Volume 1, Release 1 September 17, 2010

Published in real time by
Optimal Data Analysis, LLC

Optimal Data Analysis

A Multidisciplinary Journal Specializing in Maximum-Accuracy Methods

ISSN 2155-0182

Optimal Data Analysis (*ODA*) is a multidisciplinary journal specializing in maximum-accuracy methods, which together are subsumed as constituents of the optimal data analysis (ODA) paradigm. Articles published in *ODA* are parsed into sections. *Invited* articles are written in response to the Editor's invitation. *Review* articles present reviews of theory, method, or empirical findings relating to the ODA paradigm. *Method* articles discuss technical aspects of optimal or heuristic algorithms and analytic processes. *Versus* articles feature comparisons between alternative methodologies. *Application* articles use optimal statistical analysis methods to address applied topics in any academic discipline. *Software* articles discuss design or operation of existing or theoretical software systems which explicitly maximize (weighted) classification accuracy, and suboptimal heuristic systems which seek maximum accuracy. *Integrated System* articles discuss theoretical or existing closed-system "black-box" or "robotic" applications engineered using optimal analytic methods. *Consulting* articles highlight areas constituting consulting opportunities for proponents of optimal analytic methods.

Manuscript Submission. Electronically submit one copy of the manuscript in Microsoft Word^{TM} 97-2003 or later, to: Journal@OptimalDataAnalysis.com. Follow *Instructions to Authors* (see *Table of Contents*). Opinions and statements published are the responsibility of the authors, and do not necessarily represent policies of Optimal Data Analysis, LLC, or views of Editors.

Copyright and Permissions. Authors must obtain permission from the Editor and original author in order to reproduce an article of 500 or more words: permission is usually granted if the ODA copyright notice is included on the first page of reproduced material. Permission from ODA is waived for authors wishing to reproduce a single table or figure from an article, provided the author's permission is granted, and that credit is given to ODA as copyright holder and to the author via a citation (there is a $25 service fee for a letter from the Editor permitting such reproduction). Authors wishing to reproduce more than a single table or figure for commercial purposes must obtain permission from the Editor. Permission is waived for authors wishing to reproduce their own material for personal use, and for the photocopying of isolated journal articles for nonprofit classroom or library use by instructors and educational institutions. Access services may use unedited abstracts without permission of ODA or author. Libraries may photocopy articles beyond limits of U.S. copyright law by obtaining permission from the editor: submit requests to Journal@OptimalDataAnalysis.com.

Reprints. Authors may print copies of their own and other articles, and/or may order any requested number of reprints when they receive notification that their article has been accepted for publication, or at any subsequent time. For information please see webpage.

Single Releases, Custom Issues, and Back Volumes. For information on soft-or hard-bound copies of *single releases* (the equivalent of a print issue), *custom issues* (a subset of articles hand-selected as a special collection), or *back volumes* (all articles from an entire year), please see webpage .

Microform editions. For information regarding microform editions please contact Optimal Data Analysis.

Optimal Data Analysis (ISSN 2155-0182) is published in real time in one volume per year by Optimal Data Analysis LLC, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. There is no subscription fee, and there is no manuscript submission fee.

Articles published in *Optimal Data Analysis* utilize and discuss methods which yield maximum-accuracy.

Optimal Data Analysis

A Multidisciplinary Journal Specializing in Maximum-Accuracy Methods

Preface

Invited

Review

Method

Versus

Application

Software

Integrated System

Consulting

Other

Volume 1, Release 1
September 17, 2010

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Preface to Volume 1, Release 1

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Manucripts published in *Optimal Data Analysis* (*ODA*) are parsed in sections, more of which will be added as new domains of inquiry are discovered. Not every Release of every Issue will present articles in every section. Sections represented in this Release, and the articles they host, are briefly described below.

Invited. This section features articles written in response to the Editor's invitation. The only manuscript in this section in the first Release is written by Fred Bryant (*ODA*'s Co-Editor), a member of the faculty at Loyola University Chicago (LUC). Outside the Optimal Data Analysis company laboratory, Fred has most experience in using ODA and CTA, and teaching these methods to colleagues and students: he has been there from the beginning. Fred describes the ontogenesis and etiology of the use of optimal methods by faculty and students in the Department of Psychology at LUC.

Review. This section offers reviews of theory, method, or empirical findings relating to the ODA paradigm. The sole article in this section is written by Paul Yarnold (*ODA*'s Editor) and Robert Soltysik (*ODA*'s Co-Editor), and presents an introductory review of crucial concepts in the ODA paradigm, including both univariate and multivariable ODA methods.

Method. This section presents articles discussing technical aspects of optimal and heuristic algorithms and analytic processes. Paul and Robert lead four of the five articles in this section in this Release. The first manuscript discusses how to maximize the classification accuracy of any nonlinear model via a new optimal pruning methodology.

The second paper offers a multivariable optimal data analysis (MultiODA) formulation we developed years ago and decided to publish now, which has proven extremely powerful in the laboratory in a wide domain of frontiers.

The third article demonstrates that it is not necessary to "control" for "covariates" by forcing them into a classification model before entering the attributes of theoretical interest. Indeed, it is shown that such forced entry can result in a model that is substantially weaker in performance than another model—based on exactly the same attributes, but arranged in a different (algorithm-determined) geometry.

The fourth article, led by Barbara Maria Yarnold, discusses how UniODA may be used to maximize the accuracy of a model derived by probit analysis.

The fifth and final article in this section is a research note exploring precision and convergence properties of Monte Carlo simulation used to estimate exact Type I error.

Versus. This section features articles in which alternative methodologies (at least one of which is an optimal method) compete against each other. Paul and Robert lead four of the five articles in this section in this Release (setting the table, it is hoped, for interested others to add to the series). The first paper compares the use of aggregated (e.g., ethnicity) *vs*. referenced (e.g., white vs. African-American, white *vs*. Hispanic, etc.) categorical variables in CTA.

The second article is the first of a series comparing the findings of CTA performed manually *vs*. using automated software. The application in this article involves predicting mortality from *Pneumocystis carinii* pneumonia, a topic which has been investigated via CTA more times and over a longer period of time than any other specific area of inquiry. Dr. Charles L. Bennett, M.D., Ph.D., graciously offered Paul and Robert use of data which we previously published, generated from his NIH-funded projects which completed approximately a decade ago, for expository purposes.

The third article, led by Rachel Coakley, is the second article in the "Manual *vs*. Automated CTA" series. The original manuscript was recently published, and used a manually-derived CTA. The original model is contrasted with a new model developed using automated CTA software.

The fourth article uses Gen-UniODA to model discrimination in organizations, for an application which was problematic for log-linear model.

The fifth and final manuscript in this section for this Release reveals how ordinal data are commonly misidentified as being categorical, and incorrectly analyzed by chi-square. This paper demonstrates the appropriate, straightforward UniODA analysis.

Application. This section features articles using optimal statistical analysis methods to address applied topics in any academic discipline. The first manuscript in this section, led by Robert Soltysik, identifies and corrects paradoxical confounding present in serial meteorological measurements, then uses automated weighted CTA to predict temperature and pressure anomalies across the USA and the northern hemisphere. A heretofore unexplained recent Artic ice flux event is also demystified.

The second manuscript in this section, led by Jennifer Howard Smith, uses manually-derived CTA to model college freshman attrition. The paper is based on her dissertation, which is believed to be the first to use CTA, and represents the first CTA conducted outside the Optimal Data Analysis company laboratory.

The final paper in this section derives from Hideo Suzuki's thesis, and uses manually-derived CTA to model the development of juvenile delinquency. Hideo has used UniODA software to derive CTA models for many years, and is well-versed in traditional multivariate classification methodologies. A member of *ODA*'s Board of Editors, Hideo agreed to be the emissary of *ODA* to the nation of Japan. We wish to transliterate all articles involving optimal methods and published in Japanese, and republish them in *ODA* to further dissemination and accelerate progress in this area. Japanese journals wishing to transliterate articles originally published in *ODA* for republication should contact the Editor. In all cases, citations will credit the original work.

Software. Articles in this section discuss design or operation of existing or theoretical software systems which explicitly maximize (weighted) classification accuracy, and suboptimal heuristic systems which seek maximum accuracy. The first manuscript, led by Robert, discusses the motivation, reporting and use of automated of the CTA software which is now commercially available, including a list of the control commands and example analyses. The second article, a brief report written by Fred at the Editor's request, discusses how to use a widely-available software system to produce a data file needed to optimize the classification accuracy of a logistic regression model.

Integrated System. This section features articles which discuss theoretical or existing closed-system "black-box" or "robotic" applications which are engineered using optimal analytic methods. The sole manuscript in this section in this Release is written by William Collinge, a member of *ODA*'s Editorial Board. The manuscript discusses the alpha test of a web-based, interactive, structured patient diary using CTA to identify targetable behavioral antecedents of symptoms for individual fibromyalgia patients.

Consulting. This section features articles highlighting areas which constitute consulting opportunities for application of optimal analytic methods. The first manuscript in this section—and final manuscript in this release, is led by Fred and describes a now classic court case addressing paradoxical confounding.

Other. Finally, this section functions like a bulletin board. This Release features links to author instructions; advertiser instructions; how to obtain optimal software; how to obtain bound copies and reprints; and seven Special Calls.

Author Notes

Mail correspondence to the author at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send eMail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

The Loyola Experience (1993-2009): Optimal Data Analysis in the Department of Psychology

Fred B. Bryant, Ph.D.
Loyola University Chicago

This article traces the origins and development of the use of optimal data analysis (ODA) within the Department of Psychology at Loyola University Chicago over the past 17 years. An initial set of ODA-based articles by Loyola faculty laid the groundwork for a sustained upsurge in the use of ODA among graduate students which has lasted for more than a decade and a half. These student projects subsequently fueled an increase in ODA-based publications by other Loyola Psychology faculty, who directly supervised the various student projects. Thus, ODA initially trickled down from faculty to students, but later grew up in the opposite direction. The most frequent use of ODA in Loyola's Psychology Department has been to conduct classification tree analysis, with less common uses of ODA including optimal discriminant analysis and the iterative structural decomposition of transition tables. As more Loyola Psychology graduate students find academic jobs and continue using ODA in their research, we expect that they will replicate the Loyola experience in these new academic settings.

When you discover a new tool that you believe is superior to other tools you've used before, naturally you want not only to use the new tool, but also to tell others about it so they can enjoy its benefits too. Such has been the case in the Department of Psychology at Loyola University Chicago since early 1993, when the first version of Optimal Data Analysis (ODA) 1.0 for DOS became publicly available. The purpose of this brief article is to describe the 17-year process through which the use of ODA sprang up, took hold, and spread among graduate students and faculty in Loyola's Psychology Department.

The Early Days of ODA at Loyola

I have known Paul Yarnold and Rob Soltysik since they first began working on the problem of optimal classification in the early 1980s. I served as a beta-tester for both the original DOS-based[1] and more recent Windows-based[2] versions of the ODA software. In late 1992, I cheered from the sidelines as Paul and Rob put the finishing touches on ODA 1.0 for DOS. And when ODA 1.0 for DOS appeared in print, I wrote the first published review of the

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

new software[3] and began using ODA in my research. Later I also published the first review of ODA for Windows.[4]

Having fallen in love with the power, versatility, and elegance of ODA, I began publishing research articles using ODA as a statistical tool in 1994.[5] I first directly collaborated with departmental colleagues to use ODA in 1996, in publishing an article using optimal discriminant analysis as an alternative to Student's *t* test with two Loyola clinicians in the *Journal of Consulting and Clinical Psychology*.[6] At the same time, I continued publishing ODA-based research on my own, and began extolling the capabilities of the new ODA software to my graduate students. Interestingly, it was the graduate students, rather than the faculty, who more eagerly embraced ODA as a statistical tool in their research.

How Loyola Researchers Have Used ODA

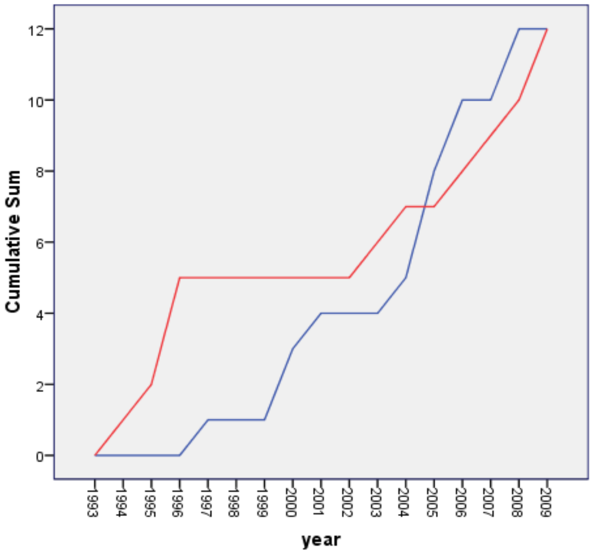
At Loyola, researchers have used ODA in multiple ways to address a wide variety of different research questions in clinical psychology, social psychology, neuropsychology, behavioral medicine, and biochemistry. Table 1 summarizes the 12 faculty publications and 12 graduate student projects (11 dissertations and 1 master's thesis) in Loyola's Psychology Department that have used ODA over the past 16 years (1994-2009).

TABLE 1: Published Journal Articles and Graduate Student Projects (Master's Theses and Dissertations) in Loyola's Psychology Department Using ODA by Year (1993-2009)

Year	Student Projects	Published Journal Articles
1993	0	0
1994	0	1
1995	0	1
1996	0	3
1997	1	0
1998	0	0
1999	0	0
2000	2	0
2001	1	0
2002	0	0
2003	0	1
2004	1	1
2005	3	0
2006	2	1
2007	0	1
2008	2	1
2009	0	2
TOTAL	12	12

Figure 1 illustrates the cumulative number of faculty publications (red) and graduate student projects (blue) from 1993 to 2009.

FIGURE 1: Loyola Psychology Department Publications (Red) and Dissertations/Theses (Blue) Using ODA From 1993-2009



Note the patterns that emerge across the 17-year span. The Loyola Psychology Department's experience with ODA originated in the early publications by department faculty. This initial set of articles laid the groundwork for a sustained upsurge in the use of ODA by Loyola graduate students over more than a decade and a half. These graduate student projects subsequently fueled the increase in ODA-based publications by other Loyola Psychology faculty, who directly supervised the various student projects. Thus, although ODA initially trickled down from faculty to students, it later grew up in the opposite direction.

Classification Tree Analysis

By far, the most frequent use of ODA at Loyola has been to conduct multiattribute classification tree analysis (CTA). For example, Loyola graduate students have used CTA to identify predictive models for discriminating students who drop out versus return to college following the first year[7], children's emotional responsiveness versus unresponsiveness during psychotherapy[8], child molesters versus non-molesters[9], positive versus nonpositive adaptation to childhood[10], convicted juvenile delinquents versus non-delinquent youth[11], positive versus negative morbidity and mortality outcomes following bone marrow transplant[12], high versus low effect sizes in a meta-analysis of methodological and intervention characteristics associated with primary prevention programs for children and adolescents[13], engaging versus not engaging in risky sexual behavior among minority adolescents[14] and adult male homosexuals[15], high versus low social competence among children with spina bifida[16], and state mental health care agency decisions to commit children to residential treatment versus foster homes.[17] In addition, department faculty and graduate students have jointly published journal articles using CTA to predict early sexual debut among adolescents[18], positive adaptation to childhood[19], psychiatric hospital admission decisions for children in foster care[20], malingering in forensic neuropsychological examinations[21], change in job status following traumatic brain injury[22], and clinically significant sexual concerns in a child welfare population.[23]

Optimal Discriminant Analysis

The next most common use of ODA in Loyola's Psychology Department has been to conduct optimal discriminant analysis, as an exact-probability alternative to parametric discriminant analysis or Student's *t* test. For example, Loyola faculty publications have used ODA in this fashion to discriminate Type As versus Type Bs using the Type A Self-Rating Inventory[5] and the Students Jenkins Activity Survey[24], males versus females in self-ratings of affective intensity[25], high- versus low-quality child therapy sessions based on therapist discourse[6], and physicians versus undergraduates

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

in levels of sympathy and empathy.[26] Layden et al. used this form of discriminant analysis to identify an optimal cut-score for using psychiatric ratings to assess toxicity in patients undergoing lithium treatment for bipolar depression.[27]

Iterative Structural Decomposition

Another Loyola dissertation in clinical psychology used ODA to conduct an analysis for which no alternative statistical test exists. In this particular project, the student had couples discuss an area of disagreement in their marriage for 15 minutes, and then used an established interaction scoring system to code these interactions. Based on existing theory, the student predicted that couples having only one depressed spouse would engage in the following sequence of behaviors: (a) depressive behavior, followed by (b) spouse's supportive behavior, followed by (c) more depressive behavior, followed by (d) spouse's incongruent behavior, followed by (e) angry/defensive behavior, followed finally by (f) spouse's critical/rejecting behavior. Following procedures outlined by Yarnold and Soltysik[2] (pp. 209-222), the data were organized into transition tables representing the frequencies of various verbal exchanges between spouses over time. Supporting the hypothesized temporal model, an iterative structural decomposition of the transition tables revealed that the data conformed to the predicted sequence of behaviors significantly more than would be expected by chance alone.[28] It is unclear how one would test the hypothesized behavioral sequence using any other inferential statistical tool.

The Future of ODA in Psychology

If the past is any indication of the future, then ODA has a bright future, not only at Loyola but elsewhere. The recent availability of ODA-based software that automatically constructs classification tree models is likely to accelerate the use of CTA across a wider variety of research disciplines. In the future, enumerated CTA models may well replace traditional hierarchically-optimal CTA models, particularly given the superior classification accuracy of the former. The automated CTA software also offers the ability to analyze class variables that have more than two levels, thereby enabling new forms of nonlinear optimal regression modeling. We can foresee a vast array of new applications for CTA, including meta-analysis, cross-cultural tests of similarities and differences, and optimal path analysis.

Obviously, it is relatively easy to export the Loyola Experience with ODA to other universities. All that is needed is a faculty member to lay the groundwork through an initial set of ODA-based publications, along with graduate students who are seeking to analyze data for their dissertation or master's thesis. As more Loyola Psychology graduates find academic jobs and continue to use ODA in their research, we expect that they will replicate the Loyola experience in these new academic settings.

I close by noting an unanticipated aspect of the Loyola experience with ODA. Namely, some of the graduate students who have used ODA in their dissertation research have later had the opportunity to teach introductory statistics in psychology at the undergraduate level, both at Loyola and at other colleges and universities. Naturally, these graduate instructors have taught their students about ODA and its statistical advantages, and these undergraduates are now approaching faculty members in psychology at Loyola and elsewhere to supervise independent research projects and honors theses that use ODA. Once again, the process of learning has come full circle, as the students themselves become teachers and disseminate statistical methods to students, faculty, and beyond.

References

[1]Soltysik R.C., & Yarnold P.R. (1993). *ODA 1.0: Optimal Data Analysis for DOS*. Chicago: Optimal Data Analysis, Inc.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

[2]Yarnold, P.R., & Soltysik, R.C. (2005). *Optimal Data Analysis: A guidebook with software for Windows*. Washington, DC: APA Books.

[3]Bryant, F.B. (1994). Analyze your data optimally using ODA 1.0. *Decision Line, 25,* 16-19.

[4]Bryant, F.B. (2005). How to make the best of your data [Review of Optimal Data Analysis]. *PsycCRITIQUES-Contemporary Psychology: APA Review of Books, 50,* Article 5 (7 pp.).

[5]Yarnold, P.R., & Bryant, F.B. (1994). A measurement model for the Type A self-rating inventory. *Journal of Personality Assessment, 62,* 102-115.

[6]Russell, R.L., Bryant, F.B., & Estrada, A.U. (1996). Confirmatory P-technique analyses of therapist discourse: High- versus low-quality child therapy sessions. *Journal of Consulting and Clinical Psychology, 64,* 1366-1376.

[7]Brockway, J.H. (1997). *Here today, gone tomorrow: Understanding freshman attrition using person-environment fit theory*. Doctoral dissertation, Loyola University Chicago (112 pp.).

[8]Elling, K.A. (2000). *Predicting children's emotional responsiveness during therapy sessions*. Doctoral dissertation, Loyola University Chicago (119 pp.).

[9]Bivens, A.J. (2001). Accurate classification of child molesters using context variation and Hierarchical Optimal Classification Tree Analysis. Doctoral dissertation, Loyola University Chicago (110 pp.).

[10]Coakley, R.M. (2004). *Constructing a prospective model of psychosocial resilience in early adolescents with spina bifida: An application of optimal data analysis in pediatric psychology*. Doctoral dissertation, Loyola University Chicago (233 pp.).

[11]Suzuki, H. (2005). *Prospectively tracing profiles of juvenile delinquents and non-delinquents: An optimal data analysis.* Master’s thesis, Loyola University Chicago (87 pp.).

[12]Hurley, C.L. (2005). *Medical, demographic, and psychological predictors of morbidity and mortality in autologous bone marrow transplant patients.* Doctoral dissertation, Loyola University Chicago (192 pp.).

[13]Wolf, J.L. (2005). *A meta-analysis of primary preventive interventions targeting the mental health of children and adolescents: A review spanning 1992—2003.* Doctoral dissertation, Loyola University Chicago (128 pp.).

[14]Kapunga, C.T. (2006). *Individual, parental and peer influences associated with risky sexual behaviors among African-American adolescents.* Doctoral dissertation, Loyola University Chicago (125 pp.).

[15]Laforce, M. (2006). *A classification profile of high-risk sexual behavior among men who have sex with men.* Doctoral dissertation, Loyola University Chicago (94 pp.).

[16]Jandasek, B.N. (2008). *Predictors of social competence in adolescents with spina bifida.* Doctoral dissertation, Loyola University Chicago (199 pp.).

[17]Snowden, J. (2008). *Predictors of stepping-up from foster homes to residential treatment: A profile of children in the child welfare system.* Doctoral dissertation, Loyola University Chicago (136 pp.).

[18]Donenberg, G.R., Bryant, F.B., Emerson, E., Wilson, H.W., & Pasch, K.E. (2003). Tracing the roots of early sexual debut among adolescents in psychiatric care. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 594-608.

[19]Coakley, R.M., Holmbeck, G.N., & Bryant, F.B. (2006). Constructing a prospective model of psychosocial adaptation in young adolescents with spina bifida: An application of optimal data

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

analysis. *Journal of Pediatric Psychology, 31,* 1084-1099.

[20]Snowden, J.A., Leon, S.C., Bryant, F.B., & Lyons, J.S. (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis. *Journal of Clinical Child and Adolescent Psychology, 36,* 8-18.

[21]Smart, C.M., Nelson, N.W., Sweet, J.J., Bryant, F.B., Berry, D.T.R., Granacher, R.P., & Heilbronner, R.L. (2008). Use of MMPI-2 to predict cognitive effort: A hierarchically optimal classification tree analysis. *Journal of the International Neuropsychological Society, 14,* 842-852.

[22]Han, S.D.. Suzuki, H., Drake, A.I., Jak, A.J., Houston, W.S., & Bondi, M.W. (2009). Clinical, cognitive, and genetic predictors of change in job status following traumatic brain injury in a military population. *Journal of Head Trauma Rehabilitation, 24*, 57-64.

[23]Lyons, A.M., Leon, S.C., Zaddach, C., Luboyeski, E.J., & Richards, M. (2009). Predictors of clinically significant sexual concerns in a child welfare population. *Journal of Child and Adolescent Trauma, 2,* 28-45.

[24]Bryant, F.B., & Yarnold, P.R. (1995). Comparing five alternative factor-models of the Student Jenkins Activity Survey: Separating the wheat from the chaff. *Journal of Personality Assessment, 64,* 145-158.

[25]Bryant, F.B., Yarnold, P.R., & Grimm, L.G. (1996). Toward a measurement model of the Affect Intensity Measure: A three-factor structure. *Journal of Research in Personality, 30,* 223-247.

[26]Yarnold, P.R., Bryant, F.B., Nightingale, S.D., & Martin, G.J. (1996). Assessing physician empathy using the Interpersonal Reactivity Index: A measurement model and cross-sectional analysis. *Psychology, Health, and Medicine, 1,* 207-221.

[27]Layden, B.L., Minadeo, N., Suhy, J., Metreger, T., Foley, K., Borge, G., Crayton, J., Bryant, F.B., & Mota de Freitas, D. (2004). Biochemical and psychiatric predictors of Li^+ response and toxicity in Li^+-treated bipolar patients. *Bipolar Disorders, 6*, 53-61.

[28]Hoffman, L.A.D. (2000). *Marital interaction and depression: A test of the interactional systems model of depression.* Doctoral dissertation, Loyola University Chicago (194 pp.).

Author Notes

The author wishes to thank Scott Leon for creative insights about the evolution of the use of ODA at Loyola. Address correspondence to the author at: Department of Psychology, Loyola University Chicago, 6526 North Sheridan Road, Chicago, IL 60626. Send eMail to: fbryant@luc.edu.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Optimal Data Analysis: A General Statistical Analysis Paradigm

Paul R. Yarnold, Ph.D., and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

Optimal discriminant analysis (ODA) is a new paradigm in the general statistical analysis of data, which explicitly maximizes the *accuracy* achieved by a model for every statistical analysis, in the context of exact distribution theory. This paper reviews optimal analogues of traditional statistical methods, as well as new special-purpose models for which no conventional alternatives exist.

Rarely does a technical report concerning an apparently focused and arcane classification methodology, such as optimal discriminant (data) analysis—ODA, stand a realistic chance of appealing to a diverse scientific community. Even more rarely, however, does one have the opportunity to report the emergence of a new paradigm in the statistical analysis of data.[1] ODA is a highly intuitive, powerful, and exact methodology for the general statistical analysis of data, and this paper reports on the emergence of this paradigm.

ODA is *the* methodology that explicitly maximizes the accuracy of any type of statistical model for the training sample—that is, for the data upon which statistical analysis is performed and upon which the statistical model is based. An increasing awareness of the intuitive appeal of maximizing accuracy (and minimizing errors), and commercial availability of dedicated software, are fueling increasingly widespread application of ODA.[1] Nevertheless, because ODA is relatively new, and therefore relatively few introductory and review resources covering the paradigm are yet widely available, this paper introduces many major concepts and methods of the ODA paradigm.

Initial Assumptions

An ODA model explicitly maximizes the number of correctly classified observations for a specific application. Observations are considered correctly classified when the model assigns them to the class of which they are, in reality, a member, and are misclassified otherwise. The number of misclassifications arising in a given analysis is referred to as the "*optimal value*." It is clear that derivation of a distribution theory for ODA requires investigation of distributions underlying optimal values. Using the *simplest possible data structure* to illustrate derivation of exact distribution theory, imagine a hypothetical application having the following three features.

First, assume one binary class variable. In ODA, a class variable is what one is trying to predict, discriminate, or classify. Examples of binary class variables might include gender (male, female), therapy (drug, placebo), or outcome (success, failure). Class variables may of

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

course consist of more than two levels, but two levels is the simplest case.

Second, assume one random continuous attribute. In ODA, an attribute is a variable that will be employed in an effort to predict the class variable. The continuity assumption implies that every observation will achieve a unique score on the attribute (no ties). Nothing is assumed about the shape of the distribution underlying scores on the attribute, but only that the scores are random—for example, uniform or normal. Single-attribute ODA analyses are referred to as univariable ODA, or UniODA. Because the present case involves a continuous attribute, we are discussing a "continuous UniODA design".

Finally, assume three observations: two from one class, and one from the other class (three observations are required because with two the problem is trivial: the mean of two observations' scores on a continuous attribute is a perfect discriminant classifier for those two observations). Though it is arbitrary, refer to these as classes "1" and "0", respectively. Hereafter, the total number of observations is referred to as *n*, and the number of observations in class *c* as n_c.

Note that only the continuity assumption is capable of being violated by "real-world" data (we return to this point later). The first (binary class variable) and third (*n* in each class) assumptions can never be violated because they exactly define the structure of the design. That is, we are considering a UniODA design with a binary class variable, and with n_1=2 and n_0=1: any deviation from this structure, such as more than two class levels or different sample sizes, simply defines another specific UniODA design.

The UniODA Model

For clarity we give an example of a two-category continuous UniODA model. Imagine that a cardiologist wished to determine if heart rate variability (HRV)—the standard deviation of one's heart rate over a 24-hour period (the continuous attribute), can discriminate patients who die (class 0) versus live (class 1). For a given sample UniODA would provide at least one optimal model, consisting of a *cutpoint* and a *direction*, which when used together explicitly maximize forecasting accuracy: percent accurate classification, or PAC. For example, a UniODA model could be: "if HRV score is greater than (direction) 12.2 (cutpoint), assign that person to class 0; otherwise, assign that person to class 1."

A UniODA model is said to be optimal because the total number of misclassifications resulting from application of the model to the data is minimized, and the number of correct classifications is maximized. In the example, no alternative combination of HRV cutpoint and direction would yield fewer misclassifications than the model which UniODA identified.

Multiple optimal models which all yield the same maximum PAC may occur for a given data set. For example, two different HRV cutpoints might result in the same overall number of misclassifications, yet one model may have greater sensitivity (ability to accurately classify members of class 1) and lower specificity (ability to accurately classify members of class 0) than the other model. In such cases, it is necessary to select one optimal model, preferably before conducting the analysis, using an appropriate decision heuristic.[1] Examples of such selection heuristics include the sensitivity or specificity heuristic (select the model having greatest sensitivity or specificity, respectively), or the balanced performance heuristic (select the model with the smallest difference between sensitivity and specificity).[1]

Exact Distribution Theory

We are now ready to derive the theoretical distribution of optimal values for a two-category continuous UniODA design with n_1=2 and n_0=1. First, it is necessary to determine the set of all possible outcomes that could occur if the attribute were continuous and random. In order to differentiate the two observations from class 1, they will be called "1A" and "1B."

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

There are six possible outcomes: one is that the value of the attribute for observation 1A is greater than that for observation 1B, which in turn is greater than that for the observation from class 0. Symbolically, {1A > 1B > 0}. The five other possible outcomes are: {1A > 0 > 1B}; {1B > 1A > 0}; {1B > 0 > 1A}; {0 > 1A > 1B}; and {0 > 1B > 1A}. Because the attribute was random, each of these six possible outcomes is equally likely, with a probability of 1/6.

Next, it is necessary to determine the optimal value for each of the six possible outcomes. This, of course, means that UniODA must be performed for each of the six possible data configurations.[1] Two of the six possible outcomes (those in which the attribute of the class 0 observation lies between the attributes of the two class 1 observations) have an associated optimal value of 1 misclassification, because at least one observation will be misclassified regardless of where the cutpoint is placed). The other four possible outcomes (in which the two class 1 observations can be perfectly separated via a cutpoint from the class 0 observation) have an optimal value of 0 misclassifications. Cumulating optimal values over the set of possible outcomes gives the theoretical distribution of optimal values for this UniODA design: the probability of an optimal value of 0 is 4/6, and the probability of an optimal value of 1 is 2/6.

Enumerating in this manner the theoretical distribution of optimal values for balanced (equal number of class 0 and 1 observations), continuous, two-tailed (no *a priori* hypothesis was specified) UniODA designs required a CRAY-2 supercomputer—which only achieved results for $n \leq 30$ due to exponential increases in the number of combinations.[2] Examination of the resulting table of optimal values for *post hoc* UniODA revealed organization which motivated discovery[3] and proof[4] of a closed-form solution for one-tailed confirmatory UniODA.

Inexact Measures

What if data aren't continuous, and there are ties—violating the continuity assumption? Discontinuity in empirical data is thought to reflect imprecise measurement, and not as compromising of theoretical probabilities[5], but this begged the question of exactly how imprecise can measurement become before the theoretical probabilities become compromised? This line of thinking naturally led to the question of what would occur for a binary attribute—and it was then that we understood that the binary attribute problem was the optimal analogue to chi-square analysis, and the continuous attribute problem was the optimal analogue to *t*-test. Proceeding with binary enumeration we found the binary and continuous distributions differ. This finding motivated two important insights.

First, there is a theoretical dimension—which we call *precision*—which may be used to describe the metric underlying the attribute for any specific UniODA problem. The precision dimension is bounded at the extremes by binary data (least precise) versus continuous data (most precise). Just as specific distribution theory can be derived for the extreme poles of the precision dimension, so too can *exact* distribution theory be derived for every specific attribute measure metric: for example, if the attribute is measured using a 7-point Likert scale, then derive distribution theory by assuming a 7-point Likert scale was used. As it is possible to derive distribution theory that assumes that the specific measure metric actually used in a given application was in fact used, distribution theory for ODA can be based strictly on structural features of a problem, and such distribution theory *will never be violated* by data for a given application.

The second insight is that UniODA is an optimal alternative to common conventional statistical methods: Student's t-test is often used to analyze data involving a binary class variable and a continuous attribute, and chi-square is often used to analyze data involving a binary class variable and a binary attribute. UniODA

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

may also be used, and exact distributions may be determined for, designs that lie anywhere on the precision dimension—anywhere between the binary and continuous polar extremes. This is not true for conventional statistical procedures.

ODA as an Alternative to Conventional Statistical Methodologies

Encouraged by early success, we began programmatic research to assess the domain of experimental designs and data configurations that may be addressed using UniODA. We next investigated multicategory problems involving class variables with more than two levels. For a continuous attribute, multicategory UniODA is analogous to oneway analysis of variance, and for a binary attribute it is analogous to log-linear analysis.[6,7]

UniODA, and other models within the ODA paradigm, clearly can be used to analyze different data configurations that are evaluated using a host of different conventional statistical methods. Why should ODA be used rather than a host of conventional methods?

First, only ODA explicitly maximizes (weighted) classification accuracy and provides a forecasting model for every application. Not only do conventional methods fail to explicitly maximize PAC, but many, such as *t*-test or chi-square, also fail to provide a forecasting model.

Second, no matter what the nature of a particular data configuration might be—for example, the number of class levels, attribute metrics, or class sample-size imbalances, the classification performance of every ODA model is summarized using a normed measure of effect strength, called effect strength for sensitivity, or ESS.[1] On this index 0 represents the level of classification performance that is expected by chance, and 100 represents perfect, errorless classification. No such intuitive, universal index can be used to compare the effect strength of different conventional methods such as analysis of variance, logistic regression, and tau.

Third, conventional methods require assumptions regarding the nature of the data. Unlike ODA—for which distribution theory is exact for every design, conventional methods are inappropriate when the data violate their assumptions. Whereas the assumptions of ODA must conform to the data, data must conform to the assumptions of conventional methods.

Finally, with ODA a *single methodology* may be *optimally applied* to analyze a *host of problems*, while with the conventional approach a *host of methods* may be *suboptimally applied* to analyze *a single problem*. ODA is therefore simultaneously more unique *and* parsimonious than conventional methods.

To illustrate the flexibility and power of ODA as a general statistics paradigm, below we describe different common data configurations and conventional methods often used in their analysis, and the corresponding ODA model.

Binary Class Variable and BinaryAttribute

The most common conventional method for analyzing data of this type is chi-square analysis: the ODA analogue is two-category binary UniODA. Chi-square is an approximate statistic that should not be used when the expected value for a given cell (cells are formed by cross-tabulating the class variable with the attribute) is less than five.[8] In contrast, binary UniODA is an exact statistic with no such restriction: one- and two-tailed estimated *p* by UniODA and Fisher's exact test are isomorphic except in a hypothetical degenerate condition.[1]

It is easy to show that UniODA may be particularly useful in small sample designs. For example, imagine a problem with $n = 6$, three observations from class 0 all scoring 0 on the attribute, and three observations from class 1 all scoring 1. Chi-square can't be used to analyze this problem, as the expected value is less than five in all four cells. When analyzed using two-tailed binary UniODA, a single optimal model (if attribute < 0.5 then class = 0; else class = 1)

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

emerged that achieved 100% PAC, $p<0.032$. No systematic review/comparison of chi-square versus binary UniODA has yet been reported.

Binary Class Variable and Multiple Binary Attributes

The most common linear methods for analyzing data of this type include log-linear or logistic regression analysis. Completely binary problems are easiest for ODA to solve, but can be problematic for conventional methods, with aspects including marginal imbalance, sparse cells, singularities, and structural zeros (some design cells don't exist), for example, rendering binary data difficult or impossible to analyze. The optimal linear analogue is binary MultiODA—a linear model which uses two or more attributes to explicitly maximize classification accuracy (discussed ahead).

For example, we reanalyzed data from a study designed to predict if 120 persons with AIDS would require home care or structured long-term care (the class variable) on the basis of three binary attributes which assessed the attitudes of patient and physician towards long-term care, and whether the patient had mental impairment.[9] The data were "ill-condiioned" and thus could not be analyzed by log-linear or logistic regression methods. MultiODA, however, found a two-attribute model that achieved 93.3% PAC in <1/20 CPU second on a 33MHz 386 microcomputer running a special-purpose ODA search algorithm (discussed ahead).

Binary Class Variable and Continuous Attribute

Among the most frequently reported of statistical tests, Student's *t*-test is a common conventional procedure for analyzing data of this type. The ODA analogue is two-category continuous UniODA.

It is easy to construct a hypothetical problem for which *t*-test fails to find a significant intergroup mean difference on the attribute, while UniODA detects nearly perfect intergroup discriminability. Imagine that ten class A observations each score a value of 0 on the attribute; nine class B observations all score 1, and a tenth class B observation scores -9. Because the mean difference on the attribute between groups is zero, *t*-test would conclude that the groups can't be discriminated whatsoever by the attribute. But, with UniODA, 95% of the observations are correctly classified—nearly perfect intergroup discriminability. Systematic research contrasting UniODA and *t*-test is not yet available.

Binary Class Variable and Multiple Continuous Attributes

Common linear methods for analyzing data in this configuration are linear discriminant analysis, logistic regression analysis, and one-way multivariate analysis of variance.[6,7] The linear ODA analogue is continuous MultiODA, but UniODA has been used with great success to maximize accuracy achieved by suboptimal models.[10,11]

Monte Carlo research is often used to contrast continuous MultiODA versus conventional statistical methods.[12,13] A difficulty with such simulation research is that the experimental data are generated using idealized routines that meet criteria—such as normally distributed data and coincident covariance, which are important for conventional statistical methodologies but which are no substitute for "real-world" data collected by naturalistic empirical observation. Our strategy has been to analyze a variety of different applications using MultiODA, and then compare the performance against suboptimal methods such as Fisher's discriminant or logistic regression analysis, using training and validity data. Early results are encouraging, but more research is needed to compare "in the field" classification performance of MultiODA versus conventional procedures.[9,14,15]

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Binary or Multicategory Class Variable and Continuous and Binary Attributes

Multinomial logistic regression analysis is a commonly employed conventional analysis for problems of this type. The linear optimal analogue is MultiODA, with weights used to reduce problem size by eliminating redundant data profiles (discussed ahead). Little research using either approach is available, and to our knowledge no prior research comparing these approaches has yet been published (until now).

Analyzing credit screening data for a British bank, our objective was to develop a model to predict credit worthiness (the class variable) for a sample of 325 credit applicants. Attributes were two binary variables and a third 4-point ordinal attribute. A nonparametric classification methodology that performed sample stratification based on a recursive chi-square procedure identified four interaction terms used as attributes in follow-up analysis. With these data logistic regression analysis and MultiODA both achieved 90.5% PAC in training analysis, but the latter model used one less term (and thus was more efficient and parsimonious) than the former model. Comparing the two models using jackknife validity analysis revealed that PAC for the MultiODA model was stable, but regressed to 83.1% for the obviously over-determined logistic regression model.

Multicategory Class Variable and Polychotomous Attribute

Common conventional methodologies for analyzing these designs include chi-square, log-linear, or multinomial logistic regression analysis. The optimal analogue is multicategory UniODA. As was true for designs that involved one binary class variable and multiple binary attributes, issues such as structural zeros, sparse cells, imbalanced marginal distributions, small samples, and multicollinearity may spell disaster for conventional designs. As discussed earlier, these are *not* problems for ODA.

It is easy to construct an example for which conventional analyses are inappropriate, but for which multicategory UniODA is ideal. For example, imagine a problem with a three-category (A, B, C) class variable, with each category having three observations. Further imagine all three class A observations scored a value of 1 on the attribute; all three class B observations scored a 2, and all three class C observations scored a 3. Although the small sample renders conventional methods inappropriate, a multicategory UniODA achieved 100% PAC, two-tailed p<0.01.

Multicategory Class Variable and Continuous Attribute

The most common conventional analysis used for such designs is oneway analysis of variance, and the optimal analogue is multicategory UniODA. As for *t*-test, distribution theory for analysis of variance is highly sensitive to assumption violations.[5] Such data can present insurmountable problems for multinomial logistic regression, because of small samples, sparse cells, and marginal imbalance, particularly when polychotomous attributes are thrown in the mix: for example the analysis will fail if a degenerate attribute—which has fewer response categories than the class variable has levels—is included in the analysis.

As an example of a three-category UniODA, imagine the following hypothetical data set, problematic for conventional methods due to the small sample, the presence of outliers, heterogeneity, the presence of zero variance for one group, and non-normality (in Table 1, *X* is the attribute).

