van Eemeren & Grootendorst, 1994). Problem validity (or problem-solving validity) refers to a procedure's capacity for contribution to the idealized goals of argumentation – that is, to resolution of disagreement on the merits of the competing positions. In commonsense terms, a procedure lacks problem validity if it leads arguers into false conclusions, false consensus, paradox, impasse, or other argumentative failure. Problem-valid procedures contribute to the quality of argumentation, either providing new ways to resolve doubt or offering new protections against missteps.

The idea of problem validity is a bridge between pragma-dialectics as a critical and empirical enterprise and a pragma-dialectical program of design. ITPs and the argumentative forms that develop around them are design solutions to recurring argumentative problems. Any newly proposed ITP or associated argument form might be an advance for argumentative practice within its field, but until its problem validity is known, it should be regarded as a potential design failure.

2. Argumentation Templates

Within expert fields of all kinds, and especially scientific fields, argumentative practice tends to stabilize around ITPs to produce stereotyped forms of demonstration and defense of claims. We will use the term "argumentation

template" to refer to these stereotyped forms. These templates function as outlines for the development of an argument, including not only formal and functional qualities captured in the notion of an argumentation scheme, but also procedural and presentational guidance for the arguer attempting to develop a case for a scientific claim, starting from scratch. Clear contemporary examples of argumentation templates are formats for writing research reports or for writing environmental impact statements.

Argumentation templates of this kind are not simply outlines for writing, however. These templates amount to a synopsis of the burden of proof to be met by empirical claims, often defining specific assurances an expert must provide in order to produce an argument that will be convincing to other experts. In scientific fields, the assurances invoked by standard templates generally involve observational and analytic steps, including laboratory procedure and statistical analysis. While the connections between specific concrete research procedures and any particular empirical claim may be quite obscure, these procedures, once widely accepted, allow individual scientists to hand off portions of the burden of proof associated with the claim and to have that burden

Among the most common of scientific handoffs are those involving statistical analysis of observational data. This handoff may occur very literally, as when the researcher delegates analysis to a statistician or to a statistically sophisticated assistant. But even when the researcher conducts his or her own analysis, an argumentative handoff often occurs through the importation of a complex but unarticulated substructure into the empirical argument. In Toulmin's terms, we would want to regard statistical tests as warrants for drawing empirical conclusions from data; but if a test is treated as a warrant, its backing is an open-ended and possibly not-fully-coherent body of statistical theory that becomes increasingly obscure as the warranting move becomes increasingly common (Gigerenzer et al, 1989, esp. pp. 106-109). Whether the handoff is literal or figurative, then, conventional statistical procedures introduce deep dependencies on authority into argumentation templates. There is efficiency in this if the procedures are good ones, but there is also the risk that the procedure will come to be treated as a black box whose workings are mysterious but whose results are accepted without question. It is quite convenient, in fact, to think of some argumentation templates as actually including black boxes that turn data into conclusions.

There is little doubt that on the whole the growth of statistics has improved our

ability to reason about both the natural world and social phenomena, and these improvements have stabilized into highly successful argumentation templates (such as the stylistic and substantive requirements of the APA Publication Manual). However, any particular proposal for statistical analysis may either improve our ability to reason or set it back in some unexpected way. In the rise of statistical thinking over the past several centuries we can see the invention of new safeguards against error, but we can also see that new fallacies get invented right along with nonfallacious moves, and that these two sometimes stabilize into widely applied templates. The emphasis within pragma-dialectics on procedure and procedural rules provides some unusual and powerful tools for examination of these argumentation templates as abstract designs for the management of doubt.

3. Evaluating Problem Validity

Central to establishing the problem validity of any argumentative structure or strategy is examination of how that structure or strategy advances or impedes the abstract goals of critical discussion. In foundational statements of pragma-dialectics, problem validity is a matter of testing a set of rules for their ability to contribute to resolution. Argumentation templates are not exactly rules in the pragma-dialectician's sense, but rather standard ways of attempting to conform with rules such as those defining the idealized practice of critical discussion. Many argumentation templates come about as ways of invoking or reporting the outcome of intersubjective testing procedures established within an expert field, and the intersubjective testing procedures, in turn, come about as ways of regulating the introduction of new assertions. We can extend the examination of problem validity to any component of argumentation that becomes part of a field's standard practice.

