Resolving Intra- Study Inconsistency

Robert D. O'Keefe
Driehaus College of Business
Kellstadt Graduate School of Business
DePaul University
1 East Jackson Blvd.
Chicago Il 60604
Rokeefe@depaul.edu

Abstract

This paper deals with the major elements of research projects as performed within the various disciplinary divisions of the behavioral sciences and related fields. It is primarily concerned with the consistency that should obtain among these elements and employs the graphical mode of balance or consistency theories as a means of illustrating various relationships that may be found among these elements.

The concept of the plausible rival hypothesis (H_p) is extended to take into account possible sources of alternative explanation existing within bodies of data related to the primary topic of research.

A researcher faced with a non-rejectable null hypothesis is advised to re-review the literature and attempt to uncover alternative theory based data that might lead to just such a prediction. This prediction is, in effect, an H_p drawn from the available conceptual evidence as distinguished from an H_p attributable to errors in design or method.

Examples are provided and the most complete of these illustrates a case of such inconsistency found in the topical area of research on *Conformity* and resolved by attention to the literature of *Inter-Personal Perception.*

Key Words: Consistency Theory, Research Design, Statistical Inference, Plausible Rival Hypothesis, Resolving Intra- Study Inconsistency

Introduction

In the research paradigm most often employed by behavioral scientists two forms of inferential evidence interact to guide the investigator's approach, design, method and, most important for the purposes of this presentation, the researcher's conclusions.

At the outset of the period of a study the researcher relies upon conceptual evidence. Essentially this means examining the available findings and conclusions of other research reports and from these extracting whatever information seems relevant to the research question or problem as hand. It is very possible that the discovery of relevant inductive evidence may of itself suggest a researchable problem.

Should the research problem be novel, or the methodological approach unique, the available conceptual evidence may be derived from intuition. However, such intuitions that may simulate efforts in a presently un-investigated or incompletely investigated behavioral domain are not without some empirical basis. In the usual case, these intuitions are rooted in one or a series of observations that occurred in a natural setting.

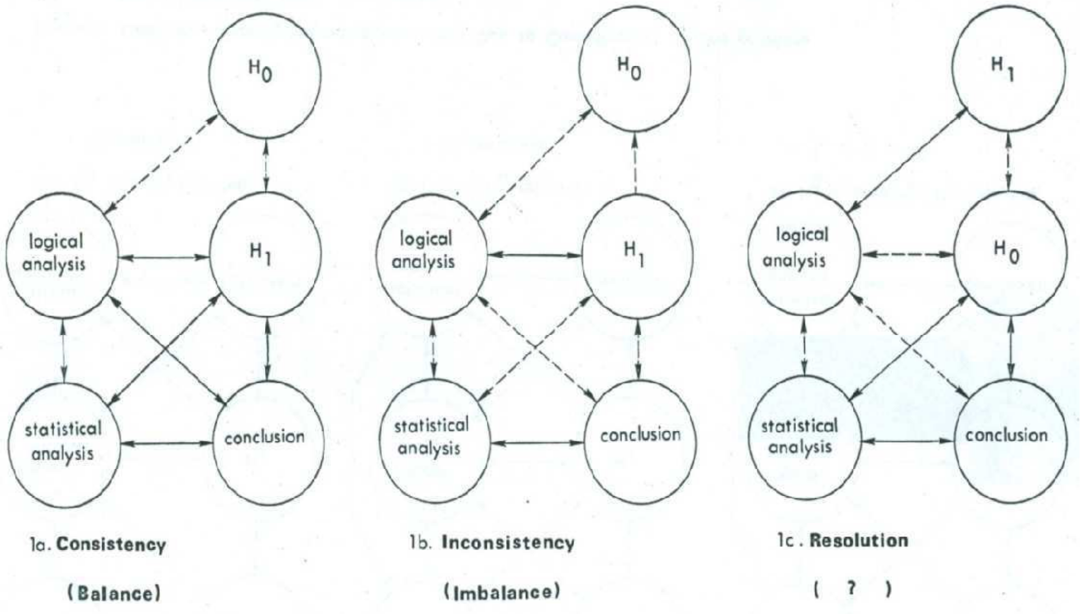
Whether the planned study is completely innovative, or one in a long series of exact or modified replications, the intuitive observations and previously reported inductions are initially submitted to a logical analysis by the researcher. It is this sort of inference process that leads to the statement of hypothesis.