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

TABLE 1: Hypothetical data set for three-category UniODA

Class	*X*	Class	*X*	Class	*X*
A	29	B	35	C	5
A	30	B	35	C	42
A	31	B	35	C	43
A	50	B	35	C	50

In this example the mean *X* of classes A, B, and C is exactly equal, so F=0. However, the UniODA model (if *X* < 33 then class = A; if *X* > 38.5 then class = C; else class = B) correctly classified 10 of the 12 data points: overall and mean PAC over all three groups is 83.3%, two-tailed *p*<0.05.

Ordered Class Variable and Continuous and/or Binary Attributes

Among the many types of nonparametric methods in use, Kendall's tau is arguably the least problematic procedure conventionally used to evaluate associations among ordinal (ranked) data.[16] Tau is a *computed* index for evaluating the relationship between *two* ordered variables: collect data, compute tau, and "it is what it is." Ahead we show that multicategory MultiODA can be used to determine criterion weights for two *or more* attributes to generate a summary score which explicitly *maximizes* tau.

Receiver Operator Curve (Signal Detection) Analysis

Bayesian classification methods are commonly used to evaluate the discriminating power of attributes.[17] Such procedures typically aim to maximize the sensitivity, specificity, or some combination of sensitivity and specificity achieved using an attribute. Since ODA models may be derived which explictly maximize sensitivity, specificity, or any weighted composite of sensitivity and specificity, either for individual attributes or for sets of attributes, we call this application "optimal signal detection analysis."

In summary, it is a common practice to employ multiple different statistical methods, each requiring data to satisfy different essential assumptions, to analyze a given sample of data in numerous "different" (actually related) ways. We recommend using a single statistical method to analyze data with one objective function in mind: maximizing accuracy. The utility of this approach will undoubtedly receive increased attention as researchers learn more about the unrivaled generalizability and power of ODA across different data configurations.

Fast MultiODA Solutions

Early research was highly productive, and new applications for UniODA models were discovered routinely as new data structures were considered.[1] As data configurations became increasingly complex, so did ODA models, and researchers began formulating and investigating optimal linear models for designs with a binary class variable and two or more ordinal and/or binary attributes: an optimal analogue of logistic regression or Fisher's discriminant analysis. These multivariable ODA models are called "MultiODA," for short.

Although UniODA problems can easily be solved for enormous samples, MultiODA problems may be computationally intractable for tiny samples, even on the fastest computers. Several procedures affording reductions of an order of magnitude or more in solution time for

MultiODA problems were recently developed, and analysis is feasible for enormous samples in favorable circumstances. Review of MultiODA here will be brief: so much work has focused on MultiODA models that a review is warranted. Below we review two fast new methods to solve MultiODA problems: MIP45 is a mixed integer formulation, and WARMACK a special-purpose search algorithm. These methods are extended for nonlinear and multicategory MultiODA.

MIP45

The first approach to computing a MultiODA solution that we shall discuss is a mixed integer linear programming formulation called MIP45, in which the discriminant function is normalized so the sum of the absolute values of the coefficients adds to one.[18] This enables one to determine, for each constraint, a lower bound for the value of the problem parameter, *M*. This is in distinction to previous formulations of this problem, where *M* is defined as "a very large number." Since the value of *M* can be kept low for each constraint, the branch-and-bound procedure can fathom branches more quickly than other formulations. Also, fewer branches need to be stored in memory, and computation time is reduced.

We compared computational resources needed to solve a problem in classification of medical residency applicants using MIP45 and a recent formulation that did not limit *M*. The problem had 3 attributes and 49 observations. Running the SAS/OR optimization package on an IBM 3090/600 mainframe computer, MIP45 solved the problem in 48 CPU seconds, versus 268 CPU seconds using the other formulation: MIP45 analyzed 2,896 branches, versus 14,549 branches using the other formulation.

MIP45 can be extended to obtain MultiODA solutions which maximize the weighted number of satisfied inequalities. As for UniODA, this is useful in two different contexts.

First, the weights may represent the return obtained in the correct classification of an observation. For example, consider the problem of predicting whether the price of a stock will rise or fall over a given time horizon, given a series of market indicators and price measurements. If the prediction is for a rise in the stock price, the stock will be purchased. Conversely, if a fall in the price is predicted, the stock will be sold short. The weighted MultiODA solution of this problem would maximize the trading return over the set of observations.

The other context in which the weighted criterion is useful occurs when the number of observations in each class differs. In this case, the weighted MultiODA solution balances the number in each class by maximizing the mean PAC over the two classes.

A useful extension of MIP45 involves fixing the sign of the discriminant coefficients (e.g., in a confirmatory design). In fact, bounds or any linear constraints on the coefficients may be imposed. Yet another type of constraint which can be modeled is any Boolean function of actual or predicted class membership among the observations. One example of this would be forcing certain observations to be classified correctly in the MultiODA solution (if this is feasible). Another example would be forcing observation A to be assigned to a certain class only if observation B is similarly classified.

Finally, a method for reducing the problem size can be applied when multiple observations share identical values for all attributes. In this case, these observations may be aggregated into a single observation, with a weight applied to the objective function. This procedure is especially useful with binary attributes: we solved binary MultiODA problems having five attributes and one *million* observations in less than ten CPU seconds on an IBM 3090/600.

WARMACK

A second approach to obtaining fast solutions to MultiODA problems involves our adaptation of a fast search algorithm initially developed by Warmack and Gonzalez (hence the origin of the name we use to refer to the method).[19] With this method we obtained a reduction of an order of magnitude or more in computation time versus the MIP45 approach.

We conducted Monte Carlo research to investigate the computer resources required by this algorithm as a function of *n*, the number of attributes, and the relative discriminability of the data. Problems having 2 attributes and 700 observations can be solved in less than one CPU minute on an IBM 3090/600. This is also true for problems with 3 attributes and 200 observations, or 4 attributes and 100 observations. Our findings show that the number of attributes exerts greater influence on computation time than *n* or relative discriminability of the data.

Extension of MultiODA to Nonlinear and Multicategory Problems

MultiODA may be extended to a large class of nonlinear separating surfaces. This is accomplished by defining attributes which are polynomial functions of the original attributes. Any nonlinear function which is linear in the parameters of the original attributes may be modeled in this manner.

It is also possible to solve multicategory problems involving more than two class levels using either MIP45 or WARMACK. There are two ways to accomplish this. If there are *k* class categories, the first method is to determine the ODA solution obtained with *k*-1 separating surfaces in parallel with each other. From a computational standpoint, this is equivalent to adding an extra attribute for each additional class.

The second method involves the determination of *k* different discriminant functions: an observation is assigned to the class for which the maximum value is obtained over these functions. If there are *p* original attributes, this is equivalent to a MultiODA problem with *p* times *k* attributes.

In conclusion, MIP45 and WARMACK make feasible the solution of much larger MultiODA problems than have been possible to solve previously, particularly for binary problems. Optimal analogues to conventional statistical methods are now available to researchers. However, ODA is far more than simply an optimal analogue to conventional statistics.

Special-Purpose ODA Models

The flexibility of the ODA methodology lends itself to special-purpose classification applications for which there are no alternative conventional statistical procedures. Indeed, the number of different ODA models that may be created is limitless, due to the inherently infinite number of possible unique classification applications. Nevertheless, below we describe some specialized ODA models that should be of great utility across a variety of applications.

Minimizing the Number of Terms in a MultiODA Solution

When performing an analysis, it is desirable to obtain a solution with as few terms as possible, in light of the principle of parsimony. This can be achieved in the context of the MIP45 formulation: an upper bound is set on the number of misclassifications, and the number of attributes used in the solution is minimized. This results in a more parsimonious model, with a corresponding increase in statistical power.

Optimal Attribute Subsets

A related problem is the determination of an optimal subset of attributes with exactly *k* members. This also is an extension of MIP45. This procedure is useful when the ratio of number of attributes to number of observations is too high to yield a meaningful model, or when redundant (multicollinear) attributes are present.

For example, we used this procedure to discriminate 15 Type A from 15 Type B (class variable) undergraduates using a subset of 20 items (attributes) from the Bem Sex-Role Inventory. With *k* specified at 2 attributes, MultiODA

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

identified a single solution that achieved 93.3% PAC; with *k* specified at 3 attributes MultiODA identified a single solution with 100% PAC. These problems required 91.9 and 73.9 CPU seconds to solve on an IBM 3090/600 computer running SAS/OR. When the attributes selected by MultiODA were evaluated using logistic regression analysis, 90% PAC was achieved for both the 2- and 3-attribute models. The best 2-attribute model identified using stepwise logistic regression achieved 90% PAC, and the best 3-attribute model achieved 93.3% PAC.

Integer-Valued Coefficients

UniODA may be used to solve MultiODA problems in which the model weights for the attributes (the discriminant coefficients) are constrained to take on a small set of values. For example, in a problem having *p* attributes, the discriminant coefficients restricted to the values 0, 1, or -1, and the threshold coefficient unconstrained, all optimal solutions may be found by solving $3^p/2$ UniODAs. In general, for *k* possible coefficient values and *p* attributes, $k^p/2$ UniODAs are solved. If *k* and *p* are relatively small, then few computational problems arise due to the fast speed of UniODA. An additional benefit of this analysis is that optimal attribute subsets of every size are evaluated. We solved a problem with 3 coefficient values, 8 attributes, and 900 observations in 716 CPU seconds on a 33Mhz 386 microcomputer.[15]

Optimal Selection of Observation Subsets with Unknown Class Membership

In some problems, observations are available for which class membership is unknown. Typically, exactly *k* of these observations are to be acted upon in some manner. The initial phase of the MultiODA approach to this problem involves partitioning observations into two sets: the decision set, consisting of observations with unknown class membership, and the evaluation set, consisting of observations with known class membership.

To illustrate this, consider the problem of selecting *k* job applicants from a pool of applicants. The attributes may reflect measures of previous employment experience and skills required to perform the job task. The evaluation set is comprised of previously hired individuals who have been measured on these attributes. Each individual in the evaluation set is weighted by a performance index, in this case a measure of job performance. The decision set is comprised of the pool of job applicants, *k* of whom are to be selected for employment, and all of whom have been measured on the attributes. MultiODA identifies a solution which maximizes the weighted number of inequalities in the evaluation set, such that exactly *k* inequalities in the decision set are satisfied.

Or, consider the problem of selecting prisoners to be released under a court mandate which requires that exactly *k* must be released, due to overcrowding. Here the decision set is the current population of prisoners, and the evaluation set are those prisoners who previously were released. The performance index, which is to be minimized, is a measure of mayhem produced by the previously released prisoners.

Other interesting applications of this method lie in the areas of market research, investment selection, and pattern recognition.

Ordered Class Variables

Another fruitful area of investigation relates to the use of MultiODA in analysis of data which have been sorted into ordered (ranked) categories. MultiODA is used to maximize the goodness of fit between the actual and predicted category assignments. Kendall’s tau is a similarity index widely used for comparison of two ranked sequences, and is proportional to the number of satisfied inequalities between paired observations. Thus, MultiODA finds a linear discriminant function which *maximizes* the value of Kendall’s tau. It is worthwhile to

2155-0182/10/$3.00

note that this situation differs from the multi-category case in that the latter corresponds to the analysis of *unordered* categories.

Optimal Nonparametric Linear Multiple Regression

A distribution-free approach to multiple linear regression is available using the Kendall's tau procedure. Initially observations are ranked according to their values on the dependent measure. MultiODA is then used to find the optimal predicted rank sequence. As a final step, an inequality-constrained multiple linear regression problem is solved for each optimal rank sequence. This quadratic program uses sum-of-squared-error as the objective function, and the inequalities corresponding to the paired observations as constraints. The linear model produced by this procedure is the model with the highest R^2 for which the value of Kendall's tau is the maximum achievable overall. If multiple optimal sequences exist, the solution with the highest R^2 is selected. We have solved such a problem with 3 independent variables and 22 observations in 49 CPU seconds on a 50 MHz 486 microcomputer.

Optimal Templates

Another interesting application of MultiODA lies in the design of optimal templates. To illustrate this, imagine an individual is given a list of questions and set of possible responses for each question, one of which is to be selected as the individual's answer to that question. Each question is answered by filling in a circle (e.g., on an "IBM answer form") corresponding to a selected answer. The class membership of each individual is known. The objective of this MultiODA procedure is to produce a template, that is, a series of holes on an opaque sheet, so that overlaying the template on the answer sheet and counting the number of filled-in holes produces a discriminant score for the individual. This score is compared to the cutpoint obtained by MultiODA in order to assign class membership to individuals. This assignment minimizes the number of classification errors.

This problem was formulated as a pure integer program. As an example, consider the application of creating a template for personnel selection purposes. A 38-item questionnaire, with each item answered as true or false, was completed by 107 employees of a corporation, 70 of whom were known desirable workers, and 37 of whom were known undesirable workers. MultiODA identified a template which resulted in 74.8% PAC, requiring 26 CPU minutes on an IBM 3090/600 running SAS/OR.

MultiODA with Boolean Attributes

The ODA approach of minimum error may also be applied to classification problems with purely logical attributes. In this case, the decision rule involved in the assignment of an observation to a class is a Boolean function of logical attributes which have been measured for that observation. We wish to find a Boolean function with at most *k* terms which minimizes the number of misclassifications. Alternatively, we may look for a function with at most *k* misclassifications which minimizes the number of logical terms. These problems can be formulated as integer programs, or solved in crude brute force manner via exhaustive enumeration.

Consider the following application as an example of this procedure. A pair of emergency physicians independently diagnosed 51 patients with hip trauma for bony abnormality. Each physician rated each patient as abnormal or normal based a measure of sound conduction, and also based on physical inspection. Presence of bony abnormality (the class variable) was independently determined radiographically. A Boolean MultiODA identified a single optimal solution that achieved 96% overall PAC. The optimal decision rule was: if either physician rates either attribute as abnormal, then classify the observation as abnormal; else classify the observation as normal.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Classification Tree Analysis

Hierarchically optimal classification tree analysis, or CTA, is an algorithm which chains UniODA analyses together so as to stratify the sample in a manner that explicitly maximizes ESS.[20] As for MultiODA, discussion of CTA lies outside the domain of this manuscript: sufficient work using CTA has accumulated so that a comprehensive review is warranted.

Summary

Research described herein, indeed the sum total of all of the world's knowledge in this field to date, merely scratches the surface of what ODA entails, what ODA offers. Although we can only imagine what we must be missing, it is clear to see that ODA is a powerful new paradigm in the statistical analysis of data. It is intuitively appealing, in the mathematical modeling of any process, that the model should make as few mistakes as possible. This is the essence of the ODA approach. Its fruitfulness, particularly in its application to the analysis of problems previously unanalyzable, is an indication of its value as a general-purpose problem-solving tool. Because ODA is inherently distribution- and metric-free, it avoids the necessity of making distributional assumptions required by conventional parametric methods. In ODA, powerful modeling capabilities of mathematical programming are joined with the inferential capabilities of statistics. Furthermore, one may combine different ODA methods so that every problem can be formulated in terms of its own unique characteristics. It thus seems appropriate to postulate that, in the area of optimal statistics, the best surely is yet to come.

References

[1]Yarnold PR, Soltysik RC. (2005). *Optimal data analysis: a guidebook with software for windows*. Washington DC, APA Books.

[2]Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, *22*, 739-752.

[3]Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: one-tailed hypotheses. *Educational and Psychological Measurement*, *54*, 646-653.

[4]Carmony L, Yarnold PR, Naeymi-Rad F (1998). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research*, *74*, 223-238.

[5]Bradley JV (1968). *Distribution-free statistical tests*. Englewood Cliffs NJ, Prentice-Hall.

[6]Grimm LG, Yarnold PR. (Eds.) (1995). *Reading and understanding multivariate statistics*. Washington DC, APA Books.

[7]Grimm LG, Yarnold PR. (Eds.) (2000). *Reading and understanding more multivariate statistics*. Washington DC, APA Books.

[8]Yarnold JK (1970). The minimum expectation of chi-square goodness-of-fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, *65*, 864-886.

[9]Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine, 10,* 601-606.

[10]Yarnold PR, Soltysik RC (1991). Refining two-group multivariable classification models using univariate optimal discriminant analysis. *Decision Sciences*, *22*, 1158-1164.

[11]Yarnold PR, Hart LA, Soltysik RC (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant

analyses. *Educational and Psychological Measurement*, *54*, 73-85.

[12]Rubin PA (1990). Heuristic solution procedures for a mixed-integer programming discriminant model. *Mangerial and Decision Economics*, *11*, 255-266.

[13]Stam A, Joachimsthaler EA (1990). A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *European Journal of Operational Research*, *46*, 113-122.

[14]Yarnold PR, Soltysik RC, Martin GJ (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, *13*, 1015-1021.

[15]Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine*, *17*, 2405-2414.

[16]Reynolds HT (1977). *The analysis of cross-classifications*. New York NY: Free Press.

[17]Kraemer HC (1992). *Evaluating medical tests: objective and quantitative guidelines*. Newbury Park CA, Sage.

[18]Soltysik RC, Yarnold PR (2010). Two-group MultiODA: mixed-integer programming solution with bounded *M*. *Optimal Data Analysis*, *1*, 30-37.

[19]Soltysik RC, Yarnold PR (1994). The Warmack-Gonzalez algorithm for linear two-category multivariable optimal discriminant analysis. *Computers and Operations Research*, *21*, 735-745.

[20]Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, *16*, 1451-1463.

Author Notes

Address correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans Street, Suite 330, San Diego, CA 92106. Send Email to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Maximizing Accuracy of Classification Trees by Optimal Pruning

Paul R. Yarnold, Ph.D., and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

We describe a pruning methodology which maximizes effect strength for sensitivity of classification tree models. After deconstructing the initial "Bonferroni-pruned" model into all possible nested sub-branches, the sub-branch which explicitly maximizes mean sensitivity is identified. This methodology is illustrated using models predicting in-hospital mortality of 1,193 (Study 1) and 1,660 (Study 2) patients with AIDS-related *Pneumocystis carinii* pneumonia.

Classification tree models typically begin with a root variable which has two eminating branches and separates the sample into two partitions. In such applications the tree model may be said to consist of two parts: the left-hand side, and the right-hand side. Extending this methodology to applications involving more than two eminating branches is straightforward: for example, with three eminating branches there are the left-hand, middle, and right-hand branches. To facilitate clarity, this article considers applications having two eminating branches.

Identifying the tree model which explicitly maximizes mean sensitivity—and thus the effect strength for sensitivity (ESS), first necessitates identifying every possible sub-branch for every branch eminating from the root variable. For example, imagine a left-hand branch consisting of three nodes: A (root), B (middle attribute) and C (at the end of the branch). This branch has two nested sub-branches: one involves only nodes A and B (C collapsed into B), and the other involves only node A (C and B collapsed into A). For clarity of exposition refer to the *left* branch with *three* attributes (A, B, C) as "L3"; to the trimmed *left* branch with *two* attributes (A, C collapsed into B) as "L2"; and to the trimmed *left* branch with *one* attribute (C and B collapsed into A) as "L1". Also imagine this hypothetical tree model had a right-hand branch consisting of two nodes: A (sides have the same root attribute) and D (at the end of the branch). The *right* branch involving *two* attributes (A, D) is called "R2", and the trimmed *right* branch with *one* attribute (D collapsed into A) is called "R1". The first step of optimal pruning involves obtaining a confusion table (rows are the actual class category, columns are the predicted class category) for all (sub)branches of the original tree model: here, for L1, L2, L3, R1, and R2.

The second step in finding the tree model having maximum sensitivity involves obtaining every unique combination of left and right (sub)branch: in the present example the six unique combinations are L1-R1, L2-R1, L3-R1, L1-R2, L2-R2 and L3-R2. Next, combine (or "integrate") the confusion tables for each of the six different combinations. Finally, the table with greatest mean sensitivity may be identified by direct observation. The optimized model is the combination of (sub)branches with associated confusion table having maximum ESS.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

An Example of Optimal Pruning: Predicting In-Hospital Mortality

The major cause of hospitalization and death for people with HIV infection early in the AIDS epidemic, *Pneumocystis carinii* pneumonia or PCP had in-hospital mortality rates as high as 60% in the 1980s.[1] Here we demonstrate pruning to maximize ESS for a model obtained via classification tree analysis (CTA) to predict in-hospital mortality due to PCP.[2] Analysis was performed for 1,193 patients (89% of the total sample) with complete data for model attributes, who were discharged alive (N=988) or who died in-hospital (N=205). Derived manually using UniODA software[3,4] the CTA model involved four attributes: alveolar-arterial oxygen gradient ($AaPo_2$) is the difference in partial pressure of oxygen between the pulmonary system and the blood (elevated values indicate more severe pneumonia); body mass index is a measure of nutritional status that is predictive of poor short- and long-term survival rates; and prior AIDS indicates if the current episode of PCP is the first clinical evidence that full-blown AIDS has developed (at the time the data were collected, patients with prior history were more likely to be severely ill, develop multiple complications of AIDS, and die). The CTA model (Figure 1) yielded ESS=21.2, a relatively weak effect.

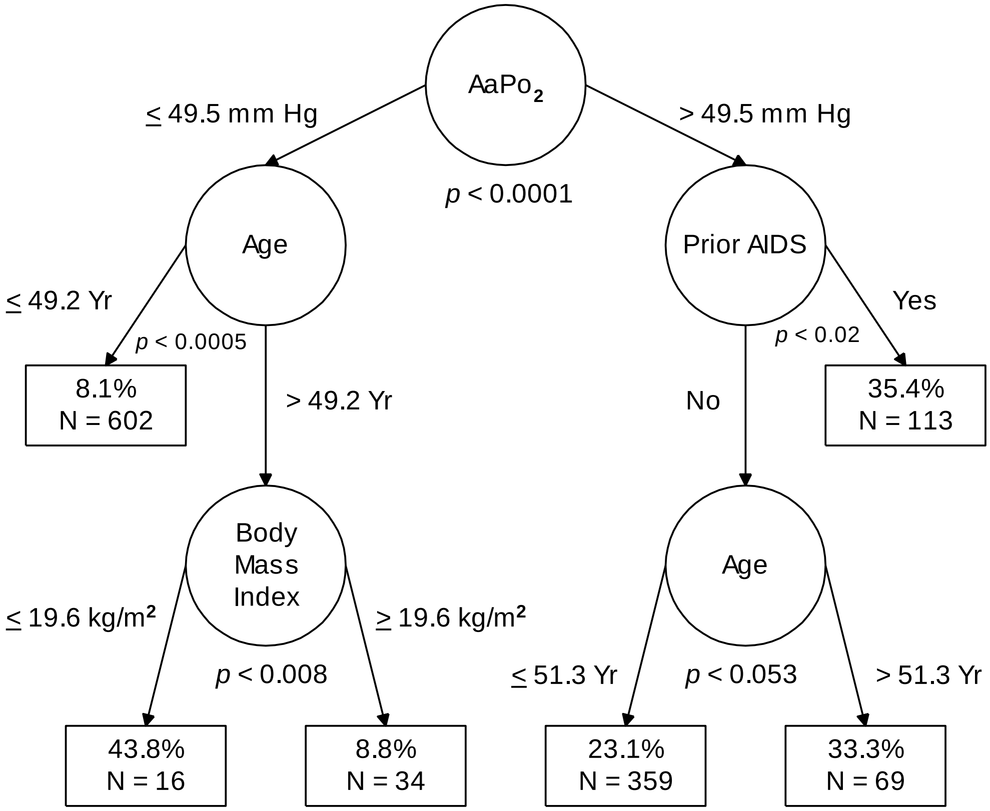
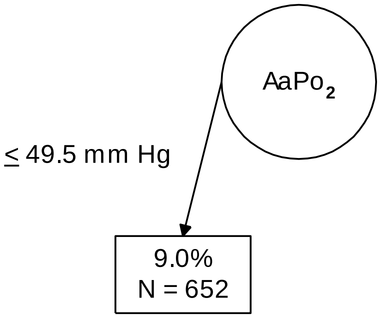


Figure 1: Initial Non-Pruned CTA Model of In-Hospital Mortality

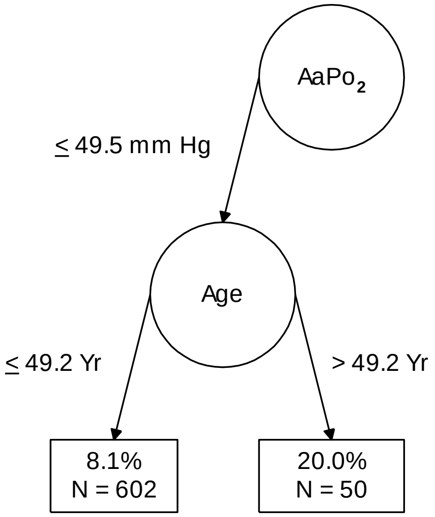
The root variable of the initial non-pruned CTA model has two eminating branches and therefore left and right sides, and each side has three nodes. For Step 1 of the optimal pruning procedure, Figure 2 gives schematic illustrations of L1-L3 and R1-R3, and their corresponding confusion tables, respectively.

Figure 2A:
L1 Sub-Branch and Confusion Table



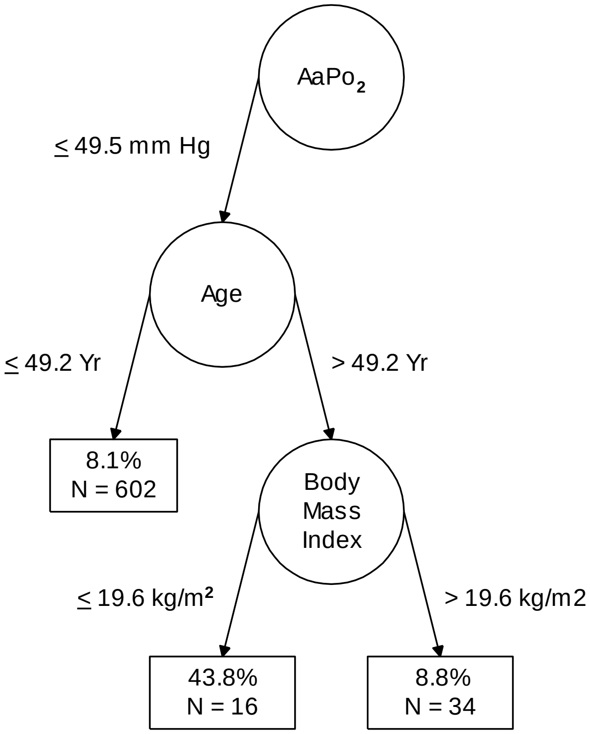
		L1 Predicted	
		Alive	Dead
Actual	Alive	593	0
	Dead	59	0

Figure 2B:
L2 Sub-Branch and Confusion Table



		L2 Predicted	
		Alive	Dead
Actual	Alive	553	40
	Dead	49	10

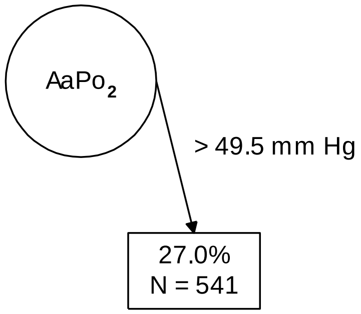
Figure 2C:
L3 Sub-Branch and Confusion Table



		L3 Predicted	
		Alive	Dead
Actual	Alive	584	9
	Dead	52	7

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

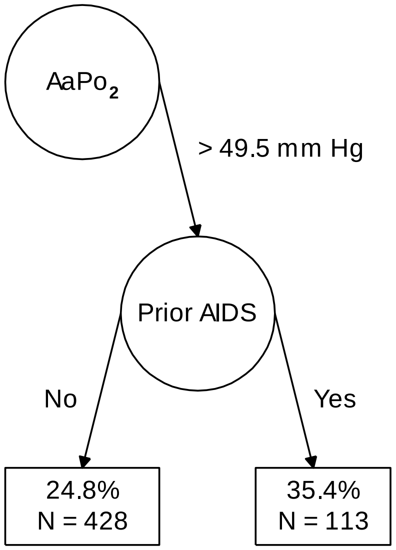
Figure 2D:
R1 Sub-Branch and Confusion Table



R1 Predicted

		Alive	Dead
Actual	Alive	0	395
	Dead	0	146

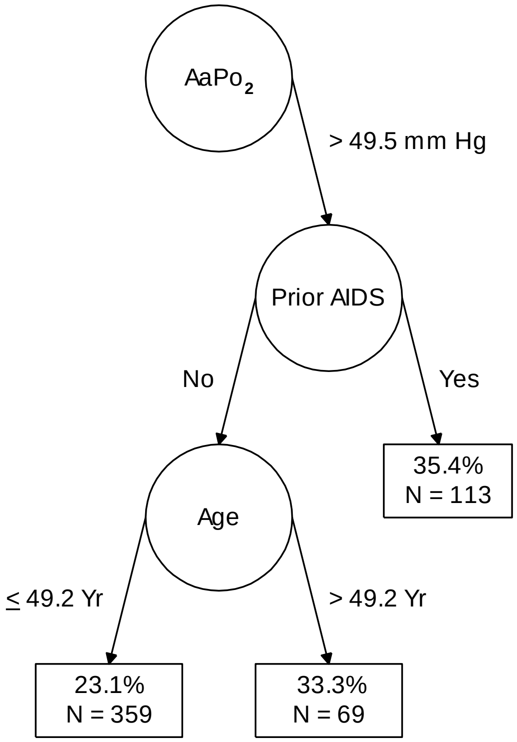
Figure 2E:
R2 Sub-Branch and Confusion Table



R2 Predicted

		Alive	Dead
Actual	Alive	322	73
	Dead	106	40

Figure 2F:
R3 Sub-Branch and Confusion Table



R3 Predicted

		Alive	Dead
Actual	Alive	276	119
	Dead	83	63

Figure 2: All Possible Sub-Branches of the Initial Non-Pruned CTA Model, and Corresponding Confusion Tables

For the final step of the optimal pruning procedure, Table 1 gives integrated confusion tables for all nine possible combinations of left (L1-L3) and right (R1-R3) sub-branches, and their associated mean sensitivity and ESS. As seen in Figure 3, the combination L1-R3 has the greatest mean sensitivity (66.9%), corresponding to optimized ESS=33.7.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Table 1: Classification Results for Every Combination of Left (L1-L3) and Right (R1-R3) Sub-Branch

Model	Confusion Table			Model	Confusion Table			Model	Confusion Table		
L3-R3		Predicted		*L3-R2*		Predicted		*L3-R1*		Predicted	
		Alive	Dead			Alive	Dead			Alive	Dead
	Alive	860	128		Alive	906	82		Alive	584	404
Actual				Actual				Actual			
	Dead	135	70		Dead	158	47		Dead	52	153
		ESS=21.2				ESS=14.6				ESS=33.7	
L2-R3		Predicted		*L2-R2*		Predicted		*L2-R1*		Predicted	
		Alive	Dead			Alive	Dead			Alive	Dead
	Alive	829	159		Alive	875	113		Alive	553	435
Actual				Actual				Actual			
	Dead	132	73		Dead	155	50		Dead	49	156
		ESS=19.5				ESS=13.0				ESS=32.1	
L1-R3		Predicted		*L1-R2*		Predicted		*L1-R1*		Predicted	
		Alive	Dead			Alive	Dead			Alive	Dead
	Alive	869	119		Alive	915	73		Alive	593	395
Actual				Actual				Actual			
	Dead	142	63		Dead	165	40		Dead	59	146
		ESS=18.7				ESS=12.1				ESS=31.2	

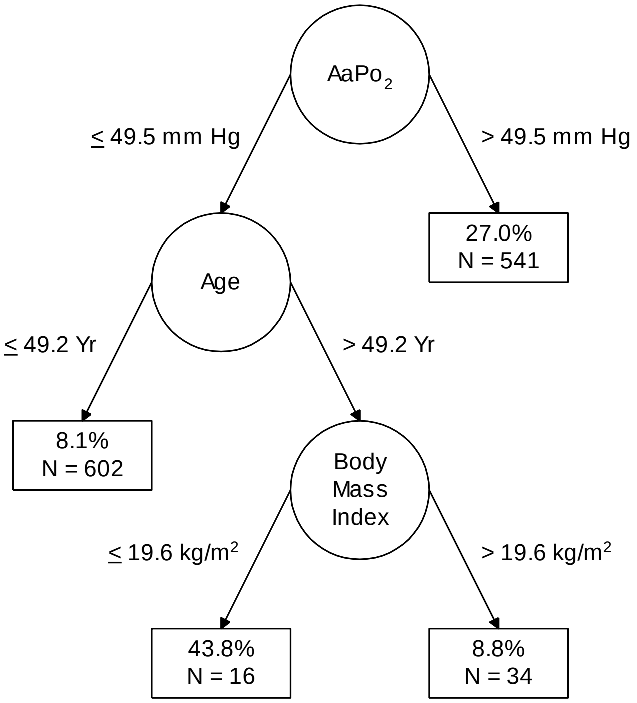


Figure 3: Optimized CTA Model

Compared to the initial CTA model, the pruned maximum sensitivity version of the CTA model (L1-R3) provided 10.4% greater mean sensitivity (60.6% versus 66.9%, respectively), corresponding to ESS values of 21.2 (relatively weak effect) versus 33.7 (moderate effect) respectively, and reflecting a 59.9% improvement in ESS for the optimized model. The optimized model used one fewer node than the non-pruned model, rendering it 98.7% more efficient than the initial model (i.e., averaging 8.4 versus 4.24 ESS-units-per-attribute, respectively).

A Second Example of Optimal Pruning: Predicting In-Hospital Mortality

The years 1995 to 1997 witnessed early adoption of highly active antiretroviral therapy for HIV, and the in-hospital morality rate from PCP had fallen to approximately ten percent.

Here we demonstrate pruning to maximize ESS for a model obtained CTA) to predict in-hospital mortality due to PCP during this time period.[1] Analysis was performed for 1,194 patients (72% of the total sample) with complete data for model attributes, who were discharged alive (N=1,054) or who died in-hospital (N=140). Derived manually using UniODA software, the CTA model involved four attributes: AaPo$_2$, albumin, and wasting (rapid decline of 20% or more in overall body weight). The non-pruned CTA model (Figure 4) yielded ESS=21.2, a relatively weak effect.

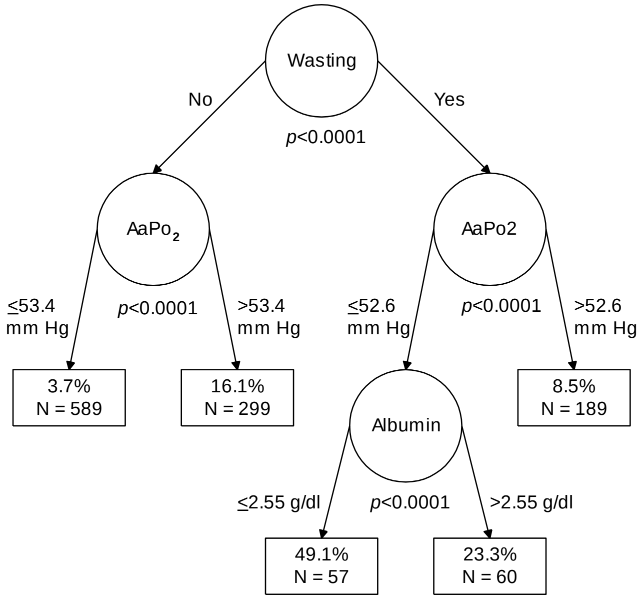


Figure 4: Initial Non-Pruned CTA Model of In-Hospital Mortality from PCP

Shown in Figure 5, the optimized model achieved 53.8% sensitivity (correct prediction of dead patients) and 84.3% specificity (correct prediction of live patients), and has much more robust endpoint denominators than did the original model. The moderate ESS=45.2 achieved by the optimized model represents a 36.6% improvement versus the ESS for the non-pruned model. And, by averaging 22.6 ESS units-per-attribute, the optimized model is 173% more efficient than the original model.

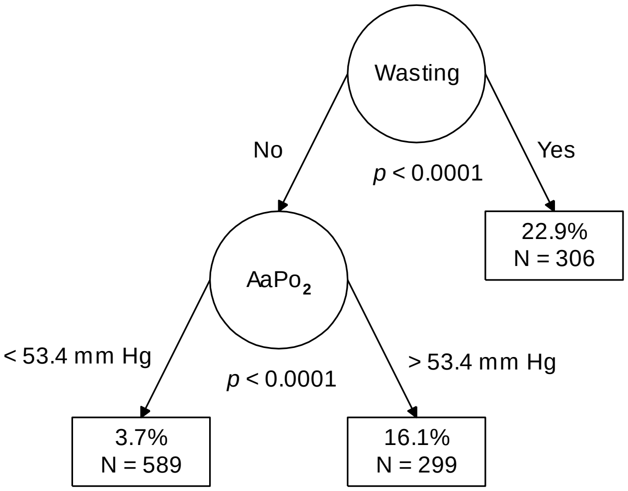


Figure 5: Optimized CTA Model of In-Hospital Mortality from PCP

Table 3 is used for assigning a severity-of-illness score to patients based on the findings of the optimized CTA model: rows are model endpoints reorganized in increasing order of percent of class 1 (dead) membership. Stage is an *ordinal index* of severity of illness, and p_{death} a *continuous index*: increasing values on these indices indicate worsening disease. Compared to Stage 1, p_{death} is 4.4-times as high in Stage 2, and 6.2 times as high in Stage 3.

Table 3: Staging Table for Predicting In-Hospital Mortality From PCP

Stage	Wasting	AaPo$_2$	N	p_{death}	Odds
1	No	≤53.4	589	0.037	1:26
2	No	>53.4	185	0.161	1:5
3	Yes	------	306	0.229	2:7

To use the table to stage disease severity for a given patient, simply evaluate fit between patient data and each stage descriptor. Begin at

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Stage 1, and work sequentially through stages until identifying the descriptor which is true for the data of the patient undergoing staging. For example, consider a hypothetical patient with no signs of wasting, and with $AaPo_2$=55.7 mm Hg. Stage 1 does not fit because the patient's $AaPo_2$ exceeds 53.4 mm Hg. However, because the patient does not show signs of wasting, and has $AaPo_2$>53.4 mm Hg, Stage 2 fits the data of this hypothetical patient.

Discussion

While there is no doubt that the methodology described here will always maximize the mean sensitivity, and therefore the ESS, of any classification model, it remains unknown with what relative frequency, and to what extent, optimal pruning will return a model which has different structure than the original non-pruned model. Furthermore, the advantage of optimal pruning has only been demonstrated for models derived using CTA, and should be generalized to models developed by other nonlinear means.

References

[1]Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, *et al*. A new preadmission staging system for predicting inpatient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early highly active antiretroviral therapy (HAART) era. *American Journal of Respiratory and Critical Care Medicine* 2000, 161:1081-1086.

[2]Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: an example of hierarchically optimal classification tree analysis. *Statistics in Medicine* 1997, 16: 1451-1463.

[3]Yarnold PR. Discriminating geriatric and non-geriatric patients using functional status information: an example of classification tree analysis via UniODA. *Educational and Psychological Measurement* 1996, 56:656-667.

[4]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

Author Notes

Mail correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Two-Group MultiODA: Mixed-Integer Linear Programming Solution with Bounded *M*

Robert C. Soltysik, M.S., and Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Prior mixed-integer linear programming procedures for obtaining two-group multivariable optimal discriminant analysis (MultiODA) models require estimation of the value of a parameter, *M*. A new formulation is presented which establishes a lower bound for *M*, which executes more quickly than prior formulations. A sufficient condition for the nonexistence of classification gaps and ambiguous solutions, optimal weighted classification, use of nonlinear terms, selecting an optimal subset of attributes, and aggregation of duplicate observations are discussed. When the design involves six or fewer binary attributes, MultiODA models may easily be obtained for massive samples.

.

Classification models derived via multivariable optimal discriminant analysis (MultiODA) are linear discriminant classifiers which explicitly maximize classification accuracy for a given sample.[1] Mixed-integer linear programming formulations for two-group MultiODA models require estimation of the value of a parameter, *M*, commonly defined as "a prohibitively large number."[2] If the estimated value of *M* is too low then suboptimal solutions may occur, and excessively large values of *M* will decrease computational efficiency and may introduce numerical (round-off) error.[3] We present a goal programming formulation which establishes a lower bound for *M*, and then we discuss a sufficient condition for the nonexistence of classification gaps and ambiguous solutions, weighted classification, the use of nonlinear terms, selection of optimal subsets of attributes, and aggregation of duplicate observations.

