A

A Priori Monte Carlo Simulation

An a priori Monte Carlo simulation is a special case of a Monte Carlo simulation that is used in the design of a research study, generally when analytic methods do not exist for the goal of interest for the specified model or are not convenient. A Monte Carlo simulation is generally used to evaluate empirical properties of some quantitative method by generating random data from a population with known properties, fitting a particular model to the generated data, collecting relevant information of interest, and replicating the entire procedure a large number of times (e.g., 10,000). In an a priori Monte Carlo simulation study, interest is generally in the effect of design factors on the inferences that can be made rather than a general attempt at describing the empirical properties of some quantitative method. Three common categories of design factors used in a priori Monte Carlo simulations are sample size, model misspecification, and unsatisfactory data conditions. As with Monte Carlo methods in general, the computational tediousness of a priori Monte Carlo simulation methods essentially requires one or more computers because of the large number of replications and thus the heavy computational load. Computational loads can be very great when the a priori Monte Carlo simulation is implemented for methods that are themselves computationally tedious (e.g., bootstrap, multilevel models, and Markov chain Monte Carlo methods).

For an example of when an a priori Monte Carlo simulation study would be useful, Ken Kelley and Scott Maxwell have discussed sample size planning for multiple regression when interest is in sufficiently narrow confidence intervals for standardized regression coefficients (i.e., the accuracy-in-parameter-estimation approach to sample size planning). Confidence intervals based on noncentral *t* distributions should be used for standardized regression coefficients. Currently, there is no analytic way to plan for the sample size so that the computed interval will be no larger than desired some specified percent of the time. However, Kelley and Maxwell suggested an a priori Monte Carlo simulation procedure when random data from the situation of interest are generated and a systematic search (e.g., a sequence) of different sample sizes is used until the minimum sample size is found at which the specified goal is satisfied.

As another example of when an application of a Monte Carlo simulation study would be useful, Linda Muthén and Bengt Muthén have discussed a general approach to planning appropriate sample size in a confirmatory factor analysis and structural equation modeling context by using an a priori Monte Carlo simulation study. In addition to models in which all the assumptions are satisfied, Muthén and Muthén suggested sample size planning using a priori Monte Carlo simulation methods when data are missing and when data are not normal—two conditions most sample size planning methods do not address.

Even when analytic methods do exist for designing studies, sensitivity analyses can be implemented within an a priori Monte Carlo simulation framework. Sensitivity analyses in an a priori Monte Carlo simulation study allow the effect of misspecified parameters, misspecified models, and/or the validity of the assumptions on which the method is based to be evaluated. The generality the a priori Monte Carlo simulation studies is its biggest advantage. As Maxwell, Kelley, and Joseph Rausch have stated, "Sample size can be planned for any research goal, on any statistical technique, in any

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.
Do not copy, post, or distribute

situation with an a priori Monte Carlo simulation study" (2008, p. 553).

Ken Kelley

See also Accuracy in Parameter Estimation; Monte Carlo Simulation; Power Analysis; Sample Size Planning

Further Readings

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*(1), 1. doi:10.1037/met0000195.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. doi:10.1111/2041-210X.12504.

Kelley, K., & Maxwell, S. E. (2008). Power and accuracy for omnibus and targeted effects: Issues of sample size planning with applications to multiple regression. In P. Alasuuta, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 166–192). Thousand Oaks, CA: Sage.

Muthén, L., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *4*, 599–620. doi:10.1207/S15328007SEM0904_8.

Abstract

An abstract is a summary of a research or a review article and includes critical information, including a complete reference to the work, its purpose, methods used, conclusions reached, and implications. For example, here is one such abstract from the *Journal of Black Psychology* authored by Timo Wandert from the University of Mainz, published in 2009 and titled "Black German Identities: Validating the Multidimensional Inventory of Black Identity."

All the previously-mentioned elements are included in this abstract: the purpose, a brief review of important ideas to put the purpose into a context, the methods, the results, and the implications of the results.

> This study examines the reliability and validity of a German version of the Multidimensional Inventory of Black Identity (MIBI) in a sample of 170 Black Germans. The internal consistencies of all subscales are at least moderate. The factorial structure of the MIBI, as assessed by principal component analysis, corresponds to a high degree to the supposed underlying dimensional structure. Construct validity was examined by analyzing (a) the intercorrelations of the MIBI subscales and (b) the correlations of the subscales with external variables. Predictive validity was assessed by analyzing the correlations of three MIBI subscales with the level of intra-racial contact. All but one prediction concerning the correlations of the subscales could be confirmed, suggesting high validity. No statistically significant negative association was observed between the Black nationalist and assimilationist ideology subscales. This result is discussed as a consequence of the specific social context Black Germans live in and is not considered to lower the MIBI's validity. Observed differences in mean scores to earlier studies of African American racial identity are also discussed.

Abstracts serve several purposes. First, they provide a quick summary of the complete publication that is easily accessible in the print form of the article or through electronic means. Second, they become the target for search tools and often provide an initial screening when a researcher is doing a literature review. It is for this reason that article titles and abstracts contain key words that one would look for when searching for such information. Third, they become the content of reviews or collections of abstracts such as PsycINFO, published by the American Psychological Association (APA). Finally, abstracts sometimes are used as stand-ins for the actual papers when there are time or space limitations, such as at professional meetings. In this instance, abstracts are usually presented as posters in presentation sessions.

Most scholarly publications have very clear guidelines as to how abstracts are to be created, prepared, and used. For example, the APA, in the *Publication Manual of the American Psychological Association*, provides information regarding the elements of a good abstract and suggestions for creating one. While guidelines for abstracts of scholarly publications (such as print and electronic journals) tend to differ in the specifics, the following four guidelines apply generally:

1. The abstract should be short. For example, APA limits abstracts to 250 words, and MEDLINE limits them to no more than 400 words. The abstract should be submitted as a separate page.
2. The abstract should appear as one unindented paragraph.
3. The abstract should begin with an introduction and then move to a very brief summary of the method, results, and discussion.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

4. After the abstract, five related keywords should be listed. These keywords help make electronic searches efficient and successful.

With the advent of electronic means of creating and sharing abstracts, visual and graphical abstracts have become popular, especially in disciplines in which they contribute to greater understanding by the reader.

Neil J. Salkind

See also American Psychological Association Style; Ethics in the Research Process; Literature Review

Further Readings

American Psychological Association. (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Fletcher, R. H. (1988). Writing an abstract. *Journal of General Internal Medicine*, *3*(6), 607–609.

Luhn, H. P. (1999). The automatic creation of literature abstracts. In I. Mani & M. T. Maybury (Eds.), *Advances in automatic text summarization* (pp. 15–21). Cambridge: MIT Press.

Accuracy in Parameter Estimation

Accuracy in parameter estimation (AIPE) is an approach to sample size planning concerned with obtaining narrow confidence intervals. The standard AIPE approach yields the necessary sample size so that the expected width of a confidence interval will be sufficiently narrow. Because confidence interval width is a random variable based on data, the actual confidence interval will almost certainly be wider or narrower than the expected confidence interval width. A modified AIPE approach allows sample size to be planned so that there will be some desired degree of assurance that the observed confidence interval will be sufficiently narrow.

AIPE and modified AIPE are "fixed *N*" procedures, in that one needs to specify parameters in order to find the sample size, which is then a fixed value. A new version of AIPE, termed sequential AIPE, is not a fixed *N* procedure but rather is sequential, where sampling of cases continues until a stopping rule is satisfied. Whereas the standard AIPE approach addresses questions such as "what size sample is necessary so that the expected width of the 95% confidence interval width will be no larger than ω," where ω is the desired confidence interval width, the modified AIPE approach addresses questions such as "what size sample is necessary so that there is γ100% assurance that the 95% confidence interval width will be no larger than ω," where γ is the desired value of the assurance parameter. Furthermore, sequential AIPE does not ask "What size sample is necessary?" but rather "Is the accuracy of the estimate sufficient for sampling to stop?" This entry further discusses the importance of confidence interval width, the origins and goals of the AIPE approach, and the subsequent development of the sequential AIPE approach.

Confidence interval width is a way to operationalize the accuracy of the parameter estimate, holding everything else constant. Provided appropriate assumptions are met, a confidence interval consists of a set of plausible parameter values obtained from applying the confidence interval procedure to data, where the procedure yields intervals such that $(1 - \alpha)$ 100% will correctly bracket the population parameter of interest, where $1 - \alpha$ is the desired confidence interval coverage. Holding everything else constant, as the width of the confidence interval decreases, the range of plausible parameter values is narrowed and thus more values can be excluded as implausible values for the parameter. In general, whenever a parameter value is of interest, not only should the point estimate itself be reported, but so too should the corresponding confidence interval for the parameter, as it is known that a point estimate almost certainly differs from the population value and does not give an indication of the degree of uncertainty with which the parameter has been estimated. Wide confidence intervals, which illustrate the uncertainty with which the parameter has been estimated, are generally undesirable. Because the direction, magnitude, and accuracy of an effect can be simultaneously evaluated with confidence intervals, planning a research study in an effort to obtain narrow confidence intervals is considered an ideal way to improve research findings and increase the cumulative knowledge of a discipline.

Operationalizing accuracy as the observed confidence interval width is not new. In fact, Jerzy Neyman (1937) used the confidence interval width as a measure of accuracy in his seminal work on the theory of confidence intervals: "the accuracy of estimation corresponding to a fixed value of $1 - \alpha$ may be measured by the length of the confidence interval" (p. 358; notation changed to reflect current usage). Statistically, accuracy is defined as the square root of the mean square error, which is a function of precision and bias. When the bias is zero, accuracy and precision are equivalent concepts.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The accuracy in parameter estimation approach is so named because its goal is to improve the overall accuracy of estimates and not just the precision or bias alone. Precision can often be improved at the expense of bias, which may or may not improve the accuracy. Thus, to not obtain estimates that are sufficiently precise but possibly more biased, the (sequential) AIPE approach sets its goal as obtaining sufficiently accurate parameter estimates as operationalized by the width of the corresponding $(1 - \alpha)$ 100% confidence interval.

Research studies are often undertaken with the goal of basing important decisions on the results. However, when an effect has a corresponding confidence interval that is wide, decisions based on such effect sizes need to be used with caution. A point estimate can be impressive according to some standard, but for the confidence limits to illustrate that the estimate is not very accurate. For example, a commonly used set of guidelines for the standardized mean difference in the behavioral, educational, and social sciences is that population standardized effect sizes of 0.2, 0.5, and 0.8 are regarded as "small," "medium," and "large" effects, respectively (Cohen, 1969, 1988). Suppose that the population standardized mean difference is thought to be medium (i.e., 0.50) based on an existing theory and a review of the relevant literature. Further suppose that a researcher planned a sample size so that there would be a statistical power of .80 when the Type I error rate is set to .05, which yields a necessary sample size of 64 participants per group (128 total). In such a situation, supposing that the observed standardized mean difference was in fact exactly 0.50, the 95% confidence interval has a lower and upper limit of .147 and .851, respectively. Thus, the lower confidence limit is smaller than "small" and the upper confidence limit is larger than "large." Although there was enough statistical power (recall sample size was planned so that power = .80 and indeed the null hypothesis of no group mean difference rejected, $p = .005$), in this case sample size was not sufficient from an accuracy perspective, as illustrated by the wide confidence interval.

Historically, confidence intervals were not often reported in applied research in the behavioral, educational, and social sciences, as well as in many other domains. Jacob Cohen (1994) once suggested researchers failed to report confidence intervals because their widths were "embarrassingly large" (p. 1002). In an effort to plan sample size so as to not obtain confidence intervals that are embarrassingly large, and in fact to plan sample size so that confidence intervals are sufficiently narrow, the (sequential) AIPE approach should be considered. The argument for planning sample size from an AIPE perspective or using sequential AIPE, in which sample size is not literally planned, but sampling continues until a stopping rule is satisfied, is due to the desire to report point estimates and confidence intervals instead of or in addition to the results of null hypothesis significance tests. This paradigmatic shift has led to (sequential) AIPE approaches to sample size planning becoming more useful than was previously the case, given the emphasis now placed on confidence intervals instead of focusing solely on the results of null hypothesis significance tests.

Whereas the power analytic approach to sample size planning has as its goal rejecting a false null hypothesis with some specified probability, the (sequential) AIPE approach is not concerned with whether or not some specified null value can be rejected (i.e., is the null value outside of the confidence interval limits), making it a fundamentally different approach than the power analytic approach. Not surprisingly, the (sequential) AIPE and power analytic approaches can suggest very different values for sample size, depending on the particular goals (e.g., desired width or desired power) specified. The (sequential) AIPE approach to sample size planning is able to simultaneously consider the direction of an effect (which is what the null hypothesis significance test provides), its magnitude ("best" and "worst" case scenarios based on the values of the confidence limits), and the accuracy with which the population parameter was estimated (via the width of the confidence interval).

Ken Kelley and Scott Maxwell first used the term "accuracy in parameter estimation" (and the acronym AIPE) as a general framework in a 2003 article where they argued for its widespread use in lieu of or in addition to the power analytic approach. However, the general idea of AIPE has appeared in the literature sporadically since at least the 1960s. In a 2000 article, James Algina and Stephen Olejnik discussed a similar approach with the goal of an estimate sufficiently close to its corresponding population value, while in 2003, Michael Jiroutek and colleagues proposed a method to simultaneously have a sufficient degree of power and confidence interval narrowness. As of 2020, the most extensive program for planning sample size from the (sequential) AIPE perspective is R (R Core Developmental Team) using the MBESS package.

Ken Kelley

See also Confidence Intervals; Effect Size, Measures of; MBESS; Power Analysis; Sample Size Planning

Further Readings

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

coefficient. *Multivariate Behavioral Research*, *35*, 119–136. doi:10.1207/S15327906MBR3501_5.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:10.1037/0003-066X.49.12.997.

Guenther, W. C. (1965). *Concepts of statistical inference*. New York: McGraw-Hill.

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics*, *59*, 580–590. doi:10.1111/1541-0420.00068.

Kelley, K. (2006–2020). Methods for the Behavioral, Educational, and Social Sciences (MBESS): An R Package [computer software and manual]. Accessible from http://cran.r-project.org/web/packages/MBESS/index.html

Kelley, K. (2007a). Methods for the behavioral, educational, and social science: An R package behavior research methods. *Behavior Research Methods*, *39*(4), 979–984. doi:10.3758/BF03192993.

Kelley, K. (2007b). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1–24.

Kelley, K. (2007c). Sample size planning for the coefficient of variation: Accuracy in parameter estimation via narrow confidence intervals. *Behavior Research Methods*, *39*(4), 755–766. doi:10.3758/BF03192966.

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2019). Sequential accuracy in parameter estimation for population correlation coefficients. *Psychological Methods*, *24*, 492–515.

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*, 226–243. doi:10.1037/met0000127.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305–321. doi:10.1037/1082-989X.8.3.305.

Kelley, K., & Maxwell, S. E. (2008). Power and accuracy for omnibus and targeted effects: Issues of sample size planning with applications to multiple regression. In P. Alasuutari, J. Brannen, & L. Bickman (Eds.), *Handbook of social research methods* (pp. 166–192). CA: Sage.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385. doi:10.1037/1082-989X.11.4.363.

Mace, A. E. (1964). *Sample size determination*. New York, NY: Reinhold.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. doi:10.1146/annurev.psych.59.103006.093735.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A*, 236, 333–380. doi:10.1098/rsta.1937.0005.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: Dorsey Press.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182. doi:10.1037/1082-989X.9.2.164.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25–32. doi:10.3102/0013189X031003025.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133. doi:10.1080/00031305.2016.1154108.

Wilkinson, L., & the American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology: Guidelines and explanations. *American Psychologist*, *54*, 594–604. doi:10.1037/0003-066X.54.8.594.

Action Research

Action research differs from conventional research methods in three fundamental ways. First, its primary goal is social change. Second, members of the study sample accept responsibility for helping resolve issues that are the focus of the inquiry. Third, relationships between researcher and study participants are more complex and less hierarchical. Most often, action research is viewed as a process of linking theory and practice in which scholar-practitioners explore a social situation by posing a question, collecting data, and testing a hypothesis through several cycles of action. The most common purpose of action research is to guide practitioners as they seek to uncover answers to complex problems in disciplines such as education, health sciences, sociology, or anthropology. Action research is typically underpinned by ideals of social justice and an ethical commitment to improve the quality of life in particular social settings. Accordingly, the goals of action research are as unique to each study as participants' contexts; both determine the type of data-gathering methods that will be used. Because action research can embrace natural *and* social science methods of scholarship, its use is not limited to either positivist or heuristic approaches. It is, as John Dewey

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

pointed out, an *attitude* of inquiry rather than a single research methodology.

This entry presents a brief history of action research, describes several critical elements of action research, and offers cases for and against the use of action research.

Historical Development

Although not officially credited with authoring the term *action research,* Dewey proposed five phases of inquiry that parallel several of the most commonly used action research processes, including curiosity, intellectualization, hypothesizing, reasoning, and testing hypotheses through action. This recursive process in scientific investigation is essential to most contemporary action research models. The work of Kurt Lewin is often considered seminal in establishing the credibility of action research. In anthropology, William Foote Whyte conducted early inquiry using an action research process similar to Lewin's. In health sciences, Reginald Revans renamed the process *action learning* while observing a process of social action among nurses and coal miners in the United Kingdom. In the area of emancipatory education, Paulo Freire is acknowledged as one of the first to undertake action research characterized by participant engagement in sociopolitical activities.

The hub of the action research movement shifted from North America to the United Kingdom in the late 1960s. Lawrence Stenhouse was instrumental in revitalizing its use among health care practitioners. John Elliott championed a form of educational action research in which the researcher-as-participant takes increased responsibility for individual and collective changes in teaching practice and school improvement. Subsequently, the 1980s were witness to a surge of action research activity centered in Australia. Wilfred Carr and Stephen Kemmis authored *Becoming Critical,* and Kemmis and Robin McTaggart's *The Action Research Planner* informed much educational inquiry. Carl Glickman is often credited with a renewed North American interest in action research in the early 1990s. He advocated action research as a way to examine and implement principles of democratic governance; this interest coincided with an increasing North American appetite for postmodern methodologies such as personal inquiry and biographical narrative.

Characteristics

Reflection

Focused reflection is a key element of most action research models. One activity essential to reflection is referred to as *metacognition,* or thinking about thinking. Researchers ruminate on the research process even as they are performing the very tasks that have generated the problem and, during their work, derive solutions from an examination of data. Another aspect of reflection is circumspection, or learning-in-practice. Action research practitioners typically proceed through various types of reflection, including those that focus on technical proficiencies, theoretical assumptions, or moral or ethical issues. These stages are also described as learning *for* practice, learning *in* practice, and learning *from* practice. Learning for practice involves the inquiry-based activities of readiness, awareness, and training engaged in collaboratively by the researcher and participants. Learning in practice includes planning and implementing intervention strategies and gathering and making sense of relevant evidence. Learning from practice includes culminating activities and planning future research. Reflection is integral to the habits of thinking inherent in scientific explorations that trigger explicit action for change.

Iterancy

Most action research is cyclical and continuous. The spiraling activities of planning, acting, observing, and reflecting recur during an action research study. Iterancy, as a unique and critical characteristic, can be attributed to Lewin's early conceptualization of action research as involving hypothesizing, planning, fact-finding (reconnaissance), execution, and analysis (see Figure 1).

These iterations comprise internal and external repetition referred to as *learning loops*, during which participants engage in successive cycles of collecting and making sense of data until agreement is reached on appropriate action. The result is some form of human activity or tangible document that is immediately applicable in participants' daily lives and instrumental in informing subsequent cycles of inquiry.

Collaboration

Action research methods have evolved to include collaborative and negotiatory activities among various participants in the inquiry. Divisions between the roles of researchers and participants are frequently permeable; researchers are often defined as both full participants and external experts who engage in ongoing consultation with participants. Criteria for collaboration include evident structures for sharing power and voice; opportunities to construct common language and understanding among partners; an explicit code of ethics and principles; agreement regarding shared ownership of data; provisions for sustainable community involvement and action; and consideration of generative methods to assess the process's effectiveness.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The collaborative partnerships characteristic of action research serves several purposes. The first is to integrate into the research several tenets of evidence-based responsibility rather than documentation-based accountability. Research undertaken for purposes of accountability and institutional justification often enforces an external locus of control. Conversely, responsibility-based research is characterized by job-embedded, sustained opportunities for participants' involvement in change; an emphasis on the demonstration of professional learning; and frequent, authentic recognition of practitioner growth.

Role of the Researcher

Action researchers may adopt a variety of roles to guide the extent and nature of their relationships with participants. In a *complete participant* role, the identity of the researcher is neither concealed nor disguised. The researchers' and participants' goals are synonymous; the importance of participants' voice heightens the necessity that issues of anonymity and confidentiality are the subject of ongoing negotiation. The *participant observer* role encourages the action researcher to negotiate levels of accessibility and membership in the participant group, a process that can limit interpretation of events and perceptions. However, results derived from this type of involvement may be granted a greater degree of authenticity if participants are provided the opportunity to review and revise perceptions through a member check of observations and anecdotal data. A third possible role in action research is the *observer participant*, in which the researcher does not attempt to experience the activities and events under observation but negotiates permission to make thorough and detailed notes in a fairly detached manner. A fourth role, less common to action research, is that of the *complete observer*, in which the researcher adopts passive involvement in activities or events, and a deliberate—often physical—barrier is placed between the researcher and the participant in order to minimize contamination. These categories only hint at the complexity of roles in action research. The learning by the participants and by the researcher is rarely mutually exclusive; moreover, in practice, action researchers are most often full participants.

Intertwined purpose and the permeability of roles between the researcher and the participant are frequently elements of action research studies with agendas of emancipation and social justice. Although this process is typically one in which the external researcher is expected and required to provide some degree of expertise or advice, participants—sometimes referred to as internal researchers—are encouraged to make sense of, and apply, a wide variety of professional learning that can be translated into ethical action. Studies such as these contribute to understanding the human condition, incorporate lived experience, give public voice to experience, and expand perspectives of participant and researcher alike.

A Case for and Against Action Research

Ontological and epistemological divisions between qualitative and quantitative approaches to research abound, particularly in debates about the credibility of action research studies. On one hand, quantitative

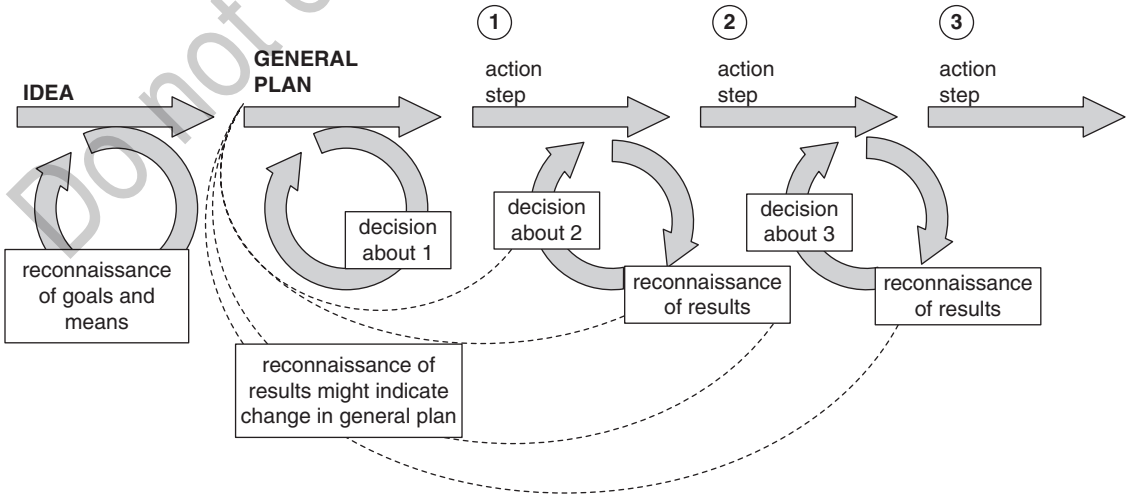

Figure 1 Lewin's Model of Action Research

Source: Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, *2,* 34–46.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

research is criticized for drawing conclusions that are often pragmatically irrelevant; employing methods that are overly mechanistic, impersonal, and socially insensitive; compartmentalizing, and thereby minimizing, through hypothetico-deductive schemes, the complex, multidimensional nature of human experiences; encouraging research as an isolationist and detached activity void of, and impervious to, interdependence and collaboration; and forwarding claims of objectivity that are simply not fulfilled.

On the other hand, qualitative aspects of action research are seen as quintessentially unreliable forms of inquiry because the number of uncontrolled contextual variables offers little certainty of causation. Interpretive methodologies such as narration and autobiography can yield data that are unverifiable and potentially deceptive. Certain forms of researcher involvement have been noted for their potential to unduly influence data, while some critiques contend that Hawthorne or halo effects—rather than authentic social reality—are responsible for the findings of naturalist studies.

Increased participation in action research in the latter part of the 20th century paralleled a growing demand for more pragmatic research in all fields of social science. For some humanities practitioners, traditional research was becoming irrelevant, and their social concerns and challenges were not being adequately addressed in the findings of positivist studies. They found in action research a method that allowed them to move further into other research paradigms or to commit to research that was clearly bimethodological. Increased opportunities in social policy development meant that practitioners could play a more important role in conducting the type of research that would lead to clearer understanding of social science phenomena. Further sociopolitical impetus for increased use of action research derived from the politicizing effects of the accountability movement and from an increasing solidarity in humanities professions in response to growing public scrutiny.

The emergence of action research illustrates a shift in focus from the dominance of statistical tests of hypotheses within positivist paradigms toward empirical observations, case studies, and critical interpretive accounts. Research protocols of this type are supported by several contentions, including the following:

- The complexity of social interactions makes other research approaches problematic.
- Theories derived from positivist educational research have been generally inadequate in explaining social interactions and cultural phenomena.
- Increased public examination of public institutions such as schools, hospitals, and corporate organizations requires insights of a type that other forms of research have not provided.
- Action research can provide a bridge across the perceived gap in understanding between practitioners and theorists.

Reliability and Validity

The term *bias* is a historically unfriendly pejorative frequently directed at action research. As much as possible, the absence of bias constitutes conditions in which reliability and validity can increase. Most vulnerable to charges of bias are action research inquiries with a low saturation point (i.e., a small *N*), limited interrater reliability, and unclear data triangulation. Positivist studies make attempts to control external variables that may bias data; interpretivist studies contend that it is erroneous to assume that it is possible to do *any* research—particularly human science research—that is uncontaminated by personal and political sympathies and that bias can occur in the laboratory as well as in the classroom. While value-free inquiry may not exist in any research, the critical issue may not be one of credibility but, rather, one of recognizing divergent ways of answering questions associated with purpose and intent. Action research can meet determinants of reliability and validity if primary contextual variables remain consistent and if researchers are as disciplined as possible in gathering, analyzing, and interpreting the evidence of their study; in using triangulation strategies; and in the purposeful use of participation validation. Ultimately, action researchers must reflect rigorously and consistently on the places and ways that values insert themselves into studies and on how researcher tensions and contradictions can be consistently and systematically examined.

Generalizability

Is any claim of replication possible in studies involving human researchers and participants? Perhaps even more relevant to the premises and intentions that underlie action research is the question, Is this *desirable* in contributing to our understanding of the social world? Most action researchers are less concerned with the traditional goal of generalizability than with capturing the richness of unique human experience and meaning. Capturing this richness is often accomplished by reframing determinants of generalization and

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

avoiding randomly selected examples of human experience as the basis for conclusions or extrapolations. Each instance of social interaction, if thickly described, represents a slice of the social world in the classroom, the corporate office, the medical clinic, or the community center. A certain level of generalizability of action research results may be possible in the following circumstances:

- Participants in the research recognize and confirm the accuracy of their contributions.
- Triangulation of data collection has been thoroughly attended to.
- Interrater techniques are employed prior to drawing research conclusions.
- Observation is as persistent, consistent, and longitudinal as possible.
- Dependability, as measured by an auditor, substitutes for the notion of reliability.
- Confirmability replaces the criterion of objectivity.

Ethical Considerations

One profound moral issue that action researchers, like other scientists, cannot evade is the use they make of knowledge that has been generated during inquiry. For this fundamental ethical reason, the premises of any study—but particularly those of action research—must be transparent. Moreover, they must attend to a wider range of questions regarding intent and purpose than simply those of validity and reliability. These questions might include considerations such as the following:

- Why was this topic chosen?
- How and by whom was the research funded?
- To what extent does the topic dictate or align with methodology?
- Are issues of access and ethics clear?
- From what foundations are the definitions of science and truth derived?
- How are issues of representation, validity, bias, and reliability discussed?
- What is the role of the research? In what ways does this align with the purpose of the study?
- In what ways will this study contribute to knowledge and understanding?

A defensible understanding of what constitutes knowledge and of the accuracy with which it is portrayed must be able to withstand reasonable scrutiny from different perspectives. Given the complexities of human nature, complete understanding is unlikely to result from the use of a single research methodology. Ethical action researchers will make public the stance and lenses they choose for studying a particular event. With transparent intent, it is possible to honor the unique, but not inseparable, domains inhabited by social and natural, thereby accommodating appreciation for the value of multiple perspectives of the human experience.

Making Judgment on Action Research

Action research is a relatively new addition to the repertoire of scientific methodologies, but its application and impact are expanding. Increasingly sophisticated models of action research continue to evolve as researchers strive to more effectively capture and describe the complexity and diversity of social phenomena.

Perhaps as important as categorizing action research into methodological compartments is the necessity for the researcher to bring to the study full self-awareness and disclosure of the personal and political voices that will come to bear on results and action. The action researcher must reflect on and make transparent, prior to the study, the paradoxes and problematics that will guide the inquiry and, ultimately, must do everything that is fair and reasonable to ensure that action research meets requirements of rigorous scientific study. Once research purpose and researcher intent are explicit, several alternative criteria can be used to ensure that action research is sound research. These criteria include the following types, as noted by David Scott and Robin Usher:

> *Aparadigmatic* criteria, which judge natural and social sciences by the same strategies of data collection and which apply the same determinants of reliability and validity
>
> *Diparadigmatic* criteria, which judge social phenomena research in a manner that is dichotomous to natural science events and which apply determinants of reliability and validity that are exclusive to social science
>
> *Multiparadigmatic* criteria, which judge research of the social world through a wide variety of strategies, each of which employs unique postmodern determinants of social science
>
> *Uniparadigmatic* criteria, which judge the natural and social world in ways that are redefined and reconceptualized to align more appropriately with a growing quantity and complexity of knowledge

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

In the final analysis, action research is favored by its proponents because it

- honors the knowledge and skills of all participants
- allows participants to be the authors of their own incremental progress
- encourages participants to learn strategies of problem solving
- promotes a culture of collaboration
- enables change to occur in context
- enables change to occur in a timely manner
- is less hierarchical and emphasizes collaboration
- accounts for rather than controls phenomena

Action research is more than reflective practice. It is a complex process that may include either qualitative or quantitative methodologies, one that has researcher and participant learning at its center. Although, in practice, action research may not often result in high levels of critical analysis, it succeeds most frequently in providing participants with intellectual experiences that are illuminative rather than prescriptive and empowering rather than coercive.