A general methodology for evaluation of problem validity would include several steps:

(1) reconstruction of the argumentative move to be evaluated, including both formal design features and informal accommodations worked out in practice;

(2) comparison of the generalized output from this move with a critical standard to identify any vulnerabilities; and

(3) investigation of how these vulnerabilities look in actual instances of argumentation.

A noteworthy feature of this methodology, and one that is particularly characteristic of pragma-dialectics, is the emphasis on examination of what

results from the practice to be evaluated. Problem validity is about the suitability of an argumentative move for advancing arguments within some practical setting. Problem validity has to do not with the qualities of individual bits of argumentation, but with pragmatic properties of rules or other agreements about how to conduct discussion.

Problem validity has some general affinities with the concept of "ecological rationality" as interpreted within the work of Gigerenzer and "the ABC Research Group" on adaptive thinking (Gigerenzer, 2000; Gigerenzer, Todd, & the ABC Research Group, 1999). Ecological rationality is reasoning that is well-adapted to the environment in which it occurs, taking advantage of the structure of that environment to gain efficiency or reliability. In the ABC research program, shortcut reasoning heuristics and rules of thumb are examined in terms of their success in supporting good decisions. A heuristic may have little or no logical defensibility but still be very successful in its actual use.

Heuristics and rules of thumb are common in all human reasoning, and are often treated analytically as fallacies and biases. But some of these heuristics can be given convincing defense as "fast and frugal" methods for making decisions. Gigerenzer and associates (1999) have shown, using computer simulation of judgments, that supposedly biased judgmental strategies are often beautifully adaptive to information environments with predictable structure. The gist of the ABC group's argument is that heuristic reasoning is not a poor substitute for either 'unbounded rationality' or 'optimization under constraints,' but an adaptive response to contexts of choice that are already structured to prefer certain kinds of strategies. Very simple and unreasonable heuristics for decisions under uncertainty can be shown to be ecologically rational, by showing that these heuristics, applied in certain environments, produce good decisions with minimum cost.

The general idea that we may adopt a broad rule for decisions based on its overall productivity has direct relevance to statistical testing, which is broadly understood by scientists themselves as adoption of a decision rule for interpretation of experimental outcomes. The idea that a rough heuristic may prove to be defensible on the same grounds has direct relevance to our specific topic, which is rules of thumb for application of statistical tests. Especially relevant, though, is the idea that we might test any decision-making strategy, including an ITP, by simulating its use in conditions controlled through explicit

modeling.

4. Rules of Thumb for Application of Statistical Tests

Much empirical work in the social sciences involves statistical tests of the differences among groups of observations. A significant result is taken as evidence of a difference, a relationship, or an effect, allowing for a very simple argumentative structure to apply in many cases:

Effect E is indicated by test T.
T rarely produces false indications when properly applied.
T has been properly applied.
Therefore (presumptively), E.

For example, an experiment on alternative teaching strategies might involve testing differences in exam scores for several groups of students, or an experiment on alternative persuasive strategies might involve testing differences in responses for several audiences. Statistical tests suitable for these purposes are well known and include *t*-tests for differences between two group means and *F*-tests for differences among three or more means.

The idea that "T rarely produces false indications when properly applied" could open a disagreement space of its own, but it rarely does within social science practice. For purposes of empirical argument within research contexts like these, a researcher who has collected observations of a certain kind may defend a claim about an effect such as a group-to-group difference simply by presenting results of a standard test such as a *t*-test or an *F*-test. The justification for the test itself is typically external to the empirical field in which the test is applied, having been delegated sometime in mid-1900s to statistics as a subfield within mathematics (Gigerenzer et al., 1989, esp. pp. 115-118). That T rarely produces false indications when properly applied is generally taken for granted, though the researcher is then under obligation to provide assurances that the test has in fact been properly applied. If these assurances can be given, letting the test function as an unquestioned black box is as reasonable as the theory backing the test.

Among the assurances a researcher must provide are assurances of the quality of measurement, the quality of the observational sample, and the fairness of the comparative design. These assurances, while interesting, have no further extension in our case study. The assurances that will concern us most are those that condition the interpretation of the results of the statistical test: those

commonly known as statistical assumptions (For an overview of these assumptions, see any good textbook treatment of the analysis of variance, such as Keppel, 1991, esp. ch. 5). The common F-test for differences among group means assumes that observations taken within the groups are drawn independently of one another from a population or more than one population whose elements have normally distributed values on the variable measured as an outcome of the experiment. These are commonly known as the independence assumption and the normality assumption, respectively. The test also assumes that if several populations are sampled, their members are equally heterogeneous. This is commonly known as the homogeneity of variance assumption. If any of the assumptions are violated, the acceptability of the statistical test itself may be called into question.