MIP45 Methodology

In a two-group linear MultiODA problem with p attributes and m observations, a set of m row vectors $\boldsymbol{a_i}$ is given, the components of which are $p = n$-1 observed values and a dummy value of unity. Each observation i is a member of either class 0 or class 1. A weight vector $\boldsymbol{x}$ is determined so that i is predicted to belong to class 0 when $\boldsymbol{a_i x} < 0$, or to class 1 when $\boldsymbol{a_i x} > 0$. Observation i is considered to be correctly classified if its predicted class membership is the same as its actual class membership, and misclassified otherwise. Solutions of interest yield maximum classification accuracy, that is, minimize the number of misclassified observations. This is achieved by determining $\boldsymbol{x^*}$ which satisfy the maximum number of inequalities in the system:

$$\boldsymbol{a_i x} < 0 \quad \text{for observations in class 0,}$$

$$\boldsymbol{a_i x} > 0 \quad \text{for observations in class 1.} \qquad (1)$$

This problem may be formulated as a mixed-integer linear programming model. To accomplish this, the strict inequalities in (1) are replaced with $\boldsymbol{a_i x} \underline{<} -\varepsilon$ or $\boldsymbol{a_i x} \underline{>} \varepsilon$, where $\varepsilon \underline{>} 0$. This is necessary due to the inability of simplex-based algorithms for mixed-integer programming to handle strict inequalities (mixed-integer techniques based upon interior-point algorithms[4] may not suffer this limitation). Letting ε be strictly positive removes the ambiguity in the classification status of observations i for which $\boldsymbol{a_i x} = 0$, but also introduces the possibility of a classification gap. It will be shown that there are conditions under which ambiguities can be removed for $\varepsilon = 0$. Consider the following model:

$$\text{MIP45:} \quad z = \min \sum_{i=1}^{m} d_i \qquad (2)$$

subject to

$$\sum_{j=1}^{n} a_{ij} (x_j^+ - x_j^-) - M_i d_i \le -\varepsilon, \ \underline{i} \in I_0 \qquad (3)$$

$$\sum_{j=1}^{n} a_{ij} (x_j^+ - x_j^-) + M_i d_i \ge \varepsilon, \ \underline{i} \in I_1 \qquad (4)$$

$$\sum_{j=1}^{n} (x_j^+ + x_j^-) = 1 \qquad (5)$$

$$x_j^+ - g_j \le 0, \ j=1,..., n \qquad (6)$$

$$x_j^- + g_j \le 1, \ j=1,..., n \qquad (7)$$

$$x_j^+, x_j^- \ge 0, \ j=1,..., n \qquad (8)$$

$$g_j \in \{0,1\}, \ j=1,..., n \qquad (9)$$

$$d_i \in \{0,1\}, \ \underline{i}=1,...,m \qquad (10)$$

where

a_{ij} is the jth component of observation $\boldsymbol{a_i}$

I_0 is the set of observations belonging to class 0

I_1 is the set of observations belonging to class 1

$$M_i = \max_j |a_{ij}| + \varepsilon \qquad (11)$$

z is the number of misclassified observations.

The weight vector $\boldsymbol{x}$ is obtained by

$$x_j = x_j^+ - x_j^-, \ j=1,..., n. \qquad (12)$$

Since constraints (6) and (7) ensure that not more than one of the x_j^+ and x_j^- are positive for any j, we can think of these values as the "positive" and "negative" parts of x_j, respectively. Note that $g_j = 1$ when $x_j > 0$ and $g_j = 0$ when $x_j < 0$. Also note that the g_j, along with (6), (7), and (9), may be dropped when $\varepsilon > 0$.

Constraint (5) normalizes $\boldsymbol{x}$ so that

$$\sum_{j=1}^{n} |x_j| = 1 \ ; \qquad (13)$$

that is, the sum of the absolute values of the discriminant weights is constrained to equal one. This normalization prevents the trivial solution $\boldsymbol{x} = \boldsymbol{0}$ (when $\varepsilon > 0$), and allows us to establish a lower bound for the M_i. It is necessary for the M_i to be large enough to force compliance of the constraints (3) and (4). This is accomplished by (11). To see this, consider constraint (4). Since $\sum_j |x_j| = 1$, it is clear that

2155-0182/10/$3.00

$$\boldsymbol{a_i x} \geq -\max_j |a_{ij}| \qquad (14)$$

and

$$\boldsymbol{a_i x} + \max_j |a_{ij}| + \varepsilon \geq \varepsilon. \qquad (15)$$

Therefore, when $d_i = 1$,

$$\boldsymbol{a_i x} + M_i d_i \geq \varepsilon. \qquad (16)$$

Because the normalization (5) requires that all optimal weight vectors ***x**** lie on a 45° properly rotated hypercube centered at the origin, this formulation is referred to as MIP45. It may be the case that more than one solution for ***d*** may be optimal for a problem. This corresponds to the existence of multiple optimal dichotomies of predicted class membership. It is also generally true that a solution space for ***x*** of positive volume exists for each dichotomy. The issue of selecting among optimal ***x**** may be addressed by a number of methods, such as linear programming[5] and *a priori* decision heuristics.[6]

Resolving Classification Gaps and Ambiguities

In the above formulation, at least n - 1 of the $\boldsymbol{a_i x^*}$ are at zero when $\varepsilon = 0$ is specified**.** From (1), it is seen that the criterion of strict separation of the classes should be met. An optimal value $z^* > 0$ in the solution of the following linear program guarantees that this separation is maintained.

LP: max $z = y$

subject to

$$\sum_{j=1}^{n} a_{ij}\,(b_j^+ - b_j^-) + \underline{y} \leq 0,\ \underline{i} \in \underline{I_0} \text{ and } \boldsymbol{a_i x^*} \leq 0 \qquad (17)$$

$$\sum_{j=1}^{n} a_{ij}\,(b_j^+ - b_j^-) - \underline{y} \geq 0,\ \underline{i} \in \underline{I_1} \text{ and } \boldsymbol{a_i x^*} \geq 0 \qquad (18)$$

$$\sum_{j=1}^{n} (b_j^+ - b_j^-) = 1 \qquad (19)$$

$$b_j^+,\ b_j^-,\ y \geq 0 \qquad (20)$$

$$b_j = b_j^+ + b_j^- . \qquad (21)$$

This LP may be executed for each optimal dichotomy. If $z^* > 0$ is obtained, ***b**** is a new discriminant vector which optimizes criterion (1). Otherwise, ambiguity remains in the classification status of observations for which $\boldsymbol{a_i b^*} = 0$: such observations should not be classified.

The advantage of establishing a lower bound for M is illustrated with an example involving discriminating between excellent versus less than excellent medical residents using information obtained during their application for residency training. Since rating applicants for residency training is a difficult, time-intensive decision-making task, a linear discriminant classifier that successfully predicts resident performance might be of great interest and utility to admissions committees.

The sample was $m = 49$ residents enrolled in a three-year internal medicine residency program.[7] The clinical performance (class) variable was based on the mean rating on an explicit 10-point scale made by residents' supervisors: a mean rating of nine or greater on this scale reflected “excellent” (or better) clinical performance (class = 1, $m_1 = 27$), and a mean rating of less than nine reflected less than excellent clinical performance (class = 0; $m_0 = 22$). The $n - 1 = 3$ application information variables (attributes) included medical board scores, faculty evaluations (a composite measure reflecting ratings of letters of recommendation and medical school grading system), and academic distinction (a composite measure reflecting honors attained in medical school and medical school status).

The computer resources required to solve this problem using MIP45 versus Stam and Joachimsthaler[8] was compared (other prior formulations were slower). For MIP45, ε was set at 0. For Stam and Joachimsthaler, values of 1, 10, 100, and 1000 were used for M, and a value of 1 for ε.[9] All formulations were solved on an IBM 3090/300 computer running SAS/OR.[10] As seen in Table 1, except when M = 1, Stam and Joachimsthaler required more computational effort (CPU time, pivots, and integer branches) than did MIP45. Using M = 1, ε = 1 in Stam and Joachimsthaler resulted in a useless solution, and using M = 10 or 100 resulted in suboptimal solutions of (3). Since a decision-maker using M = 10 or M = 100 would have no direct evidence that these solutions were suboptimal, it would also be unclear whether the solution attained by Stam and Joachimsthaler (or other unbounded formulations) using M = 1000 was optimal. In contrast, since the value of z^* attained in LP was positive, a decision-maker using MIP45 to solve this problem would be certain that the solution was unambiguously optimal: a clear advantage.

TABLE 1

Illustration of Computational Resources Needed by MIP45 Versus Stam and Joachimsthaler[8] to Solve a Problem with 49 Observations and Three Attributes, Using SAS/OR run on an IBM 3090/300 Computer

Formulation	M	ε	Objective Value	CPU Seconds	Integer Branches	Pivots
Stam	1	1	29	1.1	0	31
Stam	10	1	17	131.8	8,629	36,607
Stam	100	1	15	276.7	19,755	89,564
Stam	1000	1	14	268.4	14,549	57,351
MIP45	LB	0	14	48.0	2,896	15,333

Note: For MIP45 the M_i were set at their lower bounds (LB). For solutions resulting in the optimal value of 14 misclassifications, model coefficients for board scores and faculty evaluation were positive, and the coefficient for academic distinction was negative. For MIP45, z^* = .00439.

Weighted Classification

Rather than weighting each observation equally, we consider weighting each case in (2) by a positive scalar c_i. This is significant for two reasons. First, the c_i may represent the cost of misclassifying observation i. In this case an optimal solution would minimize the cost of misclassification (or, equivalently, maximize the return of correct classification) for the sample. Second, the c_i may represent factors which balance the number of class 0 and class 1 observations when these are not equal. In this case an optimal solution would maximize the number of

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

correct classifications weighted by population membership in each class. An example would be $c_i = 1/m_0$ for observations in class 0, and $c_i = 1/m_1$ for observations in class 1, where m_0 and m_1 are the number of observations in categories 0 and 1, respectively. This latter weighting scheme is particularly useful in badly imbalanced applications for which $m_0 >> m_1$, or visa versa: use of such "priors weights" forces the model to classify observations from both classes accurately, and inhibits the identification of degenerate models which classify all observations into a single class category.

Adding Nonlinear Terms as Attributes

Here we generalize the notion of maximum pattern classification accuracy achieved by separating hyperplanes to sets of nonlinear separating surfaces. For example, consider quadratic surfaces in p-measurement space of the form:

$$\sum_{j} a_{ij}\underline{\mathrm{x}}_{\mathrm{j}} + \sum_{k \leq p} \sum_{l \leq k} a_{ik}a_{il}x_{kl} + a_{in}x_n \qquad (22)$$

for all i. The MultiODA solution can be attained by augmenting the a_j and x in the MIP45 model by the interaction terms in (22). This solution produces a weight vector $\boldsymbol{x}$ which yields the minimum number of misclassifications achievable by a quadratic separating surface. This process may be applied to any nonlinear discriminant function which is linear in the parameters of the measurement space.

Optimal Attribute Subset Selection

In the foregoing we have assumed that all p attributes are included in the MultiODA model. However, we may wish to select a subset of $k < p$ attributes for the application of the model. For example, imagine an application involving 50 observations and ten attributes. In order to identify a model that may generalize if used to classify independent random samples, we may wish to maintain a minimum observation-to-attribute ratio of 10-to-1, so a maximum of five of the ten potential attributes may be used. Of all possible 5-attribute models, which yields maximum accuracy? Optimal attribute subset selection methodology can be incorporated in the MIP45 model by defining n zero-one variables q_j and including the following constraints:

$$\underline{\mathrm{x}}_{\underline{\mathrm{j}}}^{-} - q_j \leq 0, j=1,..., n, \qquad (23)$$

$$g_j + q_j \leq 1, j=1,..., n, \qquad (24)$$

and

$$\sum_{j=1}^{n} g_j + \sum_{j=1}^{n} q_j = k. \qquad (25)$$

In an optimal solution to such a MultiODA model, measurement $\underline{\mathrm{j}}$ is selected for inclusion only if $g_j + q_j = 1$. The number of misclassifications obtained is the fewest achievable in any k-dimensional subspace of the original p-dimensional measurement space.

Aggregation of Duplicate Observations

If duplicate observations occur in the data set (i.e., two or more observations have the same value for every attribute measurement), the following procedure may be used to aggregate the duplicate observations into a single observation, reducing the size of the overall problem. The resulting problem is equivalent to the original one, with m' observations, and objective value $z + v$.

```
$m' := m : s_0 = 0: s_1 = 0: v := 0$
for each $i = 1, …, m'$
  for each $j < i$
    if $a_i = a_j$ then
      if $i \in \underline{\mathrm{I}}_0$ then $s_0 := s_0 + c_i$ else $s_1 := s_1 + c_i$
      remove observation $i$ from list : $m' := m' - 1$
    end if
```

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

```
next $j$, $i$
for each $i = 1, …, m'$
  if $s_0 > s_1$ then
    $w_j := s_0 - s_1 : v := v + s_1$
  else if $s_1 > s_0$ then
    $w_j := s_1 - s_0 : v := v + s_0$
  else
    $v := v + s_0$ : remove observation $i$ from list :
                                        $m' := m' -1$
  end if
next $i$
```

This procedure is particularly useful when $\boldsymbol{a_j}$ is a zero-one vector (all attributes are binary). Here all the patterns lie on the vertices of the *p*-dimensional unit hypercube. If more than one pattern lies on some vertex, then by using the above procedure we may obtain a weighted MIP45 model equivalent to the original model, but with fewer constraints. If the number of original patterns *m* is large relative to the number of attributes *p*, a significant reduction in the size of the model may be obtained. For instance, regardless of the value of *m*, if p=8 then we end up with no more than $2^8 = 256$ constraints of type (6) in the model. Since the number of constraints is independent of *m*, extremely large problems may be solved with this procedure, provided *p* is moderately small.

In order to illustrate the potential solution efficiency gained by using this special purpose algorithm for problems involving entirely binary data, we ran 30 Monte Carlo experiments. In each experiment there were five binary attributes, such that the total possible number of different profiles was $2^5 = 32$. Values on each attribute were determined separately for each observation on the basis of a random uniform number between 0 and 1: numbers < 0.5 were assigned the value of 0, and numbers $\geq$ 0.5 were assigned the value of 1. We ran five balanced ($m_0 = m_1$) experiments for each total sample size of 50, 100, 1000, 10^4, 10^5, and 10^6 total observations. All formulations were solved on an IBM 3090/300 computer running SAS/OR. As seen in Table 2, as the number of observations increased: (a) the number of distinct profiles increased toward its theoretical upper bound (the theoretical upper bound was achieved in all of the problems involving 10^6 observations, and in four of the five problems involving 10^5 observations); (b) the misclassification rate increased towards its theoretical upper bound (i.e., for a balanced design with an even number of observations, the theoretical upper bound for the number of misclassifications is one less than one-half of the total number of observations); and (c) the mean number of CPU seconds required to solve the problem was approximately twenty seconds for problems with 1000 or *more* total observations.

TABLE 2

Results of Monte Carlo Experiments for Binary Data: Five Random Attributes

Number of Observations	Number of Profiles	Number (%) of Misclassifications	CPU Seconds
50	19	14 (28%)	4.5
50	23	13 (26%)	5.5
50	20	16 (32%)	8.4

50	23	14 (28%)	12.5
50	21	12 (24%)	1.7
100	30	27 (27%)	14.2
100	26	34 (34%)	9.0
100	25	43 (43%)	10.5
100	24	37 (37%)	7.5
100	20	33 (33%)	3.0
1000	32	432 (43%)	17.8
1000	30	445 (44%)	25.1
1000	31	449 (45%)	16.4
1000	31	460 (46%)	24.3
1000	31	454 (45%)	19.2
10000	29	4870 (49%)	12.3
10000	31	4838 (48%)	23.8
10000	31	4842 (48%)	24.9
10000	29	4828 (48%)	11.9
10000	31	4839 (48%)	9.2
100000	32	49545 (50%)	14.5
100000	32	49532 (50%)	21.6
100000	31	49526 (50%)	6.3
100000	32	49475 (49%)	25.2
100000	32	49376 (49%)	16.8
1000000	32	498331 (50%)	24.3
1000000	32	498759 (50%)	17.2
1000000	32	498450 (50%)	32.5
1000000	32	497861 (50%)	4.5
1000000	32	498837 (50%)	16.8

Discussion

MIP45 solves two problems common to prior goal programming formulations of two-group MultiODA: *M* is automatically set at its lower bound, and it is possible to determine whether classification gaps or ambiguities exist. Collateral benefits of MIP45 include its greater computational efficiency and solution speed relative to prior formulations, particularly for applications involving binary attributes.

This study contrasted the computational characteristics of the MIP45 formulation of the MultiODA problem to the formulation of Joachimsthaler and Stam (see Table 1). Other mixed-integer programming formulations have appeared more recently. Rubin developed a decomposition technique to solve the Multi-ODA problem.[11] Silva and Stam developed a partitioning method for MultiODA which was reported to compare favorably with MIP45.[12] Pfetsch developed a technique to optimize irreducible inconsistent subsystems (IIS) of linear inequalities in order to determine a maximum feasible subsystem of these inequalities.[13] Finally, Bremner and Chen developed a MIP formulation for the halfspace depth problem which uses IIS cuts in a branch-and-cut algorithm.[14] We eagerly anticipate computational comparisons between these formulations.

References

[1]Yarnold PR, Soltysik RC, Martin GJ. Heart rate variability and susceptibility for sudden cardiac death: an example of multivariable optimal discriminant analysis. *Statistics in Medicine* 1994, 13:1015-1021.

[2]Joachimsthaler EA, Stam A. Mathematical programming approaches for the classification problem in two-group discriminant analysis. *Multivariate Behavioral Research* 1990, 25:427-454.

[3]Gehrlein WV. General mathematical programming formulations for the statistical classification problem. *Operations Research Letters* 1986, 5:299-304.

[4]Karmarkar N. A new polynomial time algorithm for linear programming. *Combinatorica* 1984, 4:373-395.

[5]Koehler GJ, Erenguc SS. Minimizing misclassifications in linear discriminant analysis. *Decision Sciences* 1990, 21:63-74.

[6]Yarnold PR, Soltysik RC. Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences* 1991, 22:739-752.

[7]Curry RH, Yarnold PR, Bryant FB, Martin GJ, Hughes RL. A path analysis of medical school and residency performance: implications for housestaff selection. *Evaluation in the Health Professions* 1988, 11:113-129.

[8]Stam A, Joachimsthaler EA. A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *European Journal of Operational Research* 1990, 46:113-122.

[9]Bajgier SM, Hill AV. An experimental comparison of statistical and linear programming approaches to the discriminant problem. *Decision Sciences* 1982, 13:604-612.

[10]Cornell R, Luginbuhl RC, Yeo C. *SAS/OR user's guide, version 6*. SAS Institute, Durham, NC, 1989.

[11]Rubin PA. Solving mixed-integer classification problems by decomposition. *Annals of Operations Research* 1997, 74:51-64.

[12]Silva APD, Stam A. A mixed-integer programming algorithm for minimizing the training sample misclassification cost in two-group classification. *Annals of Operations Research* 1997, 74:129-157.

[13]Pfetsch ME. Branch-and-cut for the maximum feasible subsystem problem. *SIAM Journal on Optimization* 2008, 19:21-38.

[14]Bremner D, Chen D. A branch and cut algorithm for the halfspace depth problem. 2009: arXiv:0910.1923v1.

Author Notes

This research was supported in part by a grant from the NSF (SES-8822337), and an early version of the paper was invited for presentation at the TIMS/ORSA Joint National Meeting, Chicago, 1993. Appreciation is extended to the computer center of the University of Illinois at Chicago, for computer resources used in this research. Mail correspondence to authors at: Optimal Data Analysis, 1220 Rosecrans Street, Suite 330, San Diego, CA 92106. Send Email to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Unconstrained Covariates in CTA

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

In traditional statistical covariate analysis it is common practice to force entry of the covariate into the model first, to eliminate the effect of the covariate (i.e., “equate the groups”) on the dependent measure. In contrast, in CTA the covariate is treated as an ordinary attribute which must compete with other eligible attributes for selection into the model based on operator-specified options. This paper illustrates optimal covariate analysis using an application involving predicting patient in-hospital mortality via CTA.

A study of 1,641 patients hospitalized for *Pneumocystis cariini* pneumonia (PCP) used logistic regression analysis to model in-hospital mortality: after forcing a measure of severity-of-illness into the model first, PCP prophylaxis was the only attribute significantly associated with lower hospital survival.[1] During development of an enumerated model involving only these two attributes, a non-pruned[2] CTA model was identified which is analogous to the logistic regression analysis, in that both models *initially adjusted* for severity of illness. CTA analyses were performed using automated software with a minimum endpoint denominator of N=25 to ensure sufficient statistical power.[3] The optimal solution involved one parse of the root attribute (i.e., the first and second attributes entering the CTA model were both PCP severity-of-illness), so the model has three emanating branches (see Figure 1).

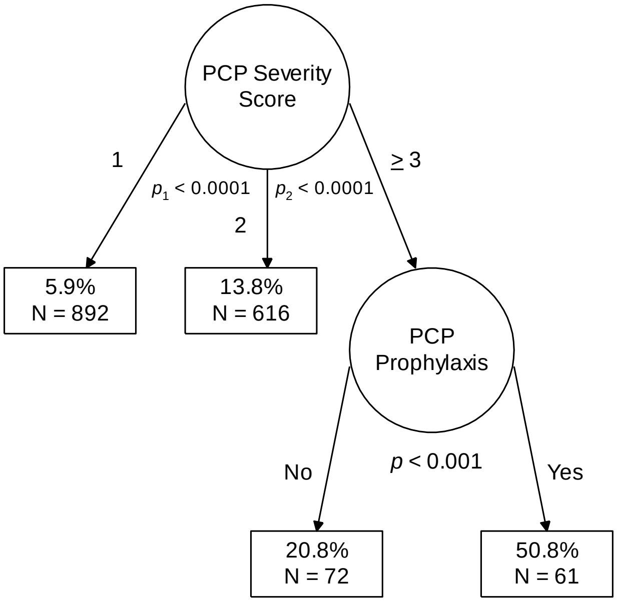


Figure 1: Algorithmic CTA Model Predicting In-Hospital Mortality, Covariate Entered First

Consistent with findings using logistic regression, this CTA model returned weak gain versus chance in predicting mortality: 97.9% of

1,457 living and 16.8% of 184 deceased patients were correctly classified: ESS=14.8, efficiency= 14.8/2 or 7.4 ESS units-per-attribute. Though the CTA *model* is weak, the right-most endpoint indicates that the combination of a PCP severity score of three or greater, and PCP prophylaxis, predicted nearly 51% mortality for 61 patients. Thus, for applications in which it is important to identify particularly vulnerable strata, a variety of different CTA models should be examined in hopes of discovering one or more of such fruitful branches (i.e., combinations).

In contrast, as illustrated in Figure 2, the enumerated CTA model obtained using the same two attributes has robust endpoint denominators; correctly classified 67.9% of the 1,457 living and 61.4% of the 184 dead patients; and obtained moderate strength (ESS=29.4) and efficiency (9.8 ESS units-per-attribute).

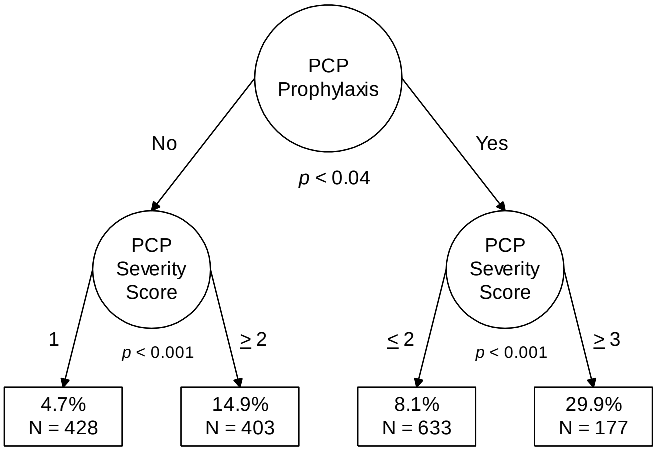


Figure 2: Enumerated CTA Model Predicting In-Hospital Mortality: Covariate Unconstrained

Table 1 gives the staging table for the enumerated CTA model, used for predicting in-hospital mortality from PCP. Table rows are model endpoints reorganized in increasing order of percent of class 1 (“dead”) membership. Stage is an *ordinal index* indicating increasing severity of illness, and p_{death} is a *continuous index* of disease severity. The 1st and 4th strata reflect a 6.4-fold difference in likelihood of dying in-hospital: compared to Stage 1, p_{death} is approximately two times higher in Stage 2, three times higher in Stage 3, and six times higher in Stage 4.

Table 1: Staging Table for Predicting In-Hospital Mortality From PCP

Stage	PCP Prophylaxis	Severity Score	N	p_{death}	Odds
1	No	1	428	0.047	1:20
2	Yes	≤ 2	633	0.081	1:11
3	No	≥ 2	403	0.149	1:6
4	Yes	≥ 3	177	0.299	3:7

Although identical attributes were used by the two CTA models and the original linear logistic regression analysis, the attributes were arranged in different geometries in the different models. Of course, an analyst’s imposition of attribute entry or sequence order in CTA, or any chained optimal analysis, should be performed on the basis of theory, that is, to directly address *a priori* hypotheses.[4] However, the present case clearly indicates the need for caution regarding unchecked rigid adherence to methodological traditions which may actually impede progress achieved using emerging and new technologies. Automated CTA software makes the comparative analysis of multiple theoretical perspectives feasible for most applications: challenging and defeating unfruitful traditions ought to make for interesting, if not exciting research.

References

[1]Curtis JR, Yarnold PR, Schwartz DN, Weinstein RA, Bennett CL (2000). Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American*

2155-0182/10/$3.00

Journal of Respiratory and Critical Care Medicine, *162*, 393-398.

[2]Yarnold PR, Soltysik RC. Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis* In press.

[3]Soltysik RC, Yarnold PR. Introduction to automated CTA. *Optimal Data Analysis* In press.

[4]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

Author Notes

Dr. Charles L. Bennett, M.D., Ph.D., graciously offered Paul and Robert use of data which we previously published, generated from his NIH-funded projects which completed approximately a decade ago, for expository purposes. Mail correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Maximizing the Accuracy of Probit Models via UniODA

Barbara M. Yarnold, J.D., Ph.D. and Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Paralleling the procedure used to maximize ESS of linear models derived using logistic regression analysis or Fisher's discriminant analysis, univariate optimal discriminant analysis (UniODA) is applied to the predicted response function values provided by a model derived by probit analysis (PA), and returns an adjusted decision criterion for making classification decisions. ESS obtains its theoretical maximum value with this adjusted decision criterion, and the ability of the PA model to return accurate classifications is optimized. UniODA-refinement of a PA model is illustrated using an example involving political science analysis of federal courts.

Probit analysis (PA) has gained in popularity as research in political science seeks increasingly accurate models of court decision-making.[1-8] For applications having a binary class variable and at two or more attributes, PA allows assessment of the independent relationship between class variable and attribute. Parameter estimates are obtained by maximum-likelihood, and indicate the amount of change in the cumulative normal probability function that is associated with a one-unit change in the attribute value. Goodness-of-fit of PA models was traditionally assessed using R^2 and chi-square, but this was criticized.[9] The supreme criterion for all classification models is their ability to make accurate predictions. PA does not explicitly maximize classification accuracy, but effect strength for sensitivity (ESS) yielded by PA models may be maximized by optimizing the models decision-making criterion.[10,11] This note illustrates the use of UniODA-refinement to optimize a model derived using PA.

Federal Court Decisions in Asylum-Related Appeals

To illustrate this method we consider the asylum-related appeals to the federal courts covering the period of 1980-1987, constituting 137 cases having complete data. The class variable indicated whether aliens won (N=59) or lost (N=78) their appeal. Six binary attributes used in PA included whether any organizations were involved in the appeal; the alien was from a country hostile to the USA; the alien was from Europe; the court was located in the Western USA; a high percentage of the judges involved in the appeal were appointed by a Democratic President; and whether there was a high level of immigrant-flow into the circuit. The resulting PA model correctly predicted 71.2% of the wins and 55.1% of losses, resulting in ESS=26.4.

UniODA was then used to optimize the model: the PA model was first used to obtain *Y** for every observation, and then UniODA was conducted on those *Y** using the original class variable coding.[14] The adjusted decision criterion for the PA model was: if $Y^*>0.025$ predict class=1 (win); otherwise predict class=0 (loss). The optimized PA model correctly predicted 64.4% of the wins and 71.8% of losses, yielding ESS=36.2, representing a 37% improvement in this index relative to the non-refined model.

Discussion

The objective of this note was to illustrate how UniODA-refinement can improve classification performance obtained by a model derived by PA. The example demonstrated a substantial increase in the level of training accuracy (ESS) achieved by the model, a finding which is common when the decision criteria of suboptimal models are optimized via UniODA-refinement.

References

[1] Aldrich J, Cnudde, C. Probing the bounds of conventional wisdom: comparison of regression, probit, and discriminant analysis. *American Journal of Political Science* 1975, 19:571-608.

[2] Yarnold BM. Federal court outcomes in asylum-related appeals 1980-1987: a highly politicized process. *Policy Sciences* 1990, 23:291-306.

[3] Yarnold BM. *Refugees without refuge: formation and failed implementation of U.S. political asylum policy in the 1980s*. University Press of America, Lanham, MD, 1990.

[4] Yarnold BM. The Refugee Act of 1980 and de-politicization of refugee/asylum admissions: failed policy implementation. *American Politics Quarterly* 1990, 18:527-536.

[5] Yarnold BM. *International fugitives: a new role for the International Court of Justice*. Praeger, New York, NY, 1991.

[6] Yarnold BM. The role of religious organizations in the sanctuary movement. In: *The role of religious organizations in social movements* (BM Yarnold, Ed.), Praeger, New York, NY, 1991.

[7] Yarnold BM. *Politics and the courts: toward a general theory of public law*. Praeger, New York, NY, 1992.

[8] Yarnold BM. *Abortion politics in the federal courts: right versus right*. Paragon, New York, NY, 1993.

[9] Hagle T, Mitchell G. Goodness-of-fit measures for probit and logit. *American Journal of Political Science* 1992, 36:762-784.

[10] Yarnold PR, Soltysik RC. Refining two-group multivariable models using Univariate optimal discriminant analysis. *Decision Sciences* 1991, 22:1158-1164.

[11] Yarnold PR, Hart LA, Soltysik RC. Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement*, 1994, 54:73-85.

[12] Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

Author Notes

Mail correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Precision and Convergence of Monte Carlo Estimation of Two-Category UniODA Two-Tailed *p*

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

Monte Carlo (MC) research was used to study precision and convergence properties of MC methodology used to assess Type I error in exploratory (*post hoc*, or two-tailed) UniODA involving two balanced (equal N) classes. Study 1 ran 10^6 experiments for each N, and estimated cumulative *p*'s were compared with corresponding exact *p* for all known *p* values. Study 2 ran 10^5 experiments for each N, and observed the convergence of the estimated *p*'s. UniODA cumulative probabilities estimated using 10^5 experiments are only modestly less accurate than probabilities estimated using 10^6 experiments, and the maximum observed error ($\underline{+}0.002$) is small. Study 3 ran 10^5 experiments for Ns ranging as high as 8,000 observations in order to examine asymptotic properties of optimal values for balanced designs.

A recursive, closed-form solution for the theoretical distribution of optimal values for one-tailed "confirmatory" UniODA of random data is discovered, and associated computation time is linear in N.[1] In applications where an *a priori* alternative hypothesis has been specified, Type I error rate or alpha (*p*) can be computed for any combination of optimal value and N. For two-tailed "exploratory" applications, a closed-form solution for the distribution of optimal values has not yet been discovered. For *post hoc* UniODA the enumerable open-form solution for the theoretical distribution of optimal values is computationally intractable for N>30, but the one-tailed solution can be used to determine the two-tailed distribution if overall classification accuracy is at least 75%.[1] Other means are needed to estimate the two-tailed distribution if overall classification is less than 75%. This study uses Monte Carlo (MC) research to assess precision and convergence properties of MC methods used to estimate *p* for UniODA.

Precision

One million MC experiments were run for each balanced design of $N\underline{<}30$. A design is balanced if the number of class 1 and 0 observations is identical for even N, or differs by one for odd N. In every experiment, the attribute was a uniform random number between zero and one.[2] For even N experiments the first N/2 observations were assigned to class 1, and the rest to class 0. For odd N experiments the first (N-1)/2 observations were assigned to class 1, and the rest to class 0. For each experiment the optimal value was determined and stored. For each N the estimated UniODA distribution was

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

cumulated after 10^6 experiments were run. To compare estimated and known distributions, cumulative *p*>0.001 were rounded up to the nearest thousandth, and cumulative *p*<0.001 were rounded up on the second significant digit.

The results suggest that MC experiments accurately estimated known *post hoc* UniODA distributions. Over all N and possible optimal values, 170 of 238 (71.4%) estimated cumulative probabilities were identical to the exact value; 237 of 238 (99.6%) of the estimates were ±0.001 of the exact value; and all estimates were ±0.002 of the exact probability.

Estimated cumulative probabilities were most accurate when the exact probability was small. For example, for optimal values with associated exact cumulative probabilities of 0.05<*p*<0.001, 45 of 50 (90%) of the estimated probabilities were identical to the corresponding exact probability; 49 of 50 (98%) of estimated probabilities were ±0.001 of exact probability; and all estimated probabilities were ±0.002 of the exact probability.

MC experiments also provided accurate estimates of exact cumulative probabilities for statistically marginal (0.05<*p*<0.10) effects: 13 of 15 (86.7%) of the estimated cumulative probabilities were identical to their corresponding exact values, and all estimated probabilities were ±0.001 of the exact probability.

Cumulating 10^6 MC experiments for a given N provides an accurate approximation of the UniODA distribution, but the computational cost is high. Accordingly, Study 2 investigated convergence properties of MC methodology and was designed to determine the number of MC experiments that is sufficient to achieve stable, accurate estimates of UniODA distributions.

Convergence

MC experiments were designed and data generated as in Study 1. For each N between 3 and 30 inclusive, 10^5 experiments were run in successive blocks of 1,000 experiments, and the UniODA distribution was cumulated at each block. Thus, 100 UniODA distributions were estimated for each N: the first based on 1,000 experiments, the second based on 2,000 experiments, and the 100th based on 10^5 experiments.

Many (56.9 percent) of the estimated *p*'s converged to their final value (i.e., their value at the end of the study) within 20,000 experiments, and most (86.3 percent) of the estimated *p*'s converged to their final value within 70,000 experiments.

After 10^5 experiments were completed, every estimated *p* in the range 0.001<*p*<0.10 was identical to the corresponding estimated *p* based on 10^6 experiments (precision study).

Consistent with the first study, known UniODA distributions were accurately modeled. For probabilities in the range 0.001<*p*<0.05: 35 of 50 (70%) estimated cumulative probabilities were identical to corresponding exact values; 49 of 50 (98%) estimated probabilities were ±0.001 of exact; and all estimated probabilities were ±0.002 of the exact value.

Thus, UniODA cumulative probabilities estimated using 100,000 MC experiments are only modestly less accurate than probabilities estimated using one million experiments, and the maximum observed error (±0.002) is small.

Asymptotic Convergence

A final study investigated convergence properties of interesting levels of classification performance for balanced two-category *post hoc* UniODA. MC experiments were designed and data generated as in Study 1. For all N between 1,000 and 8,000 inclusive, in steps of 1,000, a total of 10^5 MC experiments were run. Results of the simulation are presented in Table 1.

Tabled for the indicated value of *p* and N are the optimal value and the corresponding percentage accuracy in classification or PAC (top and bottom row, respectively). The optimal value is the maximum number of misclassifications possible to still achieve the *p* value. For example, for N=1,000 observations and *p*<0.001 a maximum of 438 misclassifications can be made, corresponding to 562 correct classifications, and thus to PAC=(562/1,000)x100%, or 56.2% (see Table 1).

Table 1: Maximum Optimal Value for 2-Tail *p* in Balanced 2-Category UniODA

	Two-Tail *p*<			
N	0.001	0.01	0.05	0.10
1,000	438	448	457	461
	56.2	55.2	54.3	53.9
2,000	912	927	939	945
	54.4	53.6	53.1	52.8
3,000	1393	1411	1425	1433
	53.6	53.0	52.5	52.2
4,000	1876	1896	1913	1922
	53.1	52.6	52.2	52.0
5,000	2361	2384	2403	2413
	52.8	52.3	51.9	51.7
6,000	2849	2874	2894	2905
	52.5	52.1	51.8	51.6
7,000	3336	3364	3386	3397
	52.3	51.9	51.6	51.5
8,000	3825	3853	3878	3890
	52.2	51.8	51.5	51.4

For *p*<0.05 a maximum of 457 misclassifications are possible, corresponding to PAC=(543/1,000)x100%, or 53.9%. For N=5,000 and *p*<0.01, a maximum of 2,384 misclassifications are possible, corresponding to PAC=[(5,000-2,384)/5,000]x100%, or 52.3%.

In balanced designs involving as few as 1,000 observations, a UniODA model performing only a modicum better than an unbiased flipped coin (i.e., obtaining at least 55.2% "heads") yields classification accuracy which is sufficient to achieve *p*<0.001. Therefore, as N increases in magnitude the significance of *p* as an index of performance rapidly diminishes to trivial levels.

References

[1]Soltysik RC, Yarnold PR. Univariable optimal discriminant analysis: one-tailed hypotheses. *Educational and Psychological Measurement* 1994, 54:646-653.

[2]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

Author Notes

Appreciation for computer resources used in this research is extended to the University of Illinois at Chicago. Mail correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 300, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

2155-0182/10/$3.00

Aggregated *vs*. Referenced Categorical Attributes in UniODA and CTA

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

Multivariable linear methods such as logistic regression analysis, discriminant analysis, or multiple regression analysis, for example, directly incorporate binary categorical attributes into their solution. However, for categorical attributes having more than two levels, each level must first be individually dummy-coded, then one level must be selected for use as a reference category and omitted from analysis. Selection of one or another level as the reference category can mask effects which otherwise would have materialized, if a different level had been chosen. Neither UniODA nor CTA require reference categories in analysis using multicategorical attributes.

Using a categorical attribute with three or more levels in a linear multivariable analysis requires separately dummy-coding each level, selecting one level as a reference category, and omitting it from analysis.[1] For example, imagine that a study assessed three ethnic categories: Navajo, Sumatran, and Inuit. Preparing this attribute for linear analysis first requires creating three new binary attributes: [a] Navajo (1) *vs*. others (0); [b] Sumatran (1) *vs*. others (0); and [c] Inuit (1) *vs*. others (0). Only two of the dummy-variables can be used as attributes in analysis, and one's choice can mask an effect depending on which class is selected as reference category. As an increasing number of polychotomous attributes are used, the associated design matrix becomes massive rapidly, increasing the likelihood of sparse or empty cells, imbalanced marginal distributions and nonnormality, toxic properties for linear methods. In addition to possibly masking effects, inducing numerical instability, undermining assumptions underlying the validity of *p*, and contributing to overdetermined models, the use of reference categories is also antithetical to the axiom of parsimony. Finally, in computer-intensive methods such as CTA, a larger number of attributes increases both memory and time resources needed to obtain an optimal solution.

In contrast, UniODA[2] and CTA[3] use aggregated multicategory attributes. Using the current example one "ethnicity" attribute having three levels (rather than three ethnicity attributes each having two levels) requires coding: Navajo (1), Sumatran (2), or Inuit (3).

This paper illustrates some advantages of using aggregated attributes in both bivariate (UniODA) and multivariable (CTA) analyses, using an application involving predicting use of mechanical ventilation for hospitalized patients with *Pneumocystis cariini* pneumonia (PCP).[4]

2155-0182/10/$3.00

UniODA

The analysis selected for exposition contrasts intubation rate for a total sample of 1,211 patients hospitalized for PCP in Chicago, Los Angeles, Miami, New York, and Seattle. The first analysis used the aggregated attribute, arbitrarily using dummy-codes of 1-5 for cities, respectively. The resulting UniODA model was: if city=Los Angeles or Chicago then predict a higher ventilation rate; otherwise predict a lower ventilation rate. This model correctly classified 54.9% of 1,418 non-ventilated, and 61.9% of 147 ventilated patients, yielding a relatively weak ESS=16.8 (p<0.0006), which was stable in jackknife validity analysis.

Using the aggregated city attribute and therefore one test of a statistical hypothesis, UniODA determined three cities have a lower ventilation rate than two other cities, and even though the effect is statistically significant and likely to cross-generalize for an independent random sample, the effect is weak, reflecting only 16.8% of the gain in accuracy theoretically possible to achieve beyond chance.

UniODA was next used to assess the ability of all five binary city attributes to predict ventilation: the test for Los Angeles (p<0.0006) alone achieved the criterion[2] for statistical significance with a weak effect of ESS=12.6. This result indicates that Los Angeles had a higher ventilation rate than the other four cities. Five tests of statistical hypotheses were conducted in reaching this conclusion, and must be accounted for in assessing the statistical significance of all hypothesis tests conducted within the study.

CTA

In the original research from which the example was drawn, ventilation was modeled by logistic regression analysis.[4] Predictive factors which emerged included a PCP severity score developed previously via CTA[6], location (Los Angeles), ethnicity (African-American), and a cytological confirmation of PCP diagnosis. For clarity in exposition, the same attributes selected by logistic regression were modeled presently. Algorithmic CTA[3] was run via ODA automated CTA software, using a minimum endpoint denominator of N=25 to ensure adequate statistical power.[7]

The first analysis used aggregated race and city attributes. The “aggregated attributes” model selected three attributes, and correctly classified 66.4% of intubated and 68.1% of non-intubated patients, yielding a moderate effect: ESS=34.5 (Figure 1).