Pamela Adams

See also Evidence-Based Decision Making; External Validity; Generalizability Theory; Mixed Methods Design; Naturalistic Inquiry

Further Readings

Berg, B. (2001). *Qualitative research methods for the social sciences*. Toronto, Ontario, Canada: Allyn and Bacon.

Carr, W., & Kemmis, S. (1986). *Becoming critical: Education, knowledge and action research*. Philadelphia: Farmer.

Dewey, J. (1910). *How we think*. Boston: D. C. Heath.

Freire, P. (1968). *Pedagogy of the oppressed*. New York: Herder & Herder.

Habermas, J. (1971). *Knowledge and human interests*. Boston: Beacon.

Holly, M., Arhar, J., & Kasten, W. (2005). *Action research for teachers: Traveling the yellow brick road*. Upper Saddle River, NJ: Pearson/Merrill/Prentice Hall.

Kemmis, S., & McTaggart, R. (1988). *The action research planner*. Geelong, Victoria, Australia: Deakin University.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, *2*, 34–46.

Revans, R. (1982). *The origins and growth of action learning*. Bromley, UK: Chartwell-Bratt.

Sagor, R. (1992). *How to conduct collaborative action research*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schön, D. (1983). *The reflective practitioner*. New York: Basic Books.

Adaptive Designs in Clinical Trials

Adaptive designs in clinical trials are derivations of randomized clinical trials that modify the randomization ratio during the study. In adjusting the randomization ratio during the study, adaptive designs in clinical trials maximize the likelihood of participants experiencing positive outcomes and minimize the likelihood of participants experiencing poor outcomes. Adaptive designs in clinical trials follow a three-step procedure. In Step 1, participants are randomly allocated to one of the interventions with the probability of allocation into each intervention being specified by the randomization ratio. In Step 2, participants' outcome and/or covariate data are collected. In Step 3, the randomization ratio is updated based on the collected outcome and/or covariate data. Steps 1 through 3 are then repeated until the desired number of participants has been randomized.

Randomization Ratios

A randomization ratio is the ratio of the probabilities of being randomly allocated to each of the intervention arms. A fixed randomization ratio is held constant throughout a study, but an adaptive randomization ratio can change during a study.

Data for Adjusting the Randomization Ratio

In adaptive designs in clinical trials, the randomization ratio is adjusted using outcome and/or covariate data. Outcome data refer to data reflecting the effectiveness of the intervention, typically collected at the end of the study. For example, a study intended to improve reading comprehension would likely use some measure of reading comprehension as the primary outcome, and data from this reading comprehension measure would be used to adjust the randomization ratio in an outcome-adaptive design.

Covariate data refer to information pertaining to the participants in the study (e.g., sex, race, age). Covariate-adaptive designs adjust the randomization ratio based on the covariate data with the goal of minimizing between-group differences. For example, a covariate-adaptive design might adjust the randomization ratio to produce similar male-to-female ratios in the intervention arms. While covariate data could solely be used to adjust the randomization ratio, a study using a fixed randomization ratio

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

would theoretically balance covariate data across intervention arms given a sufficiently large sample size. Thus, it appears most likely that studies using an adaptive design will use either outcome data to adjust the randomization ratio or a combination of outcome and covariate data to adjust the randomization ratio.

Assumptions

Adaptive designs in clinical trials make two assumptions. First, participant randomizations must be spread throughout the course of the study. Because adaptive designs in clinical trials use outcome and/or covariate data to adjust the randomization ratio for later randomizations, data must first be collected from some participants in order to benefit later randomizations of other participants. Second, the time between participant randomization and outcome data collection must be reasonably quick when using an outcome-adaptive design. When the randomization ratio is adjusted using outcome data, an extended period between participant randomization and outcome data collection would delay adjustment of the randomization ratio. This concern is less relevant for covariate-adaptive designs since covariate information can usually be collected quickly.

Families of Adaptive Designs

Since their inception, several families of statistical models have been used to define types of adaptive designs in clinical trials. There are the play-the-winner models, drop-the-loser models, biased coin designs, bandit models, random *N* models, and target R*. While an in-depth discussion of these models is beyond the scope of this entry, each of these models defines rules for adjusting the randomization ratio.

Generally, these statistical models for adaptive designs in clinical trials were created as alternatives to randomized clinical trials using fixed randomization ratios. However, the early models within each family were typically deterministic, with each participant's allocation being completely determined by the previous participant's allocation and outcome. For example, early play-the-winner models would allocate the second participant based on the outcome and/or covariate data of the first participant, the third participant based on the outcome and/or covariate data of the second participant, and so on. Notably, this is *not* random assignment, since it would be known how the next participant was to be allocated. However, later work extended these models to probabilistic allocation, whereby the previous participant's outcome and/or covariate data affected the randomization ratio for the next participant but did not determine allocation.

Advantages

Adaptive designs in clinical trials use outcome and/or covariate data to adjust the randomization ratio. Because the randomization ratio reflects the data from previous participants, the randomization ratio maximizes the likelihood of participants experiencing a positive outcome and minimizes the likelihood of participants experiencing a negative outcome. In some fields (e.g., cancer research), maximizing potential benefit and minimizing potential risk are of the utmost importance.

Limitations

Two limitations of adaptive designs in clinical trials are reduced statistical power and threats to internal validity.

Reduced Statistical Power

Because participant randomizations are correlated with the outcomes and/or covariates, adaptive designs in clinical trials have reduced statistical power compared to randomized clinical trials using fixed randomization ratios. Consequently, adaptive designs in clinical trials will require more participants to achieve the same level of statistical precision.

Threats to Internal Validity

Because adaptive designs in clinical trials spread randomizations across the duration of the study, threats to internal validity (e.g., history effect, maturation) may differentially affect participants. For example, a study that is completed over the course of an academic year may have reduced internal validity because of developmental changes that cause students randomized at the beginning of the study to be qualitatively different from students randomized at the end of the study.

Jeffrey C. Hoover

See also Bayesian Adaptive Randomization Design; Clinical Trial; Random Assignment; Research Design Principles; Risk

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Further Readings

Hu, F., & Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. Hoboken, NJ: Wiley.

Lee, J. J., Chen, N., & Yin, G. (2012). Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research*, *18*(17), 4498–4507. doi:10.1158/1078-0432.CCR-11-2555.

Rosenberger, W. F., & Lachin, J. M. (2015). *Randomization in clinical trials: Theory and practice* (2nd ed.). Hoboken, NJ: Wiley.

Yin, G. (2012). *Clinical trial design: Bayesian and frequentist adaptive methods*. Hoboken, NJ: Wiley.

Adjusted *F* Test

See Greenhouse–Geisser Correction

Adverse Event Reporting

Adverse event reporting entails documenting and reporting harmful outcomes observed during research to a designated monitoring committee or system, review board, and/or institution. Adverse events are unanticipated problems experienced during research which places human subjects or others at a greater risk of harm than was previously known, including physical, psychological, economic, or social harm. Such events may lead to a pause and/or termination of the research to fully evaluate the direct or indirect cause of the harmful events. These can be thought of primarily as harmful events or outcomes not expected by researchers, and not included in Institutional Review Board (IRB) or other review submissions. In addition, expected adverse events, if more severe than previously known, are also reported and further evaluated. Thus, unexpected harmful events, or expected adverse events exceeding harm thresholds previously known, are reportable events to the monitoring committee or system, review board, and/or institution under which the research is supported. Research (e.g., biomedical, social, psychological) under the realm of federal, state, or institutional support dealing with human subjects include adverse effect reporting. Using a few examples, this entry further explores unexpected and adverse events and the reporting of such events.

Examples

The following are examples requiring adverse event reporting:

1. A randomized clinical trial (RCT) is assessing a new antiretroviral drug for HIV/AIDS patients. The lead researcher, Jane Hitti, discovered that during the testing period, severe liver damage occurred in several research subjects in the intervention arm receiving the drug.
2. Conor Duggan and other researchers reported on a 2014 intervention study for adults with personality disorders. Investigators noticed an increase in negative mental health events (e.g., self-harm, overdoses) beyond what is normally seen in this population
3. A vaccine to protect individuals from a new virus results in what appears to be an increase in severe allergic reactions after receiving the vaccine injection, including anaphylaxis. This happened as part of a 2003 Centers for Disease Control and Prevention study as reported by Weigong Zhou and colleagues.

In the first example, the study was shut down due to greater than expected liver toxicity. In the second example, the adverse events were reported to the public funding agency's data monitoring committee—the study was not stopped, but no new research participants were recruited. In the third example, the anaphylaxis was reported to the Vaccine Adverse Event Reporting System for further evaluation and analysis; it was subsequently deemed a rare event allowing for continued vaccinations.

Unexpected and Expected Adverse Events

The most familiar adverse event reporting—what individuals read in scientific journals, or read, see, and hear on mass and social media platforms—concerns unexpected adverse events in RCTs, where research subjects are randomly assigned to a control group or an intervention group, then followed over time to evaluate change due to the intervention. Hypotheses regarding positive outcomes are offered by researchers, and any expected negative consequences are discussed, including safety measures to address adverse events. These issues are covered as part of the IRB submission for human subjects, normally submitted through a participating institution or similar entity. In the United States, to receive federal, state, or institution funding and adhere to Department of Health and Human Services

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

(DHHS) and Food and Drug Administration (FDA) guidelines, IRB approval is required—thus, researchers working with human subjects must also have access to an IRB and associated oversight.

Unexpected adverse events entail harmful outcomes to human subjects or others not expected. Such issues were not anticipated by researchers, nor documented in IRB or review submissions, nor were human subjects made aware of the possible negative outcomes. These events are then reported to the requisite monitoring committee or institution for further review. In some instances, the ongoing research is halted while review of such events are investigated. For example, in 2006, Edward Mills and colleagues documented a number of RCTs regarding treatments for HIV/AIDS (mostly antiretroviral treatments) that were stopped due to unexpected adverse events, including extreme treatment toxicity, side effects, encephalopathia, and death.

Expected adverse events are those events expected to occur during the research but are not more severe or prevalent than what was previously known. Such events are documented and addressed by researchers but do not fall directly under adverse event reporting unless they exceed what was expected. For example, an expected adverse event for medical researchers testing a new treatment for acne might include subjects experiencing dry lips and skin. Safety measures and protocols are covered in the review submission to address such physiological reactions. In another example, a psychologist might utilize personal interviews to evaluate one-night sexual encounters, yet also be aware that the interviews might "trigger" suppressed memories of sexual distress in a few of those interviewed—again, safety protocols are established for those possible cases. Essentially, if expected adverse events occur in research, then protocols are in place to address the events.

However, when expected adverse events are more prevalent and/or severe than previously known, then adverse event reporting is warranted. Using the previous examples, if medical researchers investigating an acne medication notice severe lip cracking and skin peeling (thus, not just dry lips and skin), or if the psychologist notices interview participants experiencing severe sexual distress requiring immediate counseling, then such adverse events are reportable. Research can be temporarily halted or shut down if such events exceed what was expected. For example, the study noted earlier by J. Hitti and colleagues (2004) was stopped, with the authors noting, "We observed greater than expected toxicity associated with nevirapine during the first phase of this randomized trial" (p. 774). The 1989 Cardiac Arrhythmia Suppression Trial, which focused on antiarrhythmic therapy, was selectively stopped due to increases in mortality for some of the treatment arms of the study. Both these applied examples underscore that adverse events were expected, but exceeded what was anticipated, putting human subjects at greater risk. Jesse A. Berlin and colleagues (2008) outlined some general probabilities for adverse events above what is expected at background levels.

Other Issues With Adverse Event Reporting

Non-RCT Research Designs

As noted earlier, the familiar examples regarding adverse event reporting entail RCTs. However, any prospective research design involving an intervention (e.g., vaccine, therapeutic, behavior-change regimen) can yield adverse events with human subjects. For example, trials that are one-shot case study designs (one intervention group, no follow-up) can have adverse events; without a control group, resulting adverse events can be compared to background event reporting, and if unique or adverse events exceed background levels, then such events are reported. Laboratory-based observational research such as those conducted in the social and behavioral sciences can also cause adverse events. A classic example is the Stanford Prison Study, conducted by Philip Zimbardo in the early 1970s, where even the principal investigator succumbed to the "power of the situation" in his role as prison superintendent and influenced possible adverse events experienced by study participants. The study ended early due to these events and the principal investigator requested an ethics investigation. Even field-based participant-observational research studies can cause adverse events to those being observed, as the observers' presence may negatively impact the population being evaluated.

Rare Event Occurrence

Adverse events can also be extremely rare events, but the detection or reporting may be delayed due to infrequency of observation. For example, Esther W. Chan and colleagues documented such negative events, noting serious adverse events such as death due to suicide for attention-deficit hyperactivity disorder treatments, and retinal detachment associated with the use of oral fluoroquinolones. Although not clearly detected in the original studies of these treatments, over time and with broader medication use, such events were subsequently documented and reported, calling the treatments into question.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Animal Subjects

Currently, adverse event reporting is not an absolute requirement with nonhuman subjects, although recommendations have been made (see Ferdowsian et al., 2020).

Where to Report Adverse Events

Different research protocols, IRBs, and funding institutions will require different reporting options. Also, the type of research conducted will determine where to report. In the United States, a few of the systems designed to monitor and capture adverse events include the Vaccine Adverse Event Reporting System, the Monitoring System for Adverse Events Following Immunization, and the FDA Adverse Event Reporting System designed as a database for postmarketing safety surveillance of drug and therapeutic biologic products. In addition, independent Data Safety and Monitoring Boards are often established for clinical trials.

William David Marelich

See also Belmont Report; Beneficence; Clinical Trial; Ethical Review Versus Scientific Review; Ethics in the Research Process; Informed Consent

Further Readings

Berlin, J. A., Glasser, S. C., & Ellenberg, S. S. (2008). Adverse event detection in drug development: Recommendations and obligations beyond phase 3. *American Journal of Public Health*, *98*(8), 1366–1371. doi:10.2105/AJPH.2007.124537.

Centers for Disease Control and Prevention. (1985, 25 January). Adverse events following immunization. *MMWR. Morbidity and Mortality Weekly Report*, *34*(3), 43–47.

Chan, E. W., Liu, K. Q., Chui, C. S., Sing, C. W., Wong, L. Y., & Wong, I. C. (2015). Adverse drug reactions—Examples of detection of rare events using databases. *British Journal of Clinical Pharmacology*, *80*(4), 855–861. doi:10.1111/bcp.12474.

Dixon, D. O., Weiss, S., Cahill, K., Fox, L., Love, J., McNamara, J., & Soto-Torres, L. E. (2011). Data and safety monitoring policy for National Institute of Allergy and Infectious Diseases clinical trials. *Clinical Trials*, *8*(6), 727–735. doi:10.1177/1740774511425181.

Duggan, C., Parry, G., McMurran, M., Davidson, K., & Dennis, J. (2014). The recording of adverse events from psychological treatments in clinical trials: Evidence from a review of NIHR-funded trials. *Trials*, *15*, 335. doi:10.1186/1745-6215-15-335.

Ferdowsian, H., Johnson, L., Johnson, J., Fenton, A., Shriver, A., & Gluck, J. (2020). A Belmont Report for animals? *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees*, *29*(1), 19–37. doi:10.1017/S0963180119000732.

Fiscella, K., Sanders, M., Holder, T., Carroll, J. K., Luque, A., Cassells, A., . . . Tobin, J. N. (2020). The role of data and safety monitoring boards in implementation trials: When are they justified? *Journal of Clinical and Translational Science*, *4*(3), 229–232. doi:10.1017/cts.2020.19.

Hitti, J., Frenkel, L. M., Stek, A. M., Nachman, S. A., Baker, D., Gonzalez-Garcia, A., . . . PACTG 1022 Study Team (2004). Maternal toxicity with continuous nevirapine in pregnancy: Results from PACTG 1022. *Journal of Acquired Immune Deficiency Syndromes*, *36*(3), 772–776. doi:10.1097/00126334-200407010-00002.

Phillips, R., Hazell, L., Sauzet, O., & Cornelius, V. (2019). Analysis and reporting of adverse events in randomised controlled trials: A review. *BMJ Open*, *9*(2), e024537. doi:10.1136/bmjopen-2018-024537.

Zhou, W., Pool, V., Iskander, J. K., English-Bullard, R., Ball, R., Wise, R. P., . . . Chen, R. T. (2003). Surveillance for safety after immunization: Vaccine Adverse Event Reporting System (VAERS)—United States, 1991–2001. *MMWR. Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002)*, *52*(1), 1–24.

Zimbardo, P. G. (1973). On the ethics of intervention in human psychological research: With special reference to the Stanford Prison Experiment. *Cognition*, *2*(2), 243–256. doi:10.1016/0010-0277(72)90014-5.

Akaike Information Criterion

The Akaike Information Criterion (AIC) is one of the first and among the most widely used model selection criteria. In general, a model selection criterion is a measure that summarizes how effectively a statistical model balances the competing objectives of parsimony and fidelity to the data used in its construction. AIC can be used to rank order a defined collection of candidate models and to identify a "best" model from among these candidates.

AIC was first introduced by Hirotugu Akaike in 1973 as an extension to the maximum likelihood principle, which assumes that the size and structure of a statistical model are known and that the data need only be used to estimate the associated (unknown) model parameters. Akaike described AIC as being based on an extension to this principle since the application of the criterion employs data for both the determination of the size and structure of the model and the estimation of its associated parameters. This entry provides a formal definition of AIC, contrasts the selection criterion and hypothesis-testing approaches for model selection,

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

gives practical notes for the use of AIC, and demonstrates the utility of AIC in an application based on epidemiological data.

Formal Definition

AIC serves as an asymptotically unbiased estimator of the expected Kullback–Leibler (KL) discrepancy between the model that gave rise to the data (i.e., the true or generating model) and a fitted candidate model. The KL discrepancy measures the degree of separation between two statistical models. Thus, for a sufficiently large sample, the fitted candidate model corresponding to the minimum value of AIC is ideally "closest" to the truth among the set of models under consideration.

To formally define AIC, consider a candidate collection of models $M_1, M_2, \ldots, M_L$, each with corresponding parameter vector θ_k ($k = 1, \ldots, L$). The dimension, or size, of each model describes the number of functionally independent parameters in θ_k and is denoted by d_k. We let $\hat{\theta}_k$ denote the estimator of θ_k obtained through maximizing the likelihood, $L(\theta_k \mid y)$, based on the data y. We define the AIC for model M_k as

$$\mathrm{AIC}_k = -2\log L\left(\hat{\theta}_k \mid y\right) + 2d_k,$$

where the first term, $-2\log L\left(\hat{\theta}_k \mid y\right)$, captures the goodness of fit of the model, and the second term, called the penalty term, reflects model complexity. The goodness-of-fit term decreases in value with improved fidelity of the fitted model to the data at hand. The penalty term, $2d_k$, characterizes model complexity through the model's dimension and therefore increases in value for models of larger size. Larger models are often associated with improved goodness of fit and hence smaller values for the corresponding term. However, the improvement in fit may be offset by increasingly large penalizations corresponding to greater model complexity. Conversely, smaller models are penalized less heavily for complexity but may be too simplistic to adequately accommodate the data at hand, yielding larger goodness-of-fit values. Among the candidate collection of models under consideration, the trade-off between goodness of fit and parsimony is then best optimized by the fitted model corresponding to the minimum AIC.

We may interpret AIC from a predictive standpoint by assuming interest in the prediction of a new panel of data, z, that is generated independently, but identically, to the fitting data, y. In this scenario, the prediction error of a model parameterized by θ_k can be measured by $-2\log L\left(\hat{\theta}_k(y) \mid z\right)$, where we use the notation $\hat{\theta}_k(y)$ to denote the vector of parameter estimators obtained through the fitting data y. The mean error of prediction based on averaging this quantity over the joint distribution of y and z is equivalent to the expected KL discrepancy that underlies AIC. This suggests that in using AIC to select a model, assuming the prediction of data generated independently from, but identically to, the fitting data, one is attempting to choose the model that minimizes a measure of mean prediction error. This predictive interpretation relates to the optimality property of asymptotic efficiency. In large sample settings, an asymptotically efficient selection criterion will select the fitted candidate model that yields predictors that minimize the mean squared error of prediction. In 1981, Ritei Shibata proved this property held for AIC and other selection criteria.

Underlying Assumptions

The development of AIC relies on the asymptotic, or large-sample, properties of the maximum likelihood estimator. Akaike showed that $-2\log L\left(\hat{\theta}_k \mid y\right)$ serves as an optimistically biased estimator of the KL discrepancy. The bias is optimistic in the sense that $-2\log L\left(\hat{\theta}_k \mid y\right)$ underestimates the actual discrepancy, or disparity, between the fitted candidate model and the true model. Akaike showed that this bias may be corrected asymptotically through the adjustment, $2d_k$, thus giving rise to the definition of AIC previously provided. In small-sample settings, where the sample size n is small relative to the model dimension (e.g., $d_k \gtrsim n/2$), this asymptotic bias adjustment is no longer valid. As a result, in such small-sample applications, AIC does not properly penalize model complexity and often leads to the selection of unnecessarily large and complex models that poorly optimize between goodness of fit and parsimony. Recognizing this limitation, Nariaki Sugiura proposed a "corrected" AIC in 1978 that serves as an exactly unbiased estimator of the KL discrepancy provided that the candidate collection of models consists of normal linear regression models. This criterion, AICc, was later generalized for application in the frameworks of normal nonlinear regression models and time series autoregressive models, autoregressive moving average models, vector autoregressive models, normal multivariate linear regression models, and certain generalized linear models (GLMs) and linear mixed models, thereby

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

increasing its popularity as a small-sample alternative to AIC.

In addition to the large-sample requirement, AIC is developed under the assumption that the true model is subsumed by the candidate model. The practical implication of this assumption is that a fitted candidate model is assumed to be either correctly or overspecified. A correctly specified candidate model is one that appropriately represents the true model, whereas a model that is overspecified includes all of the requisite structure of the true model along with additional features extraneous to the true model. Kei Takeuchi introduced the Takeuchi Information Criterion in 1976 as an alternative to AIC that relaxes this strong assumption. While the goodness-of-fit terms of Takeuchi Information Criterion and AIC are identical, the criteria differ in their penalty terms. In contrast to AIC, Takeuchi Information Criterion penalizes models through a complex function of the sample data. This data-dependent penalization may be substantially less biased in its ability to correct for the optimism inherent in using the empirical log-likelihood to estimate the expected KL discrepancy. However, since this penalization is determined from the data, it could be much more variable than the data-independent penalization, $2d_k$, of AIC.

Contrast With Hypothesis Testing

Prior to the advent of AIC, model selection was largely restricted to the hypothesis-testing paradigm. In this framework, pairwise comparisons among nested models are performed. Under the null hypothesis, the smaller, nested model is assumed to represent the truth. This model is rejected in favor of the larger candidate model if the data depart substantially from what would be expected if the smaller model were indeed true. To gauge the observed degree of departure, or evidence, against the null-hypothesized model, a p value is computed and compared to an arbitrarily defined threshold of statistical significance, with the most popular being the .05 threshold first suggested by Ronald Fisher in the early 20th century. Akaike's introduction of AIC offered an alternative to this framework and subsequently initiated the development of a growing collection of model selection criteria. In contrast to the hypothesis-testing framework, the use of model selection criteria does not constitute an attempt to determine which of two competing nested models represents the unknown truth but rather to assess candidate models on the basis of how effectively various models approximate the truth. Given a collection of candidate models, nested or otherwise, selection criteria allow one to determine those models that conform well to the data at hand (goodness of fit) without an unnecessary degree of structural complexity (parsimony). A favored model should also be generalizable in the sense that it should adequately describe or predict new data generated under the true model.

Practical Notes

In addition to the comparison of nonnested models, AIC can be used for the comparison of models based on different probability distributions. As an example, AIC may be used to decide between a GLM that employs a Poisson distribution for a count response and a GLM that utilizes a negative binomial distribution to account for overdispersion. If such comparisons are made, all terms in each empirical likelihood must be retained in the computation of AIC. This practice is in contrast to comparisons between models based on the same distribution, where data-independent terms may be discarded. AIC may not be used to compare models based on different transformations of the response variable.

While it is generally advised to pick the candidate model with the minimum value of AIC, there may be other, equally compelling candidate models indicated by the data. Table 1 provides guidelines that are commonly used to assess the level of empirical support for a given model based on the difference between the model's AIC value and the minimum AIC value in the candidate collection.

Table 1 Recommended Selection Guidelines Based on AIC Differences

$AIC_i - AIC_{min}$	*Level of Empirical Support for Model* i
0–2	Substantial
4–7	Considerably less
>10	Essentially none

Source: Burnham and Anderson (2002, p. 70). Reprinted with permission from Springer.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Application

To illustrate the utility of AIC, we consider modeling data on the yearly incidence (i.e., number of new cases) of AIDS in Belgium from the outset of the epidemic in 1981 to the year 1993. Figure 1 depicts a nonlinear trend in AIDS incidence that peaks in 1991 and is followed by a steady decline in new AIDS cases over subsequent years.

To properly model this nonlinear trend over time, we consider various polynomial models within the generalized linear modeling framework. Specifically, we assume that our response follows a Poisson distribution, and we specify a log-link function to relate the mean incidence to polynomials based on the yearly time index. Four polynomial models are entertained, each of increasing degree, d. We consider a linear ($d = 1$) model, quadratic ($d = 2$) model, cubic model ($d = 3$), and a model containing polynomial terms up to degree 12. Figure 2 depicts the fits of each polynomial to the data displayed in Figure 1.

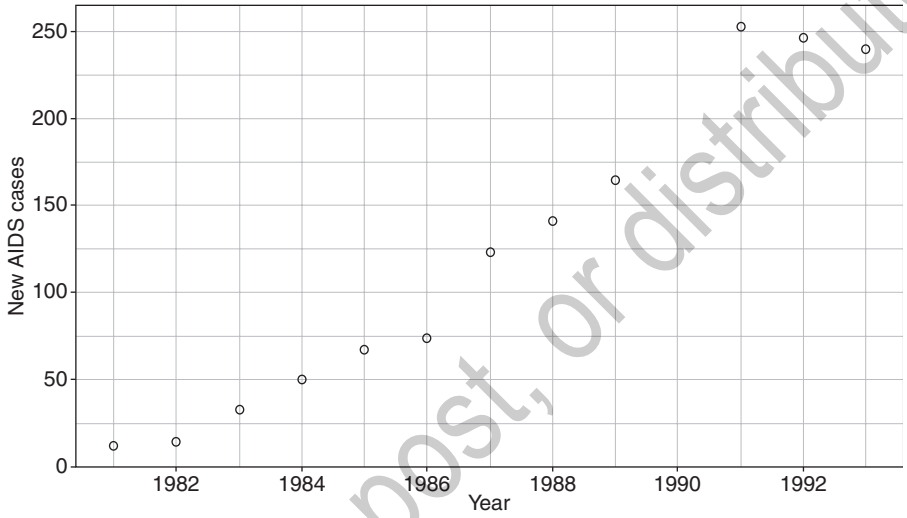

Figure 1 AIDS Incidence in Belgium, 1981–1993

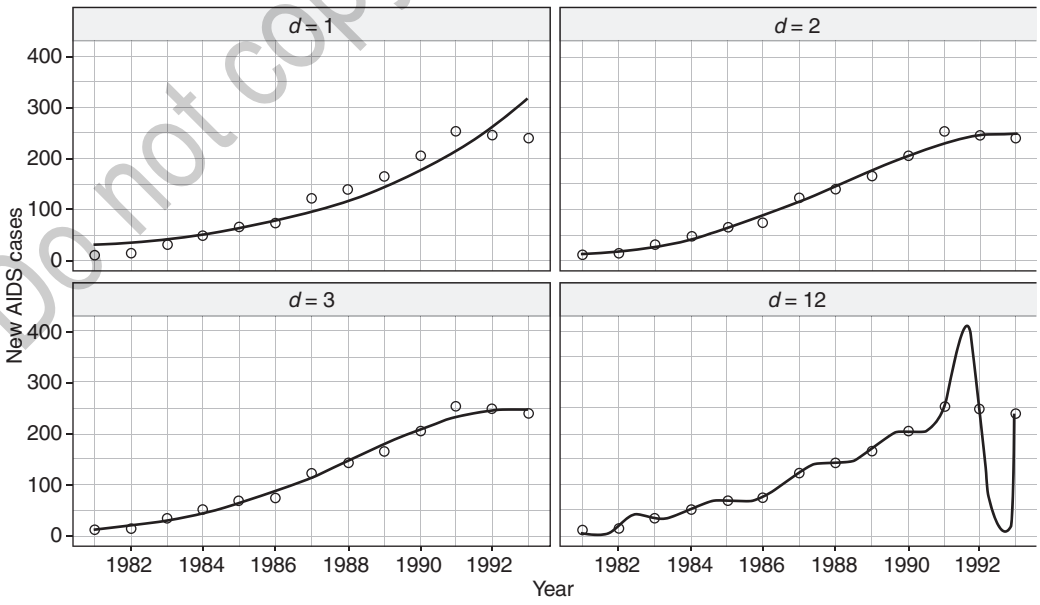

Figure 2 GLM Polynomial Fits to AIDS Incidence in Belgium, 1981–1993

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The upper left plot depicting the linear model fit ($d = 1$) only roughly approximates the trend in incidence and fails to capture the change in trend after 1991. This suggests a substantial degree of bias in the prediction of AIDS incidence over the range of years that are under consideration. The quadratic and cubic models, which are more complex than the linear model, better capture the nonlinearity in incidence over time and therefore exhibit reduced bias relative to the linear model. The most complex model ($d = 12$) yields a fitted curve that passes through each data point and is therefore not susceptible to bias problems. However, this model is beset by a high amount of variability in that the fitted curve fluctuates dramatically about the data, especially for the years 1991–1993. The fits of these models demonstrate the bias–variability trade-off, which relates directly to the optimization of goodness of fit and parsimony that serves as the impetus for the development of model selection criteria. Simpler, more parsimonious models are those with a larger amount of bias and a more minimal degree of variability. This is apparent in the simplest, linear model that depicts a smooth (low variability) fitted curve that tends to be either too low or too high to adequately characterize the true incidence values (high bias). In contrast, the most complex model ($d = 12$) yields a fitted curve that passes through every data point (low bias) but is extremely "wiggly" (high variability). The polynomial degree, model dimension, value of $-2\log L\left(\hat{\theta}_k \mid y\right)$ and AIC value for each of the four models are provided in Table 2.