Statistical evidence refers to the results of a single or series of tests appropriate to the investigator's obtained data. Quantitative analytical techniques are used in testing hypotheses

and so providing further inferential evidence in the form of a probability statement. If the obtained result i.e., the value of the test statistic falls at or beyond a certain level of confidence (conventionally p.< 05), the hypothesis as stated is accepted as confirmed. (Cohen, 1994; Dixon,1998; Iacobucci,2005)

The term confidence is particularly appropriate in this context because it seems to express exactly how the investigator feels when a carefully nurtured hypothesis is supported by the obtained statistical analysis. Probably, pleasant would be equally descriptive or euphoric, or even better, tension free.

Confirmation of the stated hypothesis indicates that there are no weak links in the inferential chain. That is, having achieved a satisfactory level of confidence in the plausibility of the stated hypothesis, the researcher derives a similarly increased confidence on his or her own powers of logical analysis. Where the consistency between prediction and outcome obtains, the final statement or conclusion is usually a positivistic paraphrase of the hypothesis. Consequently there is a perfectly symmetrical relationship among the four major elements of the study. (Thagard,2000; Winkielman, Huber, Kavanaugh,& Schwarz, 2012)



1a. **Consistency**

(Balance)

1b. **Inconsistency**

(Imbalance)

1c. **Resolution**

(?)

Figure 1. Conditions of Intra-Experimental Consistency, Inconsistency, and Resolution of Inconsistency.

A Balance Theory View of Research Study Elements

Figure 1 illustrates this relationship in the graphical mode employed by structural balance theorists. Note that 1A indicates a balanced or consistent relationship among all four elements of the research paradigm. A logical analysis of the available conceptual evidence leads to the hypothesis. Statistical analysis of the empirical data leads to the conclusion. If the statistical analysis supports or confirms the hypothesis, it also supports the logical analysis. Therefore, the logical analysis and the conclusion are equally consistent with each other.

The second graph (1B) depicts a situation in which consistency among the four elements is not so perfect. Here the logical analysis and the hypothesis are still consistent as are the statistical analysis and the conclusion. But, the conclusion is not consistent with the hypothesis (H_0). Actually this is a conventionally acceptable and a rather roundabout way of reporting that the researcher has failed to confirm the alternative (H_1), that is, the stated hypothesis.

The resolution is presented as 1C. Note that H_0 is now consistent with the statistical evidence and that the conclusion is, therefore, a statement that no difference of the required magnitude has been found.

Students of balance graphs will probably agree that 1C is more balanced that 1B but less balanced than, the best of all possible graphs, 1A. (Harary, Norman, & Cartwright 1965)

The problem seems to be that in 1C the inconsistent element is the investigator's logical analysis. This is exactly the sort of situation to which Campbell & Stanly (1963) and, to some extent Skinner (1963), refer in their discussions of behavioral research and especially the former authors' insightful description of the pain experienced by an investigator faced with the non-confirmation of a cherished hypothesis. Similar ideas are expressed by Sterling(1960) and Griffin & Ross (1991).

Campbell & Stanley go on to explain that because experimenters are biological and psychological animals, and so subject to the laws of learning, the pain or tension experienced may be associated more vividly with the research situation than with the inadequate theory that is the "true" source of frustration. It seems more likely that, since research studies can only probe a theory, a better candidate for the source of frustration and tension present in such the situation outlined above would be the investigator's own analytical and logical powers. In effect, it is, at the least, disconcerting to have to admit an error and rejecting one's own hypothesis is just such an admission. This initial frustration may be exacerbated by the thought that beyond the immediate disappointment there lies, the somewhat widely shared suspicion on the part of researchers that editors and reviewers of scholarly publications prefer positive findings and are adverse to studies which do not report that the research hypotheses were validated and the corresponding null hypotheses were rejected.(Sterling, 1959; Rosenthal,1979; Rowney & Zenisek,1980; Klayman & Ha,1987; Hubbard & Armstrong, 1992; Frick,1996; Nickerson, 2000; Wainer & Robinson, 2003)

But is this a fair appraisal? At the outset it was stated that two forms of evidence, the conceptual and statistical, should interact. It seems quite reasonable that they should, but in actuality such an interaction occurs only when the statistical evidence is consistent with the conceptual or logical evidence. In other words, only when they obtained statistical results are positive and reach an acceptable significance level.

On this point adherents of the Bayesian approach to statistical inference have vividly depicted the behavior of behavioral scientists facing their own special moment of truth. In the words of the Bayesians:

"If the null hypothesis is classically rejected, the alternative hypothesis is willingly embraced, but if the null hypothesis is not rejected it remains in a kind of limbo of suspended disbelief." (Edwards, Lindman and Savage 1963 p. 235; Edwards, 1965)

It is probably no accident that their description of classical test procedures as "curiously asymmetric" regarding rejection of one or the other hypothesis converges, if only at a nominal level, with the present discussion of imbalance or inconsistency among the intra-study elements. Bayesian statisticians object to, among other things, the testing of a rather diffuse research hypothesis, where the magnitude expected is rarely stated, against a sharply defined null hypothesis. This is the basis for their labeling of classically accepted procedures as asymmetric. (Bernardo & Smith, 2000; Trafimow,2003; Sewell, 2012)

A researcher's suspended disbelief may give way to an examination of procedures, methods, apparatus or other related considerations. It may result in an open admission of error or a partial explanation usually appended by a statement to the effect that more research is needed before the obtained outcomes are considered valid research findings. In any case, because the statistical analysis has not supported the logical analysis, the researcher actively seeks some sort of resolution for this painful inconsistency.