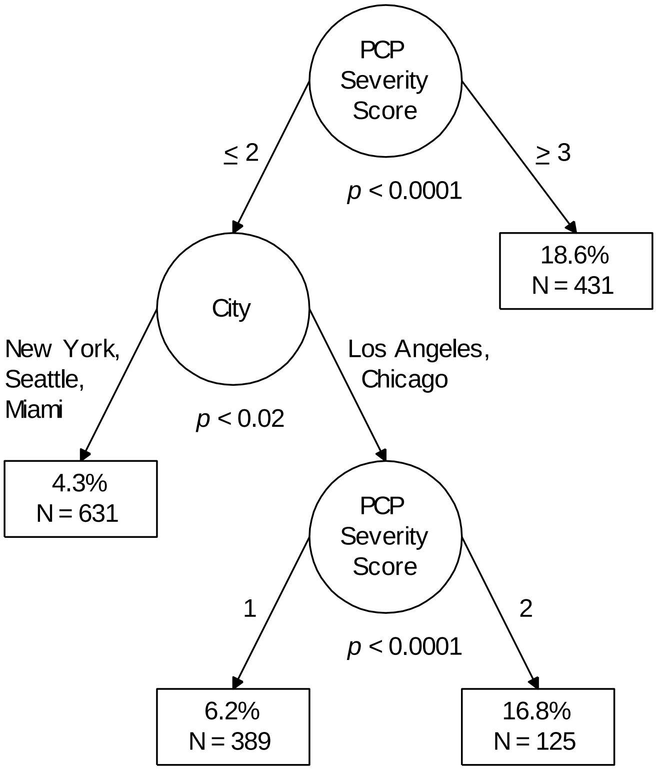


Figure 1: CTA Intubation Model using Aggregated Race and City Attributes

The second analysis used individually dummy-coded race and city attributes, although unlike linear models which require omission of a reference attribute from analysis, with CTA all of the binary attributes compete for admission to the model. The “separately coded attributes” model selected five attributes and correctly classified 78.0% of intubated and 57.0% of non-intubated patients, yielding a moderate effect: ESS=35.0 (Figure 2).

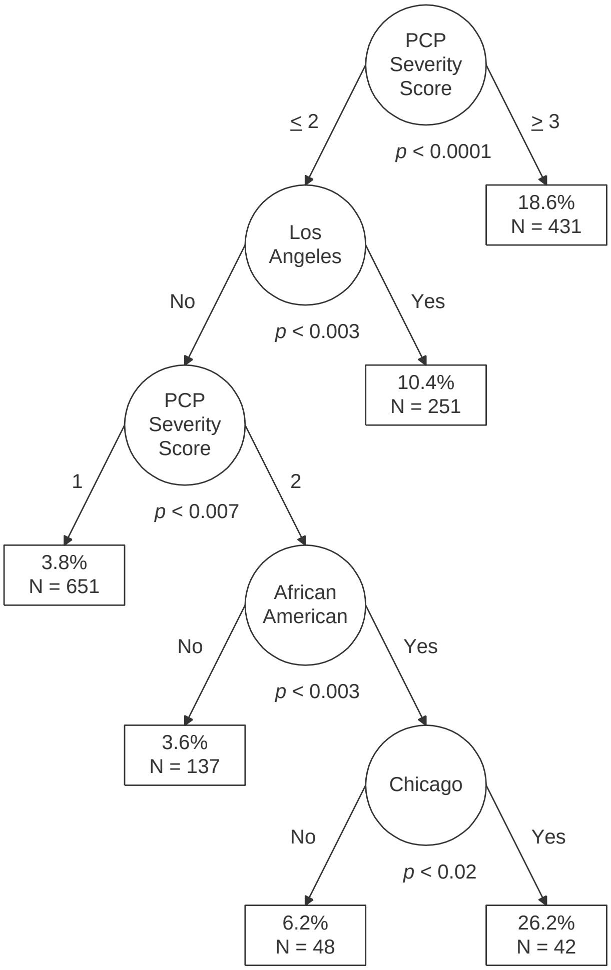


Figure 2: CTA Intubation Model using Separately Coded Race and City Attributes

The models selected the same attributes except for one separately-coded race attribute. The aggregated attributes model employed one attribute to model city, and achieved an overall model efficiency=34.5/3 or 11.5 ESS units-per-attribute. In contrast, the separately-coded attributes model used two attributes to model city, and achieved an overall model efficiency= 35.5/5 or 7.1 ESS units-per-attribute). Thus, the aggregated attributes model is 62% more efficient than the separately-coded attributes model.

Note that the final "Chicago" attribute in the separately-coded attributes CTA model was retained on the basis of model-wise Bonferroni criterion.[2] However, had one additional test of a statistical hypothesis been conducted (e.g., as in any random typical published study), then the Chicago attribute would have been pruned from the model.

Yet another advantage of parsimonious CTA models is that by having fewer endpoints into which observations are partitioned, the minimum endpoint denominators may be larger. Presently, the minimum endpoint denominator for the aggregated attributes model (N=125) is nearly three times larger than for the separately-coded attributes model (N=42). Estimates for the aggregated attributes model are thus more robust over sampling anomalies and likely to cross-generalize, especially for smaller samples.

Using a 3 GHz Intel Pentium D microcomputer, the separately-coded attributes model required 78 CPU seconds to solve, 34.5% more than the 58 CPU seconds required to solve the aggregated attributes model. These problems were relatively simple for automated CTA software to solve, so computing efficiency gained by using aggregated categorical attributes was relatively modest compared to gains obtained in complex analyses. Presently, for example, enumerated CTA models (not shown) involving aggregated (1,394 CPU seconds) or separately-coded (4,054 CPU seconds) attributes revealed a 190.8% gain in computing efficiency.

References

[1]Kleinbaum DG, Kupper LL, Nizam A, Muller KE. *Applied regression analysis and other multivariable methods* (4th Ed.). Thomson Higher Education, Belmont, Ca, 2008.

[2]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington DC, 2005.

2155-0182/10/$3.00

[3] Soltysik RC, Yarnold PR. Introduction to automated CTA software. *Optimal Data Analysis*, In press.

[4] Curtis JR, Yarnold PR, Schwartz DN, Weinstein RA, Bennett CL. Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American Journal of Respiratory and Critical Care Medicine* 2000, 162: 393-398.

[5] Yarnold PR. Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement* 1996, 56:430-442.

[6] Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, *et al*. A new preadmission staging system for predicting inpatient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine* 2000, 161:1081-1086.

[7] Yarnold PR, Soltysik RC. Statistical power analysis for UniODA and CTA. *Optimal Data Analysis*, In press.

Author Notes

Dr. Charles L. Bennett, M.D., Ph.D., generously allowed Paul and Robert to use data which we previously published, generated from his NIH-funded projects which completed approximately a decade ago, for expository purposes. Address correspondence to authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Manual *vs*. Automated CTA: Optimal Preadmission Staging for Inpatient Mortality from *Pneumocystis cariini* Pneumonia

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

Two severity-of-illness models used for staging risk of in-hospital mortality from AIDS-related *Pneumocystis cariini* pneumonia (PCP) were developed using hierarchically optimal classification tree analysis (CTA), with models derived manually via UniODA software. The first of the “*Manual vs. Automated CTA*” series, this study contrasts classification results between original models and corresponding new models derived using automated analysis. Findings provide superior staging systems which may be employed to improve results of applied research in this area.

Software designed to conduct automated CTA became commercially available in the summer of 2010.[1] Research conducted before this time obtained CTA models by a laborious manual process involving UniODA software.[2,3] Beyond obvious savings in time and labor, two primary advantages of automated CTA involve pruning.

First, Type I error for the CTA model is ensured at an investigator-specified level via a sequential Bonferroni procedure.[3] When the CTA model is derived manually, the Bonferroni procedure is conducted as best as possible as the model is grown (this becomes increasingly difficult as the model gains in complexity), as well as after the model can no longer be expanded. Attributes in close proximity to the root variable and having p near 0.05, may be forced out of the model as an increasing number of attributes load on lower branches, disrupting the model and the modeling process. When conducting automated analysis however, this recursive trimming and re-development process is user-transparent: the computer simply executes the algorithm.

Second, the automated software always conducts optimal pruning to explicitly maximize model accuracy, another process which becomes difficult to accomplish manually for complex models.[4] This paper illustrates these advantages using data previously assessed by manual CTA.

PCP in the Early AIDS Era

Research with a sample of 1,339 patients hospitalized with HIV-associated PCP between 1987 and 1990—when hospital mortality rates

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

ranged as high as 60%, is considered first.[5] With five attributes (alveolar-arterial oxygen gradient, $AaPo_2$; age—used twice; body mass index; and a binary indicator of whether a patient had prior history of AIDS) the manually-derived CTA model correctly classified 34.1% of 205 patients who died, and 87.0% of 988 living patients (146 patients had missing data on some attributes in the model), yielding a relatively weak[2] ESS= 21.2. This model offered an order-of-magnitude gain in ESS versus the best prior linear model (logistic regression), and more than doubled the ESS achieved by the best prior classification tree model (regression-based recursive partitioning).[5] This CTA model was pruned to maximize ESS, correctly classifying 74.6% of dead and 59.1% of living patients, and returning moderate ESS=33.7: a 59% improvement versus the non-optimized model.[4] Using three attributes, efficiency=11.2 ESS units-per-attribute, and thus the optimized model was 165% more efficient than the original model (4.2 ESS units-per-attribute).

Automatic CTA software was used to obtain an enumerated CTA model using the same attributes and data available for prior logistic regression and recursive partitioning analyses (see Figure 1). The enumerated CTA model had 69.5% sensitivity, 70.1% specificity, moderate ESS=39.7 (17.8% greater than for the optimized manual model), and efficiency=13.2 ESS units-per-attribute (17.9% greater than the optimized manual model). Analysis was completed in 278 CPU seconds using a 3 GHz Intel Pentium D microcomputer (used in all analyses).

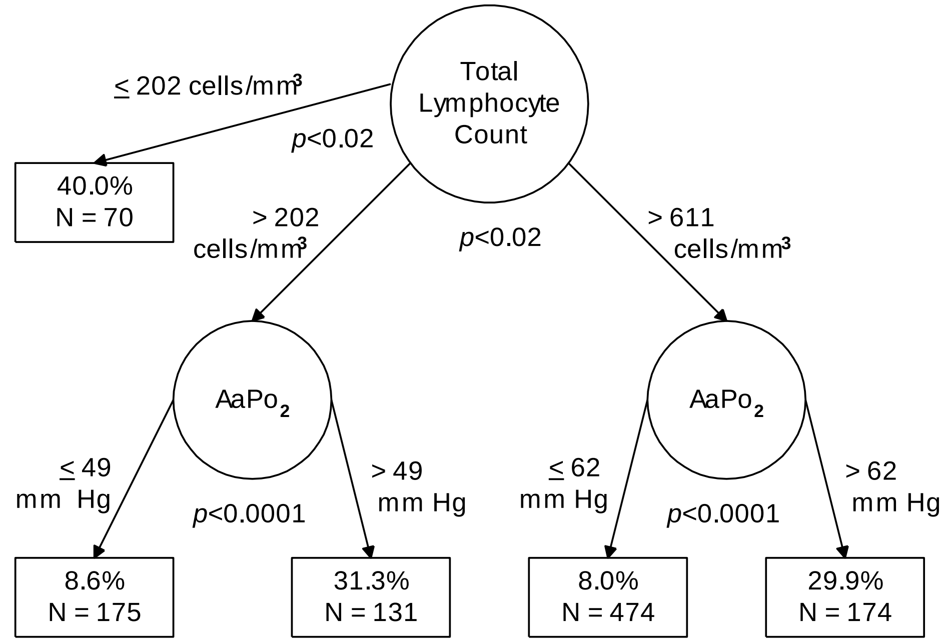


Figure 1: Enumerated CTA Model for Predicting PCP Inpatient Mortality Prior to 1995

Research in the Highly Active Antiretroviral Therapy (HAART) Era

Research investigating a sample of 1,660 patients hospitalized with HIV-associated PCP between 1995 and 1997—the period marking early adoption of non-nucleoside reverse transcriptase and protease inhibitors as HIV therapy, is considered next.[6] Using four attributes (wasting, $AaPo_2$—used twice, and Albumin, the manually-constructed CTA model correctly classified 59.4% of 128 patients who died, and

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

73.7% of 1,066 patients who lived (466 patients had missing data for model attributes), yielding moderate ESS=33.1. Pruned to maximize ESS, the two-attribute optimized model had 53.8% sensitivity (correct prediction of dead patients), 84.3% specificity (correct prediction of living patients), and moderate ESS=45.2 (the optimized model trimmed two nodes previously emanating from the right side of the root node). The optimized model thus offers a 36.6% increase in ESS versus the original model, as well as 172% greater efficiency (22.6 vs. 8.3 ESS units-per-attribute, respectively).[2,4]

An enumerated CTA model was conducted via ODA automatic CTA software, allowing a jackknife-unstable attribute to enter the model if it met the Bonferroni criterion[2] for statistical significance, and if its jackknife ESS exceeded training or jackknife ESS afforded by alternative attributes. To facilitate a direct comparison of models, the three-attribute enumerated model was developed *using the attributes selected by the manually derived model*: wasting, $AaPo_2$, and Albumin. The enumerated CTA model (see Figure 2) had 65.4% sensitivity, 88.2% specificity, a *relatively strong* ESS=53.7 (19% greater than for the optimized manual CTA model), and efficiency=17.9 ESS units-per-attribute (20.8% lower than for the optimized manual model). Analysis was completed in 101 CPU seconds.

In such "disease-staging research" it is customary to provide a *staging table*, such as in Table 1.[5] Rows in the staging table are CTA model endpoints which have been reorganized in order of increasing percent of class 1 (dead patients) membership. Stage is an *ordinal index* of severity of illness, and p_{death} is a *continuous index*: increasing values on either index indicate increasing (worsening) disease severity. The 1st and 4th strata reflect a 16-fold difference in likelihood of dying in-hospital: compared to Stage 1, p_{death} is about four times as high in Stage 2, fifteen times as high in Stage 3, and sixteen times as high in Stage 4.

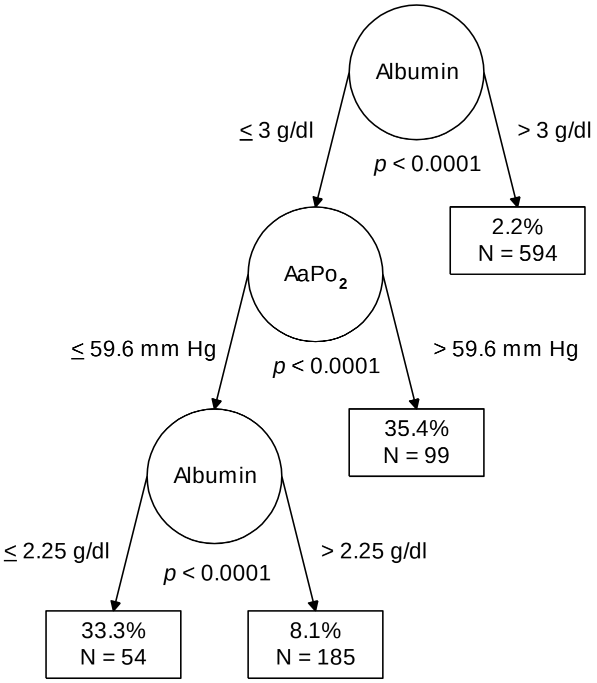


Figure 2: Enumerated CTA Model for Predicting PCP Inpatient Mortality After 1995, Based on Three Attributes

To use the table to stage disease severity for a given patient, simply evaluate fit between patient data and each stage descriptor. Begin at Stage 1, and work sequentially through stages until identifying the descriptor which is true for the data of the patient undergoing staging.

Table 1: Staging Table for Predicting In-Hospital Mortality From PCP, First Model

Stage	Albumin	$AaPo_2$	N	p_{death}	Odds
1	> 3	----	594	0.022	1:44
2	> 2.25	≤ 59.6	185	0.081	1:11
3	≤ 2.25	≤ 59.6	54	0.333	1:2
4	≤ 3	> 59.6	99	0.354	6:11

For example, imagine a patient was 54 years of age, male, morbidly obese, with albumin of 2.47 g/dl and $AaPo_2$ of 61.7 mm Hg. Here, age, gender and mass are immaterial to the staging process, because only attributes in the staging table are used in the staging process. Stage 1 does not fit, as the patient's albumin level is less than 3 g/dl. Stage 2 does not fit because the patient's $AaPo_2$ is greater than 59.6 mm Hg. Stage 3 does not fit as the patient's albumin is greater than 2.25 g/dl (when evaluating a descriptor, the first instance of inaccuracy immediately eliminates the Stage from further consideration). Because the staging table has one degree of freedom, Stage 4 must fit: the patient's albumin is less than 3 g/dl, and $AaPo_2$ is greater than 59.6 mm Hg—so Stage 4 indeed fits the data of this hypothetical patient.

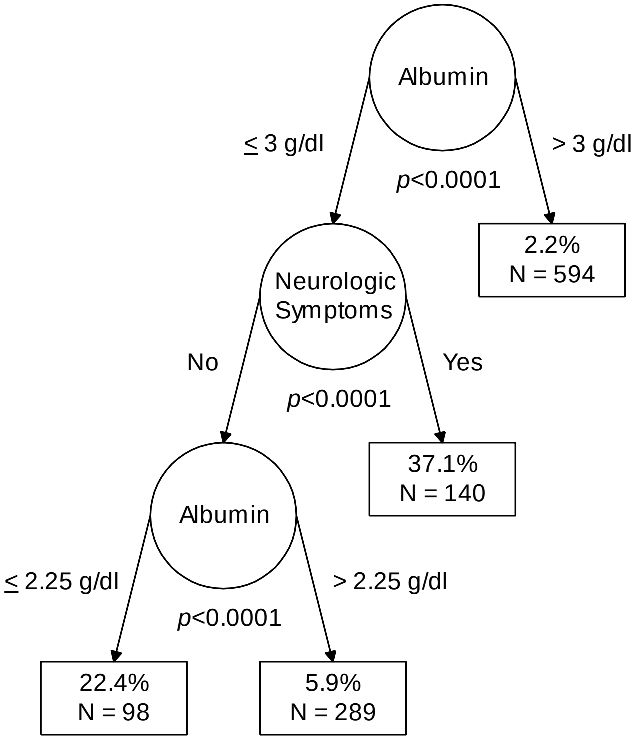


Figure 3: Algorithmic CTA Model Predicting PCP Inpatient Mortality After 1995, Using Attributes From Prior Manual Analysis

Using automated software we next ran automated *algorithmic* CTA (in which the CTA algorithm is performed with optimal parsing but without enumeration), *using all of the attributes employed in original analysis*.[6] A model having three attributes was identified (Figure 3) with 71.2% sensitivity, 83.9% specificity, a *relatively strong* ESS=55.0 (2.5% greater than for the optimized manual model), and efficiency=18.3 ESS units-per-attribute (2.4% greater than for the optimized manual model). Analysis was completed in 85 CPU seconds. The corresponding staging table is presented in Table 2.

Table 2: Staging Table for Predicting In-Hospital Mortality From PCP, Second Model

Stage	Albumin	Neurologic Symptoms	N	p_{death}	Odds
1	> 3	--------	594	0.022	1:44
2	> 2.25	No	289	0.059	1:16
3	≤ 2.25	No	98	0.224	2:7
4	≤ 3	Yes	140	0.371	3:5

An enumerated analysis was conducted next, and a CTA model emerged which yielded a relatively strong effect (ESS=61.4). However, the model included six attributes (two repeated twice), and another attribute which involved a parse. The added complexity, 100% increase in number of attributes employed in exchange for a 11.6% gain in ESS, and 44.1% decrease in efficiency associated with use of the enumerated model, argued in favor of adopting the algorithmic model in this application.

Discussion

Because of inherent importance (having already been judged worthy of publication), and to assemble a literature which may eventually be tapped to assess the magnitude of the boosted

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

ESS offered by these methods in real-world applications, all published CTA models derived manually should *minimally* be optimized using UniODA to return maximum ESS, and the pruned models should be published, as is true presently. Of course, all manually derived CTA models should be pruned to maximize ESS prior to consideration.[4] However, current state-of-the-art methodology for achieving maximum ESS involves conducting automated enumerated CTA, which is the optimal choice.

References

[1]Soltysik RC, Yarnold PR. Introduction to automated CTA. *Optimal Data Analysis* In press.

[2]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

[3]Yarnold PR. Discriminating geriatric and nongeriatric patients using functional status information: an example of classification tree analysis via UniODA. *Educational and Psychological Measurement* 1996, 56:656-667.

[4]Yarnold PR, Soltysik RC. Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis* In press.

[5]Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: an example of hierarchically optimal classification tree analysis. *Statistics in Medicine* 1997, 16: 1451-1463.

[6]Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, *et al*. A new preadmission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine* 2000, 161:1081-1086.

Author Notes

Dr. Charles L. Bennett, M.D., Ph.D., graciously permitted Paul and Robert to use data which we previously published, generated from his NIH-funded projects which completed approximately a decade ago, for expository purposes. Mail correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Manual *vs*. Automated CTA: Psychosocial Adaptation in Young Adolescents with Spina Bifida

Rachael Millstein Coakley, Ph.D., Grayson N. Holmbeck, Ph.D.,
Children's Hospital, Boston / Harvard Medical School Loyola University Chicago

Fred B. Bryant, Ph.D., and Paul R. Yarnold, Ph.D.
Loyola University Chicago Optimal Data Analysis, LLC

Compared to the manually-derived model, the enumerated CTA model was 20% more parsimonious, 3.6% more accurate and 30% more efficient, and was more consistent with *a priori* hypotheses.

A prospective study of how individual- and family-level multimethod, multi-informant attributes predict psychosocial adaptation (scholastic success, social acceptance, positive self-worth) in early adolescence was conducted for a sample of 68 families of children with spina bifida and 68 comparison families of healthy children.[1] Manually-derived CTA indicated that intrinsic motivation, estimated verbal IQ, behavioral conduct, coping style, and physical appearance best predicted psychosocial adaptation in early adolescence: health status was not a factor in the model. The model correctly classified 77.8% of the total sample, yielding ESS=55.0.

An enumerated CTA model was obtained by automated software for the same data used in manual analysis.[2] To be consistent between analyses, attributes were only allowed to enter the model if their associated ESS was stable (did not diminish) in jackknife validity analysis. The enumerated model is illustrated in Figure 1, and performance comparisons are given in Table 1.

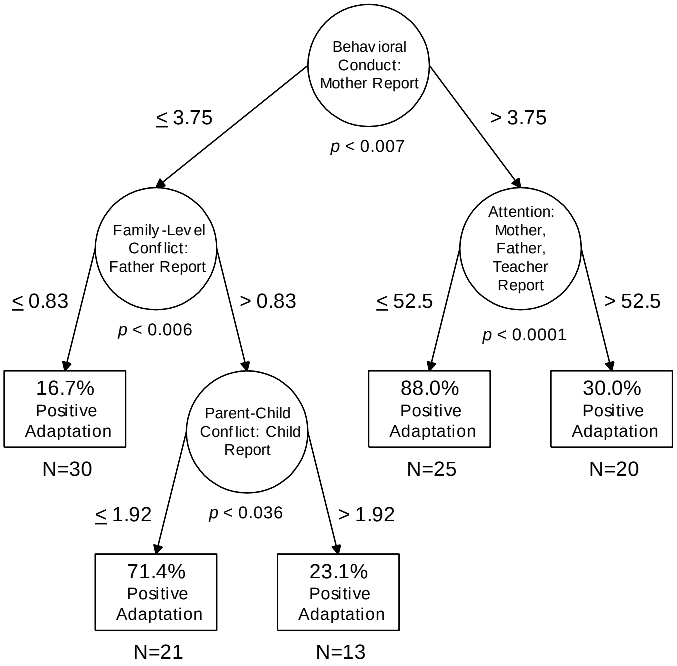

Behavioral Conduct: Mother Report
≤ 3.75
> 3.75
p < 0.007
Family-Level Conflict: Father Report
Attention: Mother, Father, Teacher Report
≤ 0.83
> 0.83
p < 0.006
≤ 52.5
> 52.5
p < 0.0001
16.7% Positive Adaptation
N=30
Parent-Child Conflict: Child Report
88.0% Positive Adaptation
N=25
30.0% Positive Adaptation
N=20
≤ 1.92
p < 0.036
> 1.92
71.4% Positive Adaptation
N=21
23.1% Positive Adaptation
N=13

Figure 1: Enumerated CTA Model Predicting Psychosocial Adaptation in Young Adolescence

Table 1: Comparing Performance of Manually-Derived *vs*. Enumerated CTA Models

		Predicted Class Status			*Predicted Class Status*		
		Manual CTA Model			Enumerated CTA Model		
		Non-Positive Adaptation	Positive Adaptation		Non-Positive Adaptation	Positive Adaptation	
Actual Class Status	Non-Positive Adaptation	40	16	71.4	49	9	84.5
	Positive Adaptation	10	51	83.6	14	37	72.6
		80.0	76.1		77.8	80.4	
	Total N Classified	117			109		
	PAC (%)	77.8			78.9		
	Model ESS	55.0			57.0		
	Number of Attributes	5			4		
	Model Efficiency	11.0			14.3		

Note: Values given to the right of the Positive Adaptation *columns* are the specificity (for non-positive adaptation) and sensitivity (for positive adaptation), and values given under the Positive Adaption *row*, beneath columns, are the negative (for non-positive adaptation) and positive (for positive adaptation) predictive values.[3] Total N classified varies as a function of missing data. PAC=percentage accuracy in classification=100% x (sum of correctly classified observations)/(total N classified).[3] ESS=effect strength for sensitivity, a normed index on which 0 is the level of classification accuracy that is expected by chance, and 100 is perfect accuracy.[3] The number of attributes in the CTA model is given, and model efficiency is defined as model ESS divided by number of attributes; is expressed in terms of mean ESS-units-per-attribute; and is a measure of the mean level of explanatory power per attribute which is used in the model—commonly, as "bang-for-the-buck".[3]

The enumerated model used four attributes rather than five as used in the manual model, and thus it was 80% as complex, or 20% more parsimonious, than the manually-derived model. Compared to the manual model the enumerated model yielded greater ESS (3.6%), PAC (1.4%), efficiency (30%), specificity (18.3%), and positive predictive value (5.7%). In contrast, the manual model had greater sensitivity (15.2%) and negative predictive value (2.8%) than the enumerated model.

The enumerated model predicted 80.4% accurately that 42.2% of the sample would have a positive adaptation, and identified 72.6% of all subjects experiencing positive adaptation. And, the enumerated model predicted 77.8% accurately that 57.8% of the sample would have a non-positive adaptation, identifying 84.5% of all subjects experiencing non-positive adaptation.

The size of sample strata identified by the enumerated model is relatively homogeneous: the largest strata (N=30, 27.5% of classified sample) is 1.3-times larger than the smallest strata (N=13, 11.9% of classified sample). And,

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

all of the attributes loading in the model influenced the classification decisions which were made for a substantial portion of the sample. The percentage of observations classified in part on the basis of their score on the attribute was: Behavioral Conduct (100% of sample); Family-Level Conflict (58.7%), Attention (41.3%) and Parent-Child Conflict (31.2%).

The automated CTA model has several important similarities to the manually-derived CTA model. First, as with the manual model, neither health status (spina bifida *vs*. able-bodied) nor socioeconomic status emerged as factors in the automated model. This suggests that both CTA models were able to identify factors that were more predictive of psychosocial adaptation than the group differences often identified in pediatric research. Second, the factor "behavioral conduct in the classroom" emerged as being highly significant in both models. This demonstrates consistency between the models and reinforces the relationship between behavioral control in the classroom and psychosocial adaptation.

There were also important differences between the two models. Counter to our original hypotheses, the *manually derived model* did *not* identify any family-level variables, *nor* did it include any variables based on mother or father report. In contrast, the *automated CTA model* supported our original hypothesis by identifying two family-level variables in the model and including three variables based in part on mother and father report. Another difference between the two models is that in the manual model all of the factors were based on characteristics of the child and two of the factors represented more internalized child qualities (i.e., intrinsic motivation, coping style). In comparison, only half of the automated model focused on child factors and these included only externalized or observable behaviors (i.e., conduct, attention).

In summary, the automated model presents a more parsimonious way of classifying this sample and supports the researchers' original hypotheses by including family-level factors and information from multiple informants (parents, teachers, child). However, it identifies a substantially different constellation of factors in the classification of psychosocial adaptation as compared to the manual model. Many theoretically important factors that emerged in the manual model that are well supported in pediatric research on psychosocial adaptation (e.g., motivation, IQ, coping style, and attractiveness) were not included in the automated model. Instead, the automated model selected a narrower constellation of factors that was highly focused on behavioral presentation and family-level conflict. These models likely represent two theoretically viable and empirically supported paths to psychosocial adaptation.

References

[1]Coakley RM, Holmbeck GN, Bryant FB. Constructing a prospective model of psychosocial adaptation in young adolescents with spina bifida: an application of optimal data analysis. *Journal of Pediatric Psychology* 2006, 31:1084-1099.

[2]Yarnold PR. Soltysik RC. Manual *vs*. automated CTA: optimal preadmission staging for inpatient mortality from *Pneumocystis carinii* pneumonia. *Optimal Data Analysis* 2010, 1:50-54.

[3]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

Author Notes

Completion of this manuscript was supported in part by research grants from the March of Dimes Birth Defects Foundation (12-FY01-0098) and the National Institute of Child Health and Human Development (RO1 HD048629). The authors wish to thank Ann Walsh Johnson, Joy Ito, Pat McGovern, Pat Braun, Caroline

Anderson, David McLone, John Lubicky, the Spina Bifida Association of Illinois, the staff of the spina bifida clinics at Children’s Memorial Hospital, Shriners Hospitals for Children, Chicago, and Loyola University of Chicago Medical Center. We also thank numerous undergraduate and graduate research assistants for help with data collection and data entry. Most importantly, we gratefully acknowledge the contributions to this study by the parents, children, teachers, and health professionals who participated over many years. Correspondence: Rachael M. Coakley, Children's Hospital Boston, Pain Treatment Service, 333 Longwood Ave, 5th Floor, Boston, MA 0215. eMail: rachael.coakley@childrens.harvard.edu.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Gen-UniODA *vs*. Log-Linear Model: Modeling Organizational Discrimination

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

An application involving a binary class variable (gender), an ordinal attribute (academic rank), and two testing periods (separated by six years) was troublesome for the log-linear model, but was easily analyzed using Gen-UniODA.

Everett[1] cross-tabulated gender and faculty rank (1=Instructor, 2=Assistant Professor; 3=Associate Professor; 4=Professor) in 1978 and 1984 at the University of New South Wales (Table 1).

Table 1: Number of Faculty by Academic Rank, Gender, and Year

	1978		1984	
Rank	Male	Female	Male	Female
1	45	28	39	28
2	176	21	114	27
3	144	6	171	18
4	127	2	121	5

Note: Adapted from Everett (1990), tabled are frequency counts.

Log-linear analysis was used to model the relative odds of men versus women at each academic rank level and across time. Analysis also included additional putative determinants of rank (unavailable for this example), including academic degree, publication level, age. Age and publication level were each split into three categories, and degree into two categories, in order to limit the number of cells in the design matrix: in light of the modest sample size, it is conceivable there could be empty cells in a complex design. Five predictor variables dictated too many interaction terms (i.e., the design matrix would be too large for the sample), so the three putative determinants were combined into a single 18-level polychotomous variable which possessed no inherent order. Examination of confidence limits suggested: "despite these suggested trends across rank and across time, none of the direct discrimination values differ significantly" (p. 383). In the final analysis which was reported, all estimates obtained by collapsed contingency (CC) table odds ratio analysis fell outside of the range of odds estimated by other methods, indicating induction of Simpson's Paradox[2]: "The underestimation is much more severe for the odds ratio CC derived from collapsing fitted subtables, further underlining problems associated with collapsing across a non-independent variable" (p. 384).

Using UniODA, in contrast, the analysis is straightforward: the objective is to determine if the relative distribution of males and females (*class variable* is gender) differs on the ordinal academic rank measure (*attribute* is rank), and if this relationship has changed across time (*generalizability* variable is year).

After year (1=1978, 2=1984) and gender (1=male, 2=female) were dummy-coded, data were analyzed using the following ODA[3] code (commands are indicated in red; non-directional exploratory analysis is conducted as no *a priori* hypothesis regarding the direction of discrimination was postulated):

```
open data;
output everett.out;
vars rank gender year;
data;
1 1 1 (repeated 45 times)
1 0 1 (repeated 28 times)
2 1 1 (repeated 176 times)
2 0 1 (repeated 21 times)
3 1 1 (repeated 144 times)
3 0 1 (repeated 6 times)
4 1 1 (repeated 127 times)
4 0 1 (repeated 2 times)
1 1 2 (repeated 39 times)
1 0 2 (repeated 28 times)
2 1 2 (repeated 114 times)
2 0 2 (repeated 27 times)
3 1 2 (repeated 171 times)
3 0 2 (repeated 18 times)
4 1 2 (repeated 121 times)
4 0 2 (repeated 5 times)
end;
class gender;
attr rank;
gen  year;
mcarlo iter 25000;
loo;
go;
```

The resulting Gen-UniODA model was: if academic rank$\leq$2 then predict gender=female (77.0% correct), otherwise predict that gender= male (60.1%). The omnibus test was statistically significant (p<0.0001), and the effect was of moderate strength (ESS=37.1), indicating the model generalized over year (UniODA models all were stable in jackknife validity analysis).

Applying the Gen-UniODA model to the 1978 data: females were 86.0% correct; males 55.1% correct; p<0.0001; ESS= 41.1. Applying the model to the 1984 data: females were 70.5% correct; males 65.6% correct; p<0.0001; ESS= 36.1. Omnibus performance values were inside the domain defined by corresponding 1978 and 1984 values: again, no evidence of potential paradoxical confounding.[2]

Discussion

Gen-UniODA found moderate evidence of gender discrimination: a greater proportion of females are Instructors or Assistant Professors, and of males are (Associate) Professors, than is expected by chance. Eyeball analysis suggests the strength of the effect may be diminishing in time, because the percent of females classified correctly by the model, and ESS, fell in 1984. In addition, relative to 1978, in 1984 the number of male professors fell 4.6% while the number of women in this rank increased by 150%. The rank of Associate Professor saw a 18.8% gain in males, compared to a 200% increase in females. There were 35.2% fewer male Assistant Professors compared with a 28.5% gain for females, and while male Instructors diminished by 13.3%, there was no change in this rank for females. Considered together these results suggest that not only is the relative standing of women increasing, but so too is the relative number of women on the faculty.

Information beyond academic rank, sex and year was all that was available for analysis presently. It will be interesting to model data such as considered presently—augmented by additional putative predictors, via MultiODA[4] (optimal analogue to log-linear model) or hierarchically optimal classification tree analysis,[5] as well as to evaluate optimization of suboptimal models identified in the present context using UniODA.[6]

References

[1]Everett JE. Discrimination measure using contingency tables. *Multivariate Behavioral Research* 1990, 25:371-386.

[2]Yarnold PR. Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement* 1996, 56:430-442.

[3]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2004.

[4]Soltysik RC, Yarnold PR. The Warmack-Gonzalez algorithm for linear two-category multivariable optimal discriminant analysis. *Computers and Operations Research* 1994, 21: 735-745.

[5]Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: an example of hierarchically optimal classification tree analysis. *Statistics in Medicine* 1997, 16:1451-1463.

[6]Yarnold PR, Hart LA, Soltysik RC. Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement* 1994, 54:73-85.

Author Notes

Address correspondence to the author at: Optimal Data Analysis LLC, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

UniODA vs. Chi-Square: Ordinal Data Sometimes Feign Categorical

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Assessed using perhaps the most widely used type of measurement scale in all science, ordinal data are often misidentified as being categorical, and incorrectly analyzed by chi-square analysis. Three examples drawn from the literature are reanalyzed.

Consisting of a relatively small number of graduated levels of the measured attribute, ordinal scales may be the most broadly employed type of measurement scale in all of science. Likert-type scales, typically involving between three and ten levels, are perhaps most common.[1] For example, one's socioeconomic status is often assessed using a three-level ordinal scale, with categories corresponding to low, middle, and upper class. Also widely used, ordinal categorical scales consist of a relatively small number of qualitative categories ordered with respect to some theoretical factor.[2] For example, at the end of a clinical trial patients might be classified as being worse, unchanged, or better: the three qualitative categories are worse, unchanged, and better; the theoretical factor is quality of clinical outcome; and the categories are ordered from lowest (worse) to highest (better) with respect to quality of clinical outcome.

Since the metric underlying the attribute is ordinal, neither chi-square (nominal data) nor *t*-test (interval data) is appropriate to assess if therapies can be discriminated on the basis of clinical outcome. Traditional methods used for analysis of ordinal data include Mann-Whitney *U* test or the log-linear model, but excessive ties compromise *U*, and maximum likelihood-based methods require large samples.[3-5] Assuming neither the absence of ties nor the presence of large samples, univariate optimal discriminant analysis (UniODA) is ideal for such designs.

Plaintiff Gender and Age

Seaman and Hill[6] analyzed data obtained by Cox and Key[7] from court records of an Ohio county, involving the frequency of plaintiffs in divorce actions cross-classified by gender (wife or husband) and age (<25, 25-34, 35-44, >44). "The hypothesis that the proportion of plaintiffs that are husbands is the same, regardless of age" (p. 454) was tested using the traditional model, homogeneity of proportions. All possible *post hoc* pairwise comparisons—involving 6 separate 2-by-2 chi-square tests, were conducted to ascertain the specific reason the omnibus test was statistically significant. Two pairwise comparisons were statistically significant: those comparing the >44 age category with the 25-34 and 35-44 categories (p's<0.05). Analysis via chi-square thus indicated a greater proportion of husband plaintiffs in the >44 age category, and a greater proportion of wife plaintiffs in the 25-34 and 35-44 age categories. No statistically significant pairwise comparisons involved the <25 age category, so this strata could not be assessed in relation to other strata in the study.

Table 1: Plaintiff Age in a Divorce Action

Age	<25	25-34	35-44	>44
Husband	8	8	6	16
Wife	18	48	22	10

Note: Adapted from Seaman and Hill (1996). Tabled are frequency counts.

After gender (1=Husband, 2=Wife) and age (1=’<25’, 2=’25-34’, 3=’35-44’, 4=’>44’) were dummy-coded, data were reanalyzed using the following ODA[5] code (commands indicated in red; non-directional exploratory analysis is conducted as no *a priori* hypothesis was made):

```
open data;
output seaman.out;
vars gender age;
data;
2 4 (repeated 10 times)
2 3 (repeated 22 times)
2 2 (repeated 48 times)
2 1 (repeated 18 times)
1 4 (repeated 16 times)
1 3 (repeated 6 times)
1 2 (repeated 8 times)
1 1 (repeated 8 times)
end;
class gender;
attr age;
mcarlo iter 25000;
loo;
go;
```

The resulting UniODA model was: if age≤35-44 then predict class=wife, otherwise predict class=husband. The model achieved a moderate ESS of 31.9 (*p*<0.0001), and results were stable in jackknife validity analysis. The model classified 88 (90%) of 98 women correctly, versus only 16 (42%) of 38 men. All subjects were classified by the ODA model, including those younger than 25 years of age.

Outcomes of Marital Therapy

Snyder, Wills and Grady-Fletcher[8] reported the following four-year termination outcomes of two different types of therapy for unhappily married couples. The expected value for both entries in the right-most column of the data table is less than five, invalidating the use of chi-square with this sparse table.[9] An omnibus chi-square statistic was given for the 2-by-3 table, then eyeball interpretation of the omnibus effect was rendered: “a significantly higher percentage of (behavior therapy couples) had experienced divorce, *p*<0.01.” Although no explanation was provided—perhaps to defeat the aforementioned minimum expectation assumption violation, the No Change (“distressed”) and Improved classes were collapsed and chi-square reported higher divorce rates for behavior therapy, *p*<0.05.

Table 2: Outcomes of Marital Therapies

Type of Therapy	Divorced	No Change	Improved
Insight	3	22	4
Behavior	12	13	1

Note: Tabled are frequency counts.

These data were analyzed by ODA code paralleling that used in the first example. The model was: if outcome=divorced then predict therapy=behavior, otherwise predict therapy= insight. The model correctly classified 90% in insight therapy, 46% in behavior therapy, and yielded a moderate ESS of 35.9 (*p*<0.006).

Strength of Gender Differences

Hyde and Plant[10] reported frequencies of five categories of Cohen's *d* measure of effect strength for representative studies of gender differences, versus studies of other effects in the field of psychology. An omnibus chi-square statistic was provided for the 2-by-5 table ($p<0.0001$): "The difference between the distributions of gender effect sizes and other effect sizes is highly significant." Pairwise comparisons to disentangle the omnibus effect were not reported. Eyeball analysis suggested: "more gender differences fall in the close-to-zero category than other effects in psychology."

Table 3: Cohen's *d* by Type of Study

Type of Study	≤0.1	≤0.35	≤0.65	≤1.0	>1.0
Gender	43	60	46	17	5
Other	17	89	116	60	20

Note: Tabled are frequency counts.