The model with the minimum AIC is the quadratic model, and thus, within the collection of models under consideration, this model provides the best balance between goodness of fit and parsimony. However, one may notice that the cubic model has an AIC value that is only slightly larger than the quadratic model. Looking to the selection guidelines provided in Table 1, we would observe that the difference between the cubic and quadratic AIC values is 1.8, suggesting empirical support for both the quadratic and cubic models. Looking to Figure 2, however, we see that the fits of the quadratic and cubic models are virtually identical. Indeed, the addition of the cubic term results in only a slight decrease in the goodness-of-fit term, meaning that the difference in AIC values is close to the difference in penalizations (of two units). Thus, one may convincingly argue in favor of the quadratic model by virtue of Occam's Razor, the philosophical principle best described by the statement, "Everything should be made as simple as possible, but not simpler."

Table 2 AIC Values and Other Relevant Modeling Quantities

Polynomial Degree (d)	*Model Dimension* (d_k)	$-2\log L\left(\hat{\theta}_k \mid y\right)$	*AIC*
1	2	162.4	166.4
2	3	90.9	96.9
3	4	90.7	98.7
12	13	81.7	107.7

Javier E. Flores and Joseph E. Cavanaugh

See also Bias; Estimation; Law of Large Numbers; Likelihood Ratio Statistic; Occam's Razor; Variance

Further Readings

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd International symposium on information theory* (pp. 267–281). Budapest: Akadémia Kiadó.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*(6), 716–723.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. New York, NY: Springer.

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York, NY: Springer.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, *68*(1), 45–54. doi:10.2307/2335804.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics*, *Theory and Methods*, *A7*, 13–26. doi:10.1080/03610927808827599.

Takeuchi K. (1976). Distributions of information statistics and criteria for adequacy of models. *Mathematical Sciences*, *153*, 12–18.

Alternating Treatments Design

Alternating treatment designs, also called multielement designs, are single-case experimental designs, characterized by a rapid and frequent alternation of conditions. The specific sequence of conditions is usually determined at random, which enhances internal validity. This design feature makes possible the use of randomization tests, which improves statistical conclusion validity. Furthermore, replication of different alternation sequences across participants in the context of the same study is possible. In this entry, the main features of alternating treatment designs are described, distinguishing different ways of determining the alternation sequence and different uses. Indications for data analysis are also provided.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Main Features

Types of Designs and Applicability

The rapid alternation of conditions that takes place in an alternating treatments design distinguishes it from single-case phase designs (e.g., multiple baseline designs and withdrawal or reversal designs), which are characterized by more consecutive measurements for the same condition. It is necessary to distinguish different situations that can be described as an "alternating treatments design."

Alternating treatment designs with restricted randomization or block randomization

Although it is possible to determine the sequence of conditions through counterbalancing, it is more common to use randomization. Two different ways of determining the alternating sequence at random need to be mentioned.

On one hand, the random alternation scheme can impose that no condition is repeated until all have been conducted. Another way of describing the same procedure is to establish blocks of the same size as the number of conditions being compared and to determine at random the order of the conditions within each block. This design is called a randomized block design (although this name could induce confusion with a group design of the same name) or an alternating treatments design with block randomization. Other terms have also been used in the literature to describe this design, for instance, "alternating treatments design with a blocked pairs random assignment procedure" or "alternating treatments design in which the order of conditions is block randomized." The randomly determined sequence using block randomization is equivalent to the *N*-of-1 trials used in the health sciences, where the several random-order blocks are called multiple crossovers.

On the other hand, the random alternation scheme can impose a limit on the consecutive administrations of the same condition. It is very common to require a maximum of two consecutive sessions per condition. Such a design is called a restricted alternating treatment design or an alternating treatments design with restricted randomization or with a semi-random order of conditions.

Alternating treatment designs with restricted randomization and with block randomization are applicable to reversible behaviors and to interventions that can be introduced and removed fast, without leaving lasting effects. Moreover, unlike ABAB and multiple baseline designs, alternating treatment designs can be used to address questions about the relative strength of two interventions, rather than only studying efficacy in comparison to a control condition. It is possible and common to alternate more than two conditions.

For both kinds of alternating treatment designs, the rapid alternation of conditions is said to take place in a "comparison phase," which may be the only phase of the alternating treatment designs. However, it is also possible (although not compulsory) to have an initial baseline phase and a final phase in which only the most effective condition is used.

Adapted alternating treatment designs

Adapted alternating treatment designs are those in which at least two independent behaviors or outcome variables are treated. These behaviors are nonreversible, and the main aim is to explore which of two effective interventions is more efficient (i.e., enables faster learning). In an adapted alternating treatment design, it is critical to have the same number of sessions per condition and the researchers typically follow a procedure equivalent to block randomization.

Addressing Threats to Validity

In an alternating treatments design, several threats to internal validity are addressed. First, in relation to history, given that there are several alternations of the condition, it is less likely that external events occur simultaneously with the change in conditions. Second, there are several effects related to potential interactions among conditions and different strategies to deal with them. In relation to order effects (also called sequence effects), the random determination of the order of the conditions reduces this possibility, as there are many possible orders (e.g., AB, AC, BA, BC, CA, and CB when comparing three conditions). In relation to carryover effects, the control condition can be alternated with the intervention conditions in the comparison phase, in order to explore whether there are any systematic changes even in the absence of an active intervention. Fourth, to reduce the possibility of multiple treatment interference (i.e., the possibility that the effect of an intervention applied in frequent alternation with another intervention is different from presenting the intervention alone), the researchers can increase the amount of time between sessions (including washout periods). Finally, in terms of quantity, five repetitions of the alternation are required for meeting current design standards for solid evidence. Similarly, at least five measurements per condition are recommended. The demonstration of experimental control is boosted by

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

replication across several participants, each with their own randomly determined sequence.

Data Analysis

This section discusses options for analyzing the data from alternating treatment designs.

The Importance of the Design for the Data Analysis

The use of randomization in the design enables applying a randomization test for data analytical purposes. A randomization test can provide a *p* value for observed difference between conditions, representing the probability of obtaining such a larger difference or a larger one only due to chance (i.e., enabling a tentative inference about causality, rather than an inference about a population of individuals or measurements).

The statistical power of the randomization test is related to the number of possible alternation sequences, and the number of sequences that can be obtained for an alternating treatment design with block randomization is less than for an alternating treatment design with restricted randomization. For instance, a sequence such as AABBAABBAABB is possible only for an alternating treatment design with restricted randomization. In contrast, the visual and certain quantitative comparisons between conditions (mentioned next in the entry) are more straightforward for an alternating treatments design with block randomization, as the logical comparison is performed within blocks. In contrast, a sequence such as AABBAABBAABB could be divided in different segments for comparing adjacent conditions (e.g., AABB-AABB-AABB, AAB-BA-AB-BA-ABB or AAB-BAA-BBA-ABB). Furthermore, some alternating treatments design with restricted randomization may entail unequal number of measurement occasions per condition, which may also be limiting for certain data analytical procedures.

Finally, for an adapted alternating treatments design, the data analytical approach can be different, as the emphasis is put on the speed at which a certain level is reached. Thus, it is common to establish a mastery criterion specifying when the acquisition of the target behaviors takes place prior to gathering the data and to assess for which of the conditions this criterion is reached faster. It is also common for the graphical representations to include data points for both conditions for the same measurement occasion (i.e., for the same value in the abscissa). This makes visual analysis more straightforward. The following text deals with data analytical procedures for alternating treatments design with restricted randomization or with block randomization.

Visual Inspection

Visual inspection is suggested as a first choice for analyzing the degree to which the data path for one condition are different from (and superior to) the data path of the other condition. The data paths are represented by lines connecting the measurements belonging to the same condition. Thus, visual analysts assess the magnitude and consistency of the separation or differentiation between these connecting lines (e.g., whether they cross or not and what is the vertical distance between them).

Quantifying Differentiation

For quantifying differentiation, a specific version of the percentage of nonoverlapping data has been proposed: the first measurement for Condition A is compared with the first measurement for Condition B, the second measurement for Condition A is compared with the second measurement for Condition B, and so forth. This technique quantifies the percentage of comparisons for which there is superiority of one condition over another. The main limitations of this approach are that (a) the measurements compared may not be adjacent and could be separated by several other data points (e.g., for the second measurement for each condition in a sequence such as AABABBAABB) and (b) some of the measurements may not be used if there is an unequal number of measurements per condition. The "visual structured criterion" assesses whether the number of comparisons for which one condition is superior can be considered to represent more than chance superiority. The assessment is performed comparing the superiority observed to the cutoff points that the researchers derived empirically.

In contrast to the percentage of nonoverlapping data and the visual structured criterion, which quantify superiority in ordinal terms, the comparison involving actual and linearly interpolated values (ALIV) assesses the magnitude of effect by focusing on the average amount of distance between the data paths. What is meant by "linearly interpolated values" are the Condition A points located on the data paths (i.e., the connecting lines) for the measurement occasions for which Condition B actually takes place and, analogously, Condition B points located on the data paths (i.e., the connecting lines) for the measurement occasions for which Condition A actually takes place.

By focusing on the comparison of data paths, the visual structured criterion and ALIV aim to mimic the

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

data analytical approach of visual analysts and to provide quantifications that would summarize the outcome of the assessment. Only ALIV quantifies the magnitude of superiority of one condition over the other.

Tentative Causal Inference

Beyond the comparison of data paths, randomization tests are the first statistical options proposed for alternating treatments designs. Randomization tests provide information about statistical significance, without relying on assumptions of random sampling or regarding the shape of the distribution of the data. The likelihood of the outcome obtained under the null hypothesis of no difference between conditions is rather based on a reference distribution based on all possible alternation sequences under the randomization scheme used. The use of randomization tests is thus justified by the random determination of the alternating sequence. The test statistic usually suggested is the mean difference, which is logical due to its frequent use as a summary measure in alternating treatments design. This analytical option, as well as the possibility to select an alternating sequence at random for design of the study, is implemented in the ShinySCDA website. However, the mean difference does not represent the comparison between data paths as well as ALIV does. A randomization test for ALIV has been shown to have adequate statistical properties. Data analysis via ALIV, alongside a graphical representation of the data, can be obtained using another website: Data Analysis Techniques for Alternating Treatments Designs.

Additional Analytical Options

Additional data analytic alternatives include piecewise or multilevel regression and local regression with nonparametric smoothers. It is not clear whether applied researchers would be able to easily use and correctly interpret the results of these latter analytical options, which is why this entry emphasizes the comparison of data paths and the use of a statistical test that is not based on large sample approximations.

Rumen Manolov

See also Experimental Design; Multiple Baseline Single Case Experimental Design; Randomization Tests; Randomized Block Design; Single-Case Research Design; Visual Analysis in Single Subject Design

Further Readings

Barlow, D. H., & Hayes S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, *12*(2), 199–210. doi:10.1901/jaba.1979.12-199.

Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single–subject research: Description of designs and discussion of problems. *Topics in Early Childhood and Special Education*, *14*(1), 119–145. doi:10.1177/027112149401400111.

Horner, R. J., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 27–51). Washington, DC: American Psychological Association. doi:10.1037/14376-002.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson.

Lanovaz, M., Cardinal, P., & Francis, M. (2019). Using a visual structured criterion for the analysis of alternating-treatment designs. *Behavior Modification*, *43*(1), 115–131. doi:10.1177/0145445517739278.

Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment*, *19*(1), 4–17. doi:10.1017/BrImp.2017.16.

Manolov, R. (2019). A simulation study on two analytical techniques for alternating treatments designs. *Behavior Modification*, *43*(4), 544–563. doi:10.1177/0145445518777875.

Manolov, R., & Onghena, P. (2018). Analyzing data from single-case alternating treatments designs. *Psychological Methods*, *23*(3), 480–504. doi:10.1037/met0000133.

Nikles, J. & Mitchell, G. (Eds.). (2015). *The essential guide to N-of-1 trials in health*. Dordrecht: Springer.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain*, *21*(1), 56–68. doi:10.1097/00002508-200501000-00007.

Wolery, M., Gast, D. L., & Ledford, J. R. (2018). Comparative designs. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 283–334). London, UK: Routledge.

Websites

Data Analysis Techniques for Alternating Treatments Designs: https://manolov.shinyapps.io/ATDesign/

ShinySCDA: https://tamalkd.shinyapps.io/scda/

Alternative Hypotheses

The alternative hypothesis is the hypothesis that is inferred, given a rejected *null hypothesis*. Also called the *research hypothesis*, it is best described as an explanation for why the null hypothesis was rejected. Unlike the null, the alternative hypothesis is usually of most interest to the researcher.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

This entry distinguishes between two types of alternatives: the *substantive* and the *statistical*. In addition, this entry provides an example and discusses the importance of experimental controls in the inference of alternative hypotheses and the rejection of the null hypothesis.

Substantive or Conceptual Alternative

It is important to distinguish between the substantive (or conceptual, scientific) alternative and the statistical alternative. The conceptual alternative is that which is inferred by the scientist given a rejected null. It is an explanation or theory that attempts to account for why the null was rejected. The statistical alternative, on the other hand, is simply a logical complement to the null that provides no substantive or scientific explanation as to why the null was rejected. When the null hypothesis is rejected, the statistical alternative is inferred in line with the Neyman–Pearson approach to hypothesis testing. At this point, the substantive alternative put forth by the researcher usually serves as the "reason" that the null was rejected. However, a rejected null does not by itself imply that the researcher's substantive alternative hypothesis is correct. Theoretically, there could be an infinite number of explanations for why a null is rejected.

Example

An example can help elucidate the role of alternative hypotheses. Consider a researcher who is comparing the effects of two drugs for treating a disease. The researcher hypothesizes that one of the two drugs will be far superior in treating the disease. If the researcher rejects the null hypothesis, he or she is likely to infer that one treatment performs better than the other. In this example, the statistical alternative is a statement about the population parameters of interest (e.g., population means). When it is inferred, the conclusion is that the two means are not equal, or equivalently, that the samples were drawn from distinct populations. The researcher must then make a substantive "leap" to infer that one treatment is superior to the other. There may be many other possible explanations for the two means' not being equal; however, it is likely that the researcher will infer an alternative that is in accordance with the original purpose of the scientific study (such as wanting to show that one drug outperforms the other). It is important to remember, however, that concluding that the means are not equal (i.e., inferring the statistical alternative hypothesis) does not provide any scientific evidence at all for the chosen conceptual alternative. Particularly when it is not possible to control for all possible extraneous variables, inference of the conceptual alternative hypothesis may involve a considerable amount of guesswork, or at minimum, be heavily biased toward the interests of the researcher.

A classic example in which an incorrect alternative can be inferred is the case of the disease malaria. For many years, it was believed that the disease was caused by breathing swamp air or living around swamplands. In this case, scientists comparing samples from two populations (those who live in swamplands and those who do not) could have easily rejected the null hypothesis, which would be that the rates of malaria in the two populations were equal. They then would have inferred the statistical alternative, that the rates of malaria in the swampland population were higher. Researchers could then infer a conceptual alternative—*swamplands cause malaria*. However, without experimental control built into their study, the conceptual alternative is at best nothing more than a convenient alternative advanced by the researchers. As further work showed, mosquitoes, which live in swampy areas, were the primary transmitters of the disease, making the swamplands alternative incorrect.

The Importance of Experimental Control

One of the most significant challenges posed by an inference of the scientific alternative hypothesis is the infinite number of plausible explanations for the rejection of the null. There is no formal statistical procedure for arriving at the correct scientific alternative hypothesis. Researchers must rely on experimental control to help narrow the number of plausible explanations that could account for the rejection of the null hypothesis. In theory, if every conceivable extraneous variable were controlled for, then inferring the scientific alternative hypothesis would not be such a difficult task. However, since there is no way to control for every possible confounding variable (at least not in most social sciences, and even many physical sciences), the goal of good researchers must be to control for as many extraneous factors as possible. The quality and extent of experimental control is proportional to the likelihood of inferring correct scientific alternative hypotheses. Alternative hypotheses that are inferred without the prerequisite of such things as control groups built into the design of the study or experiment are at best plausible explanations as to why the null was rejected, and at worst, fashionable hypotheses that the researcher seeks to endorse without the appropriate scientific license to do so.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Concluding Comments

Hypothesis testing is an integral part of every social science researcher's job. The statistical and conceptual alternatives are two distinct forms of the alternative hypothesis. Researchers are most often interested in the conceptual alternative hypothesis. The conceptual alternative hypothesis plays an important role; without it, no conclusions could be drawn from research (other than rejecting a null). Despite its importance, hypothesis testing in the social sciences (especially the softer social sciences) has been dominated by the desire to reject null hypotheses, whereas less attention has been focused on establishing that the correct conceptual alternative has been inferred. Surely, anyone can reject a null, but few can identify and infer a correct alternative.

Daniel J. Denis, Annesa Flentje Santa, and Chelsea Burfeind

See also Hypothesis; Null Hypothesis

Further Readings

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997.

Cowles, M. (2000). *Statistics in psychology: An historical perspective* (2nd ed.). Philadelphia, PA: Erlbaum.

Denis, D. J. (2001). Inferring the alternative hypothesis: Risky business. *Theory & Science*, *2*(1), 1. http://theoryandscience.icaap.org/content/vol002.001/03denis.html

Gomm, R. (2017). A positivist orientation: Hypothesis testing and the 'Scientific Method.' In D. Wyse, N. Selwyn, E. Smith, & L. E. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 213–242). Thousand Oaks, CA: Sage.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference (Part 1). *Biometrika*, *20A*(1–2), 175–240. doi:10.1093/biomet/20A.1-2.175.

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123–138). Hoboken, NJ: Wiley Blackwell.

Alters

Alters are most aptly defined as the social actors, part of one's socially constructed network, without which one's membership in, and connection to, the broader societies in which one finds himself or herself embedded would be affected. The formal study of alters within social networks, via social network analysis, first appeared within the scholarly literature (beginning in the field of sociology) in the 1930s and has since become a salient area of empirical interest within the fields of communication, psychology, anthropology, and business administration. Much research has been devoted, for example, to better understanding how certain alters provide ego (oneself) with increased opportunities for dependent variables such as information, decision-making influence, and access to job opportunities. That is, as a result of producing social network ties with alters who can provide ego with certain outcomes, these network alters become more important within ego's network as compared to others. This has salient implications for the production of social capital, best defined as the benefits linked to the accumulation of network ties to alters who can, at some point and in some way, shape, form, or provide ego with something of import.

Research indicates that such outcomes of social capital (being connected to alters who can ultimately provide fruitful rewards) can range from being the first to know about intraorganizational gossip to having access to recently disclosed stock data to emotional support. The goal, one might argue, is to form relationships with alters who can provide resources deemed advantageous by and for ego as compared to others. This entry discusses the advantages that alters can provide, reasons why alter relationships form and break apart, and how alter relationships function.

Alters, certainly, vary in the advantages (both tangible and intangible) that they can provide. Certain alters can provide advantages directly to ego. However, after engaging in social network analysis, it might very well be that it is not a direct alter who is most profitable for ego within his or her social network, but rather it is ego's alter's alter. That is to say, there is one degree of social separation between oneself and the social actor who can ultimately provide the very advantage in question.

As we likely know, we are part of a myriad of different, and likely nonoverlapping, networks of social alters. We are part of friend networks we are part of employee networks we are part of family networks we are part of college/college alumni networks we are part of Facebook networks we are part of neighborhood networks and the list goes on. While some alters might be linked to multiple networks and, as such, connect previously disconnected social alters together, much of the literature indicates that individuals purposefully create nonoverlapping social networks for purposes of exclusion.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Assume that Sara learns about a forthcoming Jon Bon Jovi concert, the tickets to which have been sold out for some time. Yet, we also learn that Gillian, a social alter part of Sara's neighborhood network, who is a social alter of Daniel's, who is a social alter of Sara's, knows of a work colleague trying to sell two tickets because her roommate was recently diagnosed with pneumonia and will not, as a result, be able to attend. Sara reaps the benefit of Gillian's work partner as a result of her connection with an alter (Daniel), which is predicated on his network connection with an alter (Gillian), which is, in the end, all based on her relationship with a colleague. The conclusion? Alters, as the literature claims, do not have to be directly connected to ego in order to provide benefits and resources. As long as ego is socially connected to alters who are connected to alters who are connected to alters who are connected to alters who can provide something deemed important and necessary, then he or she has, from a social network perspective, immersed herself in a community rife with advantage and opportunity.

Alters also vary in the relational closeness that they share with ego himself or herself. It is both cognitively and behaviorally exhausting to create, develop, and maintain *strong* relationships with all of one's alters, and as research indicates, many of the ties that social actors create are *weak*: emblematic of infrequent communication with, and less emotional connection to, the alters part and parcel of one's social network(s). That said, of course, it is important not to confuse the word *weak* with the semantic undertone of *unvaluable*. For such social activities as political rallying, health communication campaigns, and social justice movements, there is an important *strength* associated with *weak* alter ties.

Unfortunately, of course, not all relationships are able to survive for a multitude of different reasons. Alters change jobs alters engage in conflict that results in the dissolution of relationships alters decide to change networks based on mere free will alters move geographic locations and there are a multitude of additional independent variables coming to affect one's relationship with an alter. When alters, for whatever reason, depart from a larger social network, it might very well disrupt the entire social fabric of that socially constructed, socially connected community. This can be depicted by a sociogram, a pictorial representation of the ways in which a network's alters are connected to one another. Figure 1 shows an example of a sociogram.

The data used to determine these connections accrue by using what is known as the *name-generator technique*. In short, this involves asking social actors who they are likely to interact with when it comes to things such as information, gossip, entertainment, advice, and the like. If the line has two arrows, this implies reciprocity (meaning that both individuals are likely to communicate with each other), whereas a line with one arrow implies nonreciprocity (meaning that one individual is more likely to communicate with the alter as compared to the other). In the case

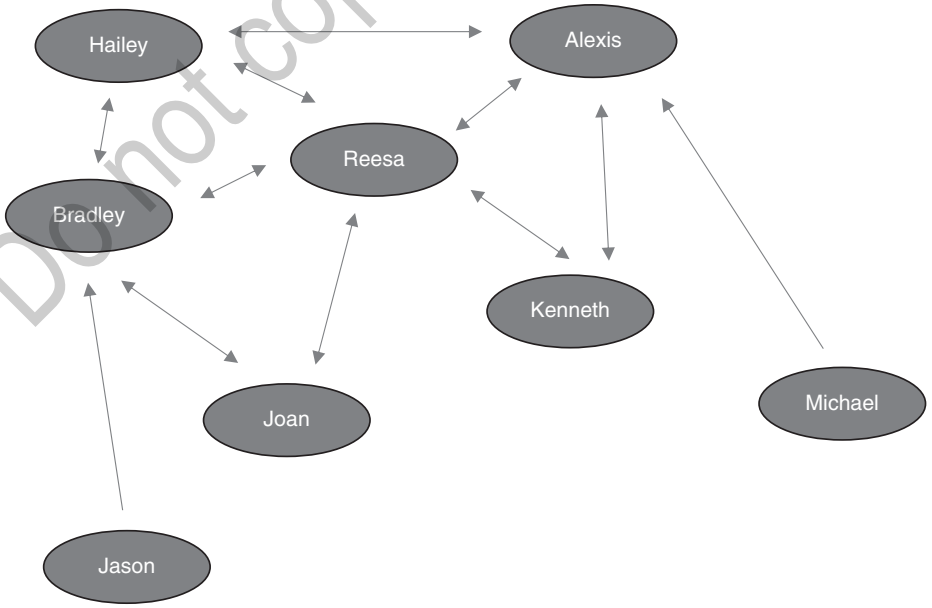

Figure 1 Eight-Person Sociogram Depicting Relationships Among Alters

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

shown in Figure 1, the question is what happens if an alter, for one reason or another, exits from this eight-person network?

As you can likely tell, Reesa is a very salient social alter within this network, as she has direct, reciprocal connections with five individuals: Hailey, Bradley, Joan, Kenneth, and Alexis. She, based on the empirical data, is the most important (central) member of this network. If she were to depart, several other social alters would be affected. For example, Reesa becomes the alter through which Alexis can communicate directly with Joan. If Reesa departs from this network, and Alexis wishes to interact with Joan, she is now forced to choose an entirely different social approach. She, in essence, would have to take an entirely different social path of connection (from Hailey to Bradley to Joan). While this might seem rather simple, research indicates that such a disruption in the configuration of alters has monumental implications for all affected social actors: especially those most important for the social connection of the network. Again, using the sociogram shown in Figure 1, if Jason and/or Michael were to depart, the implications would be much less severe, as the nonreciprocal connections that provide their connection to the network at large indicate that their departures would not be catastrophic at the macrolevel. They are both, according to the scholarly literature, considered social isolates and, as such, are likely to provide few (if any) necessary and desired resources to or for the network at large.

Corey Jay Liberman

See also Centrality; Reciprocity; Social Capital Theory; Social Network Analysis

Further Readings

Borgatti, S. P., & Everett, M. G. (1992). Notions of position in social network analysis. *Sociological Methodology*, *22*, 1–35. doi:10.2307/270991.

Marin, A. (2004). Are respondents more likely to list alters with certain characteristics? Implications for name generator data. *Social Networks*, *26*(4), 289–307. doi:10.1016/j.socnet.2004.06.001.

Morrison, E. W. (2002). Newcomers' relationships: The role of social network ties during socialization. *Academy of Management Journal*, *45*(6), 1149–1160. doi:10.5465/3069430.

Small, M. L. (2013). Weak ties and the core discussion network: Why people regularly discuss important matters with unimportant alters. *Social Networks*, *35*(3), 470–483. doi:10.1016/j.socnet.2013.05.004.

American Educational Research Association

The American Educational Research Association (AERA) is an international professional organization, based in Washington, DC, dedicated to promoting research in the field of education. Through conferences, publications, and awards, AERA encourages the scientific pursuit and dissemination of knowledge in the educational arena. Its membership is diverse, drawing from within the education professions as well as from a broader social science background.

Mission

The mission of AERA is to influence the field of education in three major areas. Firstly, through improving knowledge about education. Secondly, through promoting educational research. Finally, through encouraging the use of educational research results to make education better and thereby improve the common good.

History

The AERA publicizes its founding as taking place in 1916. However, its roots have been traced back to the beginnings of the educational administration research area and the school survey movement, both of which took place in the 1910s. This new spirit of cooperation between university researchers and public schools led the way to the founding of the National Association of Directors of Educational Research (NADER) as an interest group by eight individuals, formed within the National Education Association (NEA) Department of Superintendence in February 1915. With the creation of its first organizational constitution in 1916, NADER committed itself to the improvement of public education through applied research. NADER's two goals were to organize educational research centers at public educational settings and to promote the use of appropriate educational measures and statistics in educational research. Full membership in this new organization was restricted to individuals who directed research bureaus, although others involved in educational research could join as associate members. NADER produced its first publication, the *Educational Research Bulletin*, in 1916. Within 3 years of its founding, membership had almost quadrupled to 36 full members. In 1919, two of the founders of NADER started producing a new journal, the *Journal of*

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Educational Research (*JER*), soon to be adopted by NADER as an official publication.

With the growth in educational research programs in the late 1910s to early 1920s, NADER revised its constitution in 1921 to allow for full membership status to anyone involved in conducting and producing educational research, by invitation only, after approval from the Executive Committee. To better display this change in membership makeup, the group's name was changed in 1922 to the Educational Research Association of America (ERAA). This change in name also reflected a shift in the group's focus toward a broader representation of all who conducted educational research. This broader representation allowed for a large increase in membership, which grew to 329 members by 1931. Members were approximately two thirds from a university background and one third from the public schools. In 1928, ERAA changed its name, becoming the AERA.

After a problem involving the ownership of *JER*, AERA affiliated itself with the NEA in 1930, gaining Washington, DC, offices and support for AERA's proposed new journal. AERA underwent several changes during the 1930s. One change was the creation of their new journal, the *Review of Educational Research* (*RER*). Another was the decision to affiliate with other professional groups that shared common interests, such as the National Committee on Research in Secondary Education and the National Council of Education. The recognition of superior research articles at an annual awards ceremony was established in 1938. By 1940, AERA membership stood at 496.

Much of AERA's growth (beyond membership) has come from their journals. In 1950, the *AERA Newsletter* published its first issue. Its goal was to inform the membership about current news and events in education. Its name was changed to *Educational Researcher* in 1965. In 1963, the *American Educational Research Journal* (*AERJ*) was created to give educational researchers an outlet for publishing original research articles, as previous AERA publications focused primarily on reviews. By 1970, *RER* had changed the focus of its journal, which led to the creation of a new journal, the *Review of Research in Education* (*RRE*). In 2015, *AERA Open*, a peer-reviewed open-access journal, was created to encourage the rapid dissemination of new educational research.

AERA has been concerned with equity and civil rights both internally as an organization and externally in supplying resources for research-based policy advocacy. AERA was one of the last scholarly educational organizations to elect a person of color as president with the election of Linda Darling-Hammond in 1995, although by its 100th anniversary in 2016, it had elected nine presidents of color. In 1995, the Task Force on the Role and Future of Minorities was formed to explore ways AERA could increase the opportunities for minority engagement within the organization. The late 1990s also saw the rise of more inclusive Annual Meeting themes and the addition of a director of social justice position. AERA has experienced a growth in diversity-related Special Interest Groups (SIGs) and has established awards, fellowships, and lecture programs related to diversity.