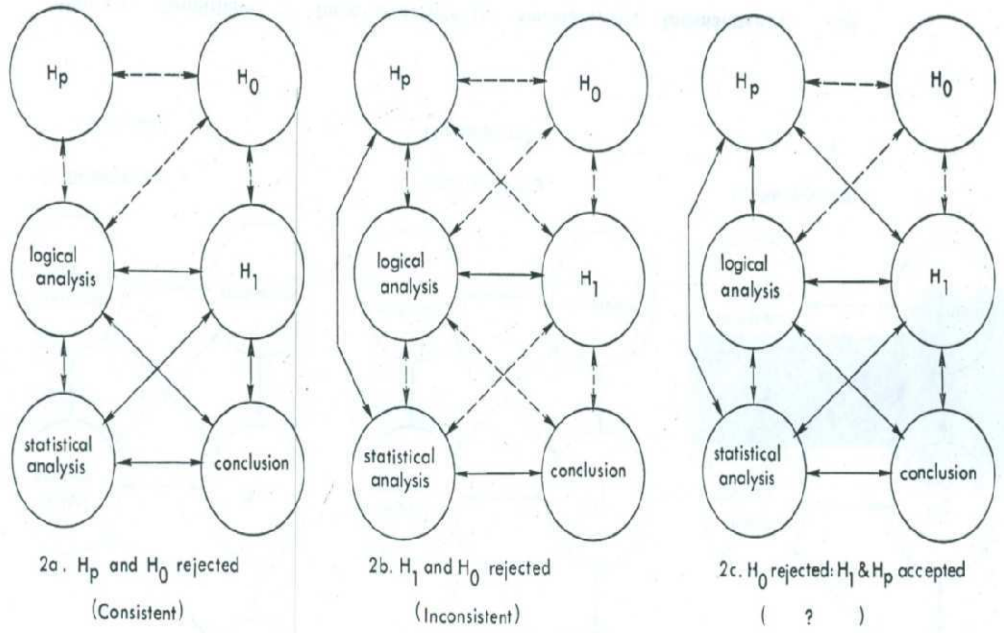
There is still another case that deserves mention. For some researchers the balanced state depicted in 1A is attained but short lived. Once again Bayesians have provided a frank summary of this situation. They describe an "interocular traumatic test" as one that makes the meaning of the collected data immediately evident to the researcher. That is, prior to any application of a statistical test, the conclusion hits him right between the eyes. However, the Bayesians caution the researchers that their enthusiast's interocular trauma may be the skeptic's random error and further advise that a little arithmetic to verify the extent of the trauma can yield peace of mind (tension reduction) at little cost. (Edwards 1965; Bradstreet1996)

The Problem of False Positives

The well known point to be underlined here is that, arithmetic or not, false positives are often accepted, reported as conclusions and interwoven into the behavioral sciences' literature.(Krantz,1999; Krueger,2001; Haller & Kraus, 2002; Liebman & Cunningham, 2009; Novella,2011) The process of identifying and extracting these false positives can be long and slow and becomes increasingly difficult since, it seems, that the original finding, whether stated with caution or not, increases in truth strength with the frequency of subsequent citations. A now classical and telling example of this phenomenon can be found in the review of research on Cognitive Dissonance provided by Chapanis and Chapanis (1964)

To further illustrate the gravity of this problem, Campbell and Stanley (1963) have compiled an inventory of factors that can jeopardize the internal and external validity of experimental, quasi-experimental and ex-post facto research designs. These factors are deservedly called plausible rival hypotheses. When left uncontrolled these factors can produce effects that may be mistakenly linked to experimental treatment(s) or to whatever research conditions have been formally hypothesized as the loci of causality. These threats to the validity of research findings are hardly an a priori compilation of compulsive dos and don'ts. Rather, their effects are based on empirical lessons derived from study and experimentation in behavioral sciences and so are comparable to a sort of vicarious avoidance learning situation for the aspirant as well as for experienced research oriented behavioral scientists. In a more global context, since empirical research has been described as an evolutionary device that aids in the selection and retention of valid scientific knowledge, attention to such sources of error is justified by an attendant gain in such selective precision. (Campbell, 1959)

An awareness of the sources of error introduces a new element into the situation. Rather than simply a sharp H_0 and a rather diffuse H_1, the investigator must consider one or a number of plausible rival hypotheses (H_{p1}, H_{p2}, …, H_{pN}). In most cases plausible rival hypotheses are eliminated through attention to design and method during the planning stages of the research project.