The statistical assumptions are very difficult to verify in any actual research situation, and for this reason researchers cannot usually provide these assurances directly. Assurances that the assumptions are met for the actual occasion of use must be obtained through examination of the same data as used in the test itself. Hence, the argumentation templates that have evolved around significance tests for group differences include specialized procedures for evaluating the reasonableness of each assumption, by testing for "violations" of various assumptions. Since the assumptions are in fact often violated, the actual use of significance tests is adjusted over time in response to decontextualized studies of the behavior of the statistical tests known as "robustness studies." The purpose of a robustness study is to determine how badly a test behaves under varied deviations from the ideal observational situation. A test that works well despite violations of assumptions is said to be robust to those violations.
The behavior of a statistical test is normally assessed in terms of its ability to control the rate at which errors of inference are made from data. "Type I error" is concluding that a difference exists when it does not, while "Type II error" is failing to find authentic differences. All sample data show differences of some kind, and the function of a statistical test of observed differences is to differentiate between differences that reflect real effects and differences that reflect only chance variation within a sample. Type I error can be set to any desired rate through designable features of tests; by broad and stable convention, Type I error is controlled at 5%. In other words, tests for all kinds of differences are structured so that, if there are no true differences to be found, the test will (falsely) find differences in no more than 5% of the cases.

Type I error (and also Type II error) may vary dramatically from what the scientist expects if the assumptions required by the test are violated – but then again, they may not. What happens to Type I error rates if the observations come from something other than a normal distribution? That is the kind of question answered by robustness studies. A test that has been shown to be robust to a certain kind of violation offers the individual researcher a boilerplate rebuttal for criticisms related to the violated assumption, which can also be woven preemptively into an argumentation template to implement a structure like the following:
Effect E is indicated by test T.
T rarely produces false indications when properly applied or in other situations
S1, …(Si), … SN.
Si obtains.
Therefore (presumptively), E.

Often, these boilerplate rebuttals get appropriated into routine scientific practice as rules of thumb. Rules of thumb are common enough in statistical reasoning that van Belle (2002) recently summarized 99 such statistical and methodological rules (e.g., "make a sharp distinction between experimental and observational studies;" "randomization [of experimental subjects into groups] puts systematic sources of variability into the error term;" "consider the size of the population affected by small effects;" and "beware of *pseudoreplication*"). van Belle provided a basis for each rule, an illustration of how it works in statistical reasoning, and extensions of the rule. Some rules of thumb were formed based on statistical and methodological theory (e.g., the principles of randomization can be traced to Fisher's, 1935, work on experimental design) and others arise from practical circumstances when statistics are applied (e.g., epidemiological work shows that small effects are important when researchers are dealing with large populations – a small effect of a disease in large number of people may still mean that many will die).

Rules of thumb related to assumptions enter social science practice through textbooks, through summaries of robustness research appearing in textbooks and research handbooks, and through explicitly argued proposals for handling specific kinds of problems. For example, various texts point out that "heterogeneity of variance" is a benign violation so long as the variance of the most heterogenous group is no more than three times the variance of the least heterogeneous group (see, e.g., Keppel, 1991). The basis for this rule of thumb is a body of robustness

studies, one showing little harm from heterogeneity on the order of 3:1, and others showing considerable harm from much larger differentials. Although the empirical analysis provided by robustness studies gives good grounds for confidence in *F*-tests performed on mildly heterogeneous groups and equally good grounds for concern about in *F*-tests performed on horrendously heterogeneous groups, the 3:1 rule of thumb is itself a product of happenstance in robustness researchers' choices of conditions to examine.

Notice that just as we can examine the behavior of a specific statistical test as it is applied in any desired conditions, we can also examine the behavior of the associated rules of thumb. So long as the rule of thumb can be stated as a decision rule applied systematically, it can be modeled using the same kinds of computer simulation methods used in robustness studies (and in studies of the ecological validity of heuristics).