For these data the exploratory hypothesis that type of study could be discriminated on the basis of effect strength was tested using priors-weighted UniODA, via ODA code paralleling that used in the prior examples. The model was: if $d\leq 0.35$ then predict gender study; otherwise, predict non-gender study. Thus, relative to other areas, gender studies have disproportionately more effect sizes in the close-to-zero (≤0.1) and next-to-close-to-zero (0.11-0.35) categories. By correctly classifying 60.2% of the gender difference studies, and 64.9% of other studies, the model yielded a moderate, jackknife-stable ESS=25.1 ($p<0.0001$).

Discussion

Initial study of the congruence between chi-square and UniODA in analysis of real-world data suggests consistent findings may often be achieved, and instances of inconsistent findings may often accompany grossly imbalanced marginals.[11] Distinct advantages versus chi-square include that, for UniODA: *directional* tests of statistical hypotheses may be conducted; the validity of exact *p* is uncompromised by sparse, empty or missing cells, small samples or imbalanced marginal distributions; and use of the normed ESS index allows direct comparison of model performance across analyses differing in number of observations, marginal imbalance, and/or number of levels for categorical class variables and/or attributes.

Optimal ordinal analysis may be generalized to designs involving class variables having more than two categories (Yarnold and Soltysik[5] discuss degenerate designs involving fewer categories for attribute than class). For example, imagine a design involving a three-category class variable—such as therapies A, B, and C, and an ordinal categorical attribute with at least three ordinal improvement categories—such as none, some, and much. A UniODA model for such a design would be of the form: if improvement=none, predict therapy=A; otherwise, if improvement=some, predict therapy=B; otherwise predict therapy=C. As is true for all ODA applications, for three-category designs: exact *p* is obtained for performance achieved by the model; mean sensitivity across therapies is translated into the normed ESS scale of effect strength; and leave-one-out (LOO) "jackknife" validity analysis is used to assess the potential generalizability of the findings were the model used to classify independent random samples.

Generalizing exact ordinal analysis to designs involving more than one assessment dimension is also straightforward, whether by linear or nonlinear methods. Imagine an application having two therapeutic strategies (class variable) and two ordinal categorical outcome scales (attributes)—one assessing degree of recovery (worse, unchanged, better), and the other assessing satisfaction (unhappy, neutral, happy). Using an optimal multivariable linear approach with these data, one could obtain a main effects model including an intercept and separate coefficients for recovery and satisfaction; a saturated model additionally including a coefficient for the recovery-by-satisfaction interaction; and a quadratic model additionally including coefficients for the squares (or higher exponents) of each main effect.[12-14] Model coefficients may be real numbers, or may be constrained to any range, even binary.[15] Structurally, these ODA models are similar to models developed via traditional multivariable techniques such as discriminant or logistic regression analysis. Functionally, however—as is constitutionally true of all ODA analyses, these models would explicitly maximize (weighted) classification accuracy achieved for the sample.[5] Using an optimal multivariable nonlinear approach with these data currently entails conducting hierarchically optimal classification tree analysis, or CTA.[16]

Regardless of choice of (non)linear method, to ensure the validity of analytic findings it is recommended that variables which truly are measured using an ordinal scale are treated as though they were in fact measured using an ordinal scale.

References

[1] Nunnally JC. *Psychometric theory*. McGraw-Hill, New York NY, 1967.

[2] Kazdin AE. *Research design in clinical psychology (2nd ed.).* Allyn & Bacon, New York NY, 1992.

[3] Mann HB, Whitney DR. On a test of whether one of two variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947, 18:50-60.Snyder, Wills, 1991

[4] Hagenaars JA. *Categorical longitudinal data: log-linear, panel, trend, and cohort analysis.* Sage, Newbury Park, CA, 1990.

[5] Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2004.

[6] Seaman MA, Hill CC. Pairwise comparisons for proportions: a note on Cox and Key. *Educational and Psychological Measurement* 1996, 56:452-459.

[7] Cox MK, Key CH. Post hoc pairwise comparisons for the chi-square test of homogeneity of proportions. *Educational and Psychological Measurement* 1993, 53:951-962.

[8] Snyder DK, Wills RM, Grady-Fletcher A. Long-term effectiveness of behavioral versus insight-oriented marital therapy: a four-year follow up study. *Journal of Consulting and Clinical Psychology* 1991, 59:138-141.

[9] Yarnold JK (1970). The minimum expectation of chi-square goodness-of-fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association, 65*, 864-886.

[10] Hyde JS, Plant EA. Magnitude of psychological gender differences: another side to the story. *American Psychologist* 1995, 50:159-161.

[11] Yarnold PR, Hart LA, Soltysik RC. Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement* 1994, 54:73-85.

[12] Yarnold PR, Soltysik RC, Martin GJ. Heart rate variability and susceptibility for sudden cardiac death: an example of multivariable optimal discriminant analysis. *Statistics in Medicine* 1994, 13:1015-1021.

[13] Soltysik RC, Yarnold PR. The Warmack-Gonzalez algorithm for linear two-category multivariable optimal discriminant analysis. *Computers and Operations Research* 1994, 21: 735-745.

[14]Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ. Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine* 1995, 10:601-606.

[15]Yarnold PR, Soltysik RC, Lefevre F, Martin GJ. Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: unit-weighted MultiODA for binary data. *Statistics in Medicine* 1998, 17:2405-2414.

[16]Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: an example of hierarchically optimal classification tree analysis. *Statistics in Medicine* 1997, 16:1451-1463.

Author Notes

Address correspondence to the author at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

The Use of Unconfounded Climatic Data Improves Atmospheric Prediction

Robert C. Soltysik, M.S., and Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

This report improves measurement properties of data and analytic methods widely used in meteorological modeling and forecasting. Paradoxical confounding is defined and demonstrated using global temperature land-ocean index data. It is shown that failure to address paradoxical confounding results in suboptimal atmospheric circulation pattern models, and correcting prior measurement and analytic deficiencies results in more accurate prediction of temperature and precipitation anomalies, and export of Arctic sea ice.

Simpson's Paradox may be the single greatest threat to the validity of quantitative analysis in all empirical science.[1] The Paradox can occur when data from two or more samples, groups or time periods are combined into a single sample: under such conditions, results obtained when analyzing the combined data may be different than when analyzing individual data sets separately. The following hypothetical example illustrates confounding for a simple correlation.

Imagine we wish to correlate sea level pressure (SLP) with thunderstorm severity rated using a scale with greater values indicating greater severity, and data collected at two locations. Location A usually has relatively low SLP and short-lived, fast-moving storms: the lower the SLP the more severe the storm. The hypothetical correlation model (r=-0.8) relating SLP and severity is indicated using arrow "**A**" in Figure 1 (individual hypothetical data points from location A are indicated as "a"): data swarm A indicates strong *negative* association.

Compared to A, Location B usually has relatively high SLP and long-lived slow-moving storms: the lower the SLP the more severe the storm. The correlation (r=-0.8) relating SLP and severity is indicated in Figure 1 by arrow "**B**" (individual hypothetical data points from location B are indicated as "b"): data swarm B indicates strong *negative* association.

When data from Locations A and B are combined, the resulting correlation model (r= 0.7) relating SLP and severity is indicated by arrow "**C**" (individual hypothetical data points for combined sample are all "a" and "b"): data swarm C indicates strong *positive* association.

In this hypothetical example, for two individual samples (Locations A and B) considered separately the analysis reveals that more severe storms are associated with *decreasing* SLP. For the combined data, the same analysis reveals that more severe storms are associated with *increasing* SLP.

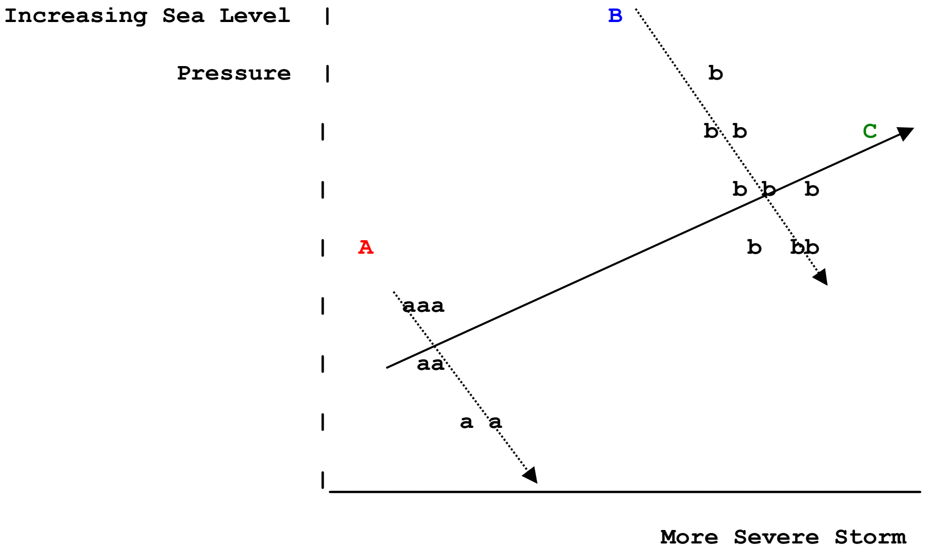


Figure 1: Hypothetical Illustration of Paradoxical Confounding

Simpson's Paradox threatens the validity of quantitative atmospheric science because nonstationarity is prevalent in longitudinal data series used in atmospheric science, such as temperature or pressure—and nonstationarity can induce Simpson's Paradox. For example, global surface temperature data clearly are nonstationary: in Figure 2, anomalies are computed relative to the period 1951-1980 (http://data.giss.nasa.gov/gistemp/).

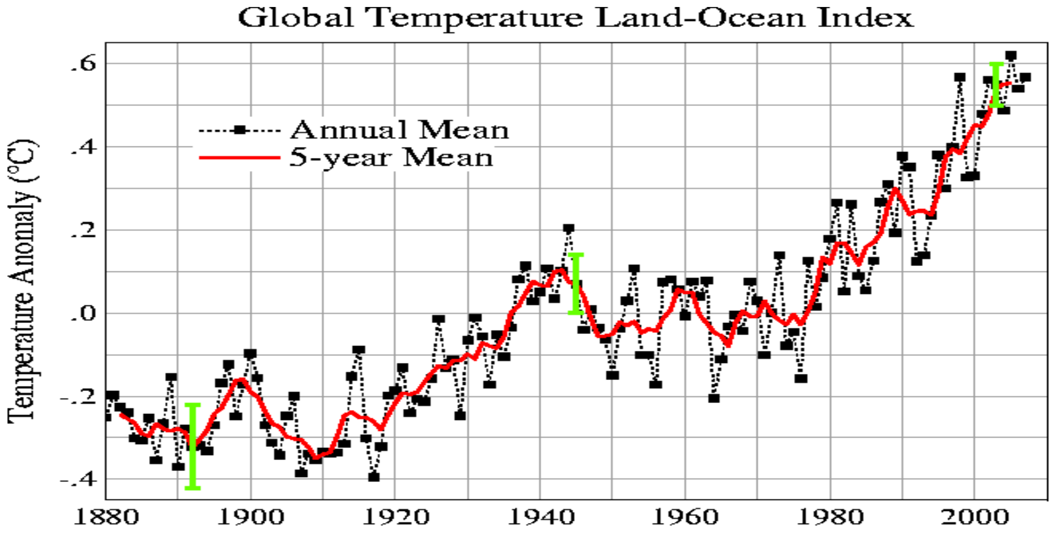


Figure 2: Mean Global Temperature Land-Ocean Index Anomaly by Year

Analysis was restricted to the time period that is the focus of most current quantitative atmospheric science, beginning in the year 1948. Eyeball inspection of Figure 2 suggests a relatively flat trajectory ("*stationary series*") through 1976, versus a steadily increasing trajectory ("*non-stationary series*") across subsequent years. Regression analyses modeling temperature anomaly (dependent measure) as a function of year (independent measure), separately by month, are summarized In Table 1: findings confirm eyeball observations, and establish the generalizability of the phenomenon to a time period more granular than is afforded by annual measurements.

Tabled for each model is the intercept as well as the value of the t-test for the two-tailed hypothesis that the value of the intercept is zero, and the associated Type I error rate. For every model, in every month, the intercept is *not* significantly different than zero for the stationary series, but *is* significantly different than zero for the nonstationary and combined series. Also tabled for each model is the slope (regression beta weight) and the value of the t-test for the two-tailed hypothesis that the value of the slope is zero, and the associated Type I error rate. Consistent with findings for intercept, for every model, in every month, the slope is *not* significantly different than zero for the stationary series, but *is* significantly different than zero for the nonstationary and combined series. Finally, Table 1 provides the percent of variance in temperature that is explained by the regression model as a function of year (R^2), and p for the regression model. If model performance for the combined sample lies outside performance results for samples considered individually, then paradoxical confounding exists: this is indicated using **red**.

Table 1: Regression Modeling of Temperature Anomaly using Year, Separately by Month: Evidence of Paradoxical Confounding

Month	Time Period	Intercept, t, p			Slope, t, p			R^2, p	
January	Stationary	559.3	0.8	0.45	-0.29	-0.8	0.46	2.1	0.45
	Non-Stationary	-3239.1	-5.3	0.0001	1.64	5.3	0.0001	49.4	0.0001
	Combined	-2114.5	-7.9	0.0001	1.08	8.0	0.0001	52.2	0.0001
February	Stationary	-140.0	-0.2	0.87	0.07	0.2	0.87	1.0	0.87
	Non-Stationary	-3842.6	-5.5	0.0001	1.95	5.6	0.0001	51.6	0.0001
	Combined	-2451.3	-8.4	0.0001	1.25	8.5	0.0001	55.3	0.0001
March	Stationary	-550.5	-0.8	0.46	0.28	0.8	0.46	2.1	0.46
	Non-Stationary	-3374.5	-5.9	0.0001	1.71	5.9	0.0001	54.9	0.0001
	Combined	-2451.8	-10.0	0.0001	1.25	10.1	0.0001	63.8	0.0001
April	Stationary	-229.4	-0.4	0.71	0.12	0.4	0.72	0.5	0.72
	Non-Stationary	-3216.2	-7.1	0.0001	1.63	7.1	0.0001	63.7	0.0001
	Combined	-2159.7	-10.3	0.0001	1.10	10.4	0.0001	65.0	0.0001
May	Stationary	-197.5	-0.3	0.75	0.10	0.3	0.75	0.4	0.75
	Non-Stationary	-2590.9	-4.9	0.0001	1.31	4.9	0.0001	45.4	0.0001
	Combined	-1845.2	-8.6	0.0001	0.94	8.7	0.0001	56.7	0.0001
June	Stationary	-145.7	-0.3	0.75	0.07	0.3	0.75	0.4	0.75
	Non-Stationary	-3291.0	-6.3	0.0001	1.67	6.4	0.0001	58.3	0.0001
	Combined	-1918.6	-9.7	0.0001	0.98	9.7	0.0001	62.0	0.0001
July	Stationary	-111.3	-0.3	0.78	0.06	0.3	0.79	0.3	0.79
	Non-Stationary	-2841.5	-4.7	0.0001	1.44	4.8	0.0001	43.8	0.0001
	Combined	-1937.1	-9.5	0.0001	0.99	9.6	0.0001	61.3	0.0001

August	**Stationary**	203.2	0.4	0.73	-0.10	-0.4	0.73	0.5	0.73
	Non-Stationary	-3492.9	-6.5	0.0001	1.77	6.6	0.0001	60.0	0.0001
	Combined	-1933.3	-8.5	0.0001	0.98	8.6	0.0001	55.8	0.0001
September	**Stationary**	3.9	0.0	0.99	-0.01	-0.0	0.99	0.1	0.99
	Non-Stationary	-3359.2	-6.3	0.0001	1.70	6.4	0.0001	58.4	0.0001
	Combined	-1888.2	-8.8	0.0001	0.96	8.8	0.0001	57.3	0.0001
October	**Stationary**	298.4	0.6	0.58	-0.15	-0.6	0.58	1.2	0.58
	Non-Stationary	-4082.0	-8.5	0.0001	2.06	8.5	0.0001	71.4	0.0001
	Combined	-1920.6	-8.5	0.0001	0.98	8.5	0.0001	55.7	0.0001
November	**Stationary**	-253.9	-0.5	.062	0.13	0.5	0.62	0.9	0.62
	Non-Stationary	-3719.7	-6.1	0.0001	1.88	6.1	0.0001	56.3	0.0001
	Combined	-2056.9	-9.1	0.0001	1.05	9.1	0.0001	58.9	0.0001
December	**Stationary**	41.4	0.1	0.95	-0.02	-0.7	0.95	0.1	0.95
	Non-Stationary	-3076.1	-5.0	0.0001	1.56	5.1	0.0001	45.1	0.0001
	Combined	-1998.4	-8.2	0.0001	1.02	8.3	0.0001	54.2	0.0001

Note: Stationary=1948–1976; Non-Stationary=1977–2007; Combined=1948-2007.

This exercise demonstrates that temperature does *not* increase between 1948 and 1976, but *does* increase thereafter; fundamentally different “statistical infrastructure” (i.e., regression models) underlies the stationary and nonstationary series; and combining data from these two series typically results in paradoxical confounding. What is the nature of the effect of this confounding? In the initial hypothetical example, the effect of the confounding was one of “direction”: the result for the combined sample was opposite in direction to results obtained for individual samples. For actual temperature data the effect of confounding is one of “magnitude”: the finding for the combined sample is in the same direction (indicating increase over time) as the finding for the nonstationary series, but the model for the combined sample misestimates the magnitude of the effect. For any month, compared to the nonstationary series, the model for the combined sample has intercept and slope coefficients with lower absolute values: models for the combined data thus underestimate the rate of change in temperature for the nonstationary series. *If Simpson’s Paradox confounds fundamental data, then models using those confounded data also are confounded.*

Measuring Atmospheric Circulation Patterns

Seminal research conducted by Barnston and Livezey used orthogonally rotated principal components analysis (PCA) of monthly mean 700 mb geopotential heights to identify the major modes of northern hemisphere upper-air variability.[2] They used combined data from the years 1950 through 1984: measurements were taken on a 358-point grid covering latitudes from 20ºN to 85ºN, and ten “robust” modes (components) were identified which persisted throughout the year. The Climate Prediction Center (CPC) performed a similar analysis of northern hemisphere 500 mb heights using data from 1950 to 2000: ten modes were identified and used to compute the values of the teleconnection indices (http://www.cpc.noaa.gov/data/teledoc/telepatcalc.shtml). Table 2 describes the ten modes of upper-air variability determined by the CPC analysis.

Table 2: Ten Modes of Upper-Air Variability Determined by the CPC Analysis

CPC Mode	Abbreviation	Description
1	NAO	North Atlantic Oscillation
2	EA	East Atlantic Pattern
3	WP	West Pacific Pattern
4	EP/NP	East Pacific / North Pacific Pattern
5	PNA	Pacific / North American Pattern
6	EA/WR	East Atlantic/West Russia Pattern
7	SCA	Scandinavia Pattern
8	TNH	Tropical / Northern Hemisphere Pattern
9	POL	Polar/ Eurasia Pattern
10	PT	Pacific Transition Pattern

Figure 3 gives the total variance in 500 mb height data that is explained by these ten modes each year. In the Figure, blue shading indicates levels of explained variation that fall below the mean. In 2003 the combined sample includes an equal number of data points from stationary (1950-1976) and nonstationary (1977-2003) series, but data from the nonstationary series dominate the combined sample by 2004. Extrapolation of earlier results suggests that increasing domination will accelerate paradoxical confounding and resulting underestimation of magnitude of effect. Note that after 2003, *performance of the quantitative model used to identify major modes of northern hemisphere upper-air variability has never been lower*.

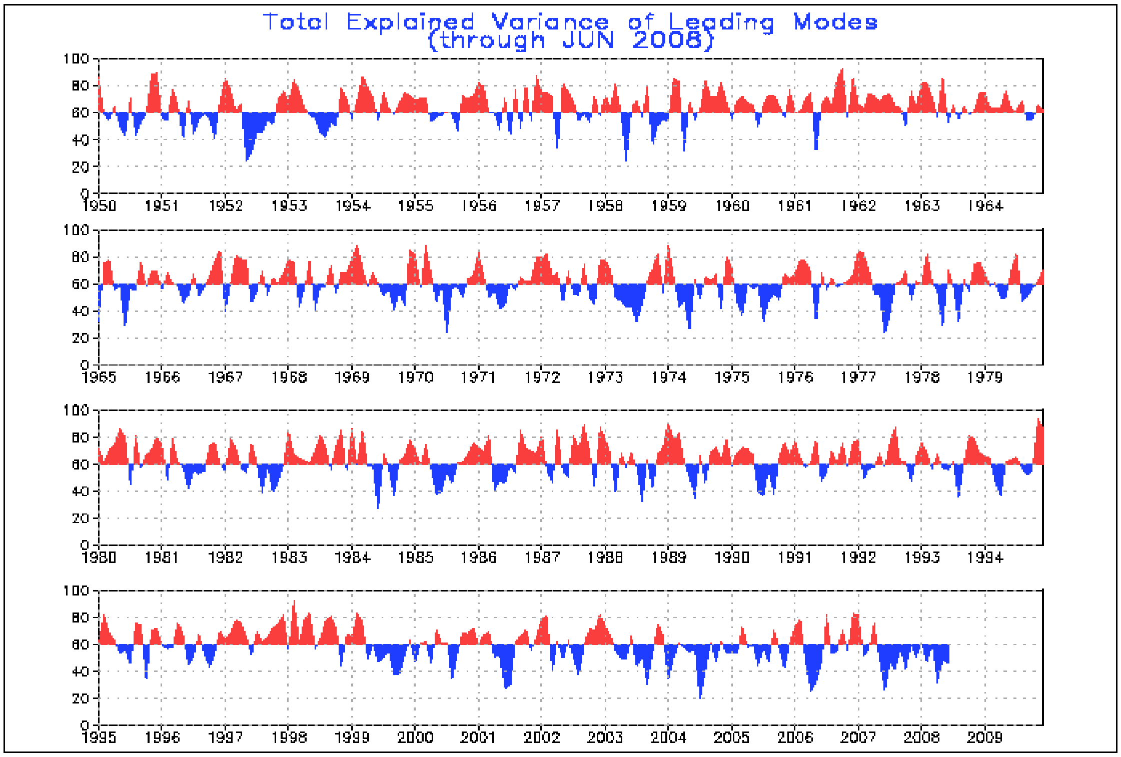


Figure 3: Variance in 500mb Height Data Explained by 10 CPC Modes, by Year

It is simple to show that this accelerating failure of the current state-of-the-art is in part attributable to paradoxical confounding. We obtained January 500 mb geopotential height data from 1948-2007 from the NCEP/NCAR Reanalysis dataset, for the full 379-point grid used in research cited earlier, separating the data into stationary (1948-1976) versus nonstationary (1977-2007) series (http://www.cdc.noaa.gov/cgi-bin/Timeseries/timeseries1.pl). We replicated prior varimax-rotated, ten-extracted-factor PCA of 500 mb height data (see Table 3). The principal component column indicates successive eigenvector (mode). For Sample, S is the stationary series, NS the non-stationary series, and C the combined S and NS data. Eigenvalue is given for each sample and mode, as is corresponding percent of total variance explained by the mode. For example, the first mode for the stationary series had an eigenvalue of 68.1, thus explaining 18.0% of the total variance of 379 measurements of 500 mb heights. Indicated using **red**, paradoxical con-founding exists when the eigenvalue for the C sample falls outside of the domain defined by the S and NS samples. Note that 80% of the modes clearly reveal paradoxical confounding: in every case except mode number 2 the effect was underestimation of explained variation.

Table 3: Replication of Prior Analysis of January 500 mb Geopotential Height Data, Separately by Series

Principal Component	Sample	Eigenvalue	Percent of Variance	Cumulative Percent Variance
1	S	68.1	18.0	18.0
	NS	75.3	19.9	19.9
	C	63.3	16.7	16.7
2	S	58.0	15.3	33.3
	NS	50.2	13.3	33.1
	C	60.0	15.8	32.5
3	S	42.0	11.1	44.4
	NS	39.1	10.3	43.4
	C	32.4	8.6	41.1
4	S	37.4	9.9	54.2
	NS	34.2	9.0	52.5
	C	29.5	7.8	48.9
5	S	24.8	6.5	60.8
	NS	27.3	7.2	59.7
	C	27.0	7.1	56.0
6	S	23.9	6.3	67.1
	NS	22.7	6.0	65.7
	C	21.0	5.5	61.5
7	S	18.6	4.9	72.0
	NS	19.6	5.2	70.8
	C	18.1	4.8	66.3

8	S	16.1	4.2	76.2
	NS	15.4	4.1	74.9
	C	13.4	3.5	69.8
9	S	13.7	3.6	79.8
	NS	15.3	4.0	78.9
	C	12.5	3.3	73.1
10	S	13.2	3.5	83.3
	NS	11.0	2.9	81.8
	C	11.4	3.0	76.2

Table 3 also provides the *cumulative* percent of total variance (of 379 variables) explained by the modes for each sample, across successive modes. Indicated using **blue**, paradoxical confounding exists when the cumulative value of this performance index for the C sample falls outside of the domain defined by the S and NS samples. *All* factors clearly reveal paradoxical confounding, and the effect was always underestimation of explained variation.

In addition to examining omnibus performance results of the current ten-mode solution, it is instructive to examine internal measurement properties of the individual modes. If the structure underlying the modes (reflected by the relationship of the 379 measurements of 500 mb heights to the mode score) is parallel, then the mode scores for the S, NS and C samples will be internally consistent (i.e., measure the same underlying construct), and a one-factor PCA of the three mode scores should explain most of the variation (theoretical maximum=100%), coefficient Alpha (positively related to the mean item-total correlation and the number of measures in the index) for the resulting factor score should be high (theoretical maximum=1.0), and the root-mean-squared-residual, or RMSR (an index of the average error in estimating the actual inter-measure correlation based on the mode structure) of the resulting factor score should be low (theoretical minimum=0). Seen below, the ten confounded current modes have poor internal measurement properties even by social science standards—for example, for personality surveys with modes measured using a fraction as many measures.[3]

Table 4: Internal Measurement Properties of Ten CPC Modes

Principal Component	Eigenvalue	Percent of Variance	Alpha	RMSR
1	1.89	63.3	0.710	0.2772
2	1.82	60.5	0.674	0.2913
3	2.22	74.1	0.825	0.1749
4	1.71	57.1	0.625	0.2744
5	1.54	51.4	0.527	0.2771

6	1.42	47.2	0.440	0.1812
7	1.45	48.5	0.469	0.3011
8	1.96	65.2	0.734	0.1805
9	1.63	54.2	0.577	0.2293
10	1.56	52.0	0.539	0.2404

Empirical results clearly demonstrate that current state-of-the-art models of modes of northern hemisphere upper-air variability are confounded by Simpson's paradox, underestimate model performance and phenomenon effect strength, and produce modes having poor measurement properties. Because data for only one month were used in this demonstration, these analyses represent a "best case scenario." Prior research first smoothed data over successive three month periods prior to conducting PCA: because the reliability of a composite exceeds the reliability of the constituents, smoothed scores will result in lower volatility (i.e., less extreme outliers) and weaker inter-measure correlations, eigenvalues, and measurement properties.

Theoretical consideration of current state-of-the-art models of modes also is not compelling. First, current modes are *non-granular*: postulating that a total of only ten modes underlie northern hemisphere upper-air variability is relatively simplistic compared with complexity underlying many large natural systems. Second, current modes are *nonparsimonious*, because computing an omnibus mode score requires (in the scoring formula) the use of all geopotential height measures. Third, low parsimony makes current mode scores robust: because many constituents (grid locations) are included in the scoring formula, positive changes in some constituents are offset by negative changes in others, so mode scores are *insensitive*. Finally, by formulation PCA is designed to produce *linear* models (modes), yet the present results failed to reveal strong linear modes as indicated by modest eigenvalues: there is therefore discordance between methodology (PCA), data (paradoxically confounded), method (how PCA was conducted), and objective (identifying psychometrically sound measures of major modes of northern hemisphere upper-air variability).

Unconfounded Measurement of Major Modes

Theoretical and empirical limitations of the original solution motivated development of a new methodology for identifying superior modes, which eliminates problems discussed earlier. Our proprietary method constitutes a theoretical shift in the way teleconnections are conceptualized, and a search algorithm. The theoretical shift necessitates an *ipsative* standardization of geopotential height data prior to conducting PCA.[4] The application of our algorithm involved searching for homogeneous spatial areas within which geopotential height measurements are highly related. Constraints included that independent application of PCA to the S, NS and C samples yields comparable, excellent macro performance (strong eigen-values) and internal measurement properties across samples, and that mode constituents are physically contiguous. Manually applied to January data the algorithm yielded 46 new modes summarized below (labels are nominal placeholders), ordered by percent of variance explained (i.e., decreasing linearity) for the stationary sample. For Sample, S=stationary, NS=nonstationary, and C=combined S and NS data. M is the number of geopotential height measures (grid locations) constituting the mode. Eigen indicates the eigenvalue of the mode for a one-factor PCA solution, and Var is the associated variance explained (100%xEigen/M). The theoretical upper-bound for internal consistency is Alpha=1, and the theoretical lower-bound for root-mean-square-error is RMSR=0. Finally, cumulative total eigenvalue, number of height measures, and total variance explained are also provided across successive modes.

Table 5: Principal Components Analysis of Unconfounded January 500 mb Geopotential Height Data, Separately by Series

							Cumulative Totals		
Mode	Sample	M	Eigen	Var	Alpha	RMSR	Eigen	M	Var
J	S	3	2.866	95.5	.977	.0331	2.866	3	95.5
	NS		2.831	94.4	.970	.0386	2.831		94.4
	C		2.844	94.8	.973	.0364	2.844		94.8
H	S	3	2.840	94.7	.972	.0412	5.706	6	95.1
	NS		2.819	94.0	.968	.0471	5.650		94.2
	C		2.827	94.2	.969	.0445	5.671		94.5
PP	S	3	2.826	94.2	.969	.0360	8.532	9	94.8
	NS		2.641	88.0	.932	.0685	8.291		92.1
	C		2.761	92.0	.957	.0476	8.432		93.7
MM	S	3	2.803	93.4	.965	.0337	11.335	12	94.5
	NS		2.743	91.4	.953	.0433	11.034		92.0
	C		2.773	92.4	.959	.0380	11.205		93.4
P	S	4	3.731	93.3	.976	.0404	15.066	16	94.2
	NS		3.575	89.4	.960	.0608	14.609		91.3
	C		3.651	91.3	.968	.0499	14.856		92.8
L	S	3	2.795	93.2	.963	.0558	17.861	19	94.0
	NS		2.729	91.0	.950	.0735	17.338		91.3
	C		2.790	93.0	.962	.0568	17.646		92.9
NN	S	3	2.793	93.1	.963	.0406	20.654	22	93.9
	NS		2.676	89.2	.939	.0562	20.014		91.0
	C		2.748	91.6	.954	.0464	20.394		92.7
M	S	4	3.724	93.1	.975	.0416	24.378	26	93.8
	NS		3.551	88.8	.958	.0603	23.565		90.6
	C		3.604	90.1	.963	.0575	23.998		92.3
Q	S	3	2.789	93.0	.962	.0541	27.167	29	93.7
	NS		2.613	87.1	.926	.0992	26.178		90.3
	C		2.707	90.2	.946	.0750	26.705		92.1
YY	S	3	2.788	92.9	.962	.0411	29.955	32	93.6
	NS		2.663	88.8	.937	.0566	28.841		90.1
	C		2.729	91.0	.950	.0474	29.434		92.0

I	S	3	2.785	92.8	.961	.0511	32.740	35	93.5
	NS		2.725	90.8	.950	.0720	31.566		90.2
	C		2.755	91.8	.955	.0612	32.189		92.0
CC	S	3	2.775	92.5	.960	.0492	35.515	38	93.5
	NS		2.653	88.4	.935	.0677	34.219		90.1
	C		2.717	90.6	.948	.0577	34.906		91.9
G	S	3	2.773	92.5	.959	.0586	38.288	41	93.4
	NS		2.802	93.4	.965	.0540	37.021		90.3
	C		2.788	92.9	.962	.0563	37.694		91.9
K	S	3	2.773	92.4	.959	.0561	41.061	44	93.3
	NS		2.672	89.1	.939	.0875	39.693		90.2
	C		2.703	90.1	.945	.0764	40.397		91.8
JJ	S	6	5.544	92.4	.984	.0348	46.605	50	93.2
	NS		5.236	87.3	.971	.0685	44.929		90.0
	C		5.360	89.3	.976	.0568	45.757		91.5
WW	S	3	2.770	92.3	.959	.0547	49.375	53	93.2
	NS		2.675	89.2	.939	.0581	47.604		89.8
	C		2.722	90.7	.949	.0483	48.479		91.5
R	S	3	2.769	92.3	.958	.0617	52.144	56	93.1
	NS		2.869	95.6	.977	.0358	50.473		90.1
	C		2.843	94.8	.972	.0422	51.322		91.6
O	S	3	2.764	92.1	.957	.0646	54.908	59	93.1
	NS		2.864	95.5	.976	.0373	53.337		90.4
	C		2.828	94.2	.970	.0468	54.150		91.8
XX	S	3	2.763	92.1	.957	.0453	57.671	62	93.0
	NS		2.730	91.0	.951	.0498	56.067		90.4
	C		2.744	91.5	.953	.0474	56.894		91.8
T	S	3	2.756	91.9	.956	.0613	60.427	65	93.0
	NS		2.694	89.8	.943	.0801	58.761		90.4
	C		2.715	90.5	.948	.0731	59.609		91.7
F	S	5	4.585	91.7	.977	.0437	65.012	70	92.9
	NS		4.393	87.9	.965	.0742	63.154		90.2
	C		4.471	89.4	.970	.0612	64.080		91.5
EE	S	3	2.749	91.6	.954	.0426	67.761	73	92.8
	NS		2.529	84.3	.907	.0898	65.683		90.0
	C		2.658	88.6	.936	.0608	66.738		91.4
2	S	3	2.743	91.4	.953	.0609	70.504	76	92.8
	NS		2.599	86.6	.923	.0844	68.282		89.8
	C		2.627	87.6	.929	.0824	69.365		91.3
B	S	6	5.472	91.2	.981	.0535	75.976	82	92.7
	NS		5.352	89.2	.976	.0773	73.634		90.0
	C		5.399	90.0	.978	.0654	74.764		91.2

ZZ	S	3	2.727	90.9	.950	.0464	78.703	85	92.6
	NS		2.787	92.9	.962	.0422	76.421		89.9
	C		2.738	91.3	.952	.0450	77.502		91.2
E	S	4	3.634	90.8	.966	.0511	82.337	89	92.5
	NS		3.526	88.2	.955	.0697	79.947		89.8
	C		3.567	89.2	.960	.0606	81.069		91.1
RR	S	3	2.723	90.8	.949	.0555	85.060	92	92.5
	NS		2.611	87.0	.925	.0813	82.558		89.7
	C		2.658	88.6	.936	.0694	83.727		91.0
D	S	3	2.721	90.7	.949	.0782	87.781	95	92.4
	NS		2.807	93.6	.966	.0521	85.365		89.9
	C		2.724	90.8	.949	.0758	86.451		91.0
C	S	4	3.605	90.1	.964	.0566	91.386	99	92.3
	NS		3.667	91.7	.970	.0500	89.032		89.9
	C		3.637	90.9	.967	.0537	90.088		91.0
U	S	3	2.703	90.1	.945	.0648	94.089	102	92.2
	NS		2.746	91.5	.954	.0624	91.778		90.0
	C		2.727	90.9	.950	.0631	92.815		91.0
LL	S	3	2.695	89.8	.943	.0603	96.784	105	92.2
	NS		2.599	86.6	.923	.0832	94.377		90.0
	C		2.680	89.3	.940	.0646	95.495		90.9
TT	S	3	2.687	89.6	.942	.0565	99.471	108	92.1
	NS		2.840	94.7	.972	.0271	97.217		90.0
	C		2.780	93.3	.964	.0345	98.275		91.0
V	S	3	2.687	89.6	.942	.0845	102.158	111	92.0
	NS		2.659	88.6	.936	.0922	99.876		90.0
	C		2.662	88.7	.937	.0914	100.937		90.9
HH	S	3	2.683	89.4	.941	.0567	104.841	114	92.0
	NS		2.567	85.6	.916	.0994	102.443		89.9
	C		2.615	87.2	.926	.0797	103.552		90.8
UU	S	3	2.681	89.4	.941	.0536	107.522	117	91.9
	NS		2.638	87.9	.931	.0757	105.081		89.8
	C		2.667	88.9	.938	.0623	106.219		90.8
GG	S	3	2.675	89.2	.939	.0627	110.197	120	91.8
	NS		2.723	90.8	.949	.0540	107.804		89.8
	C		2.714	90.5	.947	.0525	108.933		90.8
1	S	3	2.673	89.1	.939	.0603	112.870	123	91.8
	NS		2.771	92.4	.959	.0438	110.575		89.9
	C		2.747	91.6	.954	.0473	111.680		90.8
II	S	3	2.672	89.1	.939	.0578	115.542	126	91.7
	NS		2.745	91.5	.954	.0427	113.320		89.9
	C		2.706	90.2	.946	.0502	114.386		90.8

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

DD	S	4	3.562	89.1	.959	.0616	119.104	130	91.6
	NS		3.540	88.5	.957	.0588	116.860		89.9
	C		3.547	88.7	.957	.0595	117.933		90.7
VV	S	3	2.656	88.5	.935	.0715	121.760	133	91.5
	NS		2.728	90.9	.950	.0479	119.588		89.9
	C		2.717	90.6	.948	.0547	120.650		90.7
Y	S	3	2.652	88.4	.934	.0941	124.412	136	91.5
	NS		2.791	93.0	.962	.0555	122.379		90.0
	C		2.680	89.3	.940	.0864	123.330		90.7
3	S	4	3.530	88.3	.956	.0733	127.942	140	91.4
	NS		3.623	90.6	.965	.0539	126.002		90.0
	C		3.559	89.0	.959	.0671	126.889		90.6
FF	S	3	2.646	88.2	.933	.0987	130.588	143	91.3
	NS		2.701	90.0	.945	.0835	128.703		90.0
	C		2.649	88.3	.934	.0972	129.538		90.6
A	S	5	4.357	87.1	.963	.0777	134.945	148	91.2
	NS		4.538	90.8	.975	.0706	133.241		90.0
	C		4.414	88.3	.967	.0770	133.952		90.5
SS	S	3	2.603	86.8	.924	.0778	137.548	151	91.1
	NS		2.755	91.8	.955	.0471	135.996		90.1
	C		2.652	88.4	.934	.0676	136.604		90.5
BB	S	4	3.473	86.8	.949	.0656	141.021	155	91.0
	NS		3.612	90.3	.964	.0566	139.608		90.1
	C		3.514	87.8	.954	.0645	140.118		90.4

There is no evidence of paradoxical confounding (performance results for C always fall between results for S and NS), and the percentage of variance explained, Alpha, and RMSR meet psychometric criteria for “good to excellent” fit for exploratory PCA models.[1,3] We also examined internal measurement proper-ties of the individual modes via one-factor PCA of the three sample scores (S, NS, C), and analysis revealed virtually perfect measurement: for every mode, percent of total variance (of M measures) explained > 99.9%; Alpha > 0.99, and RMSR ≤ 0.0002. We attempted to model the original ten modes using the new 46 modes, and vice versa, using multiple regression analysis, but no satisfactory models were identified: the original ten modes and the new 46 modes are *not* related to each other.

Considered together these findings clearly show that the 46 new and unique modes eliminate every *empirical* problem identified for the original ten modes: there is no evidence of Simpson’s paradox (S and NS data may be combined without inducing confounding); model performance and phenomenon effect strength are not erroneously misestimated (estimates from all samples are convergent); and mode scores exhibit ideal measurement properties. The new modes also address all *theoretical* concerns identified for the original ten modes: granularity increased 4.6-fold; the new modes are parsimonious (factor weighting coefficients are all approximately one in absolute magnitude, each grid location appears on only one mode); mode scores are sensitive (composed of six or fewer strongly related grid locations, small

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

changes in geopotential heights are easily detectable); and the modes are extremely well modeled by PCA, representing a set of nearly perfectly linear measures.

Qualitative Interpretation of Ipsative Modes

Figure 4 locates the ipsative modes on a polar projection map of the northern hemisphere.