Organization of Association Governance

The organization of AERA has changed since its founding to accommodate the greater membership and its diverse interests. AERA is governed by a council, an executive board, standing committees, and award committees. AERA Council is responsible for policy setting for AERA and is formed of elected members including the president, president-elect, immediate past president, six at-large members, division vice presidents, a SIG representative, a graduate student representative, and the executive director. The council meets three to four times per year.

The Executive Board is an advisory board, which guides the president and executive director of AERA. The board meets three to four times per year. The board has managed elections, advised on the annual budget, and selected Annual Meeting sites, in addition to other needed tasks. Twenty-two current standing committees, appointed by the council, are charged with conducting specific tasks in accordance with AERA policy. These committees range in focus from the Annual Meeting Policies and Procedures Committee to the Journal Publications Committee to the Social Justice Action Committee.

Divisions

AERA has 12 scholarly or scientific areas of interest called *divisions*. Each AERA member selects one division to become a member of upon joining AERA. Members can belong to multiple divisions for an additional annual fee. Divisions hold business meetings as well as support the presentation of research in their interest area at the Annual Meeting. The 12 divisions are

Division A—Administration, Organization, & Leadership

Division B—Curriculum Studies

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Division C—Learning & Instruction

Division D—Measurement & Research Methodology

Division E—Counseling & Human Development

Division F—History & Historiography

Division G—Social Context of Education

Division H—Research, Evaluation, & Assessment in Schools

Division I—Education in the Professions

Division J—Postsecondary Education

Division K—Teaching & Teacher Education

Division L—Educational Policy & Politics

SIGs

SIGs are smaller groups within the AERA membership. These SIGs differ from divisions in that their focus tends to be on more specific topics than broad interests represented by divisions. Like divisions, SIGs hold business meetings and support presentations of research in their interest area at the Annual Meeting. There are approximately 155 SIGs registered with AERA. Membership is based on annual dues, which in 2020 ranged from US$5 to US$40. SIGs range in focus from Accreditation, Assessment, and Program Evaluation in Education Preparation (SIG #174) to Faculty Teaching, Evaluation, and Development (SIG #42) to Research on Giftedness, Creativity, and Talent (SIG #91). AERA members may join multiple SIGs. SIGs are represented by the SIG Executive Committee, and a member serves on the AERA Council and AERA Executive Board.

Membership

At over 25,000 members, AERA is among the largest professional organizations in the United States. Approximately 14% of members live outside the United States and 30% of its membership are students. Roughly 66% of its members are women. About 63% of the members hold a PhD or EdD, and 75% of its employed members work at a college/university. Membership in AERA is primarily divided among voting members and nonvoting affiliates. To be a voting member, one must either hold the equivalent of a master's degree or higher, be a graduate student sponsored by a voting member of their faculty, or be retired from a position eligible for membership. Nonvoting affiliate members are interested in educational research but do not have a high enough level of education, are undergraduate students who are sponsored by their faculty, or are non-U.S. citizens who do not meet the educational level requirement. Students pay a reduced rate for membership. Members of AERA gain many benefits including a reduced cost to attend the Annual Meeting, free membership in one division, and free subscriptions to both the *Educational Researcher* and another AERA journal of their choice.

Publications

AERA publishes seven peer-reviewed journals as well as books and e-books. AERA's peer-reviewed journals include

AERJ

AERA Open

Educational Evaluation and Policy Analysis

Educational Researcher

Journal of Educational and Behavioral Statistics

RER

RRE

AERJ focuses on original scientific research in the field of education and learning. *AERA Open* is open access, covering similar material to the *AERJ*. *Educational Evaluation and Policy Analysis* publishes original research focusing on evaluation and policy analysis issues. *Educational Researcher* publishes information that is of general interest to a broader variety of AERA members. Interpretations and summaries of current educational research as well as book reviews make up the majority of its pages. The *Journal of Educational and Behavioral Statistics* focuses on new statistical methods for use in educational and behavioral research as well as critiques of current practices. It is published jointly with the American Statistical Association. *RER* publishes a variety of reviews of previously published educational articles by interested parties from varied backgrounds. *RRE* is an annual publication that solicits critical essays on a variety of topics facing the field of education.

Annual Meetings

AERA convenes a yearly Annual Meeting as an opportunity to bring AERA's membership together to discuss and debate the latest in educational practices and research. Approximately 16,000 attendees gather to listen, discuss, and learn over a 5-day period. For the 2019 meeting, 12,560 presentation proposals were submitted, and 6,279 papers were presented in 607 sessions. Presenters are invited to contribute a full-text paper to the AERA Online Paper Repository as a way to distribute their research more broadly. This online

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

repository is available open-access to the public. In addition to presentations, business meetings, invited sessions, awards, and demonstrations are held. Many graduate student–oriented sessions are also held. Sessions focusing on educational research related to the geographical location of the Annual Meeting are presented. Another valuable educational opportunity is the many professional development and training courses offered over the span of the conference. These tend to spotlight refresher courses in statistics and research design, evaluation, or workshops on new assessment tools or classroom-based activities.

In addition to the scheduled sessions, exhibitors of software, books, and testing materials present their wares at the exhibit hall, and those seeking new jobs can met prospective employers in the career center. There are also tours of local attractions available. Each year's meeting is organized around a different theme. In 2020, the Annual Meeting theme was *The Power and Possibilities for the Public Good: When Researchers and Organizational Stakeholders Collaborate*. The Annual Meeting takes place in conjunction with the annual meeting of the National Council on Measurement in Education, which is held at the same time and place as AERA's meeting.

Other Services and Offerings

This section discusses several other AERA offerings, including the Graduate Student Council, awards, and fellowships and grants.

Graduate Student Council

Graduate students are supported through several programs within AERA, but the program that provides or sponsors the most offerings for graduate students is the Graduate Student Council. The council comprises two graduate student representatives from each division plus additional officers. Its mission is to support graduate student members during their transition to become professional researchers and/or practitioners through education and advocacy. The council sponsors many sessions at the Annual Meeting as well as publishing a newsletter multiple times per year.

Awards

AERA offers an extensive awards program, with 13 award committees overseeing the process. The recipients are announced at the President's Address during the Annual Meeting. AERA's divisions and SIGs also offer awards, which are recognized by AERA and presented during their group's business meeting. AERA's awards cover educational researchers at all stages of their career, from the Early Career award to the Distinguished Contributions to Research in Education Award. Special awards are also given for other topics including social justice issues, public service, and outstanding books.

Fellowships and Grants

AERA offers fellowships, with special fellowships focusing on minority researchers and student researchers and a variety of fellowships on diverse areas of educational research. AERA also offers several small grants to support dissertations and other research.

Carol A. Carman

See also American Statistical Association; National Council on Measurement in Education

Further Readings

Banks, J. A. (2016). Expanding the epistemological terrain: Increasing equity and diversity within the American Educational Research Association. *Educational Researcher*, *45*(2), 149–158. doi:10.3102/0013189X16639017.

Hultquist, N. J. (1976). A brief history of AERA's publishing. *Educational Researcher*, *5*(11), 9–13. doi:10.3102/0013189X005011009.

Mershon, S., & Schlossman, S. (2008). Education, science, and the politics of knowledge: The American Educational Research Association, 1915–1940. *American Journal of Education*, *114*(3), 307–340.

Websites

American Educational Research Association: http://www.aera.net

American Psychological Association Style

American Psychological Association (APA) style is a system of guidelines for writing and formatting manuscripts. APA style may be used for multiple types of manuscripts, such as theses, dissertations, reports of empirical studies, literature reviews, meta-analyses, theoretical articles, methodological articles, and case studies. APA style is described extensively in the *Publication Manual of the American Psychological Association* (APA Publication Manual). The APA Publication Manual includes recommendations on writing style,

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

grammar, and nonbiased language as well as guidelines for manuscript formatting, such as arrangement of tables and section headings. APA style is the most accepted writing and formatting style for journals and scholarly books in psychology as well as in other disciplines. The use of a single style that has been approved by the leading organization in the field aids readers, researchers, and students in organizing and understanding the information presented.

Writing Style

The APA style of writing emphasizes clear and direct prose. Ideas are to be presented in an orderly and logical manner, and writing is to be as concise as possible. APA style reinforces usual guidelines for clear writing, such as the presence of a topic sentence in each paragraph. Previous research is described in either the past tense (e.g., "Washington and Hamilton found") or past perfect tense (e.g., "researchers have argued"). Past tense is used to describe procedures and results of an empirical study conducted by the author (e.g., "participants completed a survey," "women scored higher than men"). Present tense (e.g., "these results indicate") is used in discussing and interpreting results and drawing conclusions.

Nonbiased Language

APA style guidelines recommend that authors avoid language that is biased against particular individuals or groups. The APA provides specific guidelines for describing age, gender, race, ethnicity, sexual orientation, and disability status. Preferred terms change over time and may also be debated within groups; thus, authors are advised to consult a current style manual if they are unsure of which terms are currently preferred or considered offensive. Authors may also ask participants about terms they prefer for themselves.

General guidelines for avoiding biased language include being specific, employing person-first language, using labels as adjectives instead of nouns (e.g., *gay men* rather than *gays*), and avoiding labels that imply a hierarchy or standard of judgment (e.g., *normal development*, *stroke victim*).

Formatting

The APA Publication Manual provides extensive guidelines for formatting manuscripts. These include guidelines for use of numbers, abbreviations, quotations, and headings.

Tables and Figures

In many cases, tables and figures can present numerical information more clearly and concisely than would be possible in text. Tables and figures may also allow for greater ease in comparing numerical data (e.g., the mean achievement test scores of experimental and control groups). Figures and tables should present information clearly and supplement, rather than restate, information provided in the text of the manuscript.

Headings

Headings provide the reader with an outline of the organization of the manuscript. APA style includes five heading levels. Authors are advised to begin with the highest level of heading (Level 1) and continue as needed depending on the length and complexity of the manuscript. Most manuscripts will not require all five levels of heading. Topics of equal importance should have the same level of heading throughout the manuscript (e.g., the Method sections of multiple experiments should have the same heading level for each experiment). APA style recommends that author avoid using only one headed subsection within a section.

According to APA style, headings are formatted as follows:

Level 1: Centered Bold Uppercase and Lowercase Heading
Level 2: Flush Left Bold Uppercase and Lowercase Heading
Level 3: Flush Left Bold Italic Uppercase and Lowercase Heading
Level 4: Indented Bold Uppercase and Lowercase Heading Ending with a Period.
Level 5: Indented Bold Italic Uppercase and Lowercase Heading Ending with a Period.

For Heading Levels 1 through 3, the paragraph following the heading begins on a new line. For Heading Levels 4 and 5, the paragraph following the heading begins after the period at the end of the heading.

Manuscript Sections

A typical APA style manuscript reporting on an empirical study has five sections: Abstract, Introduction, Method, Results, and Discussion.

Abstract

An abstract is a concise (typically 100–250 words) summary of the contents of a manuscript. The abstract typically includes a description of the topic or problem

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

under investigation, information about participants and research methods, and the most important findings or conclusions. The abstract of a published article will often be included in databases; this allows researchers to search for relevant studies on a particular topic.

Introduction

The introduction section introduces the reader to the question under investigation. In this section, the author describes the topic or problem, discusses existing theory and research related to the topic, and states the purpose of the current study. The introduction typically concludes with a brief statement about the present study, including the author's research questions and/or hypotheses and the ways in which these research questions and hypotheses are supported by the existing research presented in the introduction.

Method

The method section describes how the study was conducted. The method section is frequently broken up into subsections, such as participants, procedure, and measures. Procedures and measures are described such that a reader knows what would be needed to replicate the study.

Descriptions of participants typically include summaries of demographic characteristics such as participants' ages, genders, and races and/or ethnicities. Other demographic characteristics, such as socioeconomic status and education level, are reported when relevant. The method by which participants were recruited (e.g., by newspaper advertisements or through a departmental subject pool) is also included.

The procedure describes participant recruitment, any experimental manipulation(s), instructions to participants (summarized unless instructions are part of the experimental manipulation, in which case they are presented verbatim), order in which measures and manipulations were presented, and control features (such as randomization and counterbalancing).

Measures that are commercially available or published elsewhere should be referred to by name and attributed to their authors, (e.g., "Self-esteem was measured using the Perceived Competence Scale for Children (Harter, 1982)."). Measures created for the current study should be described (e.g., example items, response scales) and may be reproduced in the manuscript in a table or appendix.

Results

The results section presents and summarizes the data collected and discusses the analyses conducted and their results. Analyses are to be reported in sufficient detail to justify conclusions. All relevant analyses should be reported, even those whose results were statistically nonsignificant or that did not support the stated hypotheses.

For a quantitative study, the results section will typically include inferential statistics, such as chi-squares, F tests, or t-tests. For these statistics, the value of the test statistic, degrees of freedom, and p value should be reported (e.g., $F(1,75) = 4.60$, $p = .034$). Effect sizes for statistical tests are also frequently presented.

For a qualitative study, the results section will include descriptions of identified themes or data coding categories as well as descriptions of the data analysis procedures by which these themes or categories were identified. Qualitative research manuscripts may also include a description of the author's approach to inquiry (e.g., feminist, postmodern) and the author's own positionality relative to the research question and context.

The results section may include figures (such as graphs or models) and tables. Figures and tables will typically appear at the end of a manuscript. If the manuscript is being submitted for publication, notes may be included in the text to indicate where figures or tables are to be placed (e.g., "Insert Figure 1 here."). All figures and tables should be referenced in the manuscript text (e.g., "Scores for the intervention and control groups did not differ (see Table 2 for means).").

Discussion

In the discussion section, findings and analyses presented in the results section are summarized and interpreted. In this section, the author discusses how results relate to previously stated research questions and hypotheses. Conclusions are drawn but should remain within the boundaries of the data obtained. Ways in which the findings of the current study relate to the theoretical perspectives and prior research presented in the introduction also are typically addressed. In this section, authors also acknowledge the limitations of the current study. The discussion section may also address potential applications of the work or suggest future research.

Referring to Others' Work

It is an author's job to avoid plagiarism by noting when reference is made to another's work or ideas. This is true even when making general statements about existing knowledge (e.g., "Self-efficacy impacts many aspects of students' lives, including achievement motivation and task persistence (Bandura, 1997)."). Citations allow a reader to be aware of the original source of ideas or data and direct the reader toward sources of additional information on a topic.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

In-Text Citations

Throughout the manuscript text, credit should be given to authors whose work is referenced. In-text citations allow the reader to be aware of the source of an idea and locate the work in the reference list at the end of the manuscript. APA style uses an author–date citation method; each in-text citation includes the author's (or authors') last name(s) and the year of publication. For works with two authors, both authors' names are given. For works with three or more authors, the first author is listed by name followed by "et al.", meaning "and others." If multiple works are cited in a single in-text citation, they are listed alphabetically and separated with semicolons (e.g., Greenhoot et al., 2013; Rojas & Yoshikawa, 2017; Tucker, 2016). When a direct quotation from a source is presented, the page number of the quotation is presented along with the author and date (e.g., Bandura, 1997, p. 22).

Reference Lists

References are listed alphabetically by the last name of the first author. Citations in the reference list include names of authors, article or chapter title, journal or book title, and page numbers (if relevant). For sources accessed electronically, information regarding means of electronic access (web address or DOI) is provided. The APA Publication Manual includes guidelines for citing many different types of sources. Examples of some of the most common types of references appear below.

Book: Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman and Company.

Chapter in edited book: Rojas, N., & Yoshikawa, H. (2017). Documentation status and child development in the U.S. and Europe. In N. J. Cabrera & B. Leyendecker (Eds.), *Handbook on positive development of minority children and youth* (pp. 385–400). Springer. doi:10.1007/978-3-319-43645-6_23.

Journal article: Greenhoot, A. F., Sun, S., Bunnell, S. L., & Lindboe, K. (2013). Making sense of traumatic memories: Memory qualities and psychological symptoms in emerging adults with and without abuse histories. *Memory*, *21*(1), 125–142. doi:10.1080/09658211.2012.712975.

Article in periodical (magazine or newspaper): Tucker, J. (2016, March 9). Does social science have a replication crisis? *The Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2016/03/09/does-social-science-have-a-replication-crisis/

Research report: United States Census Bureau. (2017). Voting and registration in the election of November 2016. https://www.census.gov/data/tables/time-series/demo/voting-and-registration/p20-580.html

Meagan M. Patterson

See also Abstract; Bias; Demographics; Discussion Section; Dissertation; Methods Section; Results Section

Further Readings

American Psychological Association. (2010). *Preparing manuscripts for publication in psychology journals: A guide for new authors*. American Psychological Association. https://www.apa.org/pubs/authors/new-author-guide.pdf

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington, DC: American Psychological Association.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. doi:10.1037/0003-066X.63.9.839.

Cone, J. D., & Foster, S. L. (2006). *Dissertation and theses from start to finish: Psychology and related fields* (2nd ed.). Washington, DC: American Psychological Association.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 26–46. doi:10.1037/amp0000151.

Sternberg, R. J. (2019). *Guide to publishing in psychology journals* (2nd ed.). Cambridge, UK: Cambridge University Press.

American Statistical Association

The American Statistical Association (ASA) is a society for scientists, statisticians, and statistics consumers representing a wide range of science and education fields. Since its inception in November 1839, the ASA has aimed to provide both statistical science professionals and the public with a standard of excellence for statistics-related projects. According to ASA publications, the society's mission is "to promote excellence in the application of statistical science across the wealth of human endeavor." Specifically, the ASA mission includes a dedication to excellence with regard to statistics in practice, research, and education; a desire to

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

work toward bettering statistical education and the profession of statistics as a whole; a concern for recognizing and addressing the needs of ASA members; education about the proper uses of statistics; and the promotion of human welfare through the use of statistics.

Regarded as the second-oldest, continuously operating professional association in the United States, the ASA has a rich history. In fact, within 2 years of its founding, the society already had a U.S. president—Martin Van Buren—among its members. Also on the list of the ASA's historical members are Florence Nightingale, Alexander Graham Bell, and Andrew Carnegie. The original founders, who united at the American Education Society in Boston to form the society, include U.S. Congressman Richard Fletcher; teacher and fundraiser William Cogswell; physician and medicine reformist John Dix Fisher; statistician, publisher, and distinguished public health author Lemuel Shattuck; and lawyer, clergyman, and poet Oliver Peabody. The founders named the new organization the American Statistical Society, a name that lasted only until the first official meeting in February 1840.

In its beginning years, the ASA developed a working relationship with the U.S. Census Bureau, offering recommendations and often lending its members as heads of the census. S. N. D. North, the 1910 president of the ASA, was also the first director of the permanent census office. The society, its membership, and its diversity in statistical activities grew rapidly after World War I as the employment of statistics in business and government gained popularity. At that time, large cities and universities began forming local chapters. By its 100th year in existence, the ASA had more members than it had ever had, and those involved with the society commemorated the centennial with celebrations in Boston and Philadelphia. However, by the time World War II was well underway, many of the benefits the ASA experienced from the post–World War I surge were reversed. For 2 years—1942 and 1943—the society was unable to hold annual meetings. Then, after World War II, as after World War I, the ASA saw a great expansion in both its membership and applications to burgeoning science endeavors.

Today, ASA has expanded beyond the United States and can count 18,000 individuals as members. Its members, who represent 78 geographic locations, also have diverse interests in statistics. These interests range from finding better ways to teach statistics to problem-solving for homelessness and from AIDS research to space exploration, among a wide array of applications. As an organization, the ASA frequently produces white papers and policy recommendations to guide the use of statistics in research. One instance is a 2020 report questioning the appropriateness of relying on levels of significance when interpreting research results. Additionally, the organization frequently provides service and guidance for national and international governmental applications of statistical science. For example, the association's board published standards for statistical sampling quality for the U.S. 2020 Census.

The society comprises 24 sections, including the following: Bayesian Statistical Science, Biometrics, Biopharmaceutical Statistics, Business and Economic Statistics, Government Statistics, Health Policy Statistics, Nonparametric Statistics, Physical and Engineering Sciences, Quality and Productivity, Risk Analysis, Statistical Programmers and Analysts, Statistical Learning and Data Mining, Social Statistics, Statistical Computing, Statistical Consulting, Statistical Education, Statistical Graphics, Statistics and the Environment, Statistics in Defense and National Security, Statistics in Epidemiology, Statistics in Marketing, Statistics in Sports, Survey Research Methods, and Teaching of Statistics in the Health Sciences. Detailed descriptions of each section, lists of current officers within each section, and links to each section are available on the ASA website.

In addition to holding meetings coordinated by more than 60 committees of the society, the ASA sponsors scholarships, fellowships, workshops, and educational programs. Its leaders and members also advocate for statistics research funding and offer a host of career services and outreach projects.

Publications from the ASA include scholarly journals, statistical magazines, books, research guides, brochures, and conference proceeding publications. Among the journals available are *The American Statistician*; *Journal of Agricultural, Biological, and Environmental Statistics*; *Journal of the American Statistical Association*; *Journal of Business and Economic Statistics*; *Journal of Computational and Graphical Statistics*; *Journal of Educational and Behavioral Statistics*; *Journal of Statistics Education*; *Statistical Analysis and Data Mining*; *Statistics in Biopharmaceutical Research*; *Journal of Survey Statistics and Methodology*; and *Technometrics*.

The official website of the ASA offers a more comprehensive look at the mission, history, publications, activities, and future directions of the society. Additionally, browsers can find information about upcoming meetings and events, descriptions of outreach and initiatives, the ASA bylaws and constitution, a copy of the ethical guidelines for statistical practice prepared by the committee on professional ethics, and an organizational list of board members and leaders.

Kristin Rasmussen Teasdale

See also American Educational Research Association; American Psychological Association Style; Databases; Ethics in the Research Process; Statistic

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Further Readings

Koren, J. (1970). *The history of statistics, their development and progress in many countries*. New York, NY: B. Franklin. (Original work published 1918)

Mason, R. L. (1999). *ASA: The first 160 years*. Retrieved October 10, 2009, from http://www.amstat.org/about/first160years.cfm

Wasserstein, R. L., & Lazar, N. A. (2020). ASA statement on statistical significance and P-values. In C. W. Gruber (ed.), *The theory of statistics in psychology* (pp. 1–10). New York: Springer.

Wilcox, W. F. (1940). Lemuel Shattuck, statist, founder of the American Statistical Association. *Journal of the American Statistical Association*, *35*, 224–235.

Websites

American Statistical Association: http://www.amstat.org

Analysis of Covariance (Ancova)

Behavioral sciences rely heavily on experiments and quasi experiments for evaluating the effects of, for example, new therapies, instructional methods, or stimulus properties. An *experiment* includes at least two different treatments (conditions), and human participants are randomly assigned one treatment. If assignment is not based on randomization, the design is called a *quasi experiment*. The dependent variable or outcome of an experiment or a quasi experiment, denoted by *Y* here, is usually quantitative, such as the total score on a clinical questionnaire or the mean response time on a perceptual task. Treatments are evaluated by comparing them with respect to the mean of the outcome *Y* using either analysis of variance (ANOVA) or analysis of covariance (ANCOVA). Multiple linear regression may also be used, and categorical outcomes require other methods, such as logistic regression. This entry explains the purposes of, and assumptions behind, ANCOVA for the classical two-group between-subjects design. ANCOVA for within-subject and split-plot designs is discussed briefly at the end.

Researchers often want to control or adjust statistically for some independent variable that is not experimentally controlled, such as gender, age, or a pretest value of *Y*. A categorical variable such as gender can be included in ANOVA as an additional factor, turning a one-way ANOVA into a two-way ANOVA. A quantitative variable such as age or a pretest recording can be included as a covariate, turning ANOVA into ANCOVA. ANCOVA is the bridge from ANOVA to multiple regression. There are two reasons for including a covariate in the analysis if it is predictive of the outcome *Y*. In randomized experiments, it reduces unexplained (within-group) outcome variance, thereby increasing the power of the treatment effect test and reducing the width of its confidence interval. In quasi experiments, it adjusts for a group difference with respect to that covariate, thereby adjusting the between-group difference on *Y* for confounding.

Model

The ANCOVA model for comparing two groups at posttest *Y*, using a covariate *X*, is as follows:

$$Y_{ij} = \mu + \alpha_j + \beta(X_{ij} - \bar{X}) + e_{ij}, \quad (1)$$

where Y_{ij} is the outcome for person i in group j (e.g., $j = 1$ for control, $j = 2$ for treated), and X_{ij} is the covariate value for person i in group j, μ is the grand mean of *Y*, α_j is the effect of treatment j, β is the slope of the regression line for predicting *Y* from *X* within groups, $\bar{X}$ is the overall sample mean of covariate *X*, and e_{ij} is a normally distributed residual or error term with a mean of zero and a variance σ_e^2, which is the same in both groups. By definition, $\alpha_1 + \alpha_2 = 0$, and so $\alpha_2 - \alpha_1 = 2\alpha_2$ is the expected posttest group difference adjusted for the covariate *X*. This is even better seen by rewriting Equation 1 as

$$Y_{ij} - \beta\left(X_{ij} - \bar{X}\right) = \mu + \alpha_j + e_{ij} \quad (2)$$

showing that ANCOVA is ANOVA of *Y* adjusted for *X*. Due to the centering of *X*, that is, the subtraction of $\bar{X}$, the adjustment is on the average zero in the total sample. So the centering affects individual outcome values and group means, but not the total or grand mean μ of *Y*.

ANCOVA can also be written as a multiple regression model:

$$Y_{ij} = \beta_0 + \beta_1 G_{ij} + \beta_2 X_{ij} + e_{ij} \quad (3)$$

where G_{ij} is a binary indicator of treatment group (G_{i1}=0 for controls, G_{i2}=1 for treated), and β_2 is the slope β in Equation 3. Comparing Equation 1 with Equation 3 shows that $\beta_1 = 2\alpha_2$ and that $\beta_0 = \left(\mu - \alpha_2 - \beta\bar{X}\right)$. Centering in Equation 3 both *G* and *X* (i.e., coding *G* as 1 and +1, and subtracting $\bar{X}$ from *X*) will give $\beta_0 = \mu$ and $\beta_1 = \alpha_2$. Application of ANCOVA requires estimation of β in Equation 1. Its least squares solution is $\frac{\sigma_{XY}}{\sigma_X^2}$, the within-group covariance between pre- and posttest, divided by the within-group pretest variance, which in turn are both estimated from the sample.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Assumptions

As Equations 1 and 3 show, ANCOVA assumes that the covariate has a linear effect on the outcome and that this effect is homogeneous, the same in both groups. So there is no treatment by covariate interaction. Both the linearity and the homogeneity assumption can be tested and relaxed by adding to Equation 3 as predictors $X \times X$ and $G \times X$, respectively, but this entry concentrates on the classical model, Equation 1 or Equation 3. The assumption of homogeneity of residual variance σ_e^2 between groups can also be relaxed.

Another assumption is that X is not affected by the treatment. Otherwise, X must be treated as a mediator instead of as a covariate, with consequences for the interpretation of analysis with versus without adjustment for X. If X is measured before treatment assignment, this assumption is warranted.

A more complicated ANCOVA assumption is that X is measured without error, where *error* refers to intra-individual variation across replications. This assumption will be valid for a covariate such as age but not for a questionnaire or test score, in particular not for a pretest of the outcome at hand. Measurement error in X leads to *attenuation*, a decrease of its correlation with Y and of its slope β in Equation 1. This leads to a loss of power in randomized studies and to bias in nonrandomized studies.

A last ANCOVA assumption that is often mentioned, but not visible in Equation 1, is that there is no group difference on X. This seems to contradict one of the two purposes of ANCOVA, that is, adjustment for a group difference on the covariate. The answer is simple, however. The assumption is not required for covariates that are measured without measurement error, such as age. But if there is measurement error in X, then the resulting underestimation of its slope β in Equation 1 leads to biased treatment effect estimation in case of a group difference on X. An exception is the case of treatment assignment based on the observed covariate value. In that case, ANCOVA is unbiased in spite of measurement error in X, whether groups differ on X or not, and any attempt at correction for attenuation will then introduce bias. The assumption of no group difference on X is addressed in more detail in a special section on the use of a pretest of the outcome Y as covariate.

Purposes

The purpose of a covariate in ANOVA depends on the design. To understand this, note that ANCOVA gives the following adjusted estimator of the group difference:

$$\widehat{\Delta} = (\overline{Y}_2 - \overline{Y}_1) - \beta(\overline{X}_2 - \overline{X}_1). \quad (4)$$

In a randomized experiment, the group difference on the covariate, $\left(\overline{X}_1 - \overline{X}_2\right)$, is zero, and so the adjusted difference $\widehat{\Delta}$ is equal to the unadjusted difference $(\overline{Y}_2 - \overline{Y}_1)$, apart from sampling error. In terms of ANOVA, the mean square (*MS;* treatment) is the same with or without adjustment, again apart from sampling error. Things are different for the *MS* (error), which is the denominator of the *F* test in ANOVA. ANCOVA estimates β such that the *MS* (error) is minimized, thereby maximizing the power of the *F* test. Since the standard error (*SE*) of $\widehat{\Delta}$ is proportional to the square root of the *MS* (error), this *SE* is minimized, leading to more precise effect estimation by covariate adjustment.

In a nonrandomized study with groups differing on the covariate X, the covariate-adjusted group effect $\widehat{\Delta}$ systematically differs from the unadjusted effect $(\overline{Y}_2 - \overline{Y}_1)$. It is unbiased if the ANCOVA assumptions are satisfied and treatment assignment is random conditional on the covariate, that is, random within each subgroup of persons who are homogeneous on the covariate. Although the *MS* (error) is again minimized by covariate adjustment, this does not imply that the *SE* of $\widehat{\Delta}$ is reduced. This *SE* is a function not only of *MS* (error), but also of treatment–covariate correlation. In a randomized experiment, this correlation is zero apart from sampling error, and so the *SE* depends only on the *MS* (error) and sample size. In nonrandomized studies, the *SE* increases with treatment–covariate correlation and can be larger with than without adjustment. But in nonrandomized studies, the primary aim of covariate adjustment is correction for bias, not a gain of power.