2a. H_p and H_0 rejected (Consistent)

2b. H_1 and H_0 rejected (Inconsistent)

2c. H_0 rejected: H_1 & H_p accepted (?)

Figure 2. Conditions Existing Intra-Experimentally Due to Consideration of One or More H_ps (Plausible Rival Hypotheses).

The relationship discussed above is illustrated in Figure 2. Here Hp is introduced to illustrate that one or all of the possible intruding plausible rival hypotheses relevant to the study are rejected (2A). The second frame of the figure, (2B), illustrates a situation where a relevant H_p is acceptable as primarily responsible for the outcome attributed to, for example, an experimental treatment. In 2C H_p and H_1 are seen to interact to produce such differences as are observed. For example the correlation r_{xy} may be statistically significant as predicted by H_1, but the obtained coefficient may well be inflated due to effects predictable from a relevant H_p.

The decision as to which of these conditions a single study represents, and especially the choice between 2B or 2C, is one which can be made only after at least one exact replication of the original study in which the conditions summarized as Hp have been controlled or eliminated. The replication is depicted in 2A and each of the graphs comprising Figure 1 also assume that such factors are irrelevant.

When a researcher views the completed study as representative of 1A, there is little reason for tension. However, if at some later date a colleague, an editor, or a critical reader should detect some sort of imbalance producing error, tension is certainly aroused. The span between completion and detection, or publication and detection may range from a few days to several years and it is probably safe to say that the closer the detection to publication, the greater the tension for all concerned.

Plausible Rival Hypotheses

Thus far, the plausible rival hypothesis has been discussed from the standpoint of error in design, method, and/or other uncontrolled factors alone or in interaction. However, it seems reasonable that such hypotheses may be equally applicable to the conceptual or logical processes involved in the statement of H_1. This essentially means that for each H_1 that is accepted there may very well be a logically plausible H_{p1} of greater or equal explanatory power or value. And parallel to this, for each H_0 that is accepted or, as it is traditionally stated, fails to be rejected (Sterling 1959) there may be an H_{p0} that explains the statistical equality of the comparison groups and so can extract the researcher from a limbo of suspended disbelief and explanatory sterility. (Blalock, 1961)

To be of value such an H_p must be a careful and precise statement that leads to an expected finding of no statistical differences. A speculative discussion is not at all sufficient since the demand character of this sort of situation requires that empiricism prevail once the data are collected. In effect, the statement of the sort of plausible rival hypothesis proposed here requires a second logical analysis as careful as that which led to the original H_1.(Hagen,1997)

In the era when behavioral scientists cherished the concept of the "crucial experiment" in their thoughts and actions (Kimble 1961), an acceptable H_p was always available in the form of accepting the deductions of Theory A when the deductions of Theory B were not confirmed or vice versa. Of course, the closer the tie to either theory the more grudgingly stated the acceptance and the longer the list of demurrers and qualifications. At the present time, the "crucial experiment" is considered as something of an historical artifact. While the positions assumed by the great majority of researchers are hardly atheoretical, a great deal of work falls into the category described by Marx (1963) as functional. In brief, this means that data and theory are complementary in their relationships with neither being attributed a significant degree of precedence or value over the other. Also the trend in contemporary theoretical approaches seems more in the direction of an organized interaction between induction and deduction. (Campbell, 1961; 1963) A recent review by Alba (2012) makes essentially the same points regarding the relationship between data and theory in behavioral research.

To return to the situation at hand, H_ps that can be stated as valid explanations for observed equalities would seem to be far more utilitarian for the individual researcher and for the development of the behavioral sciences than any sort of semantic surrender to the pain inflicted by unpredicted outcomes. To plainly express the problem, when faced with a null hypothesis that cannot be rejected, the researcher can suspend judgment but should never suspend inquiry. (Smart,1964; Griffen & Ross,1991; Schmidt,1992;1996; King, Rosopa & Minium,2011)

Examining a Test Case

The stimulus for the forgoing series of comments was an article concerned with the situational and personal determinants of yielding to some form of perceived pressure, i.e., conformity. (Back and Davis 1965)

Unlike the majority of respondent samples employed in studies of this variable or trait, the respondent population described in this study was a relatively homogenous group of student nurses sharing proximal living quarters and maintaining face to face contact over an extended period of time.