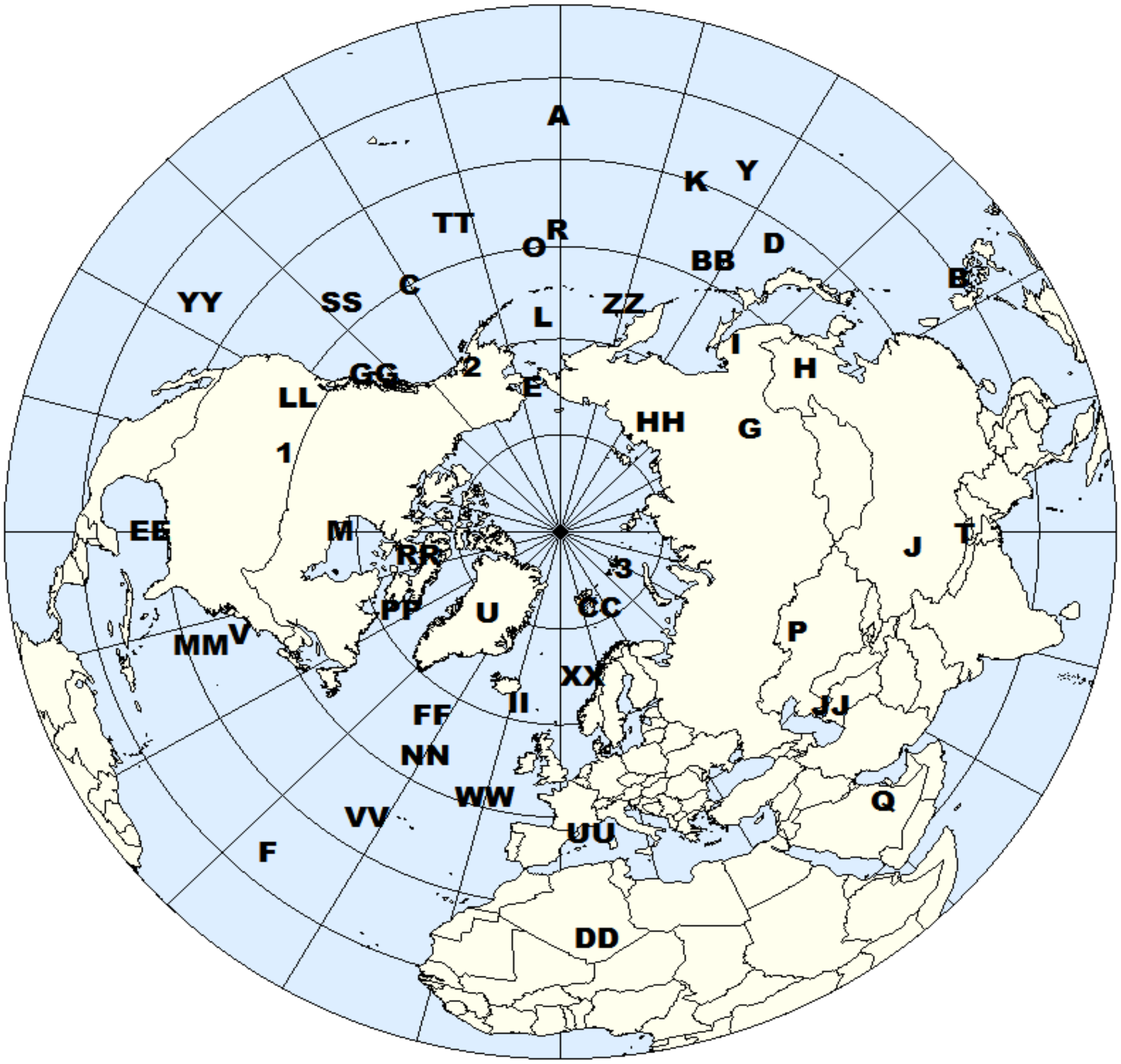


Figure 4: Polar projection Map of the Ipsative Modes

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

The principal-component-derived CPC modes of upper-air variability listed in Table 2 are highly consistent with the modes identified in the original principal components analysis[2] of 700 mb height data, and have counterparts in the ipsative modes developed presently.

The first mode, North Atlantic Oscillation (NAO), had strong positive coefficients for grid points over Greenland, corresponding to ipsative mode U. NAO also had strong negative coefficients for grid points in the North Atlantic, west of the Azores (ipsative mode VV); Manchuria (ipsative mode H); and the central plains of the US (between ipsative factors EE and 1).

The second mode, East Atlantic Pattern (EA), had strong positive coefficients for grid points over North Africa (ipsative mode DD), and in the Atlantic east of Cuba (ipsative mode F). EA also had strong negative coefficients for grid points in the North Atlantic, east of Labrador and south of Greenland (ipsative mode FF).

The West Pacific Pattern (WP) had strong positive coefficients for grid points in the Philippine Sea (ipsative mode D), and strong negative coefficients for grid points just east of Kamchatka (ipsative mode ZZ).

The East Pacific/North Pacific Pattern (EP/NP) had strong positive coefficients for grid points over southeast Alaska (between ipsative modes GG and 2). EP/NP also had strong negative coefficients for grid points in the North Pacific south of the Aleutian Islands (ipsative mode TT), and near James Bay in Canada (ipsative mode M).

The Pacific/North American Pattern (PNA) had strong positive coefficients for grid points west of Hawaii (ipsative mode A), and in the Pacific Northwest of the US (ipsative mode LL). PNA also had strong negative coefficients for grid points in the North Pacific southwest of the Aleutian Islands (ipsative mode O), and over the southeast US (ipsative mode EE).

The East Atlantic/West Russia Pattern (EA/WR) had strong positive coefficients for grid points near England (between ipsative factors II and UU), and in Siberia north of Manchuria (ipsative mode G). EA/WR also had strong negative coefficients for grid points northeast of the Caspian Sea (ipsative mode JJ).

The Scandinavian Pattern (SCA) had strong positive coefficients for grid points in Central Russia (between ipsative modes G and P), and in the North Atlantic, northwest of Spain (ipsative mode WW). SCA also had strong negative coefficients for grid points near Fin-land (between ipsative modes XX and JJ).

The Tropical/Northern Hemisphere Pat-tern (TNH) had strong positive coefficients for grid points in the North Pacific west of the Pacific Northwest of the US (ipsative mode SS), and near the Bahamas (ipsative mode MM). TNH also had strong negative coefficients for grid points near James Bay in Canada (ipsative mode M).

The Polar/Eurasia Pattern (POL) had strong positive coefficients for grid points in eastern Mongolia (near ipsative modes G and H), and strong negative coefficients for grid points in the Arctic Ocean north of eastern Siberia (ipsative mode HH).

Finally, the Pacific Transition Pattern (PT)—which did not materialize in either of the original principal component analyses for the month of January, had for the month of September strong positive coefficients for grid points over the northern plains of the US (ipsative mode 1), and west of Hawaii (ipsative mode A). PT also had strong negative coefficients for grid points in the North Pacific south of Alaska (ipsative mode C), and over the eastern US (ipsative mode V).

Predicting Temperature Anomalies

To determine whether predictive validity is augmented by nonconfounded measurement, we assessed whether statistical models that use the 46 newly discovered (*vs*. original ten) modes of northern hemisphere upper-air variability produce more accurate temperature forecasting. We used classification tree analysis, or CTA[5], to predict whether mean temperature in January, February, and March fell above or below the median temperature for the years 1950-2007, for 48 contiguous US states. Falling within the optimal data analysis paradigm, CTA explicitly maximizes model accuracy when applied to a given sample or series.[6] Proprietary software was used to automatically identify CTA models that weighted more heavily observations having greater deviations from the median temperature: of course, depending on the application, “natural weights” such as inches of rain, may be used instead of, or in conjunction with, “tailored weights” such as we used.[6] The weighted CTA algorithm was performed using three sets of attributes: *ipsative modes* (46 modes discovered presently); *published normative modes* obtained from the CPC, with PT omitted due to inactivity in January; and *computed normative modes* obtained from our replication of the CPC analysis using only January data.

The findings of these analyses are summarized in Table 6. Tabled are modes (see Table 5 for coding) emerging with p<0.05 in the weighted CTA model. The weights were determined by sorting the observations by monthly mean temperature, and adding 1.5 for every position above or below the median. WESS is a standardized measure of weighted effect strength, on which 0 is the level of weighted predictive accuracy that is expected by chance, and 100 represents errorless (perfect) weighted predictive accuracy.[6] A dash (-) indicates no solution was identified having p<0.05 for any mode; a missing row indicates no solution was identified for any data type (ipsative, published, or computed); and an asterisk (*) indicates that results for the indicated modes were identical to findings for the ipsative modes.

Models derived using ipsative modes to predict temperature anomalies in the United States convincingly and broadly outperformed corresponding models derived with normative modes, when considered from the perspective of predictive accuracy, and quantified using the standardized WESS metric:

- For a given state and month (corresponding to individual rows in Table 6), the ipsative mode model yielded the greatest WESS 117 times (91.4%), versus 5 and 6 (3.9% and 4.7%) times for published and computed normative mode models, respectively.

- In January the ipsative mode models always achieved greater WESS than the corresponding normative mode models. In February the ipsative mode models almost always (93.2% of the time) achieved greatest WESS (44 states had models based on February data), and even as the data aged substantially—for March, ipsative models usually (78.1% of the time) achieved greatest WESS (32 states had models using March data).

- For January data, using ipsative modes, all 48 states had CTA models with WESS≥90%, versus two states with CTA models involving published normative modes, and one state CTA model involving computed normative modes. For February data, using ipsative modes, a dozen states had CTA models with WESS≥90% (and three for March data), versus none using normative modes.

Table 6: Temperature Prediction via Weighted CTA by US State, for January, February, and March of 2008, Using Ipsative mode Scores, and Published and Computed Raw Mode Scores

				Published		Computed	
State	Month	Ipsative Modes	WESS	Normative Modes	WESS	Normative Modes	WESS
Alabama	Jan	B,EE,JJ,MM,2	97.43	EAWR,NAO,PNA	71.30	2,3,9	68.44
	Feb	A,C,I,EE,PP	93.80	NAO,SCA	57.74	3,6	62.59
	Mar	DD,GG	51.55	-	-	-	-
Arkansas	Jan	C,R,EE,MM,XX,2	98.54	EPNP,PNA,WP	74.63	3,5,8	80.36
	Feb	CC,DD,RR,VV	88.90	EPNP,NAO	63.35	3,5,10	79.31
	Mar	II	38.63	-	-	-	-
Arizona	Jan	C,H,U,YY,1	93.22	NAO,POL,WP	75.80	2,6	84.40
	Feb	F,II,PP	72.65	-	-	-	-
California	Jan	C,BB,GG,VV,WW,YY	98.89	PNA,WP	52.83	2,6	77.79
	Feb	RR,TT	74.87	EAWR,EPNP,PNA	76.04	-	-
Colorado	Jan	I,V,T,SS,WW	95.62	-	-	2,6	79.60
	Feb	M,O,P,Q,BB,3	91.70	-	-	-	-
	Mar	J,SS,1	72.76	NAO	39.74	1,5	57.69
Connecticut	Jan	E,K,LL,2	96.43	EA,EAWR,EPNP,NAO,WP	86.62	3,4,5	74.52
	Feb	PP,2	50.44	-	-	-	-
Delaware	Jan	V,EE,MM,2	95.15	EAWR,EPNP,NAO,WP	84.63	3,5,7	71.71
	Feb	HH,JJ,PP,SS	73.41	NAO	42.84	3	44.89
	Mar	J	37.97	-	-	-	-

Florida	**Jan**	**A,G,O,MM,PP,YY**	**98.95**	**EAWR,EPNP,PNA**	**89.19**	**2,3,6**	**82.23**
	Feb	**D,Q,CC,LL,RR**	**93.22**	**NAO**	**40.50**	**5**	**63.06**
	Mar	**K,DD,EE,GG**	**75.80**	**-**	**-**	**-**	**-**
Georgia	**Jan**	**P,EE,MM,PP,2**	**98.13**	**EAWR,EPNP,PNA**	**84.04**	**2,3,9**	**70.66**
	Feb	**A,C,H,EE,PP**	**92.34**	**NAO,SCA**	**57.16**	**3,5**	**73.47**
Iowa	**Jan**	**H,L,V,2**	**93.51**	**EPNP,SCA,WP**	**76.74**	**3,4,7,8**	**84.57**
	Feb	**D,DD,JJ**	**80.95**	**EAWR**	**49.09**	**3,7**	**44.18**
	Mar	**J,HH,LL,PP,1**	**87.26**	**PNA**	**41.15**	**-**	**-**
Idaho	**Jan**	**C,I,MM,SS,ZZ**	**94.56**	**-**	**-**	**2,3,6**	**81.59**
	Feb	**D,Q,R,BB**	**86.91**	**PNA**	**60.78**	**-**	**-**
	Mar	**D,R,Y,RR**	**93.86**	**NAO,PNA,SCA**	**83.99**	**1,5**	**63.82**
Illinois	**Jan**	**B,D,E,V,EE,WW,2**	**99.36**	**EPNP,PNA,WP**	**83.52**	**3,4,8**	**86.62**
	Feb	**D,DD,GG,PP**	**83.40**	**EAWR,NAO,SCA**	**66.04**	**-**	**-**
	Mar	**-**	**-**	**PNA**	**39.86**	**-**	**-**
Indiana	**Jan**	**D,E,K,V,EE,WW**	**96.61**	**EPNP,PNA,WP**	**82.70**	**3,5,8**	**82.35**
	Feb	**K,U,NN,RR**	**71.01**	**EAWR,NAO,POL**	**73.58**	**3**	**40.44**
	Mar	**L,II**	**57.04**	**PNA**	**39.39**	**1,10**	**57.22**
Kansas	**Jan**	**F,Q,GG,WW,1**	**96.73**	**EPNP,WP**	**59.44**	**1,3,6,9**	**69.43**
	Feb	**V,CC,FF,UU**	**80.19**	**EAWR,NAO**	**60.55**	**3,6,7,9**	**82.82**
	Mar	**D,H,FF**	**73.00**	**-**	**-**	**-**	**-**

Kentucky	Jan	E,J,V,PP,2	96.20	EAWR,EPNP,NAO	79.37	3,5	73.82
	Feb	F,I,Q,U,RR	96.55	NAO	53.36	3,6	60.14
Louisiana	Jan	U,V,EE,LL,3	96.20	NAO,PNA	69.37	1,2,6	84.22
	Feb	A,C,EE,PP	79.37	NAO	53.71	3,5,6,10	79.19
	Mar	D,DD	52.95	-	-	-	-
Massachusetts	Jan	E,I,K,LL,2	97.72	EA,EAWR,EPNP,NAO,WP	90.06	3,4,5	73.70
	Mar	-	-	-	-	2	38.92
Maryland	Jan	E,G,L,V,RR,UU	98.54	EAWR,EPNP,WP	84.28	3,5,8	71.30
	Feb	Y,RR,XX	69.96	NAO,POL	55.00	3	46.41
Maine	Jan	E,O,LL,2	95.21	EPNP,WP	61.60	3,8	65.81
	Feb	Q,RR,1	76.04	-	-	7	39.63
	Mar	Q	39.10	-	-	-	-
Michigan	Jan	D,E,GG,II	97.37	EAWR,EPNP,WP	81.71	3,5,7,8	86.56
	Feb	I,DD,GG,HH	82.76	EAWR,NAO	53.65	3,7	51.43
	Mar	J,L	57.51	PNA,SCA	59.73	2	44.18
Minnesota	Jan	C,E,CC,1,2	95.73	EAWR,EPNP,PNA,WP	88.49	4,5,8	79.78
	Feb	F,Q,NN,RR	78.08	EAWR	44.71	7,10	61.19
	Mar	J,O,1	82.70	PNA,WP	56.81	2	40.68
Missouri	Jan	D,E,F,EE,GG	94.92	EPNP,PNA,WP	85.74	3,4,7,8	93.98
	Feb	EE,RR,SS,TT,VV	93.44	EAWR,EPNP,NAO,POL	77.93	3,5,7	76.52

2155-0182/10/$3.00p

Mississippi	**Jan**	I,V,EE,2	96.20	EPNP,NAO,PNA	86.91	1,2,6	78.73
	Feb	A,C,EE,PP	79.54	NAO	52.78	3,6,10	71.95
	Mar	DD,GG	51.32	-	-	-	-
Montana	**Jan**	E,F,L,ZZ,2	96.67	EPNP,PNA,SCA,WP	84.04	2,6,9	75.45
	Feb	A,G,Q,R	85.62	PNA	47.05	7	49.80
	Mar	CC,GG,TT,3	80.60	PNA	45.35	1	39.22
North Carolina	**Jan**	E,Y,MM,XX	95.38	EAWR,EPNP,PNA	86.15	3,5	71.60
	Feb	D,T,Y,RR,VV	89.83	NAO,SCA	54.94	3,9	56.52
North Dakota	**Jan**	C,E,L,WW	96.90	EPNP,PNA,SCA,WP	91.41	1,3,5,7	80.89
	Feb	D,Q,II,RR	94.21	EAWR,PNA	61.84	7	45.35
	Mar	J,GG,1	77.91	PNA	43.83	-	-
Nebraska	**Jan**	A,V,DD,1,2	95.56	EPNP,WP	57.10	1,3,9	70.19
	Feb	Q,DD,RR,TT	86.44	EAWR	43.83	-	-
	Mar	D,LL	74.81	-	-	-	-
New Hampshire	**Jan**	E,K,JJ,LL,2	97.49	EA,EPNP,WP	71.89	3,5,7	70.72
	Feb	-	-	-	-	7	39.28
	Mar	-	-	-	-	2	40.56
New Jersey	**Jan**	E,K,H,LL	98.48	EA,EAWR,EPNP,NAO,WP	87.38	3,4,5	76.74
	Feb	Y,RR,1	70.72	-	-	3	40.68

New Mexico	**Jan**	**G,T,RR,UU,ZZ**	**97.84**	**EA,NAO**	**64.64**	**1,6**	**84.16**
	Feb	**F,G,RR,VV,1**	**88.43**	**NAO**	**43.25**	**6**	**42.84**
	Mar	**G,Y,3**	**73.52**	**-**	**-**	**-**	**-**
Nevada	**Jan**	**C,I,V,SS,ZZ**	**96.43**	**-**	**-**	**2,3,6**	**86.62**
	Feb	**RR,TT,WW**	**76.62**	**EA,PNA**	**60.43**	**-**	**-**
	Mar	**1**	**38.81**	**NAO**	**41.44**	**-**	**-**
New York	**Jan**	**II,MM,XX,2**	**97.02**	**EA,EAWR,EPNP,NAO,WP**	**89.42**	**3,4,5**	**77.79**
	Mar	**L**	**38.98**	**-**	**-**	**-**	**-**
Ohio	**Jan**	**E,L,V,RR**	**96.67**	**EAWR,EPNP,WP**	**80.65**	**3,5,8**	**79.43**
	Feb	**D,GG,HH,PP**	**81.71**	**NAO,POL**	**59.03**	**3**	**39.98**
	Mar	**L,II**	**56.22**	**-**	**-**	**1,10**	**55.93**
Oklahoma	**Jan**	**F,K,Q,DD,E,2**	**96.90**	**EA,EPNP**	**59.15**	**8**	**63.35**
	Feb	**H,EE,RR,TT,VV**	**85.86**	**EPNP,NAO**	**67.15**	**3,6,7**	**74.17**
	Mar	**D,J**	**49.09**	**-**	**-**	**-**	**-**
Oregon	**Jan**	**C,I,EE,MM,PP**	**91.88**	**NAO,PNA,WP**	**83.99**	**2,3,5**	**81.18**
	Feb	**Q,R,NN,3**	**86.15**	**PNA**	**61.72**	**1,3,7**	**63.35**
	Mar	**F,R,V,SS,2**	**82.58**	**NAO,PNA,POL**	**69.08**	**-**	**-**
Pennsylvania	**Jan**	**E,J,HH,YY**	**96.96**	**EAWR,EPNP,NAO,WP**	**85.80**	**3,5,8**	**72.36**
	Feb	**Q,RR**	**58.45**	**NAO**	**43.42**	**3,7**	**56.81**
	Mar	**L**	**39.80**	**-**	**-**	**-**	**-**

Rhode Island	Jan	E,K,LL,2	96.84	EA,EAWR,EPNP,NAO,WP	86.50	3,4,5	75.34
	Feb	G,K,2	73.52	-	-	-	-
	Mar	J,Q,CC,EE,XX	71.54	-	-	-	-
South Carolina	Jan	Q,R,MM,RR	96.73	EAWR,EPNP,PNA	85.91	2,3,9	70.89
	Feb	D,Q,JJ,RR	90.01	NAO,SCA	55.29	3,6	62.01
South Dakota	Jan	C,E,L,2	97.25	EPNP,SCA,WP	88.02	5,8	62.30
	Feb	D,Q,II,RR	92.69	EAWR	47.69	7	42.20
	Mar	D,J,DD,1	87.67	-	-	-	-
Tennessee	Jan	I,Q,V,EE,3	94.86	EAWR,EPNP,NAO,PNA	77.85	3,5	69.02
	Feb	D,T,U,RR,TT	87.38	NAO	53.13	3,6	56.75
Texas	Jan	C,EE,GG,NN,RR	92.17	NAO,PNA,POL	68.73	1,2,6	82.99
	Feb	A,M,JJ,RR,WW,3	94.62	NAO	51.96	3,5,10	74.34
	Mar	Y,FF,LL,PP	72.36	-	-	-	-
Utah	Jan	C,I,V,BB,SS,ZZ	96.32	-	-	1,2,6	84.34
	Feb	Q,CC,DD,NN	80.25	PNA	44.59	-	-
	Mar	1	41.61	NAO	43.13	1,5	58.62
Virginia	Jan	E,H,L,V,RR	97.37	EAWR,EPNP,PNA	85.68	3,5	72.06
	Feb	A,H,Y,RR,VV	92.87	NAO	49.50	3,5,9	56.98
Vermont	Jan	E,CC,JJ,LL,2	99.12	EA,EPNP,NAO,WP	73.41	3,5,7	71.30
	Mar	Q	42.72	-	-	-	-

Washington	**Jan**	**L,O,CC,EE,VV**	97.78	**EA,NAO,PNA,WP**	91.06	2,5,6	76.68
	Feb	**M,R,EE,WW**	88.37	**PNA**	67.45	1,7	58.56
	Mar	**D,H,PP,TT,XX,2**	92.93	**PNA**	57.39	-	-
Wisconsin	**Jan**	**E,M,GG,UU,ZZ**	97.84	**EAWR,EPNP,PNA**	79.31	3,5,8	75.04
	Feb	**Q,RR,ZZ,1**	74.87	**EAWR**	44.54	7	48.39
	Mar	**L,T,CC,GG,NN**	93.10	**PNA,SCA**	65.81	2	43.60
West Virginia	**Jan**	**E,H,V,EE,LL**	98.19	**EAWR,EPNP,PNA,SCA**	83.46	3,5	76.74
	Feb	**D,T,U,LL,RR,TT**	95.91	**NAO**	52.54	3	42.31
Wyoming	**Jan**	**K,DD,MM,YY,ZZ**	92.11	-	-	2,3,5	77.50
	Feb	**C,G,Q,DD**	84.57	-	-	-	-
	Mar	**D,F,LL,SS**	89.89	**NAO**	43.37	1	41.03

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

- We statistically contrasted the WESS of each pair of these three sets of factors. If no model was found, WESS was assumed to be zero. ODA was used to determine which set of modes was better at predicting whether or not the mean temperature of the states exceeded the median. The PTMP procedure[7] was used to estimate the exact Type I error of each contrast. Analyses indicated that ipsative mode models had significantly greater WESS than the published or computed normative mode models for all three months (*p*'s <0.0001), and that normative models could *never* reliably be discriminated from each other by WESS (*p*'s>0.17).

- As a test of cross-sample generalizability we also evaluated a larger field of northern hemisphere data. In the crutem3v dataset are 217 locations which have no missing data for January, February or March, for the years 1948-2007. As a test of cross-method generalizability, temperature predict-ions for each location and month were obtained using stepwise multiple regression analysis: the independent variables were the January data, and ipsative, published raw, or computed raw modes were used as dependent variables. The R^2 value for each model was determined: if no model was found, R^2 was assumed to be zero. Statistical comparison via the PTMP procedure showed that ipsative modes clearly outperformed the other modes (*p*'s<0.0001). Computed raw modes outperformed published raw modes in all cases: contrasts were statistically significant for January and February (*p*'s<0.0001), but not March (p<0.27).

Predicting Precipitation Anomalies

As a second investigation of predictive validity we assessed whether statistical models that use the ipsative modes produce more accurate precipitation forecasting. We used CTA to predict whether mean precipitation in January, February, and March fell above or below the median precipitation for the years 1950-2007, for 48 contiguous US states. As for temperature modeling, the weighted CTA algorithm was performed using three sets of attributes: the 46 newly discovered ipsative modes; published normative modes (obtained from the CPC, with PT omitted due to inactivity in January); and computed normative modes (obtained from our replication of CPC analysis using only January data). The findings of these analyses are summarized in Table 7. Tabled are modes (see Table 5 for coding) emerging with *p*<0.05 in the weighted CTA model. The weights were determined by the same method as was used in predicting temperature anomalies, but total monthly precipitation was used for the sort and median.

As when modeling temperature anomalies, models derived using ipsative modes to predict precipitation anomalies in the United States convincingly and broadly outperformed corresponding models derived by normative modes, when considered from the perspective of predictive accuracy:

- For a given state and month (corresponding to individual rows in Table 7), the ipsative mode model yielded the greatest WESS 126 times (92.6%), versus 5 (3.7%) times each for the published and computed normative mode models.

Table 7: Precipitation Prediction via Weighted CTA by US State, for January, February, and March of 2008, Using Ipsative mode Scores, and Published and Computed Raw Mode Scores

State	Month	Ipsative Modes	WESS	Published Normative Modes	WESS	Computed Normative Modes	WESS
Alabama	Jan	C,O,P,MM,NN	89.01	EA,SCA	64.47	8	39.86
	Feb	A,R,T,V,II	87.03	EA	39.45	-	-
	Mar	I,YY	59.56	-	-	-	-
Arkansas	Jan	C,R,FF,MM,YY	90.01	NAO,PNA	76.27	1,3,9	80.54
	Feb	Q	39.98	-	-	-	-
	Mar	HH	39.28	-	-	-	-
Arizona	Jan	G,LL,SS	73.47	EPNP	39.63	9	52.02
	Feb	I,J,L,1	87.14	EPNP,SCA	62.83	3,5	71.36
	Mar	G,Q,T,JJ,SS	84.51	PNA	38.11	5,7,9	57.10
California	Jan	BB,LL,NN,SS,2	94.62	EA	48.92	3,6,8	76.33
	Feb	V,SS,XX	68.79	-	-	-	-
	Mar	C,R,U,SS	84.57	NAO	44.07	-	-
Colorado	Jan	D,EE	59.44	PNA	52.48	-	-
	Feb	NN,XX	65.75	SCA	45.59	3,7	59.73
	Mar	II,SS,3	76.21	PNA	45.47	-	-
Connecticut	Jan	V,BB,XX	87.67	-	-	5	42.02
	Feb	P,HH	77.26	EAWR	43.54	10	38.92
	Mar	G,H,J	51.32	POL	44.07	-	-

Delaware	**Jan**	**B,RR**	**57.74**	**EAWR,NAO**	**51.49**	**2**	**41.44**
	Feb	**C,BB,EE**	**70.31**	**-**	**-**	**6**	**46.29**
	Mar	**CC,DD,EE,PP**	**90.77**	**NAO,WP**	**55.35**	**-**	**-**
Florida	**Jan**	**F,O,BB,CC,DD**	**92.11**	**EA**	**43.66**	**3**	**40.62**
	Feb	**T,EE,VV,2**	**94.62**	**-**	**-**	**-**	**-**
	Mar	**C,D,O,SS,TT**	**89.60**	**-**	**-**	**4,5**	**53.54**
Georgia	**Jan**	**O,MM,NN**	**73.76**	**EA**	**68.32**	**6,8**	**57.63**
	Feb	**C,J,T,SS,WW**	**91.88**	**-**	**-**	**3**	**42.02**
Iowa	**Jan**	**GG,NN**	**59.38**	**EAWR,PNA**	**60.61**	**1**	**49.68**
	Feb	**G,I,R,PP**	**77.97**	**-**	**-**	**-**	**-**
	Mar	**T,EE**	**56.52**	**-**	**-**	**-**	**-**
Idaho	**Jan**	**E,L,T,GG,WW,1**	**98.48**	**EPNP,PNA,SCA**	**75.75**	**1,6,8,9**	**86.50**
	Feb	**J,M,U,NN,XX**	**85.86**	**EA,POL**	**64.52**	**5,7**	**62.77**
	Mar	**I,U,HH,LL,3**	**89.54**	**EA,NAO,WP**	**80.77**	**2,4,5**	**84.80**
Illinois	**Jan**	**H,Q,R,MM,NN**	**92.99**	**PNA**	**50.32**	**9**	**46.05**
	Feb	**Q,U,BB,HH**	**81.06**	**-**	**-**	**7**	**39.51**
	Mar	**E,J,JJ,UU**	**87.90**	**-**	**-**	**-**	**-**
Indiana	**Jan**	**F,I,EE,HH,PP**	**91.23**	**NAO,PNA**	**72.59**	**9**	**40.56**
	Feb	**R,EE,LL,XX**	**83.46**	**-**	**-**	**4,7**	**66.45**
	Mar	**O,JJ,SS**	**74.05**	**-**	**-**	**-**	**-**

Kansas	**Jan**	**E,Y,GG,LL**	**84.72**	**-**	**-**	**3,6**	**55.91**
	Feb	**F,K,M,FF**	**78.08**	**-**	**-**	**-**	**-**
	Mar	**D,H,R,2**	**81.12**	**PNA**	**41.55**	**-**	**-**
Kentucky	**Jan**	**A,V,HH,PP**	**89.42**	**PNA,SCA**	**69.67**	**1,6**	**79.95**
	Feb	**Q,V,II,LL,TT**	**86.09**	**-**	**-**	**7**	**50.96**
	Mar	**G,NN,XX**	**75.39**	**-**	**-**	**2**	**53.83**
Louisiana	**Jan**	**H,DD,FF,WW**	**80.77**	**EA,EPNP**	**51.61**	**-**	**-**
	Feb	**C,P,T**	**71.24**	**-**	**-**	**2**	**40.09**
	Mar	**A,E,K,FF,WW**	**85.80**	**-**	**-**	**6,7**	**64.87**
Massachusetts	**Jan**	**-**	**-**	**-**	**-**	**2**	**39.45**
	Feb	**I,SS,WW,1**	**78.43**	**-**	**-**	**-**	**-**
	Mar	**C,G,HH**	**66.74**	**POL**	**50.85**	**-**	**-**
Maryland	**Jan**	**G,H,WW**	**69.73**	**-**	**-**	**-**	**-**
	Feb	**E,P,Q,YY**	**88.54**	**-**	**-**	**6**	**42.96**
	Mar	**I,HH,RR,VV**	**94.80**	**SCA,WP**	**53.19**	**-**	**-**
Maine	**Jan**	**HH,WW,YY,2**	**86.44**	**-**	**-**	**5**	**39.22**
	Feb	**J,NN,WW**	**70.25**	**-**	**-**	**1,5**	**65.40**
	Mar	**I,J,HH,SS,1**	**86.85**	**POL**	**43.78**	**-**	**-**
Michigan	**Jan**	**H,Q,T,GG,MM**	**86.97**	**PNA**	**50.38**	**1,6**	**53.95**
	Feb	**D,DD**	**68.26**	**-**	**-**	**-**	**-**

Minnesota	**Jan**	**P,FF,GG**	**77.62**	**-**	**-**	**-**	**-**
	Mar	**Q,YY,3**	**78.43**	**-**	**-**	**-**	**-**
Missouri	**Jan**	**O,Q,R,EE,SS**	**89.77**	**PNA**	**51.55**	**2,3,8**	**72.88**
	Feb	**Q,U**	**58.85**	**-**	**-**	**-**	**-**
	Mar	**L,JJ**	**62.77**	**-**	**-**	**-**	**-**
Mississippi	**Jan**	**U,V,MM,XX**	**91.12**	**EAWR**	**40.50**	**-**	**-**
	Feb	**J,NN**	**48.51**	**-**	**-**	**-**	**-**
	Mar	**CC,FF,2**	**73.12**	**-**	**-**	**-**	**-**
Montana	**Jan**	**L,V,FF,GG,VV**	**96.90**	**PNA**	**60.08**	**2,3,5,6**	**83.23**
	Feb	**M,O,P,BB**	**89.13**	**EAWR,PNA,POL**	**76.97**	**2,7**	**71.83**
	Mar	**B,H,M,Q,TT**	**85.62**	**-**	**-**	**-**	**-**
North Carolina	**Jan**	**MM**	**41.15**	**WP**	**38.98**	**-**	**-**
	Feb	**F,L,R,PP,YY**	**84.40**	**-**	**-**	**-**	**-**
	Mar	**G,EE,PP**	**73.41**	**-**	**-**	**-**	**-**
North Dakota	**Jan**	**C,D,L,HH**	**83.34**	**PNA**	**46.70**	**-**	**-**
	Feb	**L,NN,WW**	**61.72**	**-**	**-**	**-**	**-**
	Mar	**I**	**45.35**	**-**	**-**	**-**	**-**
Nebraska	**Jan**	**Q,EE,PP**	**75.86**	**-**	**-**	**9**	**39.98**
	Feb	**M,V,WW,XX**	**84.34**	**SCA**	**39.28**	**8,10**	**52.02**
	Mar	**FF,MM,NN**	**73.70**	**PNA**	**44.77**	**-**	**-**

New Hampshire	**Jan**	**Q,HH,WW,2**	**86.44**	**-**	**-**	**5**	**46.23**
	Feb	**NN,WW,2**	**70.89**	**-**	**-**	**-**	**-**
	Mar	**H,R,P,HH**	**85.74**	**POL**	**48.57**	**-**	**-**
New Jersey	**Feb**	**E,P,U,JJ**	**77.32**	**EAWR**	**40.68**	**-**	**-**
	Mar	**J,P,JJ,2**	**76.50**	**POL,SCA**	**56.52**	**-**	**-**
New Mexico	**Jan**	**O,EE,GG,LL**	**89.17**	**-**	**-**	**9**	**46.72**
	Feb	**A,O,EE,RR,WW**	**78.08**	**-**	**-**	**3,6**	**51.96**
	Mar	**Q,GG,SS**	**80.19**	**NAO,PNA**	**54.88**	**1,7**	**55.52**
Nevada	**Jan**	**U,LL,SS,YY**	**89.01**	**-**	**-**	**1**	**47.34**
	Feb	**V,DD,RR,SS,XX**	**92.69**	**-**	**-**	**-**	**-**
	Mar	**C,G,U,SS**	**72.82**	**EA,NAO**	**58.09**	**-**	**-**
New York	**Mar**	**D,H,R,HH,NN**	**87.90**	**EPNP**	**40.39**	**-**	**-**
Ohio	**Jan**	**U,BB,HH,MM**	**77.85**	**NAO,PNA,WP**	**75.39**	**1,6**	**60.08**
	Feb	**F,P,R,TT**	**95.79**	**EAWR**	**39.45**	**7**	**54.24**
	Mar	**I,SS**	**62.83**	**-**	**-**	**2,9**	**55.29**
Oklahoma	**Jan**	**D,L,EE,FF,UU**	**90.88**	**WP**	**40.68**	**6,9**	**68.73**
	Feb	**YY**	**41.03**	**-**	**-**	**1,6**	**61.48**
	Mar	**D,H,Q,II**	**86.85**	**-**	**-**	**2,5**	**62.77**
Oregon	**Jan**	**D,GG,LL,XX,YY**	**99.59**	**EPNP,PNA,SCA**	**78.61**	**1,6,8,9**	**89.66**
	Feb	**P,LL,3**	**72.36**	**EA,POL**	**74.28**	**6**	**45.18**
	Mar	**I,V,FF**	**76.91**	**EA,NAO**	**52.54**	**2**	**44.54**

Pennsylvania	Jan	J,P,U,MM	69.14	-	-	-	-
	Feb	E,Q,II,TT,WW	90.24	EAWR	40.56	2,7	52.07
	Mar	J,O,SS,XX	79.84	-	-	3	39.86
Rhode Island	Jan	JJ,LL,NN,UU	83.11	-	-	-	-
	Feb	E,P,U	86.15	EAWR	42.14	-	-
	Mar	CC	39.63	EA,POL	71.60	5,9	53.42
South Carolina	Jan	T,JJ	67.45	EA,WP	74.40	6,8	54.59
	Feb	L,R,CC,PP	75.69	-	-	-	-
South Dakota	Jan	Q,FF,TT	76.10	-	-	-	-
	Feb	A,U,LL,ZZ	87.90	-	-	-	-
	Mar	A,H,GG,WW	76.10	-	-	5,10	63.30
Tennessee	Jan	E,P,V,HH,ZZ	90.65	PNA	68.44	1,2,6	80.42
	Mar	I,M	58.27	-	-	2	42.02
Texas	Jan	L,JJ	65.81	EAWR,POL,SCA	50.15	1,6,7,9	88.90
	Feb	F,V,SS,TT,ZZ	89.95	-	-	3,7	59.15
	Mar	D,J,R,XX,2	87.61	-	-	5,7,9	77.56
Utah	Jan	J,SS,XX	77.32	PNA	40.04	1	43.13
	Feb	B,F,M,DD,XX	91.93	-	-	3	49.56
	Mar	NN,SS,WW	73.47	NAO	40.33	2	39.45
Virginia	Jan	G,I	71.71	-	-	-	-
	Feb	C,Q,NN	79.31	EA	40.09	6,8	71.42
	Mar	F,K,CC,MM,PP,RR	96.08	-	-	-	-

Vermont	**Jan**	**H,Q,V**	**77.15**	**-**	**-**	**5**	**49.74**
	Feb	**C,J,K,M,FF**	**87.03**	**-**	**-**	**-**	**-**
	Mar	**J**	**40.74**	**EPNP,WP**	**60.43**	**-**	**-**
Washington	**Jan**	**J,GG,NN,2**	**90.65**	**EA,EAWR**	**52.78**	**1,9**	**57.10**
	Feb	**-**	**-**	**EA,POL**	**54.35**	**5,6**	**61.84**
	Mar	**I,FF**	**55.58**	**EA**	**45.06**	**2,10**	**60.90**
Wisconsin	**Jan**	**A,MM,PP**	**76.97**	**PNA**	**47.63**	**1**	**48.39**
	Feb	**G,J,P,R**	**85.33**	**-**	**-**	**1,7**	**59.61**
	Mar	**Q,R,YY,1**	**83.69**	**-**	**-**	**-**	**-**
West Virginia	**Jan**	**HH,MM,3**	**81.94**	**EA,NAO,PNA**	**73.64**	**1,6,8**	**70.72**
	Feb	**A,C,Q,R**	**81.77**	**-**	**-**	**-**	**-**
	Mar	**D,G,L,M,JJ**	**94.16**	**SCA**	**38.63**	**2**	**41.26**
Wyoming	**Jan**	**T,YY,1**	**85.86**	**EA,PNA,SCA,WP**	**79.60**	**2,9**	**59.91**
	Feb	**CC,JJ,RR,WW**	**74.40**	**SCA**	**39.22**	**-**	**-**
	Mar	**D,G,BB,HH,TT**	**86.09**	**-**	**-**	**6**	**41.67**

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

- In January, ipsative mode models achieved greater WESS than corresponding normative mode models 91.3% of the time (46 states had models based on January data). Similarly, in February the ipsative mode models almost always (93.3% of the time) achieved greatest WESS (45 states had models based on February data), and even as data aged substantially—in March, ipsative models almost always (93.5% of the time) achieved greatest WESS (46 states had models based on March data).

- Using ipsative modes, for January data 12 states had CTA models with WESS$\geq$ 90%, as did 6 states for February data and 4 states for March data. Zero normative mode models achieved this level of WESS in any month modeled.

- We statistically contrasted the WESS of each pair of these three sets of modes. If no model was found, then WESS was assumed to be zero. We used ODA to determine which set of modes was better at predicting whether the mean precipitation of the states exceeded the median, or not. The PTMP procedure[7] was used to estimate the exact Type I error for each contrast. Analyses of January data (March and February had comparatively sparse data) indicated that the ipsative mode model had significantly greater WESS than the normative mode models (*p*'s<0.0002), but computed and published raw modes were indiscriminable (*p*<0.15).

Predicting Export of Arctic Sea Ice

The export of Arctic sea ice through the Fram Strait off northeast Greenland is an important factor in the freshwater balance of the North Atlantic Ocean, and affects the North Atlantic thermohaline circulation. The January monthly ice export at fluxgate *a* of the Fram Strait[8] was studied using the ipsative modes found here. The data consisted of sea ice area flux for the years 1979-2002. Kendall's tau b statistic was used to determine the correlation of modes with ice export, and the significant associations are shown in Figure 5. Negative associations were found with ipsative modes U (over Green-land), CC (near Svalbard), 3 (near Franz Josef Land), XX (off the coast of northern Norway), and SS (eastern Pacific Ocean). Positive associations were found with ipsative modes UU (Mediterranean Sea south of France), WW (North Atlantic Ocean northwest of Spain), H (over Manchuria), and BB (east of Japan).