The two purposes of ANCOVA are illustrated in Figures 1 and 2, showing the within-group regressions of outcome Y on covariate X, with the ellipses summarizing the scatter of individual persons around their group line. Each group has its own regression line with the same slope β (reflecting absence of interaction) but different intercepts. In Figure 1, of a nonrandomized study, the groups differ on the covariate. Moving the markers for both group means along their regression line to a common covariate value $\overline{X}$ gives the adjusted group difference $\widehat{\Delta}$ on outcome Y, reflected by the vertical distance between the two lines, which is also the difference between both intercepts. In Figure 2, of a randomized study, the two groups have the same mean covariate value, and so unadjusted and adjusted group difference on Y are the

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

same. However, in both figures the adjustment has yet another effect, illustrated in Figure 2. The *MS* (error) of ANOVA without adjustment is the entire within-group variance in vertical direction, ignoring regression lines. The *MS* (error) of ANCOVA is the variance of the vertical distances of individual dots from their group regression line. All variation in the *Y*-direction that can be predicted from the covariate; that is, all increase of *Y* along the line is included in the unadjusted *MS* (error) but excluded from the adjusted *MS* (error), which is thus smaller. In fact, it is only $\left(1-\rho_{XY}^2\right)$ as large as the unadjusted *MS* (error), where ρ_{XY} is the within- group correlation between outcome and covariate.

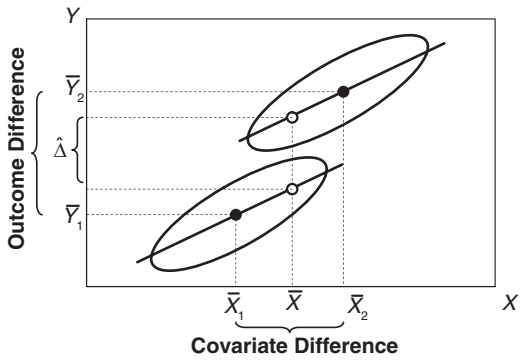

Figure 1 Adjustment of the Outcome Difference Between Groups for a Covariate Difference in a Nonrandomized Study

Notes: Regression lines for treated (upper) and untreated (lower) group. Ellipses indicate scatter of individuals around their group lines. Markers on the lines indicate unadjusted (solid) and adjusted (open) group means.

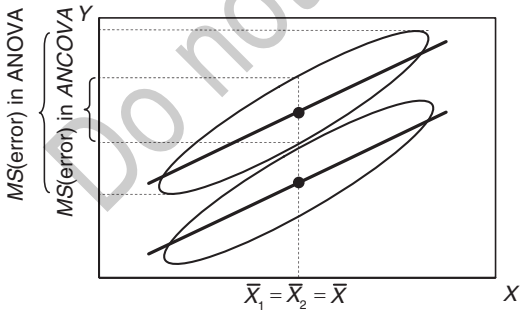

Figure 2 Reduction of Unexplained Outcome Variance by Covariate Adjustment in a Randomized Study

Notes: Vertical distance between upper and lower end of an ellipse indicates unexplained (within-group) outcome variance in ANOVA. Vertical distance within an ellipse indicates unexplained outcome variance in ANCOVA.

Using a Pretest of the Outcome as Covariate

An important special case of ANCOVA is that in which a pretest measurement of *Y* is used as covariate. The user can then choose between two methods of analysis:

1. ANCOVA with the pretest as covariate and the posttest as outcome.
2. ANOVA with the change score (posttest minus pretest) as outcome.

Two other popular methods come down to either of these two: ANCOVA of the change score is equivalent to Method 1. The Group × Time interaction test in a repeated measures ANOVA with pretest and posttest as repeated measures is equivalent to Method 2. So the choice is between Methods 1 and 2 only. Note that Method 2 is a special case of Method 1 in the sense that choosing $\beta = 1$ in ANCOVA gives ANOVA of change, as Equation 2 shows. In a randomized experiment, there is no pretest group difference, and both methods give the same unbiased treatment effect apart from sampling error, as Equation 4 shows. However, ANCOVA gives a smaller *MS* (error), leading to more test power and a smaller confidence interval than ANOVA of change, except if $\beta \approx 1$ in ANCOVA and the sample size *N* is small. In nonrandomized studies, the value for β in Equation 4 does matter, and ANCOVA gives a different treatment effect than does ANOVA of change. The two methods may even lead to contradictory conclusions, which is known as *Lord's ANCOVA paradox*. The choice between the two methods then depends on the assignment procedure. This is best seen by writing both as a repeated measures model.

ANOVA of change is equivalent to testing the Group × Time interaction in the following model (where regression weights are denoted by γ to distinguish them from the βs in earlier equations):

$$Y_{ijt} = \gamma_0 + \gamma_1 G_{ij} + \gamma_2 T_{it} + \gamma_3 G_{ij} T_{it} + e_{ijt} \qquad (5)$$

Here, Y_{ijt} is the outcome value of person *i* in group *j* at time *t*, *G* is the treatment group (0 = control, 1 = treated), *T* is the time (0 = pretest, 1 = posttest), and e_{ijt} is a random person effect with an unknown 2 × 2 within-group covariance matrix Σ of pre- and posttest measures. By filling in the 0 or 1 values for *G* and *T*, one can see that γ_0 is the pretest (population) mean of the control group, γ_1 is the pretest mean difference between the groups, γ_2 is the mean change in the control group, and γ_3 is the difference in mean change between groups. Testing the interaction effect γ_3 in Equation 5

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

is therefore equivalent to testing the group effect on change ($Y - X$). The only difference between repeated measures ANOVA and Equation 5 is that ANOVA uses (−1, +1) instead of (0, 1) coding for G and T.

ANCOVA can be shown to be equivalent to testing γ_3 in Equation 5 after deleting the term $\gamma_1 G_{ij}$ by assuming $\gamma_1 = 0$, which can be done with mixed (multilevel) regression. So ANCOVA assumes that there is no group difference at pretest. This assumption is satisfied by either of two treatment assignment procedures: (1) randomization and (2) assignment based on the pretest X. Both designs start with one group of persons so that there can be no group effect at pretest. Groups are created after the pretest. This is why ANCOVA is the best method of analysis for both designs. In randomized experiments it has more power than ANOVA of change. With treatment assignment based on the pretest such that $\bar{X}_1 \neq \bar{X}_2$, ANCOVA is unbiased whereas ANOVA of change is then biased by ignoring regression to the mean. In contrast, if naturally occurring or preexisting groups are assigned, such as Community A getting some intervention and Community B serving as control, then ANCOVA will usually be biased whereas ANOVA of change may be unbiased. A sufficient set of conditions for ANOVA of change to be unbiased, then, is (a) that the groups are random samples from their respective populations and (b) that without treatment these populations change equally fast (or not at all). The bias in ANCOVA for this design is related to the issue of underestimation of β in Equation 1 due to measurement error in the covariate. Correction for this underestimation gives, under certain conditions, ANOVA of change. In the end, however, the correct method of analysis for nonrandomized studies of preexisting groups is a complicated problem because of the risk of hidden confounders. Having two pretests with a suitable time interval and two control groups is then recommended to test the validity of both methods of analysis. More specifically, treating the second pretest as posttest or treating the second control group as experimental group should not yield a significant group effect because there is no treatment.

Covariates in Other Popular Designs

This section discusses covariates in within-subject designs (e.g., *crossovers*) and between-subject designs with repeated measures (i.e., a *split-plot design*).

A within-subject design with a quantitative outcome can be analyzed with repeated measures ANOVA, which reduces to Student's paired t test if there are only two treatment conditions. If a covariate such as age or a factor such as gender is added, then repeated measures ANOVA with two treatments comes down to applying ANCOVA twice: (1) to the within-subject difference D of both measurements (within-subject part of the ANOVA) and (2) to the within-subject average A of both measurements (between-subject part of the ANOVA). ANCOVA of A tests the main effects of age and gender. ANCOVA of D tests the Treatment × Gender and Treatment × Age interactions and the main effect of treatment. If gender and age are centered as in Equation 1, this main effect is μ in Equation 1, the grand mean of D. If gender and age are not centered, as in Equation 3, the grand mean of D equals $\beta_0 + \beta_1 \bar{G} + \beta_2 \bar{X}$, where G is now gender and X is age. The most popular software, SPSS (an IBM company, formerly called PASW® Statistics), centers factors (here, gender) but not covariates (here, age) and tests the significance of β_0 instead of the grand mean of D when reporting the F test of the within-subject main effect. The optional pairwise comparison test in SPSS tests the grand mean of D, however.

Between-subject designs with repeated measures, for example, at posttest and follow-ups or during and after treatment, also allow covariates. The analysis is the same as for the within-subject design extended with gender and age. But interest now is in the Treatment (between-subject) × Time (within-subject) interaction and, if there is no such interaction, in the main effect of treatment averaged across the repeated measures, rather than in the main effect of the within-subject factor time. A pretest recording can again be included as covariate or as repeated measure, depending on the treatment assignment procedure. Note, however, that as the number of repeated measures increases, the F test of the Treatment × Time interaction may have low power. More powerful are the Treatment × Linear (or Quadratic) Time effect test and discriminant analysis.

Within-subject and repeated measures designs can have not only between-subject covariates such as age but also within-subject or time-dependent covariates. Examples are a baseline recording within each treatment of a crossover trial, and repeated measures of a mediator. The statistical analysis of such covariates is beyond the scope of this entry, requiring advanced methods such as mixed (multilevel) regression or structural equations modeling, although the case of only two repeated measures allows a simpler analysis by using as covariates the within-subject average and difference of the original covariate.

Practical Recommendations for the Analysis of Studies With Covariates

Based on the preceding text, the following recommendations can be given: In randomized studies, covariates

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

should be included to gain power, notably a pretest of the outcome. Researchers are advised to center covariates and check linearity and absence of treatment–covariate interaction as well as normality and homogeneity of variance of the residuals. In nonrandomized studies of preexisting groups, researchers should adjust for covariates that are related to the outcome to reduce bias. With two pretests or two control groups, researchers should check the validity of ANCOVA and ANOVA of change by treating the second pretest as posttest or the second control group as experimental group. No group effect should then be found. In the real posttest analysis, researchers are advised to use the average of both pretests as covariate since this average suffers less from attenuation by measurement error. In nonrandomized studies with only one pretest and one control group, researchers should apply ANCOVA and ANOVA of change and pray that they lead to the same conclusion, differing in details only.

Additionally, if there is substantial dropout related to treatment or covariates, then all data should be included in the analysis to prevent bias, using mixed (multilevel) regression instead of traditional ANOVA to prevent listwise deletion of dropouts. Further, if pretest data are used as an inclusion criterion in a nonrandomized study, then the pretest data of all excluded persons should be included in the effect analysis by mixed regression to reduce bias.

Gerard J. P. Van Breukelen

See also Analysis of Variance (ANOVA); Covariate; Experimental Design; Gain Scores, Analysis of; Pretest–Posttest Design; Quasi-Experimental Design; Regression Artifacts; Split-Plot Factorial Design

Further Readings

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.).

Frison, L., & Pocock, S. (1997). Linearly divergent treatment effects in clinical trials with repeated measures: Efficient analysis using summary statistics. *Statistics in Medicine*, *16*, 2855–2872.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, *6*, 115–134.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Pacific Grove, CA: Brooks/Cole.

Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child & Adolescent Psychology*, *32*(3), 467–486.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–205). Boston, MA: Houghton-Mifflin.

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer.

Senn, S. J. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, *25*, 4334–4344.

Senn, S., Stevens, L., & Chaturvedi, N. (2000). Repeated measures in clinical trials: Simple strategies for analysis using summary measures. *Statistics in Medicine*, *19*, 861–877.

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*, 920–925.

Winkens, B., Van Breukelen, G. J. P., Schouten, H. J. A., & Berger, M. P. F. (2007). Randomized clinical trials with a pre- and a post-treatment measurement: Repeated measures versus ANCOVA models. *Contemporary Clinical Trials*, *28*, 713–719.

Analysis of Variance (Anova)

Usually a two-sample *t* test is applied to test for a significant difference between two population means based on the two samples. For example, consider the data in Table 1. Twenty patients with high blood pressure are randomly assigned to two groups of 10 patients. Patients in Group 1 are assigned to receive placebo, while patients in Group 2 are assigned to receive Drug A. Patients' systolic blood pressures (SBPs) are measured before and after treatment, and the differences in SBPs are recorded in Table 1. A two-sample *t* test would be an efficient method for testing the hypothesis that drug A is more effective than placebo when the differences in before and after measurements are normally distributed. However, there are usually more than two groups involved for comparison in many fields of scientific investigation. For example, extend the data in Table 1 to the data in Table 2. Here the study used 30 patients who are randomly assigned to placebo, Drug A, and Drug B. The goal here is to compare the effects of placebo and experimental drugs in reducing SBP. But a two-sample *t* test is not applicable here as we have more than two groups. Analysis

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Table 1 Comparison of Two Treatments Based on Systolic Blood Pressure Change

Treatment	
Placebo	*Drug A*
–1.3	–4.0
–1.5	–5.7
–0.5	–3.5
0.8	0.4
–1.1	–1.3
3.4	0.8
–0.8	–10.7
–3.6	–0.3
0.3	–0.5
–2.2	–3.3

of variance (ANOVA) generalizes the idea of the two-sample *t* test so that normally distributed responses can be compared across categories of one or more factors.

Since its development, ANOVA has played an indispensable role in the application of statistics in many fields, such as biology, social sciences, finance, pharmaceutics, and scientific and industrial research. Although ANOVA can be applied to various statistical models, and the simpler ones are usually named after the number of categorical variables, the concept of ANOVA is based solely on identifying the contribution of individual factors in the total variability of the data. In the above example, if the variability in SBP changes due to the drug is large compared with the chance variability, then one would think that the effect of the drug on SBP is substantial. The factors could be different individual characteristics, such as age, sex, race, occupation, social class, and treatment group, and the significant differences between the levels of these factors can be assessed by forming the ratio of the variability due to the factor itself and that due to chance only.

History

As early as 1925, R. A. Fisher first defined the methodology of ANOVA as "separation of the variance ascribable to one group of causes from the variance ascribable to other groups" (p. 216). Henry Scheffé defined ANOVA as "a statistical technique for analyzing measurements depending on several kinds of effects operating simultaneously, to decide which kinds of effects are important and to estimate the effects. The measurements or observations may be in an experimental science like genetics or a nonexperimental one like astronomy" (p. 3). At first, this methodology focused more on comparing the means while treating variability as a nuisance. Nonetheless, since its introduction, ANOVA has become the most widely used statistical methodology for testing the significance of treatment effects.

Based on the number of categorical variables, ANOVA can be distinguished into one-way ANOVA and two-way ANOVA. Besides, ANOVA models can also be separated into a fixed-effects model, a random-effects model, and a mixed model based on how the factors are chosen during data collection. Each of them is described separately.

One-Way ANOVA

One-way ANOVA is used to assess the effect of a single factor on a single response variable. When the factor is a fixed factor whose levels are the only ones of interest, one-way ANOVA is also referred to as fixed-effects one-way ANOVA. When the factor is a random factor whose levels can be considered as a sample from the population of levels, one-way ANOVA is referred to as random-effects one-way ANOVA. Fixed-effects one-way ANOVA is applied to answer the question of whether the population means are equal or not.

Table 2 Comparison of Three Treatments Based on Systolic Blood Pressure Change

Treatment		
Placebo	*Drug A*	*Drug B*
–1.3	–4.0	–7.6
0.5	–5.7	–9.2
–0.5	–3.5	–4.0
0.4	0.4	1.8
–1.1	–1.3	–5.3
0.6	0.8	2.6
–0.8	–10.7	–3.8
–3.6	–0.3	1.2
0.3	–0.5	0.4
–2.2	–3.3	–2.6

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Given k population means, the null hypothesis can be written as

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad (1)$$

The alternative hypothesis, H_a, can be written as
Ha : k population means are not all equal
In the random-effects one-way ANOVA model, the null hypothesis tested is that the random effect has zero variability.

Four assumptions must be met for applying ANOVA:

A1: All samples are simple random samples drawn from each of k populations representing k categories of a factor.

A2: Observations are independent of one another.

A3: The dependent variable is normally distributed in each population.

A4: The variance of the dependent variable is the same in each population.

Suppose, for the jth group, the data consist of the n_j measurements $Y_{j1}, Y_{j2},\ldots,Y_{n_j}, j = 1, 2,\ldots,k$. Then the total variation in the data can be expressed as the corrected sum of squares (*SS*) as follows: $TSS = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \overline{y})^2$, where $\overline{y}$ is the mean of the overall sample. On the other hand, variation due to the factor is given by

$$SST = \sum_{j=1}^{k}(\overline{y}_j - \overline{y})^2, \quad (2)$$

where $\overline{y}_j$ is the mean from the jth group. The variation due to chance (error) is then calculated as *SSE* (error sum of squares) = *TSS* – *SST*. The component variations are usually presented in a table with corresponding degrees of freedom (*df*), mean square error, and *F* statistic. A table for one-way ANOVA is shown in Table 3.

For a given level of significance α, the null hypothesis H_0 would be rejected and one could conclude that k population means are not all equal if

$$F \geq F_{k-1,n-k,1-\alpha} \quad (3)$$

where $F_{k-1,n-k,1-\alpha}$ is the $100(1-\alpha)$% point of *F* distribution with k – 1 and n – k *df*.

Two-Way ANOVA

Two-way ANOVA is used to assess the effects of two factors and their interaction on a single response variable. There are three cases to be considered: the fixed-effects case, in which both factors are fixed; the random-effects case, in which both factors are random; and the mixed-effects case, in which one factor is fixed and the other factor is random. Two-way ANOVA is applied to answer the question of whether Factor A has a significant effect on the response adjusted for Factor B, whether Factor B has a significant effect on the response adjusted for Factor A, or whether there is an interaction effect between Factor A and Factor B.

Table 3 General ANOVA Table for One-Way ANOVA (k populations)

Source	*d.f.*	*SS*	*MS*	*F*
Between	k – 1	*SST*	$MST = \frac{SST}{k-1}$	$\frac{MST}{MSE}$
Within	n – k	*SSE*	$MSE = \frac{SSE}{n-k}$	
Total	n – 1	*TSS*		

Note: n=sample size; k=number of groups; *SST*=sum of squares treatment (factor); *MST*=mean square treatment (factor); *SSE*=sum of squares error; *TSS*=total sum of squares.

All null hypotheses can be written as

1. H_{01} There is no Factor A effect.
2. H_{02} There is no Factor B effect.
3. H_{03} There is no interaction effect between Factor A and Factor B.

The ANOVA table for two-way ANOVA is shown in Table 4.

In the fixed case, for a given α, the null hypothesis H_{01} would be rejected, and one could conclude that there is a significant effect of Factor A if

$$F(Factor\ A) \geq F_{r-1,rc(n-1),1-\alpha}, \quad (4)$$

where $F_{r-1,rc(n-1),1-\alpha}$ is the $100(1-\alpha)$% point of *F* distribution with r – 1 and $rc(n-1)$ *df*.

The null hypothesis H_{02} would be rejected, and one could conclude that there is a significant effect of Factor B if

$$F(FactorB) \geq F_{c-1,rc(n-1),1-\alpha} \quad (5)$$

where $F_{c-1,rc(n-1),1-\alpha}$ is the $100(1-\alpha)$% point of *F* distribution with c – 1 and $rc(n-1)$ *df*.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Table 4 General Two-Way ANOVA Table

Source	d.f.	SS	MS	F	
				Fixed	Mixed or Random
Factor A (main effect)	$r-1$	SSR	$MSR=\frac{SSR}{r-1}$	$\frac{MSR}{MSE}$	$\frac{MSR}{MSRC}$
Factor B (main effect)	$c-1$	SSC	$MSC=\frac{SSC}{c-1}$	$\frac{MSC}{MSE}$	$\frac{MSC}{MSRC}$
Factor A XFactor B (interaction)	$(r-1)(c-1)$	$SSRC$	$MSRC=\frac{SSRC}{(r-1)(c-1)}$	$\frac{MSRC}{MSE}$	$\frac{MSRC}{MSE}$
Error	$rc(n-1)$	SSE	$MSE=\frac{SSE}{rc(n-1)}$		
Total	$rcn-1$	TSS			

Note: r=number of groups for A; c=number of groups for B; SSR=sum of squares for Factor A; MSR=mean sum of squares for Factor A; $MSRC$=mean sum of squares for the Interaction A×B; SSC=sum of squares for Factor B; MSC=mean square for Factor B; $SSRC$=sum of squares for the Interaction $A×B$; SSE=sum of squares error; TSS=total sum of squares.

The null hypothesis H_{03} would be rejected, and one could conclude that there is a significant effect of interaction between Factor A and Factor B if

$$F(Factor\,A\times Factor\,B)\geq F_{(r-1)(c-1),rc(n-1),1-\alpha}, \quad (6)$$

where $F_{(r-1)(c-1),rc(n-1),1-\alpha}$ is the $100(1-\alpha)\%$ point of F distribution with $(r-1)(c-1)$ and $rc(n-1)$ df.

It is similar in the random case, except for different F statistics and different df for the denominator for testing H_{01} and H_{02}.

Statistical Packages

SAS procedure "PROC ANOVA" performs ANOVA for balanced data from a wide variety of experimental designs. The "anova" command in STATA fits ANOVA and analysis of covariance (ANCOVA) models for balanced and unbalanced designs, including designs with missing cells; models for repeated measures ANOVA; and models for factorial, nested, or mixed designs. The "anova" function in S-PLUS produces a table with rows corresponding to each of the terms in the object, plus an additional row for the residuals. When two or more objects are used in the call, a similar table is produced showing the effects of the pairwise differences between the models, considered sequentially from the first to the last. SPSS (an IBM company, formerly called PASW® Statistics) provides a range of ANOVA options, including automated follow-up comparisons and calculations of effect size estimates.

Abdus S. Wahed and Xinyu Tang

See also Analysis of Covariance (ANCOVA); Repeated Measures Design; *t* Test, Independent Samples

Further Readings

Bogartz, R. S. (1994). *An introduction to the analysis of variance*. Westport, CT: Praeger.

Fisher, R. A. (1938). *Statistical methods for research workers* (7th ed.). Edinburgh, Scotland: Oliver and Boyd.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1998). *Applied regression analysis and other multivariable methods* (3rd ed.). Pacific Grove, CA: Duxbury Press.

Lindman, H. R. (1992). *Analysis of variance in experimental design*. New York: Springer-Verlag.

Scheffé, H. (1999). *Analysis of variance*. New York: Wiley-Interscience.

Animal Research

Animal research refers to any investigation that includes animals as the research subject or population in an investigation; therefore, there is no specific theory that

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

supports all animal research. The term *animal research* refers only to the sample, not the mechanism or the method. Animal research may be basic and/or applied, it might be direct—to better understand the animal species of interest—or it could be comparative, wherein the animal is used as a model for another species. In addition, animal research occurs in differing contexts, which might include the animal's natural habitat, like field settings, in an artificial natural habitat, like zoological gardens and aquariums, or in purely captive environs, like laboratory or agricultural settings. The empirical question and the context in which the research occurs influences the research design employed and thereby how much control is exerted on the behavior, physiology, or genetics of the animal subject or subjects. This entry discusses animal research regulations and ethics and then provides insight and examples of descriptive and experimental animal research, which are the two most commonly used designs in animal research.

Regulation and Ethics in Animal Research

Although there are several layers of animal research regulation in the United States, and these regulations and guiding principles differ internationally, most countries follow the practice referred to as the *3 Rs*: reduction, refinement, and replacement. The 3 Rs refer to limiting the number of animals subject to disturbance in research (i.e., reduction), improving procedures and protocols to minimize or eliminate pain and stress to animals (i.e., refinement), and to, wherever possible, use computer simulations or alternatives to animals in research projects (i.e., replacement). Animal research is often defined along continuums from nonintrusive to noninvasive, from noninvasive to intrusive, and from intrusive to invasive. Along with that continuum of disturbance or harm, the least intrusive is descriptive research, which occurs in the animal's natural habitat with little to no behavioral disruption. The most invasive may involve genetic or physiological manipulation through experimentation to solve applied problems like cancer, infertility, and disease.

Regardless of whether animal research is descriptive, involving no direct interaction with the animal and thereby constraining data collection to behavioral observation, or biomedical, such that the animal is the recipient of a manipulation either within a control or a treatment condition, all animal research requires ethical and regulatory oversight. In the United States, there are five formal levels of oversight governing animal research, including the Animal Welfare Act (AWA), the Public Health Service (PHS) Policy on Humane Care and Use of Laboratory Animals, Institutional Animal Care and Use Committees (IACUC), the Association for Assessment and Accreditation of Laboratory Animal Care International, and the U.S. Department of Agriculture. These five levels do not include local and state regulations, and private accrediting bodies (e.g., Association of Zoos and Aquariums) that may have their own formalized expectations and laws governing animal care.

The AWA, enacted in 1966, has been amended four times (1970, 1976, 1985, and 1991), each modification elevating the ethical standards regulating the use of animals in research, exhibition, and transport. The most significant revision occurred in 1985, resulting in the establishment of an Animal Welfare Information Center, providing a database for alternatives to animal experiments, and the formalization of an IACUC at all research facilities associated with laboratories, zoos, aquariums, and academic institutions. It is the responsibility of each facility's IACUC to review all protocols involving nonhuman, endothermic (i.e., excludes fish, amphibians, and reptiles) vertebrate animals for ethical compliance. Even field studies that do not involve invasive procedures, harm, or materially alter the behavior of an animal under study are still subject to IACUC review for exemption status.

The ethical guidelines governing the ethics evaluated through IACUC is the *Guide for the Care and Use of Laboratory Animals* (including fish, amphibians, and reptiles), referred to as *The Guide*. Published by the National Research Council and the Institute for Laboratory Animal Research, *The Guide*'s recommendations for the humane care and use of laboratory animals are enforceable through the Health Research Extension Act (HREA). HREA reflects the third level of animal ethics oversight and was passed by Congress in 1985, establishing the PHS Policy on Human Care and Use of Laboratory Animals. HREA asserts that any research facility that receives PHS funds must provide a written plan that complies with the PHS policy and with guidelines set forth in *The Guide*, which is also the basis for the fourth level of oversight governing animal research, the Association for Assessment and Accreditation of Laboratory Animal Care International, a nonprofit organization founded to promote uniform standards of animal care in U.S. laboratories including those accredited throughout the world. The final governing body in the United States, the U.S. Department of Agriculture, is responsible for the enforcement of animal ethics compliance through the AWA, specifically as it relates to animals in agriculture (i.e., livestock), domestic and companion animals (i.e., pets), and licensure of animal research facilities.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Descriptive Animal Research

The least invasive animal research is descriptive, with the most indirect animal investigation involving phylogenies. This method, referred to as the *comparative method*, examines the evolutionary origin of a species' morphological or behavioral traits. Data collected using the comparative phylogenic method may include fossil evidence, archival documents, genetic fingerprinting, bones, behavioral activity budgets, and catalogs, called *ethograms*. Although descriptive observation dates back as far as 40,000 years ago through the depiction of a bull in Lubang Jeriji Saléh cave in East Kalimantan, Borneo, the most cited descriptive animal research began in the early 1960s. Jane Goodall, one of three mentees under paleoanthropologist Louis Leakey, which included Dian Fossey, who studied mountain gorillas in Volcanoes National Park, Rwanda, and Biruté Galdikas, whose subject was orangutans in Tanjung Puting Reserve, in Indonesian Borneo. Goodall, using field observation, ethograms for cataloging behavior, and family pedigrees, established the first objective field protocols for the study of animal behavior. Simple field observation techniques are still in practice, although often in place of investigators, remote cameras, drones, and infrared capture systems are frequently in use to further limit disruption to the animal's natural behavior.

Descriptive field research of animal distribution, density, and migration may involve more intrusive methods for data collection. Territorial range, diving depths, breathing delay, travel routes, and travel speed reflect the dependent variable and may involve radio telemetry, other tracking equipment, and animal transmitters. Depending on the animal target, the transmitter attachments can vary from mildly intrusive, like those data loggers affixed to the dorsal surface of whales using a long pole and suction, or invasive, wherein the animal is darted or shot with a tranquilizer or fitted for either external transmitters through collars, tail, or ear tags, or internal implants. The transmitters may stay with the animal through their life span or break away to be collected in the field. Other descriptive data that can be collected through nonintrusive means include specimen samples from animals like shed fur, hair, scales, cuticles, carapaces, skin, fecal samples, and carcass salvage. The biological data obtained from passive field sampling can provide information about diet, endocrine function, pollutant exposure, genetic diversity, pedigree, and speciation.

In addition to descriptive research in animals' natural habitat, zoological gardens and aquariums provide an important environmental context for descriptive studies, especially those involving animal welfare, education, rehabilitation, and conservation. Beginning in the late 18th century, zoological gardens or parks (i.e., zoos) served to inspire and entertain, providing a glimpse into the diversity and scope of the world's animals. By the 20th century, the focus on entertainment was shared with education and conservation, wherein many zoos established nonprofit foundations to accept injured or orphaned animals with the goal of rehabilitation and reintroduction to their natural habitat. In these environs, descriptive animal research aims to provide approximations of a natural habitat to encourage biologically predisposed behavior that occur in natural habitats despite the artificial context. Thus, ethograms or behavioral catalogs and activity budgets may include species-typical behavior as well as stereotypies (i.e., repetition of movements or sounds). Stereotypies combined with steroid hormones and glucocorticoid function are common welfare metrics and refer to unvarying, ritualistic behavior or sequences of behavior that include self-stimulation, movement, or ingestion of inedible objects. These behavioral ethograms and activity budgets are then used to understand, improve, and implement species-specific enrichment, husbandry, and exhibit space.

Experimental Animal Research

Experiments reflect the greatest level of control in a research design, wherein the animal subject is randomly assigned to two or more treatment conditions of an independent variable. These studies most often occur in a laboratory, agricultural area, or some other captive context wherein the conditions of the independent variable and associated intervening variables such as temperature, time of day, conspecific exposure, sensory stimuli, and so on can be carefully controlled.

The dependent variable of experimental animal studies may involve behavioral, physiological, genetic, neural, or immune changes. The least disturbance among animal experiments is noninvasive behavioral studies, like those of early ethologists who studied animal social behavioral patterns and their releasing mechanisms. Examples of more invasive animal experimental research involve pathological studies and virology, wherein ultimately the animal subject must be sacrificed to evaluate the methodological outcomes. In fact, virtually all major medical treatments involve the use of animal subjects at some stage of the research. The use of animals in biomedical studies hinge on four main goals: (1) to improve our understanding of biology, (2) to better understand pathology and disease,

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

(3) to test new vaccines and treatments, and (4) to protect humans and other animals. Although initially tested using in vitro methods on isolated tissues and organs, medical trials are required by law to be ethically tested on a suitable animal model before human clinical trials. The gold standard for such research is placebo-controlled trials on animals, concluding in placebo-controlled, double-blind studies among human participants. For studies conducted to *protect humans and other animals*, these may involve commercial product testing, like those in the cosmetic industry, although greater public awareness and better ethical regulations have limited much of the use of animals in product testing.