For the most part, the study was an admirable and well executed investigation which employed a series of measures, assessed a series of related personality traits and predicted the relationships expected to occur in three distinct situations. (Campbell and Fiske 1959)
(Campbell 1960)

However, in two of three instances the researchers' statistical analyses were not completely performed. When the necessary and appropriate further steps were carried out, it became quite evident that the conclusions relevant to both these instances were invalid and as a result, the positive (statistically significant) relationships stated by the research team were, in this case, false positives. (Rosenow & Rosenthal 1996; 2000)

In the first case, the investigators reasoned that status based on competence would better insulate an individual from yielding to group opinion in an ambiguous perceptual judgment task. There is some evidence available, for example, for example, Campbell (1961) which lends credence to such a prediction and so there was very little reason to doubt the validity of the investigators' pre-experimental logical analysis.

To test their hypothesis, ratings of status based competence (Best Nurse) and on popularity (Preferred Double Date) were obtained and these were correlated with measures of yielding to perceived group pressure in a laboratory task and with self report measures of conformity with peer and authority norms.

Table 1: Obtained Correlations between Status and Conformity: Direct Tests

		Conformity		
Status Criteria		Perceptual	Peer	Authority
Best Nurse		r=-.25a (75)*	□= -.23b (76)	r_{pb}=-.25a (75)
Double Date		r= -.11	□= -.07	r_{pb}= -.13

a= p<.05
b= p<.10
*=N
	Direct Tests	T= -1.354
		NS
			T= -909
			NS
				T= -.895
				NS

On the basis of the data summarized in the upper portion of this table the researchers concluded that as predicted:

"…the nominations for best nurse (competence) were significantly correlated with conformity in the perceptual task and to the authority sources and had the expected trend on the peer index of conformity. Nominations for favorite double date were not significantly related to any of the measures." (p. 237)

The above statement was, as can be seen from the table, based on the observation that in all three of the behavioral situations one correlation coefficient is larger than its comparison statistic. In two cases the predicted relationship between competence (best nurse) and conformity was statistically significant at the p < .05 level of confidence. In the remaining instance, conformity to peer norms, the coefficient was significant at the p < .10 level and so was interpreted as a trend in the predicted direction. With regard to status based on popularity, in no case did the obtained correlation coefficient even approach the magnitude needed for statistical significance at the conventionally acceptable levels, or even a level the researchers would assume worthy of citation as a trend.

Further Analysis of the Test Case Data

However, fact of the matter is, that on the basis of evidence presented in this table, the assumptions, judgments and conclusions cited in the proceeding quote cannot be made until the obtained comparison correlation coefficients are statistically, as opposed to simply visually, compared. The necessary comparison may be performed by means of a direct statistical test such as that provided by McNemar, (1963 p.140).

As the researchers noted in the body of their report there was a correlation between the measures of popularity and conformity. This correlation was .30 (N = 75, p < .01) and, therefore, the two measures were hardly independent. Using the correlations provided in the original table and the additional correlation coefficient of .30 cited above the author performed the needed tests.

The results of these direct tests are incorporated into Table 1 and appear directly below the appropriate columns. As can be seen by the obtained statistics, there were no bases for the researchers' conclusions regarding status based on competence as opposed to status based on their operational definition of popularity.

Essentially, the non-independence of the two measures indicated that not only were some of the girls perceived to be competent and others perceived to be popular, but in addiction quite a few were rated as both popular and competent; and presumably, to complete the distribution, some may have been perceived to be neither. A comparison on this basis, four groups if possible, rather than the non-independent dichotomy might have yielded a more definitive set of relationships.

As it stands, the conclusion offered by the researchers must be labeled as invalid, and as another case of false positive existing in the literature of behavioral science. This is not, however, a case in which the root of the problem is an inadequate interpretation of available conceptual evidence. The hypothesis, as previously discussed, was sound. Here the problem resulted from a combination of factors relevant to the respondent population and an inadequate statistical analysis of the obtained data. It provides an example of intra-study inconsistency that can be resolved only through a re-analysis of the data cited above or by a complete replication with a comparable population. (Smart,1964; Hagen,1997; Krantz,1999; Wainer,1999)

In the second instance the researchers hypothesized that the greater the tendency for socially desirable responding, the more likely an individual is to yield to pressure from an attractive source. By an attractive source the researchers meant a peer with whom the respondent was friendly. Since the respondent population was a homogenous and somewhat cohesive unit, the great majority (85%) of the responses to a sociometric type rating scale indicated that the respondents regarded the other three peers present in the laboratory situation as friends or good friends. Some of these friendship choices were reciprocated and others were not. However, there were no cases in which ratings that fell into the category designated as, "not a person with whom I am close", were reciprocated.