5. Evaluating a Rule of Thumb for Non-independent Data

Independence of observations, as noted above, is one condition or rule stipulated for many statistical tests (e.g., independent samples *t*-tests, *chi*-square tests, *F*-tests for independent group means, and so on). When observations are collected in pairs or groups, it is generally acknowledged that it is inappropriate to treat them as independent. As Kenny and Judd (1986) demonstrated, treating scores for individuals within dyads or groups as independent risks bias in statistical significance tests, with the amount and direction of bias varying with the amount of dependency – that is, the size of the intraclass correlation among the participants within groups – and the experimental design. Non-independence occurs when scores are correlated and may result from natural associations between participants in a study, such as when intact dyads (e.g., parent/child, partners in a relationship, or coworkers) are used as participants. Kenny and Kashy (1991) noted that these forms of non-independence are common in research on interpersonal relationships.

Non-independence also can result from the particular circumstances of the data collection, such as when groups of participants within a study respond to the same stimuli (see Jackson & Brashers, 1994). For example, in research on social influence, it is necessary to manipulate variables by embodying the contrast of interest in concrete materials: for example, by writing a message and varying it in some respect to produce two or more versions that represent a treatment contrast. In an experiment on the effects of authority on persuasion, a variety of

messages (e.g., on AIDS, crime prevention, voting, and immigration policy) might be altered to have two versions that vary in uses of authority – for example, putting forward assertions attributed either to authentic authorities or to non-authoritative sources. In a completely randomized design, participants in the experiment read one or the other version of a message, and then complete an attitude or behavioral intention measure to determine if there are different responses to messages differing in their use of authority. Multiple replications of the treatment contrast are used to allow inference from individual messages (e.g., on AIDS or immigration) to broad, categorical differences in message strategy (e.g., to the benefit of citing authorities). But these replications are a potential source of non-independence, because subgroups of participants are responding to common stimuli. In a replicated design, where observations fall into subgroups defined by replication levels, the observations within one subgroups are more related to one another than to observations taken within other subgroups, and these relatedness can extend across the treatment levels as well (e.g., relating the individuals who got the authoritative version of the AIDS message to the individuals who got the non-authoritative version of the same message). If the replication factor is ignored and all observations classified only with respect to other factors (e.g., the authority treatment factor), then the assumption that observations are independent may be violated, because observations correlated due to common stimuli would be treated analytically as though uncorrelated. Replications, in other words, may become a "hidden factor" in a design, resulting in all subjects getting one treatment being considered one large group rather than a number of subgroups characterizable in terms of which particular experimental materials they received.

When non-independent observations are treated as though they were independent, the Type I error rate for the test is no longer known; it is no longer assured, that is, that "test T rarely produces false indications." The rate of Type I errors may be much higher than expected, a problem known as "alpha inflation" (since the rate set for Type I error is known as "alpha"). Barcikowski (1981) demonstrated through statistical simulation that treating observations from groups nested under treatments as though the observations within treatments were independent leads to substantial alpha-inflation (more Type I errors than we should expect with a set alpha-level), with the size of the alpha-inflation increasing with the size of the intraclass correlation and the number of observations per group. Kenny and Judd (1986) examined both within-group and

between-group dependencies and found that both forms of non-independence could bias a test, though the direction of bias (alpha-inflation or alpha-deflation) differs by type of non-independence.

Regardless of how observations are collected, however, an absence of correlation among observations allows the test to perform just as expected. If, despite dependent sampling, the intraclass correlation is zero, or if there are no within-group or between-group correlations, the test of differences among means will have the nominal Type I error rate. Noticing this fact, some experts have proposed rules of thumb for the handling of potentially non-independent data that allow direct application of a test when there is no evidence of non-independence but require adjustments or alternate tests when evidence of non-independence appears. In general, non-independence can be handled by taking the "hidden factors" responsible for the non-independence explicitly into account. For example, when experimental observations can be subdivided not only by treatments but also by replications, taking replications into account as a partitioning factor eliminates the non-independence among the individual observations within groups.

Kenny and Kashy (1991) described a rule of thumb for dealing with possible non-independence and for deciding what test to use to analyze data collected in pairs, structured as a two-step testing procedure. At step 1, a test for non-independence is conducted, using a very liberal criterion to avoid Type II error. At step 2, the test that is conducted depends on the outcome of the preliminary test: if the preliminary test shows no evidence of non-independence, the main analysis can be conducted as though the observations were fully independent, while if evidence of non-independence appears, some alternative form of analysis is required. Others (e.g., Forster & Dickinson, 1976) have proposed similar rules of thumb for other possible sources of non-independence.