Heide D. Island

See also Ethics in the Research Process; Field Study; Laboratory Experiments

Further Readings

Committee to Update Science, Medicine, and Animals, National Research Council. (2004). *Science*, *medicine*, *and animals*. National Academies Press. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK24656/pdf/Bookshelf_NBK24656.pdf

Guillen, J. (2013). *Laboratory animals: Regulations and recommendations for the care and use of animals in research* (3rd ed.). San Diego, CA: Academic Press.

Nordell, S. E., & Valone, T. J. (2020). *Animal behavior: Concepts*, *methods*, *and applications* (3rd ed.). New York, NY: Oxford University Press.

Rees, P. A. (2015). *Studying captive animals. A workbook of methods in behaviour, welfare, and ecology*. West Sussex, UK: Wiley.

Anonymity

In research design, anonymity refers to the concept of removing identifying features and characteristics to make the source of information unknown. The concept requires that the receiver of information is left without clues to the identity of the source because the typical identifying information is removed. This anonymity process can be completed by the sender or the source of the information. Anonymity is typically required or requested in cases where knowing the identity of the source would change the results of a study, change the treatment of the information, or change the use of any analysis. This entry discusses factors that can be used to identify research participants, the purpose of anonymity in research, how individuals' perceptions of anonymity can affect behavior, and anonymity in the online environment.

Identifying Factors

In research design, granting participants anonymity requires removing identifying factors from information. Identifying factors include (1) legal name, (2) locatability, (3) pseudonyms, (4) social categorization, (5) pattern knowledge, and (6) symbols of eligibility. Each of these must be removed from information to guarantee anonymity.

First, the primary source of identifying information is the legal name of the sender. A legal name is one that is documented through government or regulatory procedures such as a birth certificate or passport. When legal names are attached to information, they are easy to identify and trace back to the source.

Second, locatability refers to the connection of information to a specific location or place. Locatability allows the receiver to trace information to a location such as an address, a region, or nation, which provides insight into the source of the information. This type of identity can be used alongside other types to identify the source of information. For example, a geotag on a social media post can be used to identify the location of a user when posting information online.

Third, pseudonyms are nicknames that are either adopted by the source or given to the source by others. Like a legal name, pseudonyms are often documented and therefore traceable. For example, a username on a social media website is a pseudonym that can be linked to an email address or other forms of identifying information.

Fourth, social categorization is data related to the demographics or psychographics of a source of information. This includes information on gender, race/ethnicity, age, employment, income, and political or religious orientation, among other categories. Social categorization data can be traced back to a specific individual based on cross-referencing categories.

Fifth, pattern knowledge refers to the ability to cross-reference categories and content found within the information to identify the source. For example, if a data set included quotes from individuals, familiar speech patterns or phrasing could be used to identify the source, assuming the researcher has outside knowledge or interactions with the individual.

Finally, symbols of eligibility include culturally negotiated symbols. Typical symbols include wedding bands, logos, and tattoos. These symbols can be used to segment possible sources of information by matching the symbol with an individual's potential identity.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Importantly, these six factors can be used independently or in concert with each other, meaning information that is not properly anonymized can be traced back to a source if any of these factors are present. The more factors or more specific the factor, the easier it is to trace someone's identity.

Purpose of Anonymity

Anonymity is often a requirement for scholarly research, journalism, and criminal justice. There are many reasons why a source of information would wish to remain anonymous as well as why the receiver of the information would want the source to be anonymous.

In scholarly research, institutional review boards often require data be anonymized before a researcher accesses it for analysis. It is required to both protect the identity of the source and ensure more reliable analysis. Many studies, particularly those looking at sensitive information such as medical records, psychological responses, and even communicative patterns, require anonymity to protect the identity of the source. Without this anonymity, the information provided by the source might be used by individuals outside the research setting and the source might face retribution. For example, studies examining political opinions often require anonymity so that the sources' personal preferences are not exposed to unsupportive users.

Anonymity is also viewed as a way to enhance the reliability of findings. Participants who are ensured that their data and identity will be anonymous may be more likely to share authentic or accurate information without fear of exposure or retribution. For example, participants asked about their political views and who are not guaranteed anonymity may change (intentionally or unintentionally) their responses to questions to meet the expectations and biases of the researchers. However, anonymity is also a legal condition that is granted to participants after informed consent. Institutions that approve of a researcher's project and protocol also offer legal protections for participants who may share sensitive information through a study. For example, should a research study require participants to share information about illegal behaviors, the institution can provide legal protection and prevent the researcher from having to share identifying information to law enforcement. It is for this reason that institutional review boards often spend additional time reviewing proposals that may require legal interventions to protect subject anonymity.

Closely related to the reviewing of research protocols and anonymity are the actions taken when a participant becomes a whistleblower regarding unethical or illegal research activities. Institutions grant anonymity to whistleblowers to encourage their honesty and to protect them from potential backlash. In the research process, this may mean a study's subject reports unethical behaviors of researchers to an institutional review board, which then investigates the allegations. Protecting the participant's identity helps ensure that the whistleblower remains safe and feels comfortable sharing details of the problems.

Finally, anonymity is also a condition of the peer-review process or the practice of having other researchers evaluate potential conference and journal publications. In a traditional double-blind peer-review process, both the identity of the author and reviewer are kept anonymous to prevent any preexisting perceptions of the other person or the relationship between the two parties from clouding the quality of the review.

Anonymity and Behavior

The psychological state associated with perceived anonymity is linked to the reduction in inhibitions and the enhancement of some behaviors. For example, an individual may be more likely to adopt behaviors that they may otherwise avoid when in a large crowd because they are with other people and therefore less likely to be called out individually or identified. When an individual is at a concert venue, they are more likely to sing and dance along to the music, even if they would normally avoid doing those things in public. The perception of anonymity makes individuals feel as if they can act without being identified or without facing retaliation.

This psychological reaction to anonymity appears in other forms of crowds such as riots. When surrounded by a large crowd and an individual feels they cannot be singularly identified, they may be more likely to adopt violent or destructive behaviors. In these cases, the perception of anonymity is linked to the reduction of retaliation and consequences. This is often deemed a "crowd mentality" because rather than perceiving themselves as an individual, the person identifies as part of the crowd and feels they are helping to carry out the goals of the group.

When an individual is identified after perceiving themselves as anonymous, they often experience shock. The psychological feelings associated with anonymity are so powerful, that when it is violated by external identification, individuals often have a difficult time defending or acknowledging their actions.

Digital Anonymity

Anonymity online is often a challenging feature to those running digital platforms, studying digital

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

behavior, or regulating digital content as well as to users themselves. For those running digital platforms, user anonymity can make it difficult to gain insight into the types of people who regularly interact with their platforms. For example, when users create accounts, they can fill in information that is inaccurate or limit the information provided. Most platforms require verified email addresses but do not verify other identifying factors such as legal name, address, or identity characteristics. As a result, those running digital platforms are often left with unreliable or incomplete data on users.

For those managing online content, anonymity is particularly problematic when users act outside the boundaries of normal or acceptable behavior online. Trolling behavior, when an individual posts incendiary comments to motivate reactions, often requires intervention from digital platform managers. However, without insight into the identity of the user, it can be difficult to create long-term solutions such as banning the individual.

For those studying online behavior, user anonymity challenges the insights gathered from data analysis. For example, Twitter researchers who study how users adopt a specific hashtag are limited by the amount of information provided by users. Even when working directly with Twitter, identifying details are often unverified, often producing a gap in the research.

For users themselves, perceived anonymity can also challenge digital engagement and behaviors. Organizations and platforms can use myriad programs to collect identifying information—often without the users' direct knowledge. For example, one form of identity tracker still used today includes "cookie" programs allowing organizations to track how users progress through the internet and relay information about website use and browsing history back to central data warehouses. Although most users accept end-user license agreements that document and explain this tracking, many users are still unaware of how that data can be used to harvest identifiable information. Other tracking programs and software can similarly archive identifying information ranging from name, location, credit card and financial information, personal interests and preferences, and education history. Even if a user agrees to the use of cookies when going on a website, the user may still feel anonymous but is instead easily identifiable.

Anonymity online is still a relatively new topic, one that requires more research to understand how it may impact user and platform behavior. The use of digital tools to document and archive identifiable information is regularly debated by ethicists and legal scholars and will likely be an important area of scholarship in the coming years.

Alison N. Novak

See also Confidentiality; Data Mining; Interviewing; Primary Data Source; Social Network Analysis

Further Readings

Gritzalis, S. (2006). *Privacy and anonymity in the digital era*. Bingley, UK: Emerald Group.

Kennedy, H. (2006). Beyond anonymity, or future directions for internet identity research. *New Media & Society*, *8*(6), 859–876. doi:10.1177/1461444806069641.

Lancaster, K. (2017). Confidentiality, anonymity and power relations in elite interviewing: Conducting qualitative policy research in a politicised domain. *International Journal of Social Research Methodology*, *20*(1), 93–103. doi:10.1080/13645579.2015.1123555.

O'Leary, M. (2019). Moving beyond Goffman: The performativity of anonymity on SNS. *European Journal of Marketing*, *53*(1), 83–107. doi:10.1108/EJM-01–2017–0016.

Smyth, S. (2004). Researchers and their "subjects": Ethics, power, knowledge and consent. In *Researchers and their "subjects"* (1st ed., pp. 91–104). Bristol, UK: Policy Press. doi:10.2307/j.ctt1t89724.

Taylor, L. (2014). Organizational anonymity and the negotiation of research access. *Qualitative Research in Organizations and Management*, *9*(2), 98–109. doi:10.1108/QROM-10–2012–1104.

Vainio, A. (2012). Beyond research ethics: anonymity as "ontology", "analysis" and "independence." *Qualitative Research*, *13*(6), 685–698. doi:10.1177/1468794112459669.

Walford, G. (2018). The impossibility of anonymity in ethnographic research. *Qualitative Research*, *18*(5), 516–525. doi:10.1177/1468794118778606.

Wiles, C. (2008). The management of confidentiality and anonymity in social research. *International Journal of Social Research Methodology*, *11*(5), 417–428. doi:10.1080/13645570701622231.

Applied Research

Applied research is inquiry using the application of scientific methodology with the purpose of generating empirical observations to solve critical problems in society. It is widely used in varying contexts, ranging from applied behavior analysis to city planning and public policy and to program evaluation. Applied research can be executed through a diverse range of research strategies that can be solely quantitative, solely qualitative, or a mixed method research design that combines quantitative and qualitative data slices in the same project. What all the multiple facets in applied research projects share

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

is one basic commonality—the practice of conducting research in "nonpure" research conditions because data are needed to help solve a real-life problem.

The most common way applied research is understood is by comparing it to *basic research*. Basic research—"pure" science—is grounded in the scientific method and focuses on the production of new knowledge and is not expected to have an immediate practical application. Although the distinctions between the two contexts are arguably somewhat artificial, researchers commonly identify four differences between applied research and basic research. Applied research differs from basic research in terms of purpose, context, validity, and methods (design).

Research Purpose

The purpose of applied research is to increase what is known about a problem with the goal of creating a better solution. This is in contrast to basic research, in which the primary purpose is to expand on what is known—knowledge—with little significant connections to contemporary problems. A simple contrast that shows how research purpose differentiates these two lines of investigation can be seen in applied behavior analysis and psychological research. Applied behavior is a branch of psychology that generates empirical observations that focus at the level of the individual with the goal of developing effective interventions to solve specific problems. Psychology, on the other hand, conducts research to test theories or explain changing trends in certain populations.

The irrelevance of basic research to immediate problems may at times be overstated. In one form or another, observations generated in basic research eventually influence what we know about contemporary problems. Going back to the previous comparison, applied behavior investigators commonly integrate findings generated by cognitive psychologists—how people organize and analyze information—in explaining specific types of behaviors and identifying relevant courses of interventions to modify them. The question is, how much time needs to pass (5 months, 5 years, 50 years) in the practical application of research results in order for the research to be deemed basic research? In general, applied research observations are intended to be implemented in the first few years whereas basic researchers make no attempt to identify when their observations will be realized in everyday life.

Research Context

The point of origin at which a research project begins is commonly seen as the most significant difference between applied research and basic research. In applied research, the context of pressing issues marks the beginning in a line of investigation. Applied research usually begins when a client has a need for research to help solve a problem. The context the client operates in provides the direction the applied investigator takes in terms of developing the research questions. The client usually takes a commanding role in framing applied research questions. Applied research questions tend to be open ended because the client sees the investigation as being part of a larger context made up of multiple stakeholders who understand the problem from various perspectives.

Basic research begins with a research question that is grounded in theory or previous empirical investigations. The context driving basic research takes one of two paths: testing the accuracy of hypothesized relationships among identified variables or confirming existing knowledge from earlier studies. In both scenarios, the basic research investigator usually initiates the research project based on his or her ability to isolate observable variables and to control and monitor the environment in which they operate. Basic research questions are narrowly defined and are investigated with only one level of analysis: prove or disprove theory or confirm or not confirm earlier research conclusions.

The contrast in the different contexts between applied research and basic research is simply put by Jon S. Bailey and Mary R. Burch in their explanation of applied behavior research in relation to psychology. The contrast can be pictured like this: In applied behavior research, subjects walk in the door with unique family histories that are embedded in distinct communities. In basic research, subjects "come in packing crates from a breeding farm, the measurement equipment is readily available, the experimental protocols are already established, and the research questions are derivative" (p. 3).

Emphasis on Validity

The value of all research—applied and basic—is determined by its ability to address questions of internal and external validity. Questions of *internal validity* ask whether the investigator makes the correct observation on the causal relationship among identified variables. Questions of *external validity* ask whether the investigators appropriately generalize observations from their research project to relevant situations. A recognized distinction is that applied research values external validity more than basic research projects do. Assuming an applied research project adequately addresses questions of internal validity, its research conclusions are more closely assessed in how well they apply directly to solving problems.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Questions of internal validity play a more significant role in basic research. Basic research focuses on capturing, recording, and measuring causal relationships among identified variables. The application of basic research conclusions focuses more on their relevance to theory and the advancement of knowledge than on their generalizability to similar situations.

The difference between transportation planning and transportation engineering is one example of the different validity emphasis in applied research and basic research. Transportation planning is an applied research approach that is concerned with the siting of streets, highways, sidewalks, and public transportation to facilitate the efficient movement of goods and people. Transportation planning research is valued for its ability to answer questions of external validity and address transportation needs and solve traffic problems, such as congestion at a specific intersection. Traffic engineering is the basic research approach to studying function, design, and operation of transportation facilities and looks at the interrelationship of variables that create conditions for the inefficient movement of goods and people. Traffic engineering is valued more for its ability to answer questions of internal validity in correctly identifying the relationship among variables that can cause traffic and makes little attempt to solve specific traffic problems.

Research Design

Applied research projects are more likely follow a *triangulation* research design than are basic research investigations. Triangulation is the research strategy that uses a combination of multiple data sets, multiple investigators, multiple theories, and multiple methodologies to answer research questions. This is largely because of the context that facilitates the need for applied research. Client-driven applied research projects tend to need research that analyzes a problem from multiple perspectives in order to address the many constituents that may be impacted by the study. In addition, if applied research takes place in a less than ideal research environment, multiple data sets may be necessary in order for the applied investigator to generate a critical mass of observations to be able to make defensible conclusions about the problem at hand.

Basic research commonly adheres to a single-method, single-data-research strategy. The narrow focus in basic research requires the investigator to eliminate possible research variability (bias) to better isolate and observe changes in the studied variables. Increasing the number of types of data sets accessed and methods used to obtain them increases the possible risk of contaminating the basic research laboratory of observations.

Research design in transportation planning is much more multifaceted than research design in traffic engineering. This can be seen in how each approach would go about researching transportation for older people. Transportation planners would design a research strategy that would look at the needs of a specific community and assess several different data sets (including talking to the community) obtained through several different research methods to identify the best combination of interventions to achieve a desired outcome. Traffic engineers will develop a singular research protocol that focuses on total population demand in comparison with supply to determine unmet transportation demand of older people.

John Gaber

See also Planning Research; Scientific Method

Further Readings

Baily, S. J., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.

Bickman, L., & Rog, D. J. (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.

Kimmel, J. A. (1988). *Ethics and values in applied social research*. Newbury Park, CA: Sage.

Aptitudes and Instructional Methods

Research on the interaction between student characteristics and instructional methods is important because it is commonly assumed that different students learn in different ways. That assumption is best studied by investigating the interaction between student characteristics and different instructional methods. The study of that interaction received its greatest impetus with the publication of Lee Cronbach and Richard Snow's *Aptitudes and Instructional Methods* in 1977, which summarized research on the interaction between aptitudes and instructional treatments, subsequently abbreviated as ATI research. Cronbach and Snow indicated that the term *aptitude*, rather than referring exclusively to cognitive constructs, as had previously been the case, was intended to refer to any student characteristic. Cronbach stimulated research in this area in earlier publications suggesting that ATI research was an ideal meeting point between the usually distinct research traditions of correlational and experimental psychology. Before the 1977 publication of *Aptitudes and Instructional Methods,* ATI research was spurred by Cronbach and

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Snow's technical report summarizing the results of such studies, which was expanded in 1977 with the publication of the volume.

Background

When asked about the effectiveness of different treatments, educational researchers often respond that "it depends" on the type of student exposed to the treatment, implying that the treatment interacted with some student characteristic. Two types of interactions are important in ATI research: *ordinal* and *disordinal*, as shown in Figure 1. In ordinal interactions (top two lines in Figure 1), one treatment yields superior outcomes at all levels of the student characteristic, though the difference between the outcomes is greater at one part of the distribution than elsewhere. In disordinal interactions (the bottom two lines in Figure 1), one treatment is superior at one point of the student distribution while the other treatment is superior for students falling at another point. The slope difference in ordinal interactions indicates that ultimately they are also likely to be disordinal, that is, the lines will cross at a further point of the student characteristic distribution than observed in the present sample.

Research Design

ATI studies typically provide a segment of instruction by two or more instructional methods that are expected to be optimal for students with different characteristics. Ideally, research findings or some strong theoretical basis should exist that leads to expectations of differential effectiveness of the instruction for students with different characteristics. Assignment to instructional method may be entirely random or random within categories of the student characteristic. For example, students may be randomly assigned to a set of instructional methods and their anxiety then determined by some measure or experimental procedure. Or, in quasi-experimental designs, high- and low-anxiety students may be determined first and then—within the high- and low-anxiety groups—assignment to instructional methods should be random.

ATI research was traditionally analyzed with analysis of variance (ANOVA). The simplest ATI design conforms to a 2 × 2 ANOVA, with two treatment groups and two groups (high and low) on the student characteristic. In such studies, main effects were not necessarily expected for either the treatment or the student characteristic, but the interaction between them is the result of greatest interest.

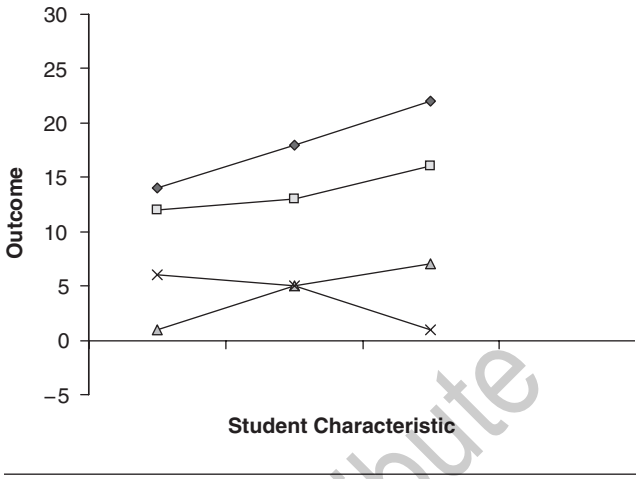

Figure 1 Ordinal and Disordinal Interactions

Cronbach and Snow pointed out that in ANOVA designs, the student characteristic examined was usually available as a continuous score that had at least ordinal characteristics, and the research groups were developed by splitting the student characteristic distribution at some point to create groups (high and low; high, medium, and low; etc.). Such division into groups ignored student differences within each group and reduced the available variance by an estimated 34%. Cronbach and Snow recommended that research employ multiple linear regression analysis in which the treatments would be represented by so-called dummy variables and the student characteristic could be analyzed as a continuous score. It should also be noted, however, that when the research sample is at extreme ends of the distribution (e.g., one standard deviation above or below the mean), the use of ANOVA maximizes the possibility of finding differences between the groups.

ATI Research Review

Reviews of ATI research reported few replicated interactions. Among the many reasons for these inconsistent findings were vague descriptions of the instructional treatments and sketchy relationships between the student characteristic and the instruction. Perhaps the most fundamental reason for the inconsistent findings was the inability to identify the cognitive processes required by the instructional treatments and engaged by the student characteristic. Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller demonstrated that when the cognitive processes involved in instruction have been clarified, more consistent ATI findings have been reported and replicated.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Later reviews of ATI research, such as those by J. E. Gustaffson and J. O. Undheim or by Sigmund Tobias, reported consistent findings for Tobias's general hypothesis that students with limited knowledge of a domain needed instructional support, that is, assistance to the learner, whereas more knowledgeable students could succeed without it. The greater consistency of interactions involving prior knowledge as the student characteristic may be attributable to some attributes of such knowledge. Unlike other student characteristics, prior domain knowledge contains the cognitive processes to be used in the learning of that material. In addition, the prior knowledge measure is likely to have been obtained in a situation fairly similar to the one present during instruction, thus also contributing any variance attributable to *situativity* to the results.

Sigmund Tobias

See also Analysis of Covariance (ANCOVA); Interaction; Reactive Arrangements

Further Readings

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116–127.

Cronbach, L. J., & Snow, R. E. (1969). *Individual differences and learning ability as a function of instructional variables*. Stanford, CA: Stanford University, School of Education.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York: Irvington Press.

Gustaffson, J., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. S. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186–242). New York: Macmillan Reference.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*, 23–31.

Tobias, S. (1976). Achievement treatment interactions. *Review of Educational Research*, *46*, 61–74.

Tobias, S. (1982). When do instructional methods make a difference? *Educational Researcher*, *11*(4), 4–9.

Tobias, S. (1989). Another look at research on the adaptation of instruction to student characteristics. *Educational Psychologist*, *24*, 213–227.

Tobias, S. (2009). An eclectic appraisal of the success or failure of constructivist instruction In S. Tobias & T. D. Duffy (Eds.), *Constructivist theory applied to education: Success or failure?* (pp. 335–350). New York: Routledge.

Aptitude-Treatment Interaction

There are countless illustrations in the social sciences of a description of a phenomenon existing for many years before it is labeled and systematized as a scientific concept. One such example is in Book II of Homer's *Iliad*, which presents an interesting account of the influence exerted by Agamemnon, king of Argos and commander of the Greeks in the Trojan War, on his army. In particular, Homer describes the behavior of Odysseus, a legendary king of Ithaca, and the behavior of Thersites, a commoner and rank-and-file soldier, as contrasting responses to Agamemnon's leadership and role as "the shepherd of the people." Odysseus, Homer says, is "brilliant," having "done excellent things by thousands," while he describes Thersites as that "who knew within his head many words, but disorderly," and "this thrower of words, this braggart." Where the former admires the leadership of Agamemnon, accepts his code of honor, and responds to his request to keep the sage of Troy, the latter accuses Agamemnon of greed and promiscuity and demands a return to Sparta.

The observation that an intervention—educational, training, therapeutic, or organizational—when delivered the same way to different people might result in differentiated outcomes, was made a long time ago, as long as the 8th century BCE, as exemplified by Homer. In attempts to comprehend and explain this observation, researchers and practitioners have focused primarily on the concept of individual differences, looking for main effects that are attributable to concepts such as ability, personality, motivation, or attitude. When these inquiries started early in the 20th century, not many parallel interventions were available. In short, the assumption at the time was that a student (a trainee in a workplace, a client in a clinical setting, or a soldier on a battlefield) possessed specific characteristics, such as Charles Spearman's *g* factor of intelligence, that could predict his or her success or failure in a training situation. However, this attempt to explain the success of an intervention by the characteristics of the intervenee was challenged by the appearance of multiple parallel interventions aimed at arriving at the same desired goal by employing various strategies and tactics. It turned out that there were no ubiquitous collections of individual characteristics that would always result in success in a situation. Moreover, as systems of intervention in education, work training in industry, and clinical fields developed, it became apparent that different interventions, although they might be focused on the same target (e.g., teaching children to read, training bank tellers

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

to operate their stations, helping a client overcome depression, or preparing soldiers for combat), clearly worked differently for different people. It was then suggested that the presence of differential outcomes of the same intervention could be explained by aptitude-treatment interaction (ATI, sometimes also abbreviated as AxT), a concept that was introduced by Lee Cronbach in the second part of the 20th century.

ATI methodology was developed to coaccount both for the individual characteristics of the intervenee and the variations in the interventions while assessing the extent to which alternative forms of interventions might have differential outcomes as a function of the individual characteristics of the person to whom the intervention is being delivered. In other words, investigations of ATI have been designed to determine whether particular treatments can be selected or modified to optimally serve individuals possessing particular characteristics (i.e., ability, personality, motivation). Today, ATI is discussed in three different ways: as a concept, as a method for assessing interactions among person and situation variables, and as a framework for theories of aptitude and treatment.

ATI as a Concept

ATI as a concept refers to both an outcome and a predictor of that outcome. Understanding these facets of ATI requires decomposing the holistic concept into its three components—treatment, aptitude, and the interaction between them. The term *treatment* is used to capture any type of manipulation aimed at changing something. Thus, with regard to ATI, treatment can refer to a specific educational intervention (e.g., the teaching of equivalent fractions) or conceptual pedagogical framework (e.g., Waldorf pedagogy), a particular training (e.g., job-related activity, such as mastering a new piece of equipment at a workplace) or self-teaching (e.g., mastering a new skill such as typing), a clinical manipulation (e.g., a session of massage) or long-term therapy (e.g., psychoanalysis), or inspiring a soldier to fight a particular battle (e.g., issuing an order) or preparing troops to use new strategies of war (e.g., fighting insurgency). *Aptitude* is used to signify any systematic measurable dimension of individual differences (or a combination of such) that is related to a particular treatment outcome. In other words, aptitude does not necessarily mean a level of general cognitive ability or intelligence; it can capture specific personality traits or transient psychological states. The most frequently studied aptitudes of ATI are in the categories of cognition, conation, and affection, but aptitudes are not limited to these three categories. Finally, *interaction* demarcates the degree to which the results of two or more interventions will differ for people who differ in one or more aptitudes. Of note is that interaction here is defined statistically and that both intervention and aptitude can be captured by qualitative or quantitative variables (observed, measured, self-reported, or derived). Also of note is that, being a statistical concept, ATI behaves just as any statistical interaction does. Most important, it can be detected only when studies are adequately powered. Moreover, it acknowledges and requires the presence of main effects of the aptitude (it has to be a characteristic that matters for a particular outcome, e.g., general cognitive ability rather than shoe size for predicting a response to educational intervention) and the intervention (it has to be an effective treatment that is directly related to an outcome, e.g., teaching a concept rather than just giving students candy). This statistical aspect of ATI is important for differentiating it from what is referred to by the ATI developers and proponents as *transaction*. Transaction signifies the way in which ATI is constructed, the environment and the process in which ATI emerges; in other words, ATI is always a statistical result of a transaction through which a person possessing certain aptitudes experiences a certain treatment. ATI as an outcome identifies combinations of treatments and aptitudes that generate a significant change or a larger change compared with other combinations. ATI as a predictor points to which treatment or treatments are more likely to generate significant or larger change for a particular individual or individuals.

ATI as a Method

ATI as a method permits the use of multiple experimental designs. The very premise of ATI is its capacity to combine correlational approaches (i.e., studies of individual differences) and experimental approaches (i.e., studies of interventional manipulations). Multiple paradigms have been developed to study ATI; many of them have been and continue to be applied in other, non-ATI, areas of interventional research. In classical accounts of ATI, the following designs are typically mentioned. In a *simple standard randomized between-persons design*, the outcome is investigated for persons who score at different levels of a particular aptitude when multiple, distinct interventions are compared. Having registered these differential outcomes, intervention selection is then carried out based on a particular level of aptitude to optimize the outcome. Within this design, often, when ATI is registered, it is helpful to carry out additional studies (e.g., case studies) to investigate the reason for the manifestation of ATI. The

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

treatment revision design assumes the continuous adjustment of an intervention (or the creation of multiple parallel versions of it) in response to how persons with different levels of aptitude react to each improvement in the intervention (or alternative versions of the intervention). The point here is to optimize the intervention by creating its multiple versions or its multiple stages so that the outcome is optimized at all levels of aptitude. This design has between- and within-person versions, depending on the purposes of the intervention that is being revised (e.g., ensuring that all children can learn equivalent fractions regardless of their level of aptitude or ensuring the success of the therapy regardless of the variability in depressive states of a client across multiple therapy sessions). In the *aptitude growth design*, the target of intervention is the level of aptitude. The idea here is that as the level of aptitude changes, different types of interventions might be used to optimize the outcome. This type of design is often used in combination with growth-curve analyses. It can be applied as either between-persons or within-person designs. Finally, a type of design that has been gaining much popularity lately is the *regression discontinuity design*. In this design, the presence of ATI is registered when the same intervention is administered before and after a particular event (e.g., a change in aptitude in response to linguistic immersion while living in a country while continuing to study the language of that country).