To test a hypothesis such as that stated above requires that there be some variation in levels of attractiveness within the respondent population. For this reason, the researchers divided the population into two units for the purpose of making the comparisons. The first of these they called "pair groups". These were dyads in which friendship designations were reciprocated. That is, if individual A named individual B as a close friend and B reciprocated this choice, the A←→ B dyad was designated a "pair group." On the other hand if C named D, but D did not reciprocate the choice or vice versa, the C→ D dyad was considered to be a "non pair" group.

Table 2: "Pair" Versus "Non-Pair" Conditions: Correlations and Direct Tests

Condition	ID-OD	Importance	Status	SD
"Pair"**	.39b	-.26d	-.35c	.31c
"Non-Pair"*	.39b	-.11	-.13	.08
Direct Test ϑ_{z1-z2}=	.00 NS	-.94 NS	-1.06 NS	1.21 NS
**N= 40	b= p <.01			
	c= p <.05			
	d = p <.10			
*N= 35				

Table 2 is a summary of the relationships reported between a number of variables and the "pair," "non-pair" conditions. On the basis of these comparisons the investigators concluded:

"The analysis with the "pair," "non-pair" subgroups brought a major surprise to light. While we had anticipated that the SD (socially desirable responding) – conformity relationship would be affected by the group composition variable, we had not anticipated that both importance (of the task) and status within the group would also be affected. In fact only other-directedness (ID-OD) continued to be related to conformity at the same level within both subgroups. All the other variables which were related in the "pair" condition were no longer even modestly related in the "non-pair" condition." (p. 234).

Once again, the researchers' statement is based upon misinterpretation of the obtained data. And again the error ensued because the analyses were not complete.(Schmidt, 1996; Sewell, (2012). The lower portion of Table 2 contains the results of the direct comparison test appropriate to the correlation coefficients presented in the researchers' original table. The test employed in this case is also taken from McNemar. (1963 p.140) The obtained test statistics (σ $z_1 - z_2$) do not support the researchers' assumption that the relationships found for the entire population changed when the population was dichotomized in the manner described above. In this case the null hypothesis cannot be rejected.

The discovery of this particular set of invalid positives is another clear case of intra-study inconsistency. As previously discussed, however, it is sometimes possible to resolve such inconsistency by examination of data related to the research project under discussion, but which falls under a subheading other than that given primacy, which in this case, was the concept of Conformity.

To be specific, it has been demonstrated that, in choice situations such as are described above that, if person A chooses person B, then A expects that the choice will be reciprocated. This means a quite predictable isomorphism of experience, in this case expectation and action would occur in the situation under consideration. (Asch, 1953) That is, if student nurse A perceived a reciprocal relationship to exist, she would behave (respond) accordingly. The characteristic equality of perceived as and responded to is certainly important in any consideration, description or explanation of human behavior (Campbell, 1963) and the explanatory potency of this isomorphic relationship is quite evident in the situation under consideration.(Griffin & Ross, 1991)

A number of available studies, for example, Taguiri (1958) and Taguiri, Bruner and Blake (1958), provide empirical evidence in support of the explanation offered above and the notable lack of statistically significant differences between "pair" and "non-pair" groups on any of the measures employed is both consistent and convergent with the findings reported in those studies.

Approaching this situation from the point of view and graphical representation of consistency and inconsistency among the elements of a research endeavor, it is evident that the researchers' original H_1 is inconsistent with the statistical evidence and, therefore, this hypothesis must be rejected. However, it is equally evident that this non-rejectable H_0 is equivalent to an H_p that can be accepted as a logically reasonable and statistically confirmed prediction. Substitution of this plausible alternative for the original, and now rejected, H_1 (See Figure 3) yields the desired mutually supportive and consistent relationship among the relevant elements of the study under discussion. (Wainer & Robinson2003)