An example of a pattern with high sea ice export is illustrated in Figure 6. The 500 mb pattern in January 1983 yielded the maximal ice export for any January in the years of 1979-2002. Low 500 mb heights extend from Green-land to Scandinavia and western Russia, and another area of low heights is found off of the Pacific coast of the USA. Areas of high 500 mb heights are seen over southwest Europe and the western Mediterranean Sea, and over Mongolia and northeast China.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

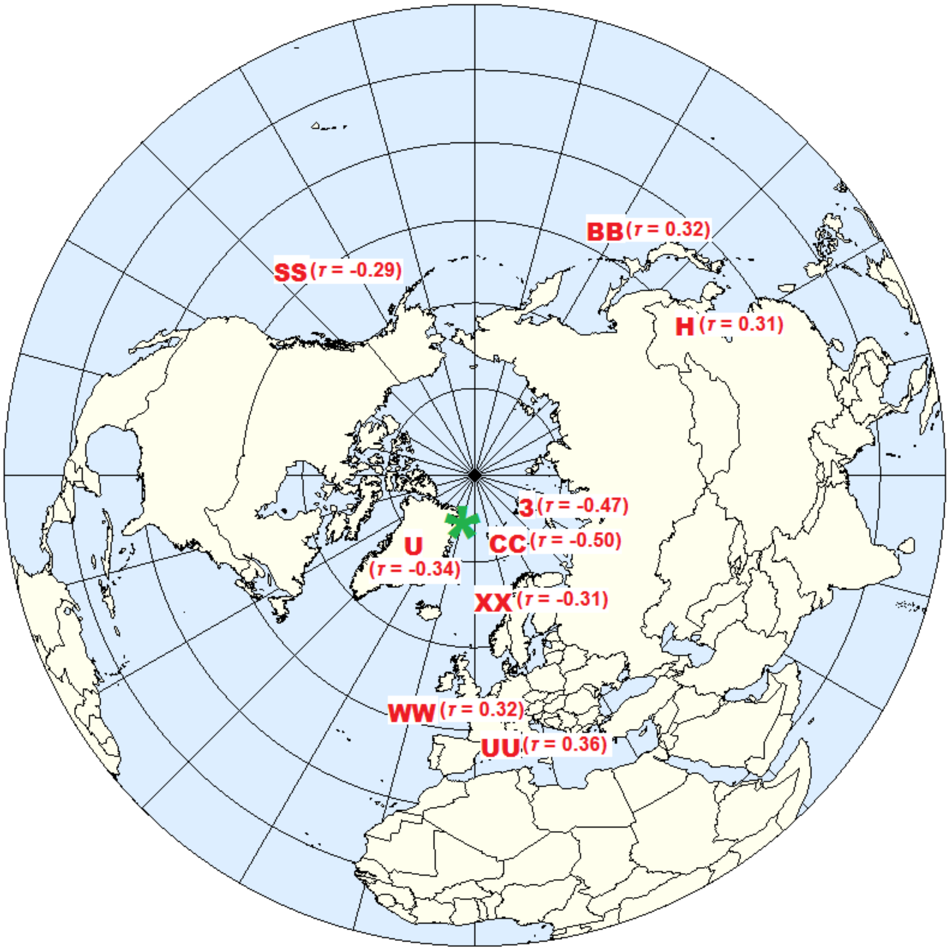


Figure 5: Ipsative modes and Kendall's Tau b Coefficients with Statistically Significant (p<.05) Associations with Ice Export at Fram Strait Fluxgate a, Indicated as *

Recent research[9] reported no correlation between SLP-based NAO and Arctic wintertime sea ice export over 1958-1977, and a positive correlation of 0.7 over 1978-1997. An eastern shift in NAO centers of variability was suggested to explain this phenomenon. However, for the 500 mb level, ipsative mode U was a stable center over Greenland, for both sets of years, 1948-1976 and 1977-2007. Mode U represents the northern center of the NAO dipole at the 500 mb level. Mode II (near Iceland) was also a stable center, coincident with the northern center of surface-level winter NAO variability: this does not support the idea of a shift at 500 mb. Furthermore, factors XX, CC and 3, located in this region, were stable in both eras and reliably associated with sea ice movement. Mode 3 is coincident with the surface center of variability in the Kara Sea, previously found to be associated with sea ice export variability.[10]

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00p

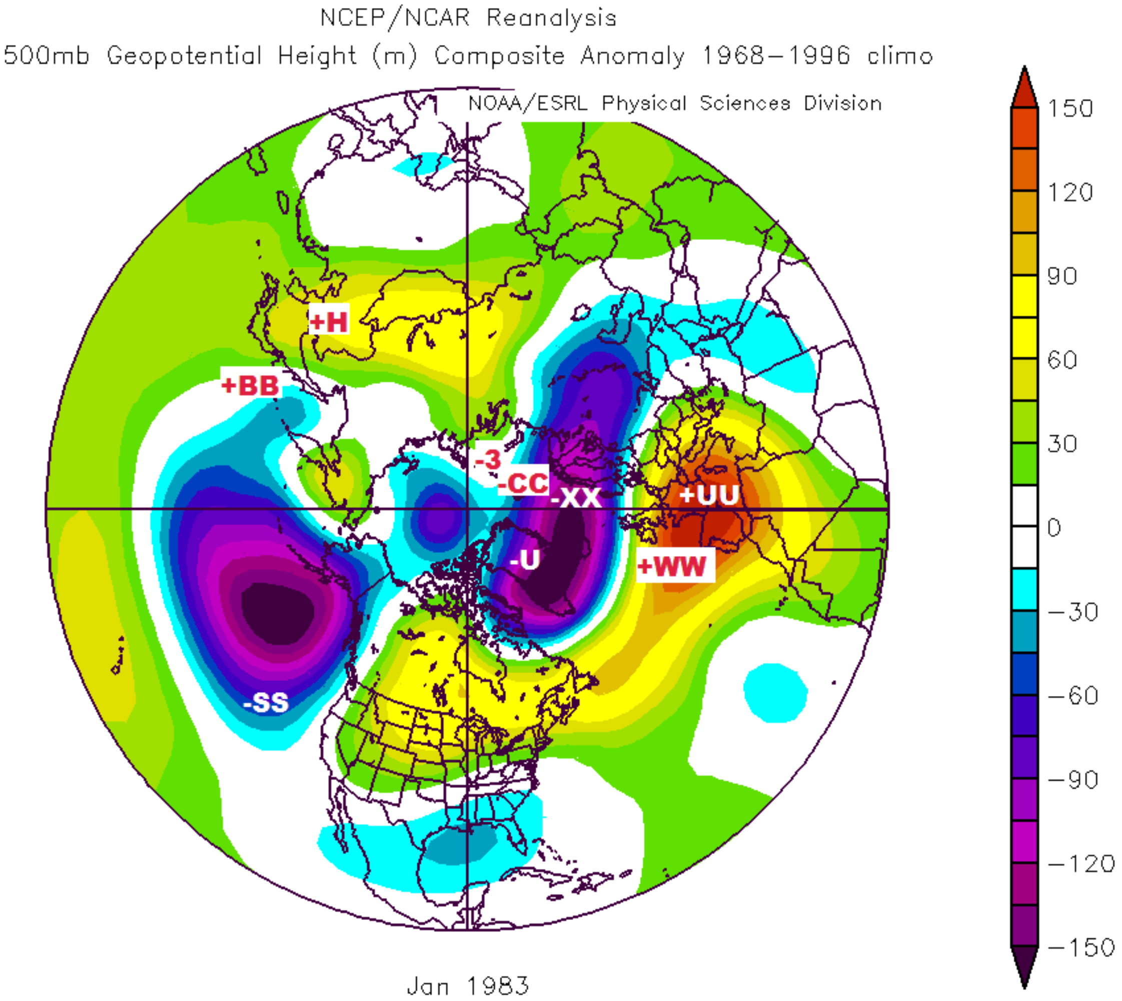

NCEP/NCAR Reanalysis
500mb Geopotential Height (m) Composite Anomaly 1968–1996 climo
NOAA/ESRL Physical Sciences Division
+H
+BB
-3
-CC
-XX
+UU
-U
+WW
-SS
150
120
90
60
30
0
−30
−60
−90
−120
−150
Jan 1983

Figure 6: 500 mb GHA for January 1983, which Entailed the Maximal January Ice Export for the Period 1979-2002: Ipsative modes are Prefixed by the Sign of their Associated Kendall's Tau b Coefficient

Epilogue

Preliminary results using uncounfounded climatic data in atmospheric prediction are very positive. An important extension of the present research is obtaining GHA modes for all months of the year. Further evaluation of optimal statistical methods used with unconfounded climatic data is warranted. Future research should use these data in applications such as, for example: predicting the ontogenesis, intensity, and path of hurricanes[11], and the ontogenesis, intensity, and location of sudden stratospheric warmings[12,13]; modeling of seasonal energy consumption and management of climate risk for energy firms[14]; forecasting and understanding the ENSO cycle (El Niño)[15], and development and evaluation of numerical weather prediction models.[16]

References

[1]Yarnold PR. Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement* 1996, 56:430-442.

[2]Barnston AG, Livezey RE. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review* 1987, 115:1083-1126.

[3]Bryant FB, Yarnold PR. Principal components, and exploratory and confirmatory factor analysis. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. APA Books, Washington, DC, 1995, pp. 99-136.

[4]Yarnold PR. Statistical analysis for single-case designs. In: FB Bryant, L Heath, E Posavac, J Edwards, E Henderson, Y Suarez-Balcazar, and S Tindale (Eds.), *Social psychological applications to social issues, volume 2: methodological issues in applied social research*. Plenum, New York, NY, 1992, pp. 177-197.

[5]Yarnold PR. Discriminating geriatric and non-geriatric patients using functional status information: an example of classification tree analysis via UniODA. *Educational and Psychological Measurement* 1996, 56:656-667.

[6]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. Washington, DC, APA Books, 2005.

[7]Cade BS, Richards JD. *User manual for blossom statistical software*. US Geological Survey, Fort Collins, CO, 2005.

[8]Kwok R, Cunningham GF, Pang SS. Fram Strait sea ice outflow. *Journal of Geophysical Research* 2004, 109:1009-1029.

[9]Holland MM. The north Atlantic oscillation–Arctic oscillation in the CCSM2 and its influence on Arctic climate variability. *Journal of Climate* 2003, 16:2767–2781.

[10]Hilmer M, Jung T. Evidence for a recent change in the link between north Atlantic oscillation and Arctic sea ice export. *Geophysical Research Letters* 2000, 27:989–992.

[11]Willoughby HE, Rappaport EN, Marks FD. *Hurricane forecasting: the state of the art*. *Natural Hazards Review* 2007, 8:45-49.

[12]Charlton AJ, Polvani LM. A new look at stratospheric sudden warming events: part I. climatology and modeling benchmarks. *Journal of Climate* 2007, 20:449-469.

[13]Charlton AJ, Polvani LM, Perlwitz J, Sassi F, Manzini E. A new look at stratospheric sudden warming events: part II. evaluation of numerical model simulations. *Journal of Climate* 2007, 20:470-488.

[14]Troccoli A (Ed.). Management of weather and climate risk in the energy industry. Proceedings of the NATO advanced research workshop on weather/climate risk management for the energy sector, Santa Maria di Leuca, Italy 6-10 October 2008. Series: *NATO science for peace and security series C: environmental security*. Springer, New York, NY, 2008.

[15]Babkina AM (Ed.). *El Niño: overview and bibliography*. Nova Science Publishers, Hauppauge, NY, 2004.

[16]Kalnay E. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, Cambridge, UK, 2002.

Author Notes

Send correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Here Today, Gone Tomorrow: Understanding Freshman Attrition Using Person-Environment Fit Theory

Jennifer Howard Smith, Ph.D., Fred B. Bryant, Ph.D.,
Applied Research Solutions, Inc. Loyola University Chicago

David Njus, Ph.D., and Emil J. Posavac, Ph.D.
Luther College Loyola University Chicago (Emeritus)

Person-Environment (PE) fit theory was used to explore the relationship between student involvement and freshman retention. Incoming freshmen (*N*=382) were followed longitudinally in a two-wave panel study, the summer before beginning college, and again during the spring of their freshman year. Involvement levels, a variety of summer and spring preferences (Ps), and spring perceptions (Es) regarding specific aspects of their college environment were assessed. Twelve PE fit indicators were derived and compared with respect to their relationship with student involvement and retention. Results indicated that involvement was linked to some PE fit indicators. Traditional parametric statistical analyses were compared with a new, nonparametric technique, Classification Tree Analysis (CTA), to identify the most accurate classification model for use in designing potential attrition interventions. Discriminant analysis was 14% more accurate than CTA in classifying returners (97% *vs*. 85%), but CTA was 962% more accurate classifying dropouts (8% *vs*. 84%). CTA identified nine clusters—five of returners and four of dropouts, revealing that different subgroups of freshmen chose to return (and stay) for different reasons. Students' end-of-the-year preferences appear to be more important than anticipated preferences, college perceptions, or PE fit levels.

People most at risk of dropping out of organizational settings are those who have been there the shortest periods of time.[1] Thus, in college settings, students most at risk of dropping out are freshmen.[2,3] Although researchers have long known about college attrition problems and have proposed a variety of theoretical models as potential remedies, little progress has been made in actually reducing student dropout rates.[2-4] The act of leaving college prior to

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

graduation is often seen as a form of failure on the part of the attritor, and not on the part of the institution. However, it may be that features of college environments may be at least partly responsible for the early withdrawal of some students.[3] This possibility makes a theory which addresses both person- and environment-focused variables (i.e., PE fit theory) potentially important in better understanding college attrition.

A large body of research has investigated the issue of college attrition, linking student departure to low levels of student integration and involvement. It is important to distinguish between two different conceptualizations of "involvement" discussed in the education literature. One way to define involvement is *behaviorally*—as the degree to which students participate in academic and social activities. Here, involvement is defined solely in terms of student behaviors (e.g., number of activities attended, frequency of participation). A second way to define involvement is *psychologically*—as students' level of perceived commitment to, or affiliation with, their university.[5,6] The present study uses only the behaviorally-based conceptualization of involvement.

Encouraging students to be involved in campus activities seems to be an effective way of positively influencing their perceptions and ultimately their persistence.[2-4,7-10] Student involvement has been shown to affect commitment to graduate; this commitment, in turn, has been linked to both intentions to remain enrolled and actual re-enrollment decisions.[2-4,11]

Calling students' freshman year a "strategic leverage point," Tinto claims that most attrition decisions arise either explicitly during the freshman year or have their roots in the first-year experience.[3] To maximize the chances for students to make a commitment to graduate, Tinto calls for an increase in freshman opportunities to engage in (formal and informal) social and academic activities. Astin's research also links college involvement to student development and college retention.[7-10,12,13] According to Astin, attritors' modal explanation for dropping out is boredom with college. Indeed, boredom may simply be another name for being uninvolved. Of course, being uninvolved may be caused by person-focused factors (e.g., student's lack of initiative), environment-focused factors (e.g., lack of college opportunities), or both.

One way to understand the interaction of person-focused and environment-focused factors on behavior is through Person-Environment (PE) fit theory. Several studies have demonstrated the relationship between the "fit" of student characteristics (P) and college attributes (E), and a plethora of educational variables including physical symptoms,[14,15] academic and social competency,[16] satisfaction,[17] academic achievement,[18] student stress and strain,[19] level of cognitive development,[20] withdrawal, alcohol consumption, anxiety, the use of mental health services, grade point average,[14] coping strategies,[21] volunteer motivation[22], school crime and misbehavior,[23] willingness to recommend their college to prospective students,[24] and retention.[25] However, few studies have investigated the direct link between PE Fit and student retention. Tinto alludes to PE fit in his retention model, but offers no specific recommendations concerning how to measure congruence between student preferences and college characteristics, nor conceptual or operational definitions of PE misfit. Empirical tests of Tinto's model also lack these components.[26] Astin also alludes to PE fit in his retention research. However, like Tinto, he does not explicitly measure PE misfit in ways recommended by congruence researchers, such as assessing PE variables on commensurate conceptually corresponding scales.

The task of validly assessing the match between personal properties and environmental features is difficult.[20,27-29] Researchers must determine which P and E variables are the most relevant to the population of interest. They also must find the best way to combine these salient dimensions into a congruence, or fit, score. Those studying PE fit must balance the two dimensions, giving equal consideration to both.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

Unfortunately, this often is not the case. Even when one is certain that this balance has been achieved, researchers must be certain that each personal variable has a commensurate environmental variable in order to justify calculating a valid PE fit score.[6,27,30-32] Whether to calculate single or multiple PE fit indicators is another important measurement issue to consider. The notion of breaking down complex environments into more manageably-sized Es can be traced to Barker[33] and Wicker,[34], and is still apparent today in studies of noisy production lines,[35] hospital wings,[36] college dormitories,[37] career counseling departments,[38,39] and classrooms.[40] A college campus may be an ideal candidate for this type of research since most university settings contain distinct sets of populations, opportunities, and values.[15,41] Tinto proposed that college environments actually are comprised of clusters of social and academic communities or subcultures.[3] If micro-environments within a school can be identified, it may be reasonable to derive PE fit indicators for each dimension, rather than to rely simply on one overall congruence score.

Researchers are far from reaching a consensus regarding how best to operationally define the PE fit construct. The most frequently used measure of congruence is the difference score, which really is an indicator of PE *misfit*.[32] P and E items are subtracted from one another, producing a “discrepancy” score. Traditionally, “Real E” items are subtracted from corresponding “Ideal P” items, with the underlying assumption that one’s actual environment typically will not exceed one’s ideal version of it. Some PE fit researchers compute the absolute value of this difference score, asserting that “P less than E” effects are similar to “E greater than P” effects.[14,25,36,42] Others, however, have preserved the direction of PE incongruence by eliminating the absolute value sign.[23,31,43-45]

It is crucial that the personal (P) and environmental (E) components comprising the congruence construct are carefully defined. Researchers, however, disagree on how best to do this. Examples of P conceptualizations are diverse and include dimensions such as: ideals,[19] expectations,[37] values,[46,47] needs,[11,48,49] interests,[18,50,51] personalities,[52] choices,[50] and demographic information.[7]

Researchers have conceptualized the environmental (E) component of PE congruence a variety of ways as well. Some define environments phenomenologically, by assessing occupants’ images of a setting, rather than assessing a setting’s objective features. Advocates of this approach believe that perceptions have real consequences.[3,24] From this perspective, university settings are defined in terms of their perceived “climates”.[48,49] A second E conceptualization defines college environments in terms of the aggregate of students’ characteristics.[5,6,50,53] Environments from this perspective are defined by who their occupants are (e.g., choice of major, ability levels, and ethnic backgrounds), rather than by what their occupants perceive.

A third way to conceptualize college environments is by the activities that occur on campus. Behaviorally-based E conceptualizations are concerned with what students and faculty actually do, rather than what perceptions they share or what characteristics they possess.[1,3,4,7,8,10] From this perspective both the opportunity for activities and the activities themselves combine to represent the E component.

Measures of student-college congruence will differ depending on which of these P and E conceptualizations are used to derive the congruence construct. Using the image-based E, PE fit assesses whether an institution lives up to the reputation or mystique surrounding it. Using the “characteristics-based” E, PE fit represents how closely each student matches the attributes of the student body majority. However, using the third, “behaviorally-based” conceptualization of “E,” PE Fit assesses the match between students’ preferences for involvement, and the actual opportunities to become involved in college.

If environments can be defined both subjectively (e.g., climates) and objectively

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

(e.g., aggregate characteristics), so can congruence measures. According to French, "subjective" PE fit reflects the match between people's preferences regarding their self-concept and their setting, and their beliefs about these attributes.[31] "Objective" PE fit, on the other hand, uses information that is independent of the biases underlying human perceptions. Actual attributes of both the person (e.g., knowledge, abilities) and the environment (e.g., policies, activities) interact to produce these PE fit indicators.

Some researchers have expressed a concern about the potential for excess error within subjective PE fit variables, claiming that an over-reliance on perceptual data may lead to the attenuation of true effects.[19] They argue that any one person's assessment of the actual environment (the E component) will contain associated error variance resulting from personal biases and the lack of relevant environmental information.[6,27] For example, students are often unaware of, or even denied access to, information concerning specific activities and interactions occurring on their campus. This lack of knowledge may add error to E scores and attenuate the true effects of PE congruence.

In response to these concerns, some researchers have suggested that the measurement gap between objective and subjective reality be narrowed.[42] Tracey and Sherry proposed that a more accurate measure of the actual environment is the *mean* of all respondents' "Real E" ratings. They claim that these environmental "consensus" scores are highly reliable because they are unlikely to be affected by individual variation. They also claim that these more objective congruence measures possess more construct validity, for they better represent the discrepancy between ideal and actual settings.

Tracey and Sherry used this technique to examine the relationship between PE fit and student strain in a college residence hall. They asked residents to describe the preferred characteristics (P) of a residence hall and then to describe the actual characteristics (E) of their own residence hall. In addition to creating subjective discrepancy scores by subtracting each participant's P score from her E score, Tracey and Sherry also created an objective PE fit indicator by computing the mean of all floormates' E scores and subtracting this measure of central tendency from each P score. It was found that discrepancy scores based on a consensus of E were more highly correlated with student stress and strain than respondents' own "subjective" PE fit scores. The superior strength of using the mean of "Real E" scores has been demonstrated in other studies investigating student-college congruence.[16] However, advocates of these "objective" measures of PE fit are not without their critics. Edwards is leery of congruence meas-ures that hold one element constant, such as when the mean of "actual" ratings is used to represent E.[54,55] He argues that when PE fit is computed this way, discrepancy scores merely represent the variance attributable to one element (e.g., P), and thus do *not* represent PE congruence at all.

Besides determining *how* to measure PE fit, another unresolved issue involves *when* to measure congruence. The traditional approach to measuring PE fit is to ask respondents to provide both their personal preferences (P) and their environmental descriptions (E) concurrently.[16,35,46] While this strategy is convenient (i.e., requiring only one data collection session), this design may suffer from a number of conceptual and methodological problems, such as restriction in range due to natural attrition. Individuals who experience PE misfit over time either exit or adapt to their environments, thus spuriously shrinking the range of the personal characteristics remaining and reducing the measure's predictive power.[14,15,56] Selective attrition results, leaving only those most congruent, and presumably those most productive and satisfied, to occupy the setting, and to complete researchers' measures. This may pose a problem, since most participants of PE fit studies are individuals who have occupied their settings the longest.[29] Individuals with considerable experi-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

ence and familiarity with a setting (e.g., tenured employees, seniors in college) are likely to possess synchronized preferences and perceptions. These members are typically few in number and may comprise an unrepresentative sample.[5] Range restriction problems also raise the issue of external validity threats. If tenured occupants possess a unique set of similar characteristics, results from any one PE fit study may be lacking with respect to generalizability.[57] One way to remedy this problem is to examine longitudinally populations that recently have entered an environment. College freshmen may serve as an ideal group for this approach.

Instead of measuring congruence at one point in time, several researchers have begun to utilize longitudinal research strategies to better understand *degrees* of, or changes in, PE fit. This nonconcurrent approach to measuring PE fit, although more time consuming, offers many benefits. For instance, these designs enable researchers to assess occupants' desires and perceptions both before and after they are influenced by the impact of their environments. If planned carefully, nonconcurrent designs are also able to include both congruent and incongruent individuals in their pool of respondents. Additionally, these designs also allow for different PE fit scores both *before* (e.g., "Anticipatory PE fit") and *after* ("Present PE fit") individuals enter and familiarize themselves with a setting to be calculated.[14,46]

Statistical Analysis Options

One goal of this project was to describe and classify as accurately as possible two groups of freshmen—those who returned as sophomores and those who did not—using PE fit variables and involvement indices. Two statistical techniques were compared with respect to their ability to accuracy classify returners and attritors. In addition to a traditional discriminant analysis (DA), an alternative statistical technique also was performed on the data. Optimal Data Analysis (ODA) is a unique nonparametric approach to statistical classification that explicitly maximizes the average *p*ercentage *a*ccuracy in *c*lassification (PAC) across groups in a sample.[58] ODA works by finding an optimal classification solution which consists of a cutpoint (the point that lies midway between successive observations that are from different groups) and a direction, which is analogous to the "sign" of a conventional statistic like a correlation. ODA finds the cutpoint and direction combination such that no other combination can result in fewer misclassifications: by definition, the resulting model is always optimal.[58]

A special application of ODA, hierarchically optimal classification tree analysis (hereafter referred to as CTA) was used in the present study, to distinguish returners from attritors. CTA is an *iterative* ODA procedure that constructs a classification tree which hierarchically maximizes the mean percent accuracy in classification (mean PAC) for a sample.[58] CTA is accomplished after several steps. First, a stopping rule is determined *a priori* (e.g., experimentwise Type I error of $p<0.05$). Second, ODA is performed for every attribute (predictor) separately, using the total sample. The attribute yielding the greatest standard effect size is then chosen and the cases are split according to this model's cutscore and direction on the attribute having greatest effect strength (the model will likely be imperfect, making both correct and incorrect classifications). Third, ODA is performed again using all of the attributes, but only on a *subset* of the sample—the respondents who were predicted to be in one class only (e.g., dropouts) in an attempt to improve classification for this partition only. If a new attribute is found to improve the predictive value it is added to that particular "branch" of the classification tree. If not, the branch ends there. The classification tree "grows" until a sufficient number of attributes is found that best describes each subset of the sample. Branches are then "pruned" (i.e., nodes are removed) if their Type I error exceeds a set criterion, or if the branches do not enhance the model's overall mean PAC.[58.59]

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

Traditional DA assumes that a set of attributes is equally relevant and meaningful to all members of a particular sample.[59] CTA, in contrast, creates separate discriminant functions for different subsets of the sample while describing clusters of individuals that share the same common pathway. For example, it may be that students choose to leave or to remain for different reasons. One segment of the freshman class may return for social reasons, while another segment may return for academic reasons. These specialized student clusters, which would be overlooked with traditional DA, may help to identify unique sets of "at-risk" freshmen.

Another advantage of CTA is freedom from the restrictive assumptions underlying parametric tests. DA requires that several assumptions be satisfied, such as independence, linearity, and distributions that are normal, in order for the estimated Type I error rate to be valid.[61] In contrast, for CTA "*p*" (i.e., the probability of making a Type I error) is *exact* and always valid, because it is based solely on the structural features of a particular data set. [58]

Because bias may enter a classification solution if the coefficients used to assign a participant to a particular group are derived using that person's data, it is important to perform leave-one-out (LOO) validity analysis (also called the jackknife procedure).[58] This procedure is then repeated, holding a different case out each time, for every case. An advantage of CTA is that LOO analysis is performed at every step in the analysis.

Purpose and Hypotheses

This study was conducted with three purposes in mind. The main purpose of this study was to assess the degree to which involvement in college activities was associated with first year students' PE fit levels, and the degree to which these PE fit levels impacted their decisions to return as sophomores. A second purpose was to determine the relative contributions that different PE fit derivations make in explaining student involvement and attrition. Finally, this study sought to compare traditional multivariate statistical strategies with nonparametric optimal analyses. Based on previous empirical tests of PE fit theory and college retention models, these three goals resulted in the following six predictions.

1. The first hypothesis addressed the dimensionality of the PE fit construct, and predicted that student "Ideals" (Ps) with respect to college environment preferences would be multidimensional, and thus multiple PE fit indicators would be derived—one per dimension. It also was expected that these dimensions would be stable over time, from summer until spring.

2. The second hypothesis addressed the relationship between students' participation in college activities and their subsequent PE congruence levels. It was hypothesized the more that students participated in college activities, the greater would be their degree of PE fit.

3. The third hypothesis addressed the relationship between PE fit and retention decisions. It was proposed that students with greater PE fit would be more likely to return for their sophomore year than students with more incongruent levels.

4. In-coming freshmen may not be as certain of their college environment preferences prior to beginning college, so the fourth hypothesis predicted "Present" PE fit (*Posttest* Ideals minus Posttest Reals) scores would be a better predictor of return status, and a better criterion of college involvement, than "Anticipatory" PE fit (*Pretest* Ideal minus Posttest Real).

5. Because it is likely that no one student can accurately describe all dimensions of a college environment, "Objective" PE fit (Posttest Ideals minus the *mean* of Posttest Reals) was hypothesized to be a better predictor of return status, and a better criterion for college involvement, than "Subjective" PE fit (individual Posttest Ideals minus individual Posttest Reals).

6. Lastly, it was proposed that PE congruence measures would be more strongly related to college involvement and retention deci-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

sions than either college preferences (P) or college perceptions (E) alone.

Method

Participants. In-coming freshmen from a large Midwestern Catholic university were surveyed during summer registration sessions, and again during the spring of their freshman year either in residence halls (for on-campus students) or by postal mail (for commuters). A total of 1,108 freshmen of the 1,186 students comprising the freshman class (93.4%) completed summer questionnaires, and 420 of these freshmen (38%) completed spring questionnaires (12 additional students completed the posttest, but not the pretest.) Of the 420 spring participants, 382 placed a confidential identification number on both questionnaires, allowing their summer and spring responses to be linked and compared. Data from these 382 "pretest-posttest" students were subsequently used to test the hypotheses; they represented 34.5% of the original sample.

Procedure and Instruments. Pretest data were obtained during summer registration sessions before the students' first semester. Posttest data were obtained at the end of participants' freshman year. Social security numbers were used to match students' pretest and posttest responses. The confidential treatment of responses was clearly emphasized to participants and was strictly enforced.

Pretest. In an attempt to increase the response rate, pretest data were collected during summer orientation sessions. All but 78 students who comprised the freshman class (1,108 of 1,186) gathered in groups of approximately 200 in a university auditorium the first morning of their respective registration sessions (numerous sessions were held throughout the summer). After completing math placement exams, freshmen completed the PE fit pretest questionnaire.

Pretest items assessed respondents' college preferences. These items represented "anticipated" ideals (Ps), since they were completed before students actually experienced college life. Participants evaluated various features of a college environment using 7-point scales, ranging from "very undesirable" to "very desirable."

The pretest questionnaire contained 46 items which were either created specifically for this college environment or were borrowed from past PE congruence instruments. Eleven items were chosen to correspond to the various components of a new university program designed to encourage freshman participation and to enhance freshman retention implemented that year. For example, freshmen were asked to indicate how desirable it would be to go on a retreat, to use electronic-mail to communicate with faculty, and to go to the symphony or theater. Fourteen items corresponded to activities common to any university setting, such as voting in a campus election, or attending a social event. Twenty-one items were borrowed and modified from the Organizational Culture Profile Item Set.[46] This set of items tapped students' preferences for certain environmental "presses" or images. For example, freshmen were asked to indicate how desirable it would be for their college environment to be rule-oriented, to be supportive, to foster independence, and to allow them time to themselves.

Posttest. The posttest questionnaire was distributed in the spring of respondents' first year, approximately 9 months after the pretest. Students residing on-campus were given posttest questionnaires in their residence halls. Commuter students were surveyed via the mail.

Respondents rated the same set of college dimensions that were included in the pretest questionnaire with the exception of three items ("reward minimal effort with high grades;" "reward good performance with high grades;" "have the same classmates in several of my courses") which were eliminated due to the findings of an exploratory principal components analysis which are discussed below. However, unlike the pretest instrument which contained only items assessing college ideals ("Anticipatory" Ps), the posttest instrument contained both

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

college preference ("Present" P) and college perception (i.e., "Real" E) items presented on commensurate scales.

For preference (P) ratings, students were asked to indicate the degree to which they desired various college attributes, and the degree to which they would desire participating in a variety of college activities (1=not at all; 7= very much). For perception (E) ratings, students were asked to indicate the extent to which each attribute accurately described their college impressions and experiences (1=not at all; 7=very much). Anchors differed depending on whether E items were presented as continuous (1=never; 7=very often) or discrete (yes/no) variables.

Attributes

Three major groups of attributes were measured to test the specified hypotheses.

Student Involvement. Sixteen "Real" (E) items were combined to create an involvement index which assessed the extent to which students participated in both academic activities (e.g., speaking up in class; seeking out one's advisor) and social activities (e.g., attending a cultural event; being active in campus politics) during their first year. Psychologically-based aspects of involvement, such as students' commitment to the university, were not assessed.

Five of the 16 involvement items tapped activities that could be done repeatedly throughout one's freshman year (e.g., chat with an instructor, go to church with friends), and were rated on 7-point scales ranging from "never" to "very often." The remaining 11 items included events that, for the most part, students would engage in only once or twice during the school year (e.g., go on a retreat, dine with a professor). To indicate whether or not they engaged in these activities, students circled either "Yes" or "No."

To create an overall index of involvement for each student, the sum for each of the two sets of items was converted to standard (z) scores, and multiplied by the number of items comprising those sets (5 and 11, respectively). These scores were then added together and divided by 16 to create an overall standardized involvement index.

PE Fit. Derivation of PE fit indicators was complex, and involved four steps. First, two principal components analyses were performed on the summer and spring sets of Ideal data to determine the dimensionality of student college preferences (Ps). Three factors were revealed and named "College Image," "Student Experience," and "Traditional-Catholic." E items were then categorized on the basis of these factors so that PE fit scores could be derived (see Results).

The second step involved computing PE Fit indicators as difference scores. PE fit indicators were computed at the factor level only.[31] However, in contrast to French's congruency formula, the absolute values of these differences were used so that specific multivariate statistical analyses could be performed.[31] Thus, for the present study, PE fit was calculated as the absolute value of the difference between the sum of student preference (P) items and the sum of the commensurate set of student perception (E) items for each of the three dimensions: PE Fit=$|\Sigma P - \Sigma E|$. These differences were then divided by the number of commensurate pairs in each of the three factors (16, 13, and 8 items, respectively). The magnitude of absolute difference scores increases as P and E ratings become increasingly discrepant, so small congruence scores represent greater PE fit.

Because several authors suggest different ways to derive PE fit scores, the third step involved deriving four distinct kinds of discrepancy scores (Table 1).[19,30,44] First, to determine the degree of congruence for students who had not yet experienced college life, "Anticipatory" PE fit scores were computed by taking the difference between *pretest* Ideal ratings and *posttest* Real ratings. Second, to determine students' level of congruence at the end of their first year, "Present" PE fit scores were derived by computing the difference between *posttest* Ideal ratings and *posttest* Real ratings.

Table 1: PE Fit Components and Derivations

Component	Operational Definition
Anticipatory Personal Preferences (P)[a]	Pretest Ideal items
Present Personal Preferences (P)	Posttest Ideal items
Actual Environmental Properties (E)	Posttest Real items

Type of PE Fit[b]	Derivation of Difference Score[c]
Anticipatory Subjective PE Fit	Pretest Ideals minus Posttest Reals
Anticipatory Objective PE Fit	Pretest Ideals minus (mean) Posttest Reals
Present Subjective PE Fit	Posttest Ideals minus Posttest Reals
Present Objective PE Fit	Posttest Ideals minus (mean) Posttest Reals

Note: [a]This construct was assessed during summer orientation sessions. All other attributes were derived using data collected at the end of respondents' first year. [b]These variables were computed for each of the three dimensions (College Image, Student Experience, and Traditional-Catholic). [c]All PE fit derivations used the absolute value of the differences.

The third and fourth types of PE fit indicators differed with respect to how the E attributes were computed. “Subjective” congruence scores were derived by taking the difference between each freshman’s set of (posttest) Ideal and Real scores. “Objective” fit scores were computed by replacing respondents’ individual Real scores with the *mean* of all students’ Real rating. Crossing Anticipatory and Present congruence measures with Subjective and Objective measures, a total of four PE fit indicators resulted: (a) Anticipatory Subjective PE Fit; (b) Present Subjective PE Fit; (c) Anticipatory Objective PE fit; and (d) Present Objective PE fit.

The final fourth step in the derivation of PE fit indicators involved computing congruence scores across the three dimensions revealed in the first step. The four PE fit indicators derived for each of these factors resulted in a total of 12 types of PE fit indicators (see Table 2).

Return Status. Retention information was obtained via the university’s Department of Institutional Research. Respondents failing to return for the sophomore year were classified as attritors, regardless of the reason for departure.

Results

Pretest-Posttest Respondents vs. Pretest-Only Respondents. Analyses comparing respondents who completed only the pretest with respondents who completed both measures were performed. Summer Ideal responses, as well as additional demographic and academic information, were compared. Because comparisons are meaningful only for students who had the *opportunity* to complete both measures, 44 students who completed the fall semester but who did not re-enroll for the spring semester were omitted from these analyses.

Results revealed that pretest-posttest and pretest only students were comparable on several important dimensions. For instance, these groups did not differ greatly with respect to attrition rates (10.5% *vs*. 13.7%, respectively), nor did they differ statistically with respect to anticipatory preferences on the three PE fit dimensions (*p*s>0.05, mean effect size=0.10). These groups also did not have different expectations regarding first-semester GPAs (3.51 *vs*. 3.57, respectively, effect size=0.04), or first-year cumulative GPAs (3.61 for both groups).

Table 2: Descriptive Statistics for PE Fit Indicators

Objective PE Fit[a]

Student Image	College Behavior	Traditional-Catholic
Anticipatory PE Fit[c]	Anticipatory PE Fit	Anticipatory PE Fit
M=0.88 sd=0.47 (378)	M=1.69 sd=0.87 (376)	M=0.82 sd=0.59 (378)
Present PE Fit[d]	Present PE Fit	Present PE Fit
M=0.88 sd=0.47 (360)	M=1.64 sd=0.91 (358)	M=0.88 sd=0.66 (345)

Subjective PE Fit[b]

Student Image	College Behavior	Traditional-Catholic
Anticipatory PE Fit	Anticipatory PE Fit	Anticipatory PE Fit
M=0.97 sd = 0.74 (342)	M=1.72 sd=0.94 (347)	M=0.88 sd=0.64 (338)
Present PE Fit	Present PE Fit	Present PE Fit
M=0.82 sd = 0.68 (344)	M=1.61 sd=0.94 (345)	M=0.73 sd=0.62 (337)

Note: M=mean; sd=standard deviation. Smaller means indicate smaller discrepancy scores and greater PE fit. Numbers in parentheses indicate the sample sizes. [a]Objective PE fit scores were derived from Individual "Ideals" and the mean of "Reals". [b]Subjective PE fit scores were derived from Individual "Ideals" and Individual "Reals." [c]Anticipatory PE fit scores were derived from Summer "Ideals" and Spring "Reals." [d]Present PE fit scores were derived from Spring "Ideals" and Spring "Reals."

However, some important differences were revealed. Although pretest-posttest and pretest-only students possessed similar GPA *expectations*, they did statistically differ in the GPAs they later *earned*. Students who completed both measures earned higher fall GPAs (3.06 *vs*. 2.97, $t(989)=2.15$, $p<0.032$), higher spring GPAs (3.06 *vs*. 2.89, $t(1017)=3.62$, $p<0.0001$), and higher first-year cumulative GPAs (3.07 *vs*. 2.94, $t(1009)=3.23$, $p<0.001$). However, the effect sizes corresponding to these differences were small (0.19, 0.28, 0.30, respectively, mean effect size= 0.26). Additionally, both gender and place of residence impacted whether or not students participated in both waves of the study. A greater percentage of women comprised the pretest-posttest group (72.5%) than the pretest-only group (57.3%). Freshmen residing off-campus were also less likely to complete both measures.

Tests of Hypotheses

Dimensionality of PE Fit. To determine whether college preferences, and the PE fit construct, were uni- or multi-dimensional, a principal components factor analysis with varimax rotation was performed on the Present Ideal data. Only participants providing both pretest and posttest information were used (n=382). Six Present Ideal items ("Is easy-going;" "Is unpredictable;" "Fosters risk-taking;" "Work under pressure;" "Rewrite a paper/Redo a project;" and "Use e-mail to communicate with faculty and classmates") did not have factor loading exceeding 0.30, and therefore were not included in the factor solution.

A total of three dimensions meaningfully described the Present Ideal data (Table 3). The first factor, labeled "College Image," reflected a set of variables which described environmental features emanating from students' impressions of what a college should be like. The factor included items such as "fosters independence," "is highly organized," and "is distinctive/different from other colleges," and closely resembled Pace and Stern's impression-based definition of a college environment's "perceived climate".[49]

Table 3: Item Loadings for Present Ideal Factors

Item	Factor 1: College Image	Loading
Is supportive		0.68
Is people-oriented		0.65
Is highly organized		0.63
Fosters independence		0.62
Is effort-oriented		0.61
Allows you time to yourself		0.60
Fosters social responsibility		0.60
Is academically demanding		0.56
Fosters social interactions		0.56
Demands good performance from you		0.53
Fosters friendships in the classroom		0.53
Fosters friendships in residence halls		0.49
Lead an active social life		0.48
Identify yourself as a [college name] student		0.40
Is distinctive/different from other college environments		0.38
Is competitive		0.35

Item	Factor 2: Student Experience	Loading
Speak before a group of your peers about a topic important to you		0.72
Attend a professor's presentation as a part of a faculty lecture series		0.60
Imagine yourself president of a club or organization		0.60
Chat with an instructor outside of class		0.60
Share ideas/Speak up in class		0.59
Become active in political groups on campus		0.59
Eat dinner with a professor		0.58
Volunteer in the local community		0.56
Go to a subsidized cultural event (such as the symphony or theater)		0.51
Vote in a campus election		0.50
Go on a retreat		0.42
Encourages volunteering to meet local community needs		0.36
Seek out your advisor for advice		0.35

Item	Factor 3: Traditional-Catholic	Loading
Go to mass/church with your friends		0.66

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

Emphasizes a Catholic/Jesuit mission	0.62
Emphasizes a single set of values throughout the university	0.52
Attend a Pep-Rally before a game	0.50
Is rule-oriented	0.48
Go to a planned social event in your residence hall	0.46
Is team-oriented	0.44
Is grade-oriented	0.40

Note: Displayed items include only Present Ideal items with factor loadings>0.30. For factors 1, 2 and 3, respectively: Chronbach's alpha=0.85, 0.83, and 0.78; eigenvalue=8.19, 3.10, and 2.27.