ATI as a Theoretical Framework

ATI as a theoretical framework underscores the flexible and dynamic, rather than fixed and deterministic, nature of the coexistence (or coaction) of individual characteristics (i.e., aptitudes) and situations (i.e., interventions). As a theory, ATI captures the very nature of variation in learning—not everyone learns equally well from the same method of instruction, and not every method of teaching works for everyone; in training—people acquire skills in a variety of ways; in therapy—not everyone responds well to a particular therapeutic approach; and in organizational activities—not everyone prefers the same style of leadership. In this sense, as a theoretical framework, ATI appeals to professionals in multiple domains as it justifies the presence of variation in outcomes in classrooms, work environments, therapeutic settings, and battlefields. While applicable to all types and levels of aptitudes and all kinds of interventions, ATI is particularly aligned with more extreme levels of aptitudes, both low and high, and more specialized interventions. The theory of ATI acknowledges the presence of heterogeneity in both aptitudes and interventions, and its premise is to find the best possible combinations of the two to maximize the homogeneity of the outcome. A particular appeal of the theory is its transactional nature and its potential to explain and justify both success and failure in obtaining the desired outcome. As a theoretical framework, ATI does not require the interaction to either be registered empirically or be statistically significant. It calls for a theoretical examination of the aptitude and interventional parameters whose interaction would best explain the dynamics of learning, skill acquisition and demonstration, therapy, and leadership. The beneficiaries of this kind of examination are of two kinds. First, it is the researchers themselves. Initially thinking through experiments and field studies before trying to confirm the existence of ATI empirically was, apparently, not a common feature of ATI studies during the height of their popularity. Perhaps a more careful consideration of the "what, how, and why" of measurement in ATI research would have prevented the observation that many ATI findings resulted from somewhat haphazard fishing expeditions, and the resulting views on ATI research would have been different. A second group of beneficiaries of ATI studies are practitioners and policy makers. That there is no intervention that works for all, and that one has to anticipate both successes and failures and consider who will and who will not benefit from a particular intervention, are important realizations to make while adopting a particular educational program, training package, therapeutic approach, or organizational strategy, rather than in the aftermath. However, the warning against embracing panaceas, made by Richard Snow, in interventional research and practice is still just a warning, not a common presupposition.

Criticism

Having emerged in the 1950s, interest in ATI peaked in the 1970s and 1980s, but then dissipated. This expansion and contraction were driven by an initial surge in enthusiasm, followed by a wave of skepticism about the validity of ATI. Specifically, a large-scope search for ATI, whose presence was interpreted as being marked by differentiated regression slopes predicting outcomes from aptitudes for different interventions, or by the significance of the interaction terms in analysis of variance models, was enthusiastically carried out by a number of researchers. The accumulated data, however, were mixed and often contradictory—there were traces of ATI, but its presence and magnitude were not consistently identifiable or replicable. Many reasons have been mentioned in discussions of why ATI is so elusive: underpowered studies, weak theoretical conceptualizations of ATI, simplistic research designs, imperfections

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

in statistical analyses, and the magnitude and even the nonexistence of ATI, among others. As a result of this discussion, the initial prediction of the originator of ATI's concept, Lee Cronbach, that interventions designed for the average individual would be ultimately replaced by multiple parallel interventions to fit groups of individuals, was revised. The "new view" of ATI, put forward by Cronbach in 1975, acknowledged that, although in existence, ATI is much more complex and fluid than initially predicted and ATI's dynamism and fluidity prevent professionals from cataloging specific types of ATI and generalizing guidelines for prescribing different interventions to people, given their aptitudes. Although the usefulness of ATI as a theory has been recognized, its features as a concept and as a method have been criticized along the lines of (a) our necessarily incomplete knowledge of all possible aptitudes and their levels, (b) the shortage of good psychometric instruments that can validly and reliably quantify aptitudes, (c) the biases inherent in many procedures related to aptitude assessment and intervention delivery, and (d) the lack of understanding and possible registering of important "other" nonstatistical interactions (e.g., between student and teacher, client and therapist, environment and intervention). And yet ATI has never been completely driven from the field, and there have been steady references to the importance of ATI's framework and the need for better-designed empirical studies of ATI.

Gene × Environment Interaction

ATI has a number of neighboring concepts that also work within the general realm of qualifying and quantifying individual differences in situations of acquiring new knowledge or new skills. Among these concepts are learning styles, learning strategies, learning attitudes, and many interactive effects (e.g., aptitude-outcome interaction). Quite often, the concept of ATI is discussed side by side with these neighboring concepts. Of particular interest is the link between the concept of ATI and the concept of Gene × Environment interaction (G × E). The concept of G × E first appeared in nonhuman research but gained tremendous popularity in the psychological literature within the same decade. Of note is that the tradition of its use in this literature is very similar to that of the usage of ATI; specifically, G × E also can be viewed as a concept, a method, and a theoretical framework. But the congruence between the two concepts is incomplete, of course; the concept of G × E adopts a very narrow definition of aptitude, in which individual differences are reduced to genetic variation, and a very broad definition of treatment, in which interventions can be equated with live events. Yet an appraisal of the parallels between the concepts of ATI and G × E is useful because it captures the field's desire to engage interaction effects for explanatory purposes whenever the explicatory power of main effects is disappointing. And it is interesting that the accumulation of the literature on G × E results in a set of concerns similar to those that interrupted the golden rush of ATI studies in the 1970s.

Yet methodological concerns aside, the concept of ATI rings a bell for all of us who have ever tried to learn anything in a group of people: what works for some of us will not work for the others as long as we differ on even one characteristic that is relevant to the outcome of interest. Whether it was wit or something else by which Homer attempted to differentiate Odysseus and Thersites, the poet did at least successfully make an observation that has been central to many fields of social studies and that has inspired the appearance of the concept, methodology, and theoretical framework of ATI, as well as the many other concepts that capture the essence of what it means to be an individual in any given situation: that individual differences in response to a common intervention exist. Millennia later, it is an observation that still claims our attention.

Elena L. Grigorenko

See also Effect Size, Measures of; Field Study; Growth Curve; Interaction; Intervention; Power; Within-Subjects Design

Further Readings

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116–127.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Press.

Dance, K. A., & Neufeld, R. W. J. (1988). Aptitude-treatment interaction research in the clinical setting: A review of attempts to dispel the "patient uniformity" myth. *Psychological Bulletin*, *104*(2), 192–213.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*, 199–208.

Grigorenko, E. L. (2005). The inherent complexities of gene-environment interactions. The *Journals of Gerontology: Series B*, *60*, 53–64.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Snow, R. E. (1984). Placing children in special education: Some comments. *Educational Researcher*, *13*, 12–14.
Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting & Clinical Psychology*, *59*, 205–216.
Spearman, C. (1904). "General intelligence" objectively determined and measured. The *American Journal of Psychology*, *15*, 201–293.
Violato, C. (1988). Interactionism in psychology and education: A new paradigm or a source of confusion? *Journal of Educational Thought*, *22*, 4–20.

Argument-Based Approach to Validity

Assessments (measures and tests) are used to assign values to attributes of people (or groups of people or objects). The assessment typically involves a limited sample of the person's behavior collected under standardized conditions, while the attribute of interest is defined much more broadly. The need for validation arises from this gap between the observed assessment results (i.e., a score) and the more general claims associated with the attribute being assessed (e.g., level of reading achievement). Validation examines how well the claims based on the assessment scores are justified.

Rather than defining validity as an abstract property of an assessment, the argument-based approach treats validation as an evaluation of the plausibility of the interpretation and uses proposed for the assessment scores. It employs two kinds of arguments in doing so. The *interpretation/use argument (IUA)* specifies a chain or network of inferences and supporting assumptions leading from the assessment results to the proposed interpretation and use. The *validity argument* evaluates the plausibility of the IUA in terms of its completeness (how well it represents the proposed interpretation and use), its coherence (how well it hangs together), and the plausibility of its inferences and assumptions. This entry further describes the argument-based approach to validity and discusses its historical antecedents, the roles of the IUA and the validity argument, and the use of inferences, arguments, and supporting evidence. It then looks at the development of the IUA, assessment, and validity argument and determination of the necessary and sufficient conditions for validity.

The argument-based approach is very general in that it is applicable to any interpretation or use of assessment scores, but it is contingent in that the evidence required for validation depends on the claims being made. Once the interpretation and use are specified as an IUA, the requirements for validation can be specified. That is, the IUA provides a framework for validating the proposed score interpretations and uses.

The validity argument subjects the IUA to challenges by questioning its assumptions and, where appropriate, by subjecting specific assumptions to empirical checks. Validation is never final or absolute because any of the claims being made can be overturned by new evidence, but an IUA that survives all reasonable challenges can be accepted as valid and used to support score-based claims about individuals or groups in the population. An exception to a standard interpretation or score use may be claimed on the basis of unusual circumstances (e.g., for a person with a disability), but in general, if the IUA represents the proposed interpretation and uses accurately, and its inferences and assumptions are supported by appropriate evidence, it can be considered valid.

Historical Antecedents

Validity analyses have been applied to a wide range of possible score interpretations and uses, and many different kinds of evidence and analyses have been employed in evaluating the claims being made. In the early 20th century, psychological assessments were evaluated in terms of how well they reflected the attribute of interest (e.g., intelligence) and in terms of the consistency of scores across replications of the assessment on the same people. Consistency across replications was evaluated under the heading of *reliability*, and the appropriateness of the interpretation was evaluated under the heading of *validity*. As new types of educational achievement tests (e.g., objective tests) were introduced early in the 20th century, they were evaluated in terms of how well their content matched the content of instruction and in terms of their reliability. These evaluations are generally considered under the heading of *content validity*.

By the 1920s, test scores were also being used to predict outcomes, or criteria, of interest (e.g., success on a job) as a basis for selection and placement decisions. These predictive uses of assessments were evaluated by developing a criterion measure of the desired outcome and investigating how well the assessment scores predicted the criterion. This type of evaluation was considered under the heading of *predictive validity* or *criterion validity*.

In the early 1950s, Lee Cronbach and Paul Meehl borrowed the notion of a theoretical construct defined in terms of its role in a theory from then-current models in the philosophy of science. In their construct-validity

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

model, the construct was defined implicitly in terms of its role in a theory, and the theory and the construct interpretation were evaluated together by examining whether the relationships implied by the theory were satisfied when the construct was estimated by assessment scores. If the theory implied that certain relationships should exist between the construct of interest and other variables, and these relationships were confirmed when the assessment scores were used to estimate the construct, both the theory and the construct interpretation of the scores would be supported. If some of the anticipated relationships were not confirmed, doubt would be cast on either the theory or the interpretation of the assessment scores. The construct validity model was quite elegant, but it was hard to apply in practice because it required a well-established theory to define the construct, and such theories were not generally available.

By the 1970s, a number of distinct validity models, including the criterion, content, and construct models (plus many more specific models), were in widespread use, generating considerable conceptual and practical ambiguity about what was required for adequate validation. In response, Samuel Messick developed a unified model of validity based on a generalized version of the construct model. Messick's model emphasized the need for multiple lines of evidence for the proposed construct interpretation and the need to evaluate the consequence of assessment uses. Like the original version of the construct model, Messick's unified construct-based model was hard to apply.

In the late 1980s and 1990s, Cronbach, Michael Kane, Lorrie Shepard, and Lyle Bachman proposed argument-based approaches as a more practical framework for validity. Rather than defining validity in terms of theory-based construct interpretations, these argument-based approaches focused on the specific inferences and assumptions inherent in the proposed interpretations and uses of the scores.

IUA and Validity Argument

As noted earlier, the argument-based approach to validity employs two kinds of arguments. The IUA specifies the inferences and assumptions inherent in the proposed interpretations and uses of the scores, and the validity argument evaluates the IUA in terms of its overall plausibility. The IUA lays out the interpretations and uses of the scores as a network of inferences and decisions and makes a preliminary case for their plausibility. The IUA plays the role that a scientific theory plays in the original construct model proposed by Cronbach and Meehl, but it allows for a wide range of possible interpretations ranging from simple generalizations of the assessment observations (e.g., in a performance test) to ambitious theoretical interpretations, as well as a range of intended uses of the scores. The more complex and ambitious score interpretations and uses generally involve more inferences and assumptions than the simpler interpretations.

The validity argument evaluates the claims in the IUA. The proposed interpretations and uses are considered valid to the extent that the IUA accurately represents the claims based on the scores and that its assumptions are adequately supported by appropriate evidence. Different interpretations and uses will rely on different inferences and assumptions and, therefore, will require different kinds of evidence for their evaluation. More complex IUAs will usually require more evidence than less complex IUAs.

Inferences, Assumptions, and Supporting Evidence

The structure and content of the validity argument will vary from case to case depending on the structure and content of the IUA. If the IUA involves only a few inferences and assumptions, the validity argument might rely on a few types of evidence (i.e., the evidence needed to evaluate the inferences and assumptions in the IUA), and if these inferences and assumptions are highly plausible a priori, the IUA might not require much empirical support for its validation. On the other hand, if the IUA includes a number of questionable assumptions, its evaluation might require an extensive body of empirical evidence.

A *scoring inference* assigns a score to each person's responses to the tasks included in the assessment and combines these task scores into an observed score for the person. The scoring rule would typically be based on the judgment of experts who develop and review the scoring criteria. If the assessment performances are evaluated by raters who apply the scoring criteria to each performance, analyses of the accuracy and consistency with which the scoring rule is applied (e.g., interrater reliability studies) would generally be expected. If the scoring procedures rely on statistical models (e.g., for scaling), the assumptions built into these models would be examined.

A *generalization inference* extends the interpretation from the person's observed score, based on a limited sample of observations, to the expected score over repeated assessments involving different, allowable conditions of observation (e.g., different occasions or raters). The generalization inference does not change the value of the score, but it does broaden its interpretation. Generalization from results on a particular application of the assessment to the expected value

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

over repeated application of the assessment under comparable conditions is a statistical inference that relies on evidence that the sampling is representative of the universe and that the sample size is large enough to ensure that the sampling error is minimal. Empirical analyses of this inference are generally discussed under the headings of reliability or generalizability theory. The IUA should indicate the range of conditions of observation (e.g., items, contexts, occasions, raters) over which the score is expected to be relatively invariant.

An *extrapolation inference* extends the interpretation to the larger domain of observations associated with the attribute of interest. Assessments are generally standardized in various ways to promote fairness and reliability. The standardized observations are drawn from a relatively narrow slice of the behavioral domain associated with the attribute and, therefore, are not representative of the attribute. Extending the interpretation to the much larger and less well-defined domain associated with the attribute generally relies on analyses that indicate that the tasks in the assessment depend on skills or propensities that are central to the conception of the attribute.

It may also be possible to empirically examine the relationships between the assessment scores and more direct measures of the attribute. For example, scores on a multiple-choice test of English composition skills might be compared to ratings of essays written by the same students. If the assessment is representative of the attribute of interest (e.g., using an essay test to assess skill in writing essays), the extrapolation would not require much support, but if the assessment is substantially different from the attribute (e.g., using a multiple-choice test to assess writing), it is important to rule out the possibility that the scores are overly dependent on the assessment format.

Predictive inferences use assessment scores to predict future behavior in some context (e.g., in an instructional program). These inferences tend to rely mainly on statistical evidence that there is a positive relationship between a person's assessment score and their standing on some criterion measure (e.g., college GPA).

Theory-based inferences rely on a theory, indicating that certain relationships should hold. For example, in Cronbach and Meehl's original version of the construct model, the constructs were assumed to be embedded in theoretical networks that could provide support for various inferences. If the theory's implications are consistent with observations based on the assessment, the theory and the interpretations of the assessment scores in terms of the theory are both supported. In particular, if the theory suggests that the construct should be related to another variable in a particular way, the assessment scores should relate to that variable in that way. If the theory implies that the construct should be largely independent of another variable, the assessment scores should be independent of that variable.

Inferences from score interpretation to score-based decisions generally involve claims that these decisions will lead to positive outcomes and will avoid negative consequences in most cases. Decisions that achieve their intended goals without serious side effects are considered acceptable, and those that do not achieve their intended goals or have serious side effects are not acceptable. A testing program with some negative consequences may be considered acceptable if it achieves its intended outcomes, but a program with serious negative consequences is not likely to be accepted.

In evaluating assessment programs, improvements in intended outcomes, like worker productivity, are important, but fairness to individuals and groups (e.g., racial-ethnic groups, language groups, genders) is equally important. Note that negative consequences always count against the decision, but they do not necessarily count against the underlying interpretation unless they indicate some defect in the underlying interpretation. The fact that an arithmetic test is not useful in predicting success in medical school does not count against its validity as a measure of arithmetic.

With the argument-based approach, validation requires the specification of the inferences and assumptions inherent in the proposed interpretation and use of the scores and then the evaluation of these inferences and assumptions in a validity argument. Because the claims vary from one interpretation/use to another, the evidence required for validation depends on the claims being made. To support the proposed interpretations and uses, the validity argument needs to justify the IUA as a whole and to rule out plausible alternative interpretations. In this process, the most doubtful parts of the argument deserve the most scrutiny because the intended interpretation or use can be undermined by a failure of any part of the IUA, even if the rest of the argument is highly plausible.

Developing the IUA, Assessment, and Validity Argument

In the early stages of assessment development, the goal is to outline a plausible IUA and an assessment that work together to support the proposed interpretation and use of the scores. Early on, any weaknesses that are identified in the IUA or the assessment tend to be addressed by adjusting the assessment or the IUA or both, but at some point the focus should shift to a more critical stance. Before the assessment is used operationally, the proposed

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

IUA should be subjected to critical review, with a particular focus on subjecting its most questionable assumptions to empirical challenges.

The requirement that the inferences and assumptions be explicitly stated and evaluated in the IUA provides some protection against inappropriate interpretations and uses of the resulting scores. To the extent that the IUA is clearly stated, gaps and inconsistencies are harder to ignore. However, just as we don't want to omit inferences or assumptions that should be in the IUA, we don't want to include inferences or assumptions that are not necessary; doing so could lead to a false-negative conclusion about validity. A proposed interpretation or use that has undergone a critical evaluation of its coherence and of the plausibility of its inferences and assumptions can be provisionally accepted as being valid, with the understanding that new evidence could lead to a reconsideration of this conclusion.

Necessary and Sufficient Conditions for Validity

The argument-based approach to validation does not specify any particular kind of interpretation or use for scores, but it does require that the claims based on the scores be clearly stated and adequately supported. More ambitious interpretations require more support, but a clear specification of the proposed interpretation and use also puts limits on the evidence required for validation. If neither the interpretation nor use requires a particular inference or assumption, there is no need to investigate that inference or assumption. For example, essentially all score interpretations require generalizability over some conditions of observation (e.g., over raters, or items, or occasions); if the scores did not generalize at all, they would simply be summaries of the observed responses. However, if the attribute being assessed (e.g., mood) is not assumed to be invariant over occasions, there would be no reason to require that the assessment scores be generalizable over occasions. Hence, adequate generalizability, or reliability, over a particular kind of condition of observation is necessary for validity if and only if the interpretation or use of the scores requires generalization over that kind of condition. The IUA specifies necessary and sufficient conditions for accepting the proposed score interpretations and uses as valid.

The extent to which the IUA fully represents the proposed interpretation and use of the scores is always questionable, and the adequacy of the evidence for various inferences and assumptions in the IUA is subject to debate. Nevertheless, if the IUA accurately reflects the intended interpretations and uses of the scores, and the validity argument supports the IUA, the proposed interpretation and use can be considered valid, at least for the time being. The criteria for accepting the validity of a proposed interpretation and use are essentially the same as the criteria for accepting a scientific theory. In both cases, we never achieve certainty, but with reasonable effort, we can achieve a high degree of confidence in the theory or in the interpretations and uses.

Concluding Remarks

The evidence required for argument-based validation depends on the proposed interpretation and use of the scores. Score interpretations that make very modest claims (e.g., claims about the level of skill in the activities included in the assessment) do not require much evidence for validation. Ambitious interpretations and uses can require an extended research program for their validation. If the IUA is coherent and complete, and all of its inferences and assumptions are supported by appropriate evidence, the proposed interpretations and uses can be considered valid. If the IUA is incomplete, or some of its inferences or assumptions are not plausible, some parts of the proposed interpretation and/or some proposed uses would be rejected.

A failure to specify the proposed interpretations and uses clearly and in some detail makes it difficult to develop a fully adequate validation effort because implicit inferences and assumptions could be overlooked. An IUA that understates the intended interpretation and use begs at least some questions, and as a result, the validation effort will not adequately evaluate the actual interpretation and use. An IUA that overstates the interpretation and use (by including some inferences or assumptions that are not required for the actual interpretation and use) will make validation more difficult and may lead to an erroneous conclusion that the scores are not valid for the interpretation and use.

The goal is to evaluate the plausibility of the claims being based on the assessment scores. Validation may not be easy, but it is generally possible to do a reasonably good job of validation with a manageable level of effort.

Michael T. Kane

See also Construct Validity; Content Validity; Generalizability Theory; Predictive Validity; Validity of Measurement

Further Readings

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Assessment validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological assessments. *Psychological Bulletin*, *52*, 281–302. doi:10.1037/h0040957.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64), Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of assessment scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

Assent

The term *assent* refers to the verbal or written agreement to engage in a research study. Assent is generally applicable to children between the ages of 8 and 18 years, although assent may apply to other vulnerable populations also.

Vulnerable populations are those composed of individuals who are unable to give consent due to diminished autonomy. Diminished autonomy occurs when an individual is incapacitated, has restricted freedom, or is a minor. Understanding the relevance of assent is important because without obtaining the assent of a participant, the researcher has restricted the freedom and autonomy of the participant and in turn has violated the basic ethical principle of respect for persons. Assent with regard to vulnerable populations is discussed here, along with the process of obtaining assent and the role of institutional review boards in the assent process.

Vulnerable Populations

Respect for persons requires that participants agree to engage in research voluntarily and have adequate information to make an informed decision. Most laws recognize that a person 18 years of age or older is able to give his or her informed *consent* to participate in the research study. However, in some cases individuals lack the capacity to provide informed consent. An individual may lack the capacity to give his or her consent for a variety of reasons; examples include a prisoner who is ordered to undergo an experimental treatment designed to decrease recidivism, a participant with an intellectual developmental disorder or an older adult with dementia whose caretakers believe an experimental psychotherapy group may decrease his or her symptoms. Each of the participants in these examples is not capable of giving permission to participate in the research because he or she either is coerced into engaging in the research or lacks the ability to understand the basic information necessary to fully consent to the study.

State laws prohibit minors and incapacitated individuals from giving consent. In these cases, permission must be obtained from parents and court-appointed guardians, respectively. However, beyond consent, many ethicists, professional organizations, and ethical codes require that *assent* be obtained. With children, state laws define when a young person is legally competent to make informed decisions. Some argue that the ability to give assent is from 8 to 14 years of age because the person is able to comprehend the requirements of the research. In general, however, it is thought that by the age of 10, children should be able to provide assent to participate. It is argued that obtaining assent increases the autonomy of the individual. By obtaining assent, individuals are afforded as much control as possible over their decision to engage in the research given the circumstances, regardless of their mental capacity.

Obtaining Assent

Assent is not a singular event. It is thought that assent is a continual process. Thus, researchers are encouraged to obtain permission to continue with the research during each new phase of research (e.g., moving from one type of task to the next). If an individual assents to participate in the study but during the study requests to discontinue, it is recommended that the research be discontinued.

Although obtaining assent is strongly recommended, failure to obtain assent does not necessarily preclude the participant from engaging in the research. For example, if the parent of a 4-year-old child gives permission for the child to attend a social skills group for socially anxious children, but the child does not assent to treatment, the child may be enrolled in the group without his or her assent. However, it is recommended that assent be obtained whenever possible. Further, if a child does not give assent initially, attempts to obtain assent should continue throughout the research. Guidelines also suggest that assent may be overlooked in cases in which the possible benefits of the research outweigh the costs. For example, if one wanted to study the effects of a lifesaving drug for children and the child refused the medication, the benefit of saving the child's life outweighs the cost of not obtaining assent. Assent may be overlooked in cases in which assent of the participants is not feasible, as would be the case of a researcher interested in studying children who died as a result of not wearing a seatbelt.

Obtaining assent is an active process whereby the participant and the researcher discuss the requirements of the research. In this case, the participant is active in

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

the decision making. *Passive consent*, a concept closely associated with assent and consent, is the lack of protest, objection, or opting out of the research study and is considered permission to continue with the research.

Institutional Review Boards

Institutional review boards frequently make requirements as to the way assent is to be obtained and documented. Assent may be obtained either orally or in writing and should always be documented. In obtaining assent, the researcher provides the same information as is provided to an individual from whom consent is requested. The language level and details may be altered in order to meet the understanding of the assenting participant. Specifically, the participant should be informed of the purpose of the study; the time necessary to complete the study; as well as the risks, benefits, and alternatives to the study or treatment. Participants should also have access to the researcher's contact information. Finally, limits of confidentiality should be addressed. This is particularly important for individuals in the prison system and for children.

Tracy J. Cohn

See also Debriefing; Ethics in the Research Process; Informed Consent; Interviewing

Further Readings

Belmont report: Ethical principles and guidelines for the protection of human subjects of research. (1979). Washington, DC: U.S. Government Printing Office.

Grisso, T. (1992). Minors' assent to behavioral research without parental consent. In B. Stanley & J. E. Sieber (Eds.), *Social research on children and adolescents: Ethical issues* (pp. 109–127). Newbury Park, CA: Sage.

Miller, V. A., & Nelson, R. M. (2006). A developmental approach to child assent for nontherapeutic research. *Journal of Pediatrics*, 25–30.

Ross, L. F. (2003). Do healthy children deserve greater protection in medical research? *Journal of Pediatrics*, *142*(2), 108–112.

Association, Measures Of

Measuring association between variables is very relevant for investigating *causality*, which is, in turn, the sine qua non of scientific research. However, an association between two variables does not necessarily imply a causal relationship, and the research design of a study aimed at investigating an association needs to be carefully considered in order for the study to obtain valid information. Knowledge of measures of association and the related ideas of *correlation*, *regression*, and *causality* are cornerstone concepts in research design. This entry is directed at researchers disposed to approach these concepts in a conceptual way.

Measuring Association

In scientific research, association is generally defined as the statistical dependence between two or more variables. Two variables are associated if some of the variability of one variable can be accounted for by the other, that is, if a change in the quantity of one variable conditions a change in the other variable.

Before investigating and measuring association, it is first appropriate to identify the types of variables that are being compared (e.g., nominal, ordinal, discrete, continuous). The type of variable will determine the appropriate statistical technique or test that is needed to establish the existence of an association. If the statistical test shows a conclusive association that is unlikely to occur by random chance, different types of regression models can be used to quantify how change in *exposure* to a variable relates to the change in the *outcome* variable of interest.

Examining Association Between Continuous Variables With Correlation Analyses

Correlation is a measure of association between two variables that expresses the degree to which the two variables are rectilinearly related. If the data do not follow a straight line (e.g., they follow a curve), common correlation analyses are not appropriate. In correlation, unlike regression analysis, there are no dependent and independent variables.

When both variables are measured as discrete or continuous variables, it is common for researchers to examine the data for a correlation between these variables by using the *Pearson product-moment correlation coefficient* (*r*). This coefficient has a value between −1 and +1 and indicates the strength of the association between the two variables. A *perfect* correlation of ±1 occurs only when all pairs of values (or *points*) fall exactly on a straight line.

A positive correlation indicates in a broad way that increasing values of one variable correspond to increasing values in the other variable. A negative correlation indicates that increasing values in one variable corresponds to decreasing values in the other variable. A correlation value close to 0 means no association between the variables. The *r* provides information about the strength of the correlation (i.e., the nearness

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

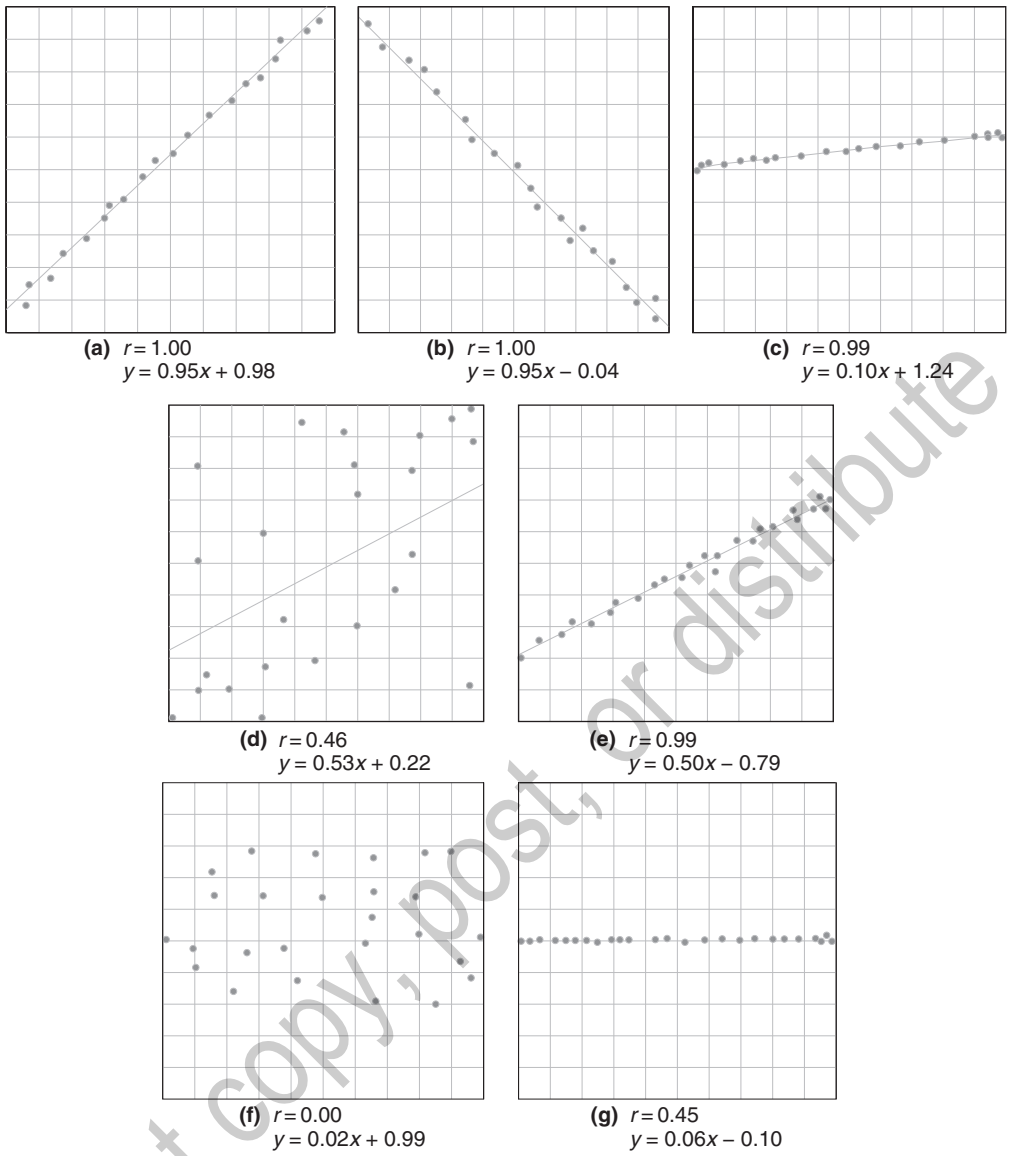

Figure 1 Scatterplots for Correlations of Various Magnitudes

Notes: Simulation examples show perfect positive (a, c) and negative (b) correlations, as well as regression lines with similar correlation coefficients but different slopes (a, c). The figure also shows regression lines with similar slopes but different correlation coefficients (d, e and f, g).