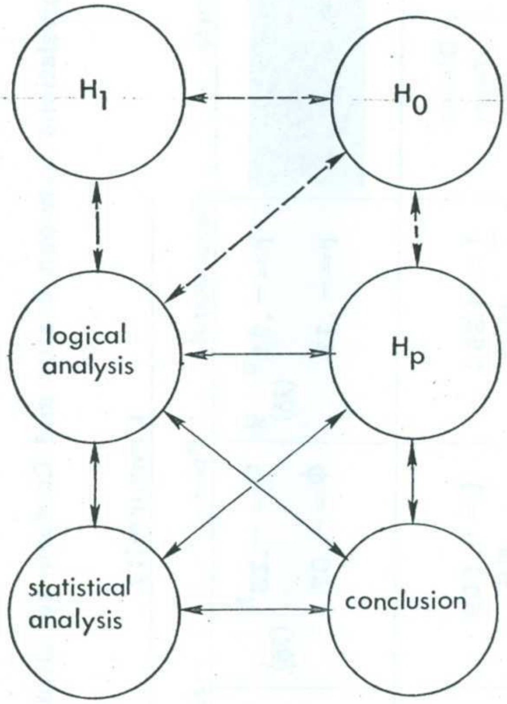


Figure 3. **Resolution Employing a Conceptual H_p.**

It seems that reciprocity of choice was important only for the investigators, who had the opportunity to examine the entire set of choices. Had the respondents been given the knowledge that some of their choices had not been reciprocated perhaps, as in other studies of this nature, the original hypothesis would have been confirmed. (Campbell, 1961) In the absence of this information the respondents acted on the basis of their individual expectations: and, since the great majority of choices fell within the categories designated as "close friend" or "friend," it follows that almost all of those present were perceived as and, therefore, responded to as attractive peers.

Concluding Comments

The resolution of inconsistency in this situation exemplifies Blalock's (1961) contention that most of the variables, which concern behavioral science, are linked in a complex causal network and this complexity must be reflected in the researcher's criterion of causation.

In an insightful article Clifford Geertz , (1961) cited Levi-Strauss' comment that: "scientific explanation does not consist, as we have been led to imagine, in the reduction of the complex to the simple. Rather, it consists in a substitution of a more intelligible complexity for one which is less so." Further, Geertz remarks that: "Whitehead once offered the natural sciences the maxim: 'Seek simplicity and distrust it'; to the behavioral sciences he might have well offered: "Seek complexity and order it'."

The approach discussed in this paper is in the spirit of Geertz' suggested maxim. The behavior of the respondent population is no less complex that it was, only somewhat more intelligible because the alternative hypothesis offered is somewhat more plausible.

The maxim is equally applicable to the task of integrating the ever -growing mass of behavioral data and interpretive prescriptions into some universally acceptable explanatory paradigms. If the behavioral sciences are to achieve such a paradigmatic state, it would seem that the logical place to begin would be at the level of the individual study.

References

Alba, J. W. (2012). In defense of bumbling. *Journal of Consumer Research. 38, 4,* 981-987

Asch, S.E. (1952). *Social Psychology.* New York: Prentice Hall.

Back, K.W., and Davis, K.E. (1965). Some personal and situational factors relevant to the consistency and prediction of conforming behavior. *Sociometry*, 23, 227-240.

Bernardo, J. & Smith, A.F. M. (2000). *Bayesian Theory*, New York NY: Wiley.

Blalock, H.M. Jr. (1961). Evaluating the relative importance of variables. *.American Sociological Review,* 26, 6,866-874.

Bradstreet, T.E. (1996). Teaching introductory statistic courses so that non-statisticians experience statistical reasoning. *The American Statistician*, 50, 1, 69-70.

Campbell, D.T. (1961). Conformity in psychology's theories of acquired behavioral dispositions In I.A. Berg and D.M. Bass (Eds.), *Conformity and Deviation* (pp.101-142). New York: Harper & Bros.

Campbell, D.T. (1959). Methodological suggestions from a comparative psychology of knowledge processes. *Inquiry,* 2, 152-182.

Campbell, D.T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist,* 15, 546-533.

Campbell, D.T. (1963). Social attitudes and other acquired behavioral dispositions. In S. Koch (Ed.), *Psychology: A Study of a Science.* Vol. 6, *Investigations of Man as a Socius*, (pp.94-172). New York: McGraw-Hill.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook of Research on Teaching* (pp.171-246). New York: Rand McNally & Co.

Chapanis, N. P., & Chapanis, A. (1964). A. Cognitive dissonance: Five years later. *Psychological Bulletin,* 61, 1-22.

Cohen, J., (1994). The earth is round (p< .05). *American Psychologist*, 45, 1304-1.

Dixon, P. (1998). Why scientists value p. values. *Psychonomic Bulletin & Review*, 5, 390-96.

Edwards,W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin,* 63, 6, 400-402.

Edwards, W., Lindman, H., & Savage, L.J.(1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*. 1, 379-390.

Geertz, C. (1966). The impact of the concept of culture on the concept of man. *Bulletin of* the *Atomic Scientists,* XXII, 4, 2-8.

Grant, D.A. (1963). Classical and operant conditioning. In A.W. Melton (Ed.), *Categories of Human Learning* (pp.3-29). New York: Academic Press.

Griffin, D.W. & Ross L. (1991). Subjective construal, social inference and human misunderstanding *Advances in Experimental Social Psychology*,24 319-59.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, 52, 15-24.

Haller,H. & Kraus, S. (2002). Misinterpretation of significance: A problem students share with their teachers. *American Psychology*, 49, 997-1003.