Evaluating this rule of thumb is not quite as straightforward as evaluating a statistical test, since the rule of thumb depends on modelling a judgment and not just a distribution of outcomes. An annoying feature of rules of thumb is that they tend not to be applied with complete consistency, but with a certain amount of opportunism varying according to the individual taste of the researcher. Nevertheless, if we want to evaluate the rule of thumb itself, and not the behavior of the individual researcher, we may make some progress by formalizing the rule and modelling what would happen if it were applied with complete consistency within a community of researchers.

Adapting methods common in robustness studies, we developed a simulation of two kinds of situations producing non independent data:
(1) situations in which all of the members of a subgroup are assigned together to one treatment condition in an experiment, and
(2) situations in which the members of a subgroup are divided between two treatment conditions. A complete technical report of the simulations is available elsewhere (Jackson & Brashers, 1993).

Very briefly, though, the simulation involved random generation of data with specific features, and application of testing strategies to these data to produce empirical Type I error rates. Varying the size of the simulated experiments (number of groups and number of observations per group) and the magnitude of the intraclass correlations, we built into the simulation three contrasting analytic strategies: an unconditional test treating all observations as independent, a conditional testing strategy that models the consistent application of the rule of thumb described above, and an unconditional test in which the source of non-independence (e.g., subgroups) is included as an explicit factor. Using computer algoritms based on SAS functions, we ran thousands of simulated experiments of each type and size, tabulating the frequency with which each testing strategy produced a statistically significant result.

Consistent with earlier findings, the unconditional test was biased, with the magnitude (and direction) of bias determined by the magnitude and form of non-independence and by study size. Type I error was enormously inflated under some conditions that are actually fairly common in social science research. Using the conditional testing strategy, this bias was substantially reduced, but not eliminated. The reason for this is that the test for non-independence may fail to detect the non-independence, even when it is built into the composition of the observations to be analyzed (a problem of Type II error). The “presumption” is misplaced in any such testing strategy, since the data are presumed to be independent unless it is shown that they are dependent. An unconditional test built on a presumption of non-independence among observations within subgroups behaves exactly as it should, producing significant results in 5% of all experiments.

Jackson and Brashers (1993) noted that any procedure constructed in this way will be vulnerable to the same “fallacy of misplaced presumption.” If group effects are present in the population, any test conducted ignoring the group effect will be

biased, so we should treat related observations as dependent whenever we are not confident that group effects are *absent*. But the testing strategy above generates individual as the unit of analysis whenever we are not confident that group effects are *not present*. The presumption should favor treating group data as dependent (since this results in an unbiased test regardless of the size of the group effect), but the policy outlined awards the presumption to treating grouped data as independent by requiring positive evidence of group effects to generate the choice of group as the unit of analysis. While preferable to an incorrect test applied unconditionally, the conditional testing strategy is inferior to a consistent policy of conducting a test that allows interdependence among observations.

We could describe this fallacy of misplaced presumption in more familiar terms as a version of argument *ad ignorantium*, since independence is considered to have been established through absence of clear evidence of non-independence. Structurally, the argument form looks something like the following:
E is indicated by T.
T rarely produces false indications when properly applied.
T is properly applied if no assumptions are violated.
No assumptions are (known to be) violated.
Therefore, E.

But the fallacy of misplaced presumption differs from an *ad ignortium* form arising from simply ignoring the possibility of non-independence. Its defining difference is in the practical decision to treat data as independent whenever a test for dependence fails to show "indications" of dependence.

The simulation methods used to evaluate the research policy suggested by the rule of thumb can be adapted to evaluation of individual empirical arguments. The observational and analytic choices can be modeled by creating a simulated experiment of the same size and design and randomly generating many repetitions of the experiment with varied assumptions about the underlying process. Brashers (1994) showed this method in his critical examination of research practices in communication and psychology, modelling dozens of studies making varied analytic decisions about experimental replication factors. For example, Brashers simulated the procedures of Fein and Hilton's (1992) study of consistency between attitudes toward groups and attitudes toward individual members of those groups. Fein and Hilton used the two-step testing strategy to decide whether to include experimental replications as an explicit factor or to

“hide” the factor and treat all observations within groups as independent. The initial test showed no significant effects involving the replications factor, so following the policy suggested by the rule of thumb would mean going forward with analysis ignoring the potential non-independence among observations sharing assignment to the same replication. Using evidence from the published results to set upper bounds for certain kinds of dependency, Brashers showed Fein and Hilton’s testing strategy to involve much more than 5% chance of Type I error.