Do not copy, post, or distribute

of the points to a straight line). Figure 1 gives some examples of correlations, correlation coefficients, and related regression lines.

A condition for estimating correlations is that both variables must be obtained by random sampling from the same population. For example, one can study the correlation between height and weight in a sample of children but not the correlation between height and three different types of diet that have been decided by the investigator. In the latter case, it would be more appropriate to apply a regression analysis.

The Pearson correlation coefficient may not be appropriate if there are outliers (i.e., extreme values). Therefore, the first step when one is studying correlations is to draw a scatterplot of the two variables to examine whether there are any outliers. These variables should be standardized to the same scale before they are plotted.

If outliers are present, *nonparametric types of correlation coefficients* can be calculated to examine the linear association. The *Spearman rank correlation coefficient*, for example, calculates correlation coefficients

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

based on the ranks of both variables. *Kendall's coefficient of concordance* calculates the concordance and discordance of the observed (or ranked) exposure and outcome variables between pairs of individuals.

When the variables one is investigating are nominal or have few categories or when the scatterplot of the variables suggests an association that is not rectilinear but, for example, quadratic or cubic, then the correlation coefficients described above are not suitable. In these cases other approaches are needed to investigate the association between variables.

Chi-Square Tests of Association for Categorical Variables

A common method for investigating "general" association between two categorical variables is to perform a *chi-square test*. This method compares the observed number of individuals within cells of a cross-tabulation of the categorical variables with the number of individuals one would expect in the cells if there was no association and the individuals were randomly distributed. If the observed and expected frequencies differ statistically (beyond random chance according to the chi-square distribution), the variables are said to be associated.

A chi-square *test for trend* can also examine for linear association when the exposure category is ordinal. Other statistical tests of association include measurements of agreement in the association, such as the kappa statistic or McNemar's test, which are suitable when the study design is a matched case-control.

Quantifying Association in General and by Regression Analyses in Particular

In *descriptive* research, the occurrence of an outcome variable is typically expressed by group measurements such as averages, proportions, incidence, or prevalence rates. In *analytical* research, an association can be quantified by comparing, for example, the absolute risk of the outcome in the exposed group and in the nonexposed group. Measurements of association can then be expressed either as *differences* (difference in risk) or as *ratios*, such as relative risks (a ratio of risks) or odds ratios (a ratio of odds), and so forth. A ratio with a numerical value greater than 1 (greater than 0 for differences) indicates a positive association between the exposure variable and the outcome variables, whereas a value less than 1 (less than 0 for differences) indicates a negative association. These measures of association can be calculated from cross-tabulation of the outcome variable and exposure categories, or they can be estimated in regression models.

General measures of association such as correlation coefficients and chi-square tests are rather unspecific and provide information only on the existence and strength of an association. *Regression analysis*, however, attempts to model the relationship between two variables by fitting a linear equation to observed data in order to quantify, and thereby predict, the change in the outcome of interest with a unit increase in the exposure variable. In regression analysis, one variable is considered to be an *explanatory variable* (the exposure), and the other is considered to be a *dependent variable* (the outcome).

The method of least squares is the method applied most frequently for fitting a regression line. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. When a point is placed exactly on the fitted line, its vertical deviation is 0. Deviations are also known as *residuals* or *errors*. The better an explanatory variable predicts the outcome, the lower is the sum of the squared residuals (i.e., residual variance).

A simple *linear regression model*, for example, can examine the increase in blood pressure with a unit increase in age with the regression model

$$Y = a + bX + e,$$

where X is the explanatory variable (i.e., age in years) and Y is the dependent variable (i.e., blood pressure in mm Hg). The slope of the line is b and represents the change in blood pressure for every year of age. Observe that b does not provide information about the strength of the association but only on the average change in Y when X increases by one unit. The strength of the association is indicated by the correlation coefficient r, which informs on the closeness of the points to the regression line (see Figure 1). The parameter a is the intercept (the value of y when $x = 0$), which corresponds to the mean blood pressure in the sample. Finally, e is the residual or error.

A useful measure is the square of the correlation coefficient, r^2, also called the *coefficient of determination*, and it indicates how much of the variance in the outcome (e.g., blood pressure) is explained by the exposure (e.g., age). As shown in Figure 1, different bs can be found with similar rs, and similar bs can be observed with different rs. For example, many biological variables have been proposed as risk factors for cardiovascular diseases because they showed a high b value, but they have been rejected as common risk factors because their r^2 was very low.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Different types of regression techniques are suitable for different outcome variables. For example, a *logistic regression* is suitable when the outcome is binary (i.e., 0 or 1), and logistic regression can examine, for example, the increased probability (more properly, the increase in log odds) of myocardial infarction with unit increase in age. The *multinomial regression* can be used for analyzing outcome with several categories. A *Poisson regression* can examine how rate of disease changes with exposure, and a *Cox regression* is suitable for survival analysis.

Association Versus Causality

Exposure variables that show a statistical relationship with an outcome variable are said to be *associated* with the outcome. It is only when there is strong evidence that this association is causal that the exposure variable is said to *determine* the outcome. In everyday scientific work, researchers apply a pragmatic rather than a philosophical framework to identify causality. For example, researchers want to discover modifiable causes of a disease.

Statistical associations say nothing by themselves on causality. Their causal value depends on the knowledge background of the investigator and the research design in which these statistical associations are observed. In fact, the only way to be completely sure that an association is causal would be to observe the very same individual living two parallel and exactly similar lives except that in one life, the individual was exposed to the variable of interest, and in the other life, the same individual was not exposed to the variable of interest (a situation called *counterfactual*). In this *ideal design*, the two (hypothetical) parallel lives of the individual are exchangeable in every way except for the exposure itself.

While the ideal research design is just a chimera, there are alternative approaches that try to approximate the ideal design by comparing similar groups of people rather than the same individual. One can, for example, perform an experiment by taking random samples from the same population and randomly allocating the exposure of interest to the samples (i.e., *randomized trials*). In a randomized trial, an association between the average level of exposure and the outcome is possibly causal. In fact, random samples from the same population are—with some random uncertainty—identical concerning both *measured* and *unmeasured* variables, and the random allocation of the exposure creates a counterfactual situation very appropriate for investigating causal associations. The randomized trial design is theoretically closest to the ideal design, but sometimes it is unattainable or unethical to apply this design in the real world.

If conducting a randomized trial is not possible, one can use *observational* designs in order to simulate the ideal design, at least with regard to measured variables. Among observational approaches it is common to use *stratification*, *restriction*, and *multiple regression* techniques. One can also take into account the propensity for exposure when comparing individuals or groups (e.g., *propensity scores techniques*), or one may investigate the same individual at two different times (*case crossover design*). On some occasions one may have access to *natural experiments* or *instrumental variables*.

When planning a research design, it is always preferable to perform a *prospective* study because it identifies the exposure before any individual has developed the outcome. If one observes an association in a *cross-sectional* design, one can never be sure of the direction of the association. For example, low income is associated with impaired health in cross-sectional studies, but it is not known whether bad health leads to low income or the opposite. As noted by Austin Bradford Hill, the existence of a *temporal relationship* is the main criterion for distinguishing causality from association. Other relevant criteria pointed out by this author are *consistency*, *strength, specificity*, *dose-response relationship*, *biological plausibility*, and *coherence*.

Bias and Random Error in Association Studies

When planning a study design for investigating causal associations, one needs to consider the possible existence of *random error*, *selection bias*, *information bias*, and *confounding*, as well as the presence of *interactions* or *effect modification* and of *mediator* variables.

Bias is often defined as the lack of *internal validity* of the association between exposure and outcome variable of interest. This is in contrast to *external validity*, which concerns generalizability of the association to other populations. Bias can also be defined as nonrandom or *systematic* difference between an estimate and the true value of the population.

Random Error

When designing a study, one always needs to include a sufficient number of individuals in the analyses to achieve appropriate *statistical power* and ensure that *conclusive estimates* of association can be obtained. Suitable statistical power is especially relevant when it comes to establishing the *absence* of association between two variables. Moreover, when a study involves a large number of individuals, more information is available. More information lowers the random error, which in turn increases the *precision* of the estimates.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Selection Bias

Selection bias can occur if the sample differs from the rest of the population and if the observed association is *modified* by a third variable. The study sample may be different from the rest of the population (e.g., only men or only healthy people), but this situation does not necessarily convey that the results obtained are biased and cannot be applied to the general population. Many randomized clinical trials are performed on a restricted sample of individuals, but the results are actually generalizable to the whole population. However, if there is an *interaction* between variables, the *effect modification* that this interaction produces must be considered. For example, the association between exposure to asbestos and lung cancer is much more intense among smokers than among nonsmokers. Therefore, a study on a population of nonsmokers would not be generalizable to the general population. Failure to consider interactions may even render associations spurious in a sample that includes the whole population. For example, a drug may increase the risk of death in a group of patients but decrease this risk in other different groups of patients. However, an overall measure would show no association since the antagonistic directions of the underlying associations compensate each other.

Information Bias

Information bias simply arises because information collected on the variables is erroneous. All variables must be measured correctly; otherwise, one can arrive at imprecise or even spurious associations.

Confounding

An association between two variables can be confounded by a third variable. Imagine, for example, that one observes an association between the existence of yellow nails and mortality. The causality of this association could be plausible. Since nail tissue stores body substances, the yellow coloration might indicate poisoning or metabolic disease that causes an increased mortality. However, further investigation would indicate that individuals with yellow nails were actually heavy smokers. The habit of holding the cigarette between the fingers discolored their nails, but the cause of death was smoking. That is, smoking was associated with both yellow nails and mortality and originated a confounded association (Figure 2).

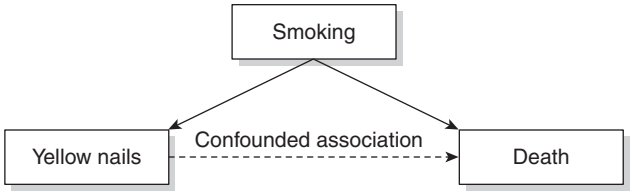

Figure 2 Deceptive Correlation Between Yellow Nails and Mortality

Note: Because smoking is associated with both yellow nails and mortality, it originated a confounded association between yellow nails and mortality.

Mediation

In some cases an observed association is mediated by an *intermediate* variable. For example, individuals with low income present a higher risk of early death than do individuals with high income. Simultaneously, there are many more heavy smokers among people with low income. In this case, heavy smoking mediates the effect of low income on mortality.

Figure 3 Smoking Acts as a Mediator Between Income and Early Death

Note: Heavy smoking mediated the effect of low income on mortality.

Distinguishing which variables are confounders and which are mediators cannot be done by statistical techniques only. It requires previous knowledge, and in some cases variables can be both confounders and mediators.

Directed Acyclic Graphs

Determining which variables are confounders, intermediates, or independently associated variables can be difficult when many variables are involved. Directed acyclic graphs use a set of simple rules to create a visual representation of direct and indirect associations of covariates and exposure variables with the outcome. These graphs can help researchers understand possible causal relationships.

Juan Merlo and Kristian Lynch

See also Bias; Cause and Effect; Chi-Square Test; Confounding; Correlation; Interaction; Multiple Regression; Power Analysis

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Further Readings

Altman, D. G. (1991). *Practical statistics for medical research*. New York: Chapman & Hall/CRC.

Hernan, M. A., Hernandez-Diaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, *155*(2), 176–184.

Hernan, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, *17*(4), 360–372.

Hill, A. B. (1965). Environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300.

Jaakkola, J. J. (2003). Case-crossover design in air pollution epidemiology. *European Respiratory Journal*, *40*, 81s-85s.

Last, J. M. (Ed.). (2000). *A dictionary of epidemiology* (4th ed.). New York: Oxford University Press.

Liebetrau, A. M. (1983). *Measures of association*. Newbury Park, CA: Sage.

Lloyd, F. D., & Van Belle, G. (1993). *Biostatistics: A methodology for the health sciences*. New York: Wiley.

Oakes, J. M., & Kaufman J. S. (Eds.). (2006). *Methods in social epidemiology*. New York: Wiley.

Rothman, K. J. (Ed.). (1988). *Causal inference*. Chestnut Hill, MA: Epidemiology Resources.

Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology*, *33*, 635–648.

AUDITING

The term "auditing" refers to a systematic review of processes involved in decisions or actions. Typically this is done to ensure conformation with accepted standards or to validate the accuracy of results. In qualitative research, auditing serves a comparable purpose and can be a valuable means of demonstrating the rigor of an investigation. Such a review offers a strong defense against criticisms that are sometimes posed in regard to qualitative research, such as questions regarding the researcher's neutrality. Auditing of the study, therefore, is a useful means of supporting the credibility and trustworthiness of findings and interpretations in qualitative research.

There continues to be debate over the most appropriate ways to demonstrate credibility or rigor in a qualitative study. Auditing is not an essential part of the process of qualitative research. It can, however, be a useful mechanism to address quality aspects of a study. Many variations of auditing are available for qualitative researchers to apply in their projects. It is important that plans for an audit are addressed early in the design of a project so that the process can be incorporated in the manner that is most appropriate to each study. This entry describes the ways in which auditing can be conducted in qualitative research, including both internal and external audits and the timing of the auditing process. It also reviews the materials needed for an audit trail.

Methods of Auditing in Qualitative Research

Auditing of a qualitative study involves oversight and, at minimum, review of the conduct of the study and the conclusions developed by investigators. There are numerous ways in which an audit can be carried out in a qualitative study. Variations include who serves as auditor, when the auditing process is initiated, how often auditing occurs, and the extent of the actual audit.

Internal Auditing

Auditing can be conducted on an internal basis in which members of the research team provide a system of checks and balances for each other. This process can promote consistency in the research process and can serve to identify, and subsequently decrease, the bias of any particular team members involved in the research. An internal audit can involve an exchange of documentation for review by other members of the team who can examine decisions and analytic processes associated with the research. An internal audit may be very useful in multisite studies where it is important to ensure consistency in the research process across the various settings. These activities enhance the research but may not provide sufficient evidence of rigor as typically sought through a more formal or external audit.

External Auditing

Auditing conducted on an external basis involves formal and systematic review carried out by people with no vested interest or involvement in the conduct of the research. An external auditor typically is a researcher knowledgeable in the processes of qualitative research and may or may not have expertise in the subject matter involved in the research. In the typical qualitative study, auditing can be accomplished quite easily by enlisting the assistance of an experienced yet objective colleague and the investigator presenting and defending decision making to that individual. The colleague also can review raw data, notes, logs, journals,

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

and other materials associated with the study. This process is may be referred to as peer review, although it accomplishes the same goal as an audit.

Timing of an Audit

The actual process of auditing can be initiated at any point in a study. Formative and ongoing auditing occurs while the study is conducted. Auditing also may be carried out on a summative basis at or near the conclusion of the study. Engaging auditors early in the process enables them to provide valuable monitoring throughout various phases of the research. The auditor may even be involved at the initial conceptual stages of the research, providing oversight and reflexive commentary as initial decisions are made regarding the design of the study. Such involvement, however, increases the risk that the auditors might become less neutral themselves due to their engagement with the project and the researcher or research team. Including auditors later in the process may allow for greater neutrality on the part of the auditors. Later involvement, however, creates a greater burden on the researchers to familiarize the auditors with the study and its processes as the auditors will be not be aware of the various nuances and twists that have occurred consistent with the emergent design typical of qualitative research. In the initial contracting with auditors, the researcher should be very clear and detailed about the desired level of involvement and the expectations placed on the auditor or auditing team.

Elements Needed for an Audit

Auditing cannot be accomplished unless there is an appropriate array of materials available for review. The collection of documentation compiled for this purpose during a qualitative study is referred to as an audit trail. Halpern identified six categories of documentation needed to constitute an audit trail. These include raw data, data reduction and analysis products, data reconstruction and synthesis products, notes regarding the processes of the study documentation of the intents and prejudgments or inclinations of the researcher, and information about any instruments used in the study. An organized system of note keeping is essential for this process. For ease of maintaining the audit trail, materials can grouped into, at minimum, raw data and field notes providing detail of actual encounters with participants; methodological notes regarding data collection processes, interview guides, other instrumentation, changes in an emergent design; and analytic memos or notes to capture ideas generated during the process of data analysis.

Beth L. Rodgers

Note: Reprinted from Rogers, B. L., Auditing. In Given, L. M., (Ed.), *The SAGE Encyclopedia of Qualitative Research Methods* (Vol. 1, pp. 42-43). Thousand Oaks, CA: Sage. doi:10.4135/9781412963909.n24.

Further Readings

Halpern, E. S. (1983). *Auditing naturalistic inquiries: The development and application of a model*. Unpublished doctoral dissertation, Indiana University, Indiana.

Lincoln, Y. S., & Guba, E. G. (1985) *Naturalistic inquiry*. CA: Sage.

Patton, M. Q. (2001). *Qualitative research and evaluation methods* (3*rd* ed.). Thousand Oaks, CA: Sage.

Rodgers, B. L., & Cowles, K. V. (1993). The qualitative research audit trail: a complex collection of documentation. *Research in Nursing and Health*, *16*, 219–226.

Schwandt, T. A., & Halpern, E. S. (1988). *Linking auditing and metaevaluation: enhancing quality in applied research*. CA: Sage.

AUTOCORRELATION

Autocorrelation describes sample *or* population observations or elements that are related to each other across time, space, or other dimensions. Correlated observations are common but problematic, largely because they violate a basic statistical assumption about many samples: independence across elements. Conventional tests of statistical significance assume simple random sampling, in which not only each element has an equal chance of selection but also each combination of elements has an equal chance of selection; autocorrelation violates this assumption. This entry describes common sources of autocorrelation, the problems it can cause, and selected diagnostics and solutions.

Sources

What is the best predictor of a student's 11th-grade academic performance? His or her 10th-grade grade point average. What is the best predictor of this year's crude divorce rate? Usually last year's divorce rate. The old slogan "birds of a feather flock together" describes a college classroom in which students are

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

about the same age, at the same academic stage, and often in the same disciplinary major. That slogan also describes many residential city blocks, where adult inhabitants have comparable incomes and perhaps even similar marital and parental status. When examining the spread of a disease, such as the H1N1 influenza, researchers often use epidemiological maps showing concentric circles around the initial outbreak locations.

All these are examples of correlated observations, that is, autocorrelation, in which two individuals drawn from a classroom or other group resemble each other more than two individuals drawn from the total population of elements by means of a simple random sample. Correlated observations occur for several reasons:

- Repeated, comparable measures are taken on the same individuals over time, such as many pretest and posttest experimental measures or panel surveys, which reinterview the same individual. Because people remember their prior responses or behaviors, because many behaviors are habitual, and because many traits or talents stay relatively constant over time, these repeated measures become correlated for the same person.
- Time-series measures also apply to larger units, such as birth, divorce, or labor force participation rates in countries or achievement grades in a county school system. Observations on the same variable are repeated on the same unit at some periodic interval (e.g., annual rate of felony crimes). The units transcend the individual, and the periodicity of measurement is usually regular. A lag describes a measure of the same variable on the same unit at an earlier time, frequently one period removed (often called t 1).
- Spatial correlation occurs in cluster samples (e.g., classrooms or neighborhoods) Physically adjacent elements have a higher chance of entering the sample than do other elements. These adjacent elements are typically more similar to already sampled cases than are elements from a simple random sample of the same size.
- A variation of spatial correlation occurs with contagion effects, such as crime incidence (burglars ignore city limits in plundering wealthy neighborhoods) or an outbreak of disease.
- Multiple (repeated) measures administered to the same individual at approximately the same time (e.g., a lengthy survey questionnaire with many Likert-type items in *agree-disagree* format).

Autocorrelation Terms

The terms *positive* or *negative autocorrelation* often apply to time-series data. Societal inertia can inflate the correlation of observed measures across time. The social forces creating trends such as falling marriage rates or rising gross domestic product often carry over from one period into the next. When trends continue over time (e.g., a student's grades), positive predictions can be made from one period to the next, hence the term *positive autocorrelation*.

However, forces at one time can also create compensatory or corrective mechanisms at the next, such as consumers' alternating patterns of "save, then spend" or regulation of production based on estimates of prior inventory. The data points seem to ricochet from one time to the next, so adjacent observations are said to be negatively correlated, creating a cobweb-pattern effect.

The order of the autocorrelation process references the degree of periodicity in correlated observations. When adjacent observations are correlated, the process is *first-order autoregression*, or AR (1). If every other observation, or alternate observations, is correlated, this is an AR (2) process. If every third observation is correlated, this is an AR (3) process, and so on. The order of the process is important, first because the most available diagnostic tests and corrections are for the simplest situation, an AR (1); higher order processes require more complex corrections. Second, the closer two observations are in time or space, the larger the correlation between them, creating more problems for the data analyst. An AR (1) process describes many types of autocorrelation, such as trend data or contagion effects.

Problems

Because the similarity among study elements is more pronounced than that produced by other probability samples, each autocorrelated case "counts less" than a case drawn using simple random sampling. Thus, the "real" or corrected sample size when autocorrelation is present is smaller than a simple random sample containing the same number of elements. This statistical attenuation of the casebase is sometimes called the *design effect*, and it is well known to survey statisticians who design cluster samples.

The sample size is critical in inferential statistics. The N comprises part of the formula for estimates of sample variances and the standard error. The standard error forms the denominator for statistics such as t tests. The N is also used to calculate degrees of freedom

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

for many statistics, such as *F* tests in analysis of variance or multiple regression, and it influences the size of chi-square.

When autocorrelation is present, use of the observed *N* means overestimating the effective *n*. The calculated variances and standard errors that use simple random sampling formulae (as most statistics computer programs do) are, in fact, too low. In turn, this means *t* tests and other inferential statistics are too large, leading the analyst to reject the null hypothesis inappropriately. In short, autocorrelation often leads researchers to think that many study results are statistically significant when they are not.

For example, in ordinary least squares (OLS) regression or analysis of variance, autocorrelation renders the simple random sampling formulae invalid for the error terms and measures derived from them. The true sum of squared errors (σ) is now inflated (often considerably) because it is divided by a fraction:

$$\sigma = \Sigma v^2 / (1 - \rho^2),$$

where v is the random component in a residual or error term.

Rho

When elements are correlated, a systematic bias thus enters into estimates of the residuals or error terms. This bias is usually estimated numerically by rho (ρ), the intraclass correlation coefficient, or the correlation of autocorrelation. Rho estimates the average correlation among (usually) adjacent pairs of elements. Rho is found, sometimes unobtrusively, in many statistics that attempt to correct and compensate for autocorrelation, such as hierarchical linear models.

An Example: Autocorrelation Effects on Basic Regression Models

With more complex statistical techniques, such as regression, the effects of ρ multiply beyond providing a less stable estimate of the population mean. If autocorrelation occurs for scores on the dependent variable in OLS regression, then the regression residuals will also be autocorrelated, creating a systematic bias in estimates of the residuals and statistics derived from them. For example, standard computer OLS regression output will be invalid for the following: the residual sum of squares, the standard error of the (regression) estimate, the *F* test, R^2 and the adjusted R^2, the standard errors of the *B*s, the *t* tests, and significance levels for the *B*s.

As long as residuals are correlated only among themselves and *not* back with any of the predictor variables, the OLS regression coefficient estimates themselves should be unbiased. However, the *B*s are no longer *best* linear unbiased estimators, and the estimates of the statistical significance of the *B*s and the constant term are inaccurate. If there is positive autocorrelation (the more usual case in trend data), the *t* tests will be inappropriately large. If there is negative autocorrelation (less common), the computer program's calculated *t* tests will be too small. However, there also may be autocorrelation in an independent variable, which generally aggravates the underestimation of residuals in OLS regression.

Diagnosing First-Order Autocorrelation

There are several ways to detect first-order autocorrelation in least squares analyses. Pairs of adjacent residuals can be plotted against time (or space) and the resulting scatterplot examined. However, the scatterplot "cloud of points" mentioned in most introductory statistics texts often resembles just that, especially with large samples. The decision is literally based on an "eyeball" analysis.

Second, and more formally, the statistical significance of the number of positive and negative runs or sign changes in the residuals can be tested. Tables of significance tests for the runs test are available in many statistics textbooks. The situation of too many runs means the adjacent residuals have switched signs too often and oscillate, resulting in a diagnosis of negative autocorrelation. The situation of too few runs means long streams of positive or negative trends, thus suggesting positive autocorrelation. The number of runs expected in a random progression of elements depends on the number of observations. Most tables apply to relatively small sample sizes, such as $N<40$. Since many time series for social trends are relatively short in duration, depending on the availability of data, this test can be more practical than it initially appears.

One widely used formal diagnostic for first-order autocorrelation is the Durbin-Watson *d* statistic, which is available in many statistical computer programs. The *d* statistic is approximately calculated as $2(1-\rho)$ where $\rho e_t e_{t-1}$ is the intraclass correlation coefficient. The e_t can be defined as adjacent residuals (in the following formula, v represents the true random error terms that one really wants to estimate):

$$e_t = \rho e_{t-1} + v_t$$

Thus *d* is a ratio of the sum of squared differences between adjacent residuals to the sum of squared

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

residuals. The *d* has an interesting statistical distribution: Values near 2 imply $\rho = 0$ (no autocorrelation); *d* is 0 when ρ=1 (extreme positive autocorrelation) and 4 when ρ=1 (extreme negative autocorrelation). In addition, *d* has two *zones of indecision* (one near 0 and one near 4), in which the null hypothesis $\rho = 0$ is neither accepted nor rejected. The zones of indecision depend on the number of cases and the number of predictor variables. The *d* calculation cannot be used with regressions through the origin, with standardized regression equations, or with equations that include lags of the dependent variable as predictors.

Many other computer programs provide iterative estimates of ρ and its standard error, and sometimes the Durbin–Watson *d* as well. Hierarchical linear models and time-series analysis programs are two examples. The null hypothesis $\rho = 0$ can be tested through a *t*-distribution with the ratio

$$\rho / \mathrm{se}_{\rho}.$$

The *t* value can be evaluated using the *t* tables if needed. If ρ is not statistically significant, there is no first-order autocorrelation. If the analyst is willing to specify the positive or negative direction of the autocorrelation in advance, one-tailed tests of statistical significance are available.

Possible Solutions

When interest centers on a time series and the lag of the dependent variable, it is tempting to attempt solving the autocorrelation problem by simply including a lagged dependent variable (e.g., y_{t-1}) as a predictor in OLS regression or as a covariate in analysis of covariance. Unfortunately, this alternative creates a worse problem. Because the observations are correlated, the residual term *e* is now correlated back with y_{t-1}, which is a predictor for the regression or analysis of covariance. Not only does this alternative introduce bias into the previously unbiased *B* coefficient estimates, but using lags also invalidates the use of diagnostic tests such as the Durbin-Watson *d*.

The first-differences (Cochrane-Orcutt) solution is one way to correct autocorrelation. This generalized least squares (GLS) solution creates a set of new variables by subtracting from each variable (not just the dependent variable) its own $t-1$ lag or adjacent case. Then each newly created variable in the equation is multiplied by the weight (1-ρ) to make the error terms behave randomly.

An analyst may also wish to check for higher order autoregressive processes. If a GLS solution was created for the AR (1) autocorrelation, some statistical programs will test for the statistical significance of ρ using the Durbin-Watson *d* for the reestimated GLS equation. If ρ does not equal 0, higher order autocorrelation may exist. Possible solutions here include logarithmic or polynomial transformations of the variables, which may attenuate ρ. The analyst may also wish to examine econometrics programs that estimate higher order autoregressive equations.

In the Cochrane-Orcutt solution, the first observation is lost; this may be problematic in small samples. The Prais-Winsten approximation has been used to estimate the first observation in case of bivariate correlation or regression (with a loss of one additional degree of freedom).

In most social and behavioral science data, once autocorrelation is corrected, conclusions about the statistical significance of the results become much more conservative. Even when corrections for ρ have been made, some statisticians believe that R^2s or η^2s to estimate the total explained variance in regression or analysis of variance models are invalid if autocorrelation existed in the original analyses. The explained variance tends to be quite large under these circumstances, reflecting the covariation of trends or behaviors.

Several disciplines have other ways of handling autocorrelation. Some alternate solutions are paired *t* tests and multivariate analysis of variance for either repeated measures or multiple dependent variables. Econometric analysts diagnose treatments of higher order periodicity, lags for either predictors or dependent variables, and moving averages (often called ARIMA). Specialized computer programs exist, either freestanding or within larger packages, such as the Statistical Package for the Social Sciences (SPSS; an IBM company, formerly called PASW® Statistics).

Autocorrelation is an unexpectedly common phenomenon that occurs in many social and behavioral science phenomena (e.g., psychological experiments or the tracking of student development over time, social trends on employment, or cluster samples). Its major possible consequence—leading one to believe that accidental sample fluctuations are statistically significant—is serious. Checking and correcting for autocorrelation should become a more automatic process in the data analyst's tool chest than it currently appears to be.

Susan Carol Losh

See also Cluster Sampling; Hierarchical Linear Modeling; Intraclass Correlation; Multivariate Analysis of Variance (MANOVA); Time-Series Study

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Further Readings

Bowerman, B. L., O'Connell. R., & Koehler, A. (2004). *Forecasting, time series, and regression* (4th ed.). Pacific Grove, CA: Duxbury Press.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., . . . & Hooten, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, *98*(3), 632–646. doi:10.1002/ecy.1674.

Jebb, A. T., & Tay, L. (2017). Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods*, *20*(1), 61–94. doi:10.1177/1094428116668035.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Marasinghe, M. G., & Koehler, K. J. (Eds.). (2018). Beyond regression and analysis of variance. In *Statistical data analysis using SAS* (pp. 529–619). Cham, Switzerland: Springer.

Ostrom, C. W. Jr. (1990). *Time series analysis: Regression techniques* (2nd ed.). Thousand Oaks, CA: Sage.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.