Harary, F., Norman, R.Z., and Cartwright, D. (1965). *Structural Models: An Introduction to the Theory of Directed Graphs*. New York: John Wiley and Sons,

Hubbard, R. & Armstrong, J. S. (2006). Why we don't really know what "statistical significance" means: A major educational failure. *Journal of Marketing Education* 23,2, 114-120.

Hubbard, R. & Armstrong, J.S. (1992). Are null results becoming an endangered species in marketing? *Marketing Letters*, 3, 127-136.

Iacobucci, D. (2005). On p-values. *Journal of Consumer Research*. 32,1, 6-1.

Kimble, G.A. (1961). *Hilgard and Marquis' Conditioning and Learning* (pp.226-234). New York: Appleton-Century-Crofts.

King, B.M., Rosopa, P. & Minium, E. (2011). *Statistical Reasoning in the Behavioral Sciences*. New York: Wiley.

Klayman, J. & Ha,Y.W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, 94, 211-228.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 44, 1372-81.

Krueger,J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, 56, 1, 16-26.

Liebman,M.D. & Cunningham, W.A. (2009). Type I and type II error concerns in research: rebalancing the scale. *Social Cognitive and Affective Neuroscience*, 4, 4,423-428.

Marx, M.H. (1963). The general nature of theory construction. In M.H. Marx (Ed.), *Theories in Contemporary Psychology* (pp.4-46). New York: The MacMillan Co.

McNemar, Q. (1963). *Psychological Statistics* (p.140). New York: John Wiley and Sons, Inc.,

Skinner, B.F.(1963). The flight from the laboratory. In M.H. Marx (Ed.), *Theories in Contemporary Psychology* (pp.323-338). New York: The MacMillan Co.

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

Novella, N. (2011) Statistical errors in mainstream journals. *Science-Based Medicine: Exploring Issues and Controversies in the Relationship between Science and Medicine.* http://www.sciencebasedmedicine.org/index.php/statistical-errors-in-mainstream-journals/,1-19.

Pyszczynski, T. & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis testing model. In L. Berkowitz (Ed.) *Advances in Experimental and Social Psychology*, 20, 294-340. San Diego CA: Academic Press.

Rosenow, R. L. & Rosenthal, R. (2000). Contrasts and correlations in effect size estimation. *Psychological Science*, 11, 446-53.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86,638-641.

Rosnow, R.L. & Rosenthal, R. (1996). Computing contrasts ,effect sizes, and counternulls on other peoples' published data: General procedures for research consumers. *Psychological Methods*, 1,4, 331-40.

Rowney,J. & Zenisek, T.J. (1980). Manuscript characteristics influencing reviewers' decisions. *Canadian Psychology*, 21, 17-21.

Schmidt, F, (1992). What do data really mean? Research findings, meta- analysis and cumulative knowledge in psychology. *American Psychologist*, 47,1173-81.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1,2, 115-29.

Sewell, M. (2012). Statistical inference (and what is wrong with classical statistics). *Working Paper, Cambridge University UK.Stats.org.UK*. 1-40

Smart, R. (1964). The importance of negative results in psychological research. *Canadian Psychologist*, 5, 225-232

Sterling,T.D. (1959). Publication decisions and their possible effects on influences drawn from tests of significance-or vice versa. *Journal of the American Statistical Association*, 54, 30-34.

Sterling, T.C. (1960). "What's so peculiar about accepting the null hypothesis? *Psychological Reports,* 7, 363-364.

Taguiri, R.(1958). Social preference and its perception. In R. Taguiri & L. Petrullo (Eds.), *Person Perception and Interpersonal Behavior* (pp.316-336), Stanford: Stanford University Press.

Taguiri,R., Bruner, J.S., and Blake, R.R. (1958). On the relations between feelings and perception of feelings among members of small groups. In E. E. Macoby, T.M. Newcomb and E.L. Hartley (Eds), *Readings in Social Psychology* 3rd.ed. (pp.110-116). New York: Holt, Rhinehart &Winston.

Thagard, P. (2000). *Coherence in Thought and Action.* Cambridge MA: MIT Press.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Baye's theorem. *Psychological Review*, 110, 526-535.

Wainer, H. (1999). One cheer for the null hypothesis significance testing. *Psychological Methods*. 4,2, 212-13.

Wainer, H.& Robinson, D.H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, 32, 22-30.

Wilkinson, L. & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54,8, 594-604.

Winkielman, P., Huber, D.E., Kavanaugh, L. & Schwarz, N. (2012) Fluency of consistency : When thoughts flow nicely and flow smoothly. In B. Gawronski & F. Strack (Eds) *Cognitive Consistency: A Fundamental Principle in Social Cognition.* (pp.89-111). New York: Guilford Press.