6. Conclusion

In the social and behavioral sciences, statistical tools and techniques figure heavily in empirical argumentation templates. But empirical social science, despite its visible adherence to templates incorporating formal requirements of proof, is far less formal in its methodology than is commonly noticed. A careful and rigorous enforcement of statistical standards of proof in empirical demonstration is blended with a casual and pragmatic acceptance of rules of thumb and other ad hoc solutions to problems of application. In itself, this is no critique of empirical argumentation; these rules of thumb may be quite reasonable, but that must be shown.

We might speculate that statistical rules of thumb are highly disciplined versions of fast and frugal heuristics, not defensible in the abstract, but effective and efficient in practice. Unfortunately, this save is not possible for the argumentative move examined in this study, since regardless of whether dyadic and grouped data are mostly independent or mostly interdependent, nothing much is gained by applying this rule of thumb.

Our point, however, is not merely to mount an objection to a particular rule of thumb, nor to suggest that we always avoid rules of thumb. Rather, what we have tried to show is an approach to the investigation of problem validity within disciplined argument fields. Other rules of thumb for statistical reasoning will fare differently when evaluated for their contributions to empirical argumentation. As it happens, though, in the case examined here, there is a readily available analytic strategy that can be shown to be uniformly acceptable, regardless of whether data show clear evidence of non-independence. In challenging the problem validity of one strategy, we also vouch for the problem validity of an alternative.

REFERENCES

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, 6, 267-285.
van Belle, G. (2002). Statistical rules of thumb. New York: John Wiley & Sons.
Brashers, D. E. (1994). *A critical review of the design and analysis of experiments using replications factors.* Unpublished dissertation, University of Arizona.
van Eemeren, F. H., & Grootendorst, R. (1984). *Speech acts in argumentative discussions*. Dordrecht, The Netherlands: Foris.
van Eemeren, F. H., & Grootendorst, R. (1994). Rationale for a pragma-dialectical perspective. In F. H. van Eemeren & R. Grootendorst (Eds.), *Studies in pragma-dialectics* (pp. 11-28). Amsterdam: International Centre for the Study of Argumentation.
van Eemeren, F. H., Grootendorst, R., Jackson, S., & Jacobs, S. (1993). *Reconstructing argumentative discourse*. Tuscaloosa, AL: U. of Alabama Press.
Fein, S., & Hilton, J. L. (1992). Attitudes towards groups and behavioral intentions towards individual group members: The impact of nondiagnostic information. *Journal of Experimental Social Psychology*, 28, 101-124.
Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F', and min F'. *Journal of Verbal Learning and Verbal Behavior*, 15, 135-142.
Gaskins, R. H. (1992). *Burdens of proof in modern discourse*. New Haven: Yale U. Press.
Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford: Oxford U. Press.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge U. Press.
Gigerenzer, G., Todd, P., & the ABC Research Group (1999). *Simple heuristics that make us smart.* Oxford: Oxford U. Press.
Jackson, S., & Brashers, D. E. (1993, May). *Assuming independence when dependence isn't evident: A fallacy of misplaced presumption.* Paper presented at the meeting of the International Communication Association, Washington, D. C.
Jackson, S., & Brashers, D. E. (1994). M > 1: Analysis of treatment x replication designs. *Human Communication Research*, 20, 356-389.
Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence

assumption in analysis of variance. *Psychological Bulletin*, 99, 422-431.
Kenny, D. A., & Kashy, D. A. (1991). Analyzing interdependence in dyads. In B. M. Montgomery & S. Duck (Eds.), *Studying interpersonal interaction* (pp. 275-285). New York: Guilford.
Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
McCloskey, D. N. (1985). *The rhetoric of economics. Madison*, WI: U. of Wisconsin Press.
Toulmin, S. E. (1958). *The uses of argument*. London: Cambridge U. Press.
Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning*. Mahwah, NJ: Erlbaum.
Walton, D. N. (1997). *Appeal to expert opinion: Arguments from authority*. University Park, PA: Pennsylvania State U. Press.