The second factor represented respondents' preferences regarding academic and social experiences. Included in this dimension were "action" items, rather than "image" items like those comprising the first factor. This factor was labeled "Student Experience" and included items such as "share ideas/speak up in class," "volunteer in the local community," and "seek out your advisor for advice." This factor closely resembled Astin's behaviorally-based definition of "college environment".[9,10,12]

The third and final dimension combined both "image" and "behavior" items to reflect what seem to be respondents' preferences for a conservative college experience. Traditional college attributes as well as features related to religiously affiliated schools comprised this factor labeled "Traditional-Catholic" and included items such as "emphasizes a single set of values throughout the university," "is rule-oriented," and "attend a pep-rally before a big game." Correlations among these three college dimensions were positive (College Image and Student Experience, r=0.45; College Image and Traditional-Catholic, r=0.40; and Student Experience and Traditional-Catholic, r=0.41, all ps<0.01).

To test the stability of this three-factor solution, a principal components factor analysis with varimax rotation also was performed on the Anticipatory Ideal items. This factor solution was then compared to the factor structure resulting from the Present Ideal data using Coefficients of Congruence (COC). Results comparing the two three-factor solutions revealed that the underlying factor structures of the two data sets were highly congruent. The highest COC was between summer and spring Student Experience dimensions (0.96), with the College Image dimension also showing comparable factor structures (0.93). The Traditional-Catholic dimensions were least congruent, but the degree of factor correspondence was still high (0.70).

Because PE fit scores involve the difference between commensurate "Ideal" and "Real" scores, only one of these two factor solutions were used to compute the discrepancy scores. The dimensions resulting from the posttest data were chosen for two reasons. First, although the two sets of three-factor solutions displayed comparable internal consistencies (Cronbach alphas=0.84, 0.83, 0.81 for summer factors *vs*. Cronbach alphas=0.85, 0.83, 0.71 for respective spring factors), the Present Ideal factors account for a larger percentage of the variance (36.5% *vs*. 34.8%) in their respective data set.

The second reason for choosing the Present Ideal factors involved students' degree of familiarity with their college setting. After having experienced a college environment for nine months, students should be better able to describe their college preferences than before starting school. Spring factors thus served as the basis from which PE fit scores were derived.

Student Involvement and PE Fit. To test the prediction that highly involved freshmen would possess more congruent PE fit levels, correlations were calculated between the involvement index and eight PE fit indicators (the involvement index was derived using 16 Student

Experience Real items: thus, the four congruence measures related to the Student Experience dimension were not included in these analyses due to the violation of the independence assumption). Supporting predictions, involvement level was significantly correlated with five of eight PE fit indicators (Table 4). However, although statistically significant, involvement accounted for little of the variance in any of the congruence measures: R^2 ranged from 2.4% for Anticipatory Subjective College Image, to 4.3% for Anticipatory Objective College Image. Degree of college involvement was related to three of four Subjective PE fit indicators and two of four Objective PE fit indicators. High involvement was associated with more congruent Subjective PE fit. However, contrary to predictions, highly involved freshmen were more likely to possess less congruent Objective PE fit levels.

Table 4: Correlations Between PE Fit Scores and Student Involvement

Objective PE Fit[a]	r	r^2	Effect Size (d)
College Image Fit (A)[c]	0.207**	0.043	0.424
College Image Fit (P)[d]	0.188*	0.035	0.381
Traditional-Catholic Fit (A)	0.064	0.004	0.127
Traditional-Catholic Fit (P)	0.002	0.000	0.004
Subjective PE Fit[b]			
College Image Fit (A)	-0.153*	0.024	0.314
College Image Fit (P)	-0.176*	0.031	0.358
Traditional-Catholic Fit (A)	-0.021	0.000	0.042
Traditional-Catholic Fit (P)	-0.170*	0.029	0.346

Note: Student Experience PE fit scores were excluded from analyses due to the independence assumption violation with the involvement variable. All analyses were performed with and without involvement items in the PE fit indicators: significance levels did not change. A single asterisk (*) indicates p<0.05 at the generalized (per-comparison) criterion, and double asterisks (**) indicate p<0.05 at the experimentwise criterion.[58] Derived from: [a]Individual "Ideals" and mean of respondents' "Reals"; [b]Individual "Ideals" and Individual "Reals"; [c]summer "Ideals" and spring "Reals"; and [d]spring "Ideals" and spring "Reals."

PE Fit and Retention. To test the prediction that PE fit scores would help to distinguish returners from dropouts, linear DA and CTA were performed. PE fit scores served as attributes, and return status as the class variable. None of the 12 PE fit variables (four fit indices across each of three dimensions: Student Image, College Behavior, Traditional-Catholic) qualified for DA or CTA analysis.

Additional Analyses

Because the attribute set outlined above did not adequately classify returners from dropouts, further analyses were performed in which several predictor variables were used. CTA and stepwise DA were performed. For CTA all single-item Ideal and Real variables were used, as was the involvement index and the Ideal, Real, and PE fit factors. For DA only the set of single item variables was used because the inclusion of construct-level variables would violate the independence assumption underlying this procedure.

Stepwise DA Model. The DA resulted in a linear model that distinguished returners from dropouts (canonical R=0.39, $\chi^2(7)$=46.53, p< 0.0001). Seven predictors combined to yield a significant discriminant function after 7 steps (Table 5). The loading matrix of correlations between predictors and the discriminant function suggest that together, three variables discriminated respondents on the basis of return status (predictors having loadings less than 0.50 were not interpreted[62]).

The best predictors for distinguishing returners from attritors assessed how organized and how competitive respondents perceived their college environment to be at the end of their freshman year. Dropouts described their college environment as more organized than returners (means=5.18 *vs.* 4.87, respectively), but less competitive than returners (means=4.65 vs. 5.52, respectively). One posttest preference rating also contributed to the classification model. Returners and dropouts differed in the degree to which they wanted to identify themselves as members of their college community, with returners possessing stronger desires (means=5.88 vs. 5.17, respectively).

Table 5: Standardized Canonical Discriminant Function Coefficients for Stepwise DA

Step	Item[a]	Coefficient[b]	Wilks Lambda
1	competitive environment (Real)	0.59	0.96
2	fosters risk-taking (Ideal)	0.31	0.94
3	highly organized college (Real)	-0.57	0.91
4	identify self as college member (Ideal)	0.53	0.89
5	team-oriented college (Ideal)	-0.32	0.87
6	fosters risk-taking (Real)	0.39	0.86
7	attend pep-rally (Ideal)	-0.33	0.85

Note: [a]All items included in the solution were assessed during the spring of students' freshman year. No summer (i.e., "anticipatory") items significantly contributed to the discriminant function. [b]Standardized canonical discriminant function coefficients.

Although the model classified almost all of the returners correctly, it performed poorly in its classification of dropouts. Group PACs for returners and attritors were 97.2% and 17.9%, respectively. The mean PAC across both groups of returners and dropouts was 57.6% (Table 6).

Table 6: DA Classification Results

Actual		Predicted Group		
Group	N	Dropouts	Returners	
Dropouts	39	7	32	7.9%
Returners	324	9	315	97.2%
		43.8%	90.8%	

Note: ESS=5.1 (weak effect).

CTA Model. CTA yielded a different solution, outperforming DA especially with respect to classifying attritors. The CTA model correctly classified 84% of dropouts and 85% of returners, with an overall mean PAC of 84.5% (see Table 7).

Table 7: CTA Classification Results

Actual		Predicted Group		
Group	N	Dropouts	Returners	
Dropouts	31	26	5	83.9%
Returners	317	48	269	84.9%
		35.1%	98.2%	

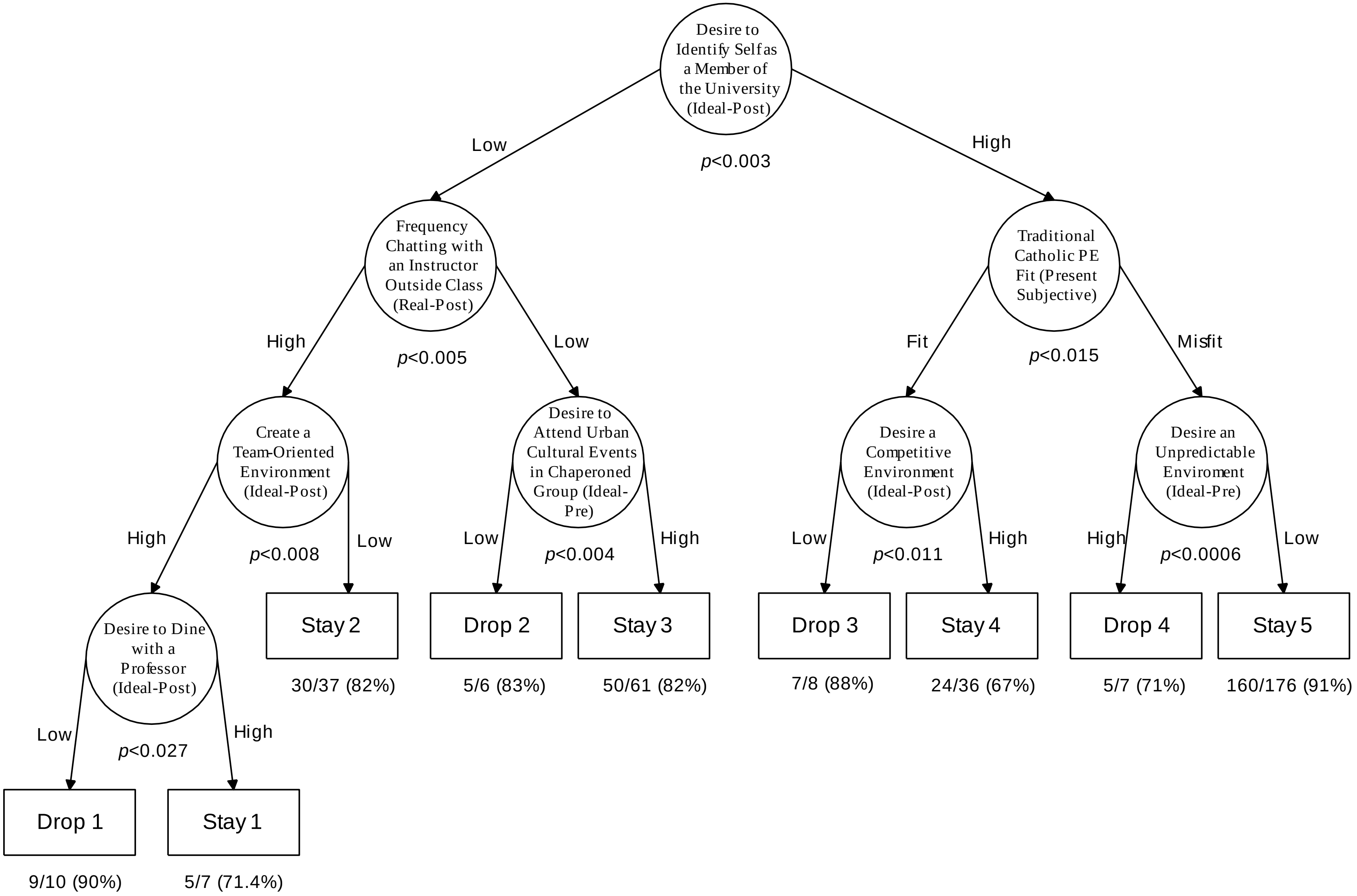
Note: ESS=68.8 (relatively strong effect).

Presented in Figure 1, CTA also revealed that different groups of dropouts left, and different groups of returners stayed, for different reasons. The CTA model revealed *four* clusters of dropouts and *five* clusters of returners.

Four common pathways through the measured attributes described the participants who did not return to the university for their sophomore year. As seen, dropouts on Path 1 ("Drop 1" in Figure 1), "Small Dose Participators" possessed little desire to identify themselves as a university member (≤0.5), chatted frequently with instructors outside of class (>3.5), desired a team-oriented environment (>5.5), but did not desire to dine with instructors (≤4.5).

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

Figure 1: CTA Model for Classifying Dropouts and Returners


Desire to Identify Self as a Member of the University (Ideal-Post)
p<0.003
Low
High
Frequency Chatting with an Instructor Outside Class (Real-Post)
p<0.005
High
Low
Traditional Catholic PE Fit (Present Subjective)
p<0.015
Fit
Misfit
Create a Team-Oriented Environment (Ideal-Post)
p<0.008
High
Low
Desire to Attend Urban Cultural Events in Chaperoned Group (Ideal-Pre)
p<0.004
Low
High
Desire a Competitive Environment (Ideal-Post)
p<0.011
Low
High
Desire an Unpredictable Enviroment (Ideal-Pre)
p<0.0006
High
Low
Desire to Dine with a Professor (Ideal-Post)
p<0.027
Low
High
Drop 1
9/10 (90%)
Stay 1
5/7 (71.4%)
Stay 2
30/37 (82%)
Drop 2
5/6 (83%)
Stay 3
50/61 (82%)
Drop 3
7/8 (88%)
Stay 4
24/36 (67%)
Drop 4
5/7 (71%)
Stay 5
160/176 (91%)

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

Dropouts on Path 2 (Drop 2), "Involvement Avoiders," also possessed little desire to identify themselves as a university member (≤5.5), but rarely chatted with their instructors outside of class (≤3.5). "Involvement Avoiders" also indicated during summer registration that they were not interested in attending urban cultural events in a chaperoned group (≤4.5).

Dropouts on Path 3 (Drop 3), "Congruent Non-Competitors," differed from the first two clusters. These students *did* want to identify themselves as a university member (>5.5). Although this cluster of dropouts possessed strong Traditional-Catholic PE fit (≤0.19), they did not desire a competitive college environment (≤5.5).

The final set of Path 4 dropouts (Drop 4), "Incongruent Thrill-Seekers," were similar to those on Path 3 in that they desired to identify themselves as university members. However, these attritors revealed incongruent Traditional-Catholic PE fit levels (>0.19), and possessed pre-enrollment desires to attend a college with an unpredictable environment (>5.5).

The PACs for Paths 1, 2, 3, and 4 classifying dropouts were 90% (9/10), 83.3% (5/6), and 88% (7/8), and 71% (5/7), respectively.

Five common pathways were used to classify students who chose to return to the university as sophomores.

Path 1 returners (Stay 1), "Large-Dose Participants," possessed little desire to identify themselves as a university member (≤5.5), chatted frequently with their instructors outside of class (>3.5), desired a team-oriented environment (>5.5), and also desired to dine with their instructors (>4.5).

Returners on Path 2 (stay 2), "Academically Involved Independents," were similar to those on Path 1 in that they possessed little desire to identify themselves as a university member (≤5.5) and chatted frequently with their instructors outside of class (>3.5). However, they differed from "Large Dose Participants" in that they did *not* desire a team-oriented college environment (≤5.5).

Returners on Path 3 (Stay 3), "Culture Seekers," also possessed little desire to identify themselves as a university member (≤5.5), and indicated that they did *not* often chat with their instructors outside of class (≤3.5). However, "Culture Seekers" indicated during summer reistration sessions a desire to attend urban cultural events with classmates and faculty members (>4.5).

Returners on Path 4 (stay 4), "Congruent Competitors," *did* want to identify themselves as a university member (>5.5), possessed good Traditional-Catholic PE fit (≤5.5), and desired a competitive college environment (>5.5).

Finally, returners on Path 5 (Stay 5), "Incongruent Routine-Seekers," wanted to identify themselves as university members (>5.5), possessed little Traditional-Catholic PE fit (>0.19), and did not desire a unpredictable environment (≤5.5).

The PACs for these five pathways were 71.4% (5/7); 81.8% (30/37); 82.0% (50/61); 66.7% (24/36); and 90.9% (160/176), respectively.

Objective vs. Subjective PE Fit. It was predicted that Objective PE fit scores would be more closely related to involvement, and would better predict students' return status, than Subjective PE fit scores. Results did not support these predictions. No Objective PE fit score contributed to the understanding of student retention and attrition. Only one subjectively derived congruence measure (Present Traditional-Catholic PE Fit) assisted in classifying returners and attritors, but only for the expanded ODA-CTA model.

A surprising pattern emerged when the involvement index was correlated with both Subjective and Objective PE fit indicators. The relationship between Subjective PE fit and involvement was in the opposite direction of the relationship between Objective PE fit and involvement. As predicted, highly involved students tended to have more congruent subjectively derived PE fit scores. However, contrary to predictions highly involved students tended to have more incongruent PE fit scores when this variable was computed using the *mean* of all respondents' Real scores. Thus, it appears that the direction of the relationship between student involvement and PE congruence may be contingent upon how the PE fit scores were derived. This unexpected relationship might best be explained by measurement artifacts, rather than true effects (discussed below).

Anticipatory vs. Present PE Fit. It was hypothesized that Present PE fit scores would better predict return status and be more closely associated with students' involvement levels than Anticipatory PE fit scores. The logic behind this prediction was that first-year students would have a better understanding of what they desired in a university after having experienced college life for two semesters.

Results revealed that Present congruence measures were only slightly better than Anticipatory congruence measures with respect to involvement and return status. Three *Present* PE fit scores, but only two *Anticipatory* PE fit scores, were associated with students' level of participation in college activities (see Table 4). With respect to return status, the only congruence measure that was included in any of the classification models was Present Subjective Traditional-Catholic, derived from posttest items (see Figure 1).

PE Fit vs. P and E Variables. It was hypothesized that PE fit difference scores would outperform P (Ideal) and E (Real) scores alone. Results did not support this prediction. Student involvement was more highly correlated with the P factors and E factors than with the PE fit factors (see Table 8). To test the relationship between P and E dimensions and retention, MANOVAs and discriminant analyses were performed, using the six Ideal (P) and three Real (E) factors in place of the PE Fit indicators to test for group differences between returners and non-returners. P and E factors did not improve the accuracy in classifying freshman returners from dropouts.

Table 8: Correlations Between Student Involvement and Ideal (P) and Real (E) Factors

Ideal (P) Dimension	r	r^2	Effect Size (d)
College Image (A)[a]	0.250^{**}	0.063	0.519
College Image (P)[b]	0.210^{**}	0.044	0.429
Student Experience (A)	0.348^{**}	0.121	0.742
Student Experience (P)	0.439^{**}	0.190	0.969
Traditional-Catholic (A)	0.357^{**}	0.127	0.763
Traditional-Catholic (P)	0.401^{**}	0.161	0.876
Real (E) Dimension			
College Image	0.293^{**}	0.086	0.613
Traditional-Catholic	0.539^{**}	0.291	1.280

Note: The Student Experience Real factor was excluded from these analyses due to the independence assumption violation between this variable and the involvement attribute. All analyses were performed with and without involvement items in the Real and Ideal factors: significance levels did not change. Double asterisks (**) indicate $p<0.05$ at the experimentwise criterion.[58] [a]Anticipatory (derived from summer items). [b]Present (derived from spring items).

Additionally, three CTA and three DA procedures were run—each containing the two P (Anticipatory and Present) and one E factor corresponding to the three college dimensions (College Image, Student Experience, Traditional-Catholic). Neither CTA nor DA procedures generated a classification solution with respect to return status when Real and Ideal factors replaced PE fit factors. However, as discussed above, when ancillary analyses expanded discriminant procedures to include single-item P and E variables, preferences and perceptions outperformed PE fit scores in distinguishing freshman returners from non-returners.

Discussion

The PE Fit literature has linked student-college congruence to a host of desirable educational variables (e.g., academic achievement, perceived competency), yet has virtually ignored attrition and retention variables. The present study attempted to merge the separate retention and PE Fit paradigms, by investigating the relationships among involvement, student-college congruence, and withdrawal decisions for one population of college freshmen over a period of one year.

Although most PE fit indicators were linked to student involvement levels, the correlations between separate P and E factors and involvement were stronger. The variable most highly correlated with student involvement measured students' perceptions (E) regarding the Traditional-Catholic nature of their college. Students who believed that the "press" of their college environment emphasized religious values, grades, and school rules, were most likely to participate in campus activities. Highly involved students also seemed to have *desired* these characteristics, since the variable correlated next highly with involvement was the Traditional-Catholic P factor.

It appears that the relationship between involvement and student-college congruence was contingent upon the way that the PE Fit indicator was derived. When subjective congruence scores were used, the relationship between these PE fit indicators and involvement was as predicted; the greater students' level of involvement, the greater the match between students' preferences and perceptions. However, when objective congruence scores were used, greater student participation resulted in more discrepant congruence scores.

One explanation for this change in direction may lie in the relationship between involvement and the Ideal (P) component of the PE fit score. By using the average "Real" rating across all respondents to derive Objective PE fit scores, any variability related to the E component of congruence was lost. Thus, variability in objectively derived PE fit scores was due to differences in student preferences (P items) only. This was not the case with subjectively derived congruence scores in which both P and E responses were free to vary.

In this study, involvement was, in fact, positively correlated with all six Ideal ratings (*r*s ranged from 0.21 to 0.44, all *p*s<0.01, mean effect size=0.72). Thus, the relationship between Objective PE fit and involvement may simply have represented a measurement artifact. Because students with the highest college standards (P ratings) were likely to have been the same students who frequently participated in college activities, it was made to appear that greater participation was linked to greater (objective) incongruence.

This is consistent with Edwards' assertion that PE fit measures must allow both the P and E components to contribute to the total variability.[54,55] When only one component is permitted to vary, Edwards claims that PE fit is no longer being assessed. Since this may have been the case in the present study, all analyses using Objective PE fit scores should be rendered suspect.

So, how is it that several congruence researchers have demonstrated that Objective PE fit was superior to Subjective PE fit in their studies? The answer may simply be they have not. A closer examination of these studies revealed that measurement problems suggested by Edwards may also explain these findings as well. For instance, Tracey and Sherry studied the relationship between Objective PE fit, Subjective PE fit, and student distress.[19] They found that objective measures of congruence were more highly correlated with distress than Subjective PE fit measures. However, this was only the case when students' Ideal (P) ratings *also* were negatively correlated with distress. When distress and college preferences were positively related, Subjective PE fit scores were more highly correlated with college distress than Objective PE fit. Thus, Tracey and Sherry's findings may suffer from the same problems as those found in the present study.

Although many studies suggest that the congruence between preferences (Ps) and perceptions (Es) is superior to either component alone in predicting behavior, studies do exist that refute this claim.[63,64] The present study might be included in this group since no classification model differentiated returners from attritors when psychometrically constructed PE fit indicators were used as predictors.

When exploratory analyses were expanded to include student preferences and perceptions measured at the individual item level, the present study supports the notion that P and E components may be more important in classifying returners from attritors than congruence measures that combined these components. Only one of the 12 PE fit indicators significantly classified returners from non-returners, and this was only for the expanded CTA model. Present Subjective Traditional-Catholic PE fit scores assisted in the classification of two clusters of dropouts and two clusters of returners. No congruence score was included in the traditional discriminant function. All other variables in both models were either P or E items.

Ideal and Real factors differed in their contribution to the classification models. Although the DA solution was comprised of both P and E variables, the CTA model was comprised almost completely of P variables. The only E item in the classification tree assessed the frequency of student-teacher interactions outside of the classroom.

The time of the year in which P variables were assessed also made a difference. The majority of the DA and the CTA items comprising these classification solutions contained responses that were assessed in the spring of respondents' freshman year. Spring preferences were better predictors of college retention than previous summer preferences perhaps because in their second semester, students did not have to speculate about aspects of college life they had yet to experience.

The CTA model may be consistent with Tinto's theory that links freshman involvement with retention.[3] According to Tinto, different types of involvement are critical at different points in time. Upon arriving to campus, the social sphere is critical to students, as they seek to find a support network. However, the focus soon switches to the academic sphere once freshmen begin their second month of college. After the first few weeks on campus, classrooms become first year students' "gateways to [future] involvement" in other social and academic arenas (p. 134). Here, fledgling students learn to engage in both formal and informal activities with both faculty and peers. Thus, according to Tinto, the quality of the learning experience (e.g., contact with, and helpfulness of, faculty and classmates) is not freshmen's first priority when they arrive on campus, but soon becomes the crucial predictor of their overall satisfaction with the college experience.

The left side of the CTA model (see Figure 1) seemed to reflect this emphasis on informal academically-oriented interactions. All behaviorally-based items in the CTA model involved informal interactions with faculty members. Both brief (chat with instructor) and extended (dine with professor; attend a cultural event) faculty interactions helped to distinguish returners from non-returners. Thus, it appears that student-teacher interactions may have been more important for enhancing freshman retention than purely social peer-only interactions.

Although the left side of the CTA model contained mostly behaviorally-based variables, the right side of the tree contained image-based preferences in addition to a Traditional-Catholic congruence variable. This side, then, reflected retention decisions based on the value-system of one's institution (Traditional-Catholic congruence) as well as the degree of thrill-seeking "press" that was thought to exist on campus. Interestingly, this "thrill-seeking" component was similar to the most important items in the traditional DA classification model. In that model, perceptions regarding how "competitive" and "organized" their college was contributed greatly to the differentiation of dropouts from attritors. However, unlike the CTA model, no behaviorally-based items were included in the DA model. These findings emphasize one of CTA's major strengths. Clusters of respond-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

ents that would not have been found with one linear discriminant function, were revealed with CTA.

Although results from these models are interesting, three important limitations must be noted. First, both the CTA and the DA classification solutions yielding a solution on the basis of retention were exploratory. Only after the psychometrically derived constructs were unable to distinguish attritors from returners, were individual "ideal" and "real" items included in the analyses.

Second, although the CTA model held up under LOO (jackknife) tests for overfitting, neither model was able to be cross-validated using a training sample, for which group membership was known, and a holdout sample, for which group membership was predicted, and later compared to reality. Although the pretest sample size was large enough to divide, the posttest sample size was not. Future studies that intend to follow freshmen students longitudinally should focus on increasing the response rate in spring phases of data collection. Special efforts also should be made to encourage commuting freshmen and freshmen who are struggling academically to participate, since these groups were somewhat under-represented in this study.

Finally, neither classification model was able to classify students on the basis of return status better than simply relying on the base rates. Because the vast majority of freshmen did return to campus for their sophomore year, simply using the classification rule, "Predict all students to return" would have resulted in a classification accuracy of close to 90%. Neither the DA model nor the CTA model could beat this rule.

However, it is important to note that the beating the base rates may not be a relevant criterion with which to base the adequacy of the classification models in this study. Because exploring the perceptions and behaviors of students most at-risk of dropping out is of utmost importance to college administrators, finding the model that most accurately classifies this "vulnerable" group may be more important than finding the model that most accurately classifies all students (dropouts and returners). The expanded CTA model was able to do just that.

The relationship between PE fit and retention might have been stronger if the reasons driving students' decisions to exit or remain in their academic setting were assessed. Factors impacting one's decision to leave college are both numerous and complex. Researchers have discussed several kinds of dropouts, including temporary or permanent; voluntary or involuntary; and attrition for academic or social reasons.[3,7,65] Additionally, leaving college may not necessarily result in negative outcomes if, for instance, one's experience with a university results in highly aversive outcomes, and better options exist elsewhere.[66] It may be that PE fit levels impact only certain kinds of attrition.

Future researchers might want to fine-tune the return-status variable to better assist college personnel in stream-lining their retention efforts. Reasons for dropping could be assessed using an exit interview or written questionnaire at the time of departure. An interesting and potentially important future study could combine the use of exit interviews with CTA techniques to better understand freshman attrition. If reasons for leaving differed among the different "clusters" of attritors, CTA models could be used as diagnostic tools for college admissions directors and administrators.

There are four important findings that may be of interest for those in the business of enhancing freshman involvement and retention. First, it may be important to encourage both students and faculty to seek each other out when they are not in the classroom. Behaviorally-based items that helped to distinguish returners from non-returners included, not peer-interactions, but different types of faculty-student interactions.

Second, in addition to desires for interactions with faculty members, students' *images* of their college are also important to students. The

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

value system that a college promotes, as well as the competitiveness and predictability of its climate, all appear to be important components in the understanding of student retention. These factors may help to impact how much of a college "member" students feel they are.

Third, college preferences may be more important than college perceptions in classifying freshmen on the basis of return status. It also may matter when researchers document these college desires. If students really do not know what they want in a college until they have occupied it for some time, administrators may want to wait until the spring of students' freshman year to assess college preferences and perceptions.

Finally, there appears to be specific statistical analysis which is ideally suited for the task of understanding college student attrition. CTA was far superior in classifying dropouts than traditional discriminant analysis techniques (84% *vs*. 18%). This finding is important since attritors comprise the group about which college administrators are most concerned. Additionally, CTA was able to identify unique clusters of dropouts (and returners) implying that, indeed, students choose to leave their colleges for a plethora of reasons. This ability to refine our understanding of college attrition may be an important first step in actually reducing the number of students who choose this route.

References

[1]Schneider, B. (1987). E = f (P, B): The road to a radical approach to person-environment fit. *Journal of Vocational Behavior, 31,* 353-361.

[2]Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45,* 89-125.

[3]Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition.* Chicago: University of Chicago Press.

[4]Chapman, D. W., & Pascarella, E. T. (1983). Predictors of academic and social integration of college students. *Research in Higher Education, 19,* 295-322.

[5]Moos, R. H. (1976). *The human context: Environmental determinants of behaviors.* New York: Wiley.

[6]Moos, R. H. (1979). *Evaluating educational environments.* San Francisco: Jossey-Hill.

[7]Astin, A. W. (1975*). Preventing students from dropping out.* San Francisco: Jossey-Bass Publishers.

[8]Astin, A. W. (1985). Involvement: The cornerstone of excellence. *Changes,* 35-39.

[9]Astin, A. W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education.* NY: American Council on Education: Macmillan.

[10]Astin, A. W. (1993). *What matters in college?* San Francisco: Jossey-Bass Publishers.

[11]Stern, G. (1970). *People in context: Measuring person-environment congruence in education and industry.* New York: Wiley.

[12]Astin, A. W. (1968). *The college environment.* Washington, D.C.: American Council on Education.

[13]Astin, A. W. (1977). *Four critical years.* San Francisco: Jossey-Bass Publishers.

[14]Cook, J. R. (1987). Anticipatory person-environment fit as a predictor of college student health and adjustment. *Journal of College Student Personnel, 28,* 394-398.

[15]Nielsen, H. D., & Moos, R. H. (1977). Student-environment interaction in the development of physical symptoms. *Research in Higher Education, 6,* 139-156.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3

[16]Janosik, S., Creamer, D. G., & Cross, L. H. (1988). The relationship of residence halls' student-environment fit and sense of competence. *Journal of College Student Development, 29,* 320-326.

[17]Pervin, L. A. (1967). Satisfaction and perceived self-environment similarity: A semantic differential study of student-college interaction. *Journal of Personality, 35,* 623-634.

[18]Reuterfors, D. L., Schneider, L. J., & Overton, T. D. (1979). Academic achievement: An examination of Holland's congruency, consistency, and differentiation predictions. *Journal of Vocational Behavior, 14,* 181-189.

[19]Tracey, T. J., & Sherry, P. (1984). College distress as a function of person-environment fit. *Journal of College Student Personnel, 25,* 436-442.

[20]Hadley, T., & Graham, J. W. (1987). The influence of cognitive development on perceptions of environmental press. *Journal of College Student Personnel, 28,* 388-394.

[21]Eagan, A. E., & Walsh, W. B.(1995). Person-environment congruence and coping strategies. *The Career Development Quarterly, 43,* 246-256.

[22]Sergent, M. T., & Sedlacek, W. E. (1990). Volunteer motivation across student organizations: A test of person-environment fit theory. *Journal of College Student Development, 31,* 255-261.

[23]Kulka, R. A., Klingel, D. M., & Mann, D. W. (1980). School crime and disruption as a function of student-school fit: An empirical assessment. *Journal of Youth and Adolescence, 9,* 353-370.

[24]Treadway, D. M. (1979). Use of campus-wide ecosystem surveys to monitor a changing institution. In L. A. Huebner (Ed.). *Redesigning campus environments: New directions for student services* (No. 8). San Francisco: Jossey-Bass.

[25]Pervin, L. A., & Rubin, D. B. (1967). Student dissatisfaction with college and the college dropout: A transactional approach. *The Journal of Social Psychology, 72,* 285-295.

[26]Fox, R. N. (1986). Application of a conceptual model of college withdrawal to disadvantaged students. *American Educational Research Journal, 23,* 415-424.

[27]Caplan, R. D. (1987). Person-environment fit theory and organizations: Commensurate dimensions, time perspectives, and mechanisms. *Journal of Vocational Behavior, 31,* 248-267.

[28]Osipow, S. H. (1987). Applying person-environment theory to vocational behavior. *Journal of Vocational Behavior, 31,* 333-336.

[29]Schneider, B. (1987). The people make the place. *Personnel Psychology, 40,* 437-453.

[30]Chatman, J. A. (1989). Improving interactional organizational research: A model of person-organization fit. *Academy of Management Review, 14,* 333-349.

[31]French, J. R. P., Jr., Rodgers, W., & Cobb, S. (1974). Adjustment as person-environment fit. In G. V. Coelho, D. A. Hamburg, & J. E. Adams (Eds.), *Coping and Adaptation.* New York: Basic Books.

[32]Rounds, J. B., Dawis, R. V., & Lofquist, L. H. (1987). Measurement of person-environment fit and prediction of satisfaction in the theory of work adjustment. *Journal of Vocational Behavior, 31,* 297-318.

[33]Barker, R. G. (1963). On the nature of the environment. *Journal of Social Issues, 19,* 26-27.

[34]Wicker, A. W. (1972). Processes which mediate behavior-environment congruence. *Behavioral Science, 17*, 265-277.

[35]MacDonald, N., & Ronayne, T. (1989). Jobs and their environments: The psychological impact of work and noise. *The Irish Journal of Psychology, 10,* 39-55.

[36]Nehrke, M. F., Morganti, J. B., Cohen, S. H., Hulicka, I. M., Whitbourne, S. K., Turner, R. R., & Cataldo, J. F. (1984). Differences in person-environment congruence between microenvironments. *Canadian Journal on Aging, 3,* 117-132.

[37]Moos, R. H., & Gerst, M. (1974). *University residence environment scale manual.* Palo Alto, CA: Consulting Psychologists Press.

[38]Huebner, L. A. (1975). *An ecological assessment: Person-environment fit.* Unpublished doctoral dissertation. Colorado State University.

[39]Salamone, P. R., & Daughton, S. (1984). Assessing work environments for career counseling. *Vocational Guidance Quarterly, 33,* 45-54.

[40]Moos, R. H., & Trickett, E. J. (1974). *Classroom environment scale manual.* Palo Alto, CA: Consulting Psychologists Press.

[41]Evans, N. J. (1983). Environmental assessment: Current practices and future directions. *Journal of College Student Personnel, 24,* 293-299.

[42]Pervin, L. A. (1967). A twenty-college study of student x college interaction using TAPE (Transactional Analysis of Personality and Environment): Rationale, reliability, and validity. *Journal of Educational Psychology, 58,* 290-302.

[43]Boxx, W. R., Odom, R. Y., & Dunn, M. G. (1991). Organizational values and value congruency and their impact on satisfaction, commitment, and cohesion. *Public Personnel Management, 20*, 195-205.

[44]Caplan, R. D., & Van Harrison, R. (1993). Person-environment fit theory: Some history, recent developments, and future directions. *Journal of Social Issues, 49,* 253-275.

[45]Kaldenberg, D. O., & Becker, B. W. (1992). Workload and psychological strain: A test of the French, Rodgers, and Cobb hypothesis. *Journal of Organizational Behaviors, 13,* 617-624.

[46]O'Reilly, C. A., III, Chatman, J., Y Caldwell, D. F. (1991). People and organizational culture: A profile comparison approach to assessing person-organizational fit. *Academy of Management Journal, 34,* 487-516.

[47]Posner, B. Z. (1992). Person-organization values congruence: No support for individual differences as a moderating influence. *Human Relations*, *45*, 351-361.

[48]Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

[49]Pace, C. R., & Stern, G. G. (1958). An approach to the measurements of psychological characteristics of college environments. *Journal of Educational Psychology, 49*, 269-277.

[50]Holland, J. L. (1987). Some speculation about the investigation of person-environment transactions. *Journal of Vocational Behavior, 31,* 337-340.

[51]Spokane, A. R., & Derby, D. P. (1979). Congruence, personality pattern, and satisfaction in college women. *Journal of Vocational Behavior, 15,* 36-42.

[52]Pervin, L. A. (1968). Performance and satisfaction as a function of individual-environment fit. *Psychological Bulletin, 69,* 56-68.

[53]Astin, A. W., & Panos, R. J. (1969). *The educational and vocational development of students.* Washington, D.C.: American Council on Education.

[54]Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology, 46,* 641-665.

[55]Edwards, J. R., & Perry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36,* 1577-1613.

[56]Finney, H. C. (1967). *Development and change of political liberalism among Berkeley undergraduates.* Unpublished Doctoral Dissertation, University of California, Berkeley.

[57]Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation.* Chicago: Rand-McNally.

[58]Yarnold, P.R., Soltysik, R.C. (2005). *Optimal data analysis: A guidebook with Software for Windows*. Washington DC: APA Books.

[59]Yarnold, P. R. (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via uniODA. *Educational and Psychological Measurement, 56,* 656-667.

[61]Silva, A. P. D., & Stam, A. (1995). Discriminant analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics.* (pp. 277-318). Washington, D.C.: American Psychological Association.

[62]Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). Northridge, CA: HarperCollins.

[63]Bretz, R. D., Jr., Ash, R. A., & Dreher, G. F. (1989). Do people make the place? An examination of the Attraction-Selection-Attrition hypothesis. *Personnel Psychology, 42,* 561-581.

[64]Witt, P. H., & Handal, P. J. (1984). Person-Environment fit: Is satisfaction predicted by congruency, environment, or personality? *Journal of College Student Personnel, 25,* 503-508.

[65]Jacoby, B. (1989). *The student as commuter: Developing a comprehensive institutional response.* Washington, D.C.: The George Washington University.

[66]Louis, M. R. (1990). Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings. *Administrative Science Quarterly, 25,* 226-251.

Author Notes

This research derives from Jennifer Howard Smith's doctoral dissertation under the direction of Emil J. Posavac, with assistance from dissertation committee members John D. Edwards, R. Scott Tindale and Eugene B. Zechmeister, and statistical consultation from Paul R. Yarnold and Fred B. Bryant. eMail: Jennifer@appliedresearchsolutionsinc.com.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Tracing Prospective Profiles of Juvenile Delinquency and Non-Delinquency: An Optimal Classification Tree Analysis

Hideo Suzuki, Ph.D., Fred B. Bryant, Ph.D., and John D. Edwards, Ph.D.
Loyola University Chicago

This study explored multiple variables that influence the development of juvenile delinquency. Two datasets of the National Youth Survey, a longitudinal study of delinquency and drug use among youths from 1976 and 1978, were used: 166 predictors were selected from the 1976 dataset, and later self-reported delinquency was selected from the 1978 dataset. Optimal data analysis was then used to construct a hierarchical classification tree model tracing the causal roots of juvenile delinquency and non-delinquency. Five attributes entered the final model and provided 70.37% overall classification accuracy: prior self-reported delinquency, exposure to peer delinquency, exposure to peer alcohol use, attitudes toward marijuana use, and grade level in school. Prior self-reported delinquency was the strongest predictor of later juvenile delinquency. These results highlight seven distinct profiles of juvenile delinquency and non-delinquency: lay delinquency, unexposed chronic delinquency, exposed chronic delinquency, unexposed non-delinquency, exposed non-delinquency, unexposed reformation, and exposed reformation.

The Federal Bureau of Investigation (FBI) reported that more than 1.5 million juveniles under the age of 18 were arrested in 2003, suggesting that about 16.3% of all individuals arrested were juveniles.[1] As a result, youth violent crime is often considered to be a major problem in the United States.[2] In addition, research indicates that a delinquent criminal career increases the potential to commit crime in adulthood.[3-11] For these reasons, juvenile delinquency and its causes have been major topics in the study of crime.[12]

Some scholars have focused on situational factors as underlying determinants of criminal behavior.[13-16] For example, because crime rates are generally high in areas of poverty, it has been argued that poor socialization (i.e., failure to teach skills to achieve middle-class success) provided by lower-class parents is a predictor of delinquency.[17] With poor socialization, lower-class adolescents feel frustrated and develop a unique subculture for their values.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

From the general view of conventional groups, this is referred to as a delinquent subculture, and youths belonging to this subculture are socially labeled as delinquent gangs. Moreover, a delinquent subculture often develops in socially disorganized areas.[18] Social disorganization is said to exist[12] when: “institutions of social control... have broken down and can no longer carry out their expected or stated functions” (p. 168). Adolescents living in socially disorganized areas have limited conventional opportunities, such as well-paying jobs or educational opportunities, which adolescents eventually perceive as an unequal distribution of power, a disjunction existing between aspirations and expectations, or a discrepancy between expectations and achievement.[18] To achieve their goals under such limited conventional opportunities, some adolescents seek alternative but illegal ways and thereby become involved in a deviant subculture.

Although prior research[17-18] addressed the general relationship between social class and delinquency, not all lower-class youths automatically engage in illegal behaviors. As an alternative conceptual viewpoint, social learning theory argues that crime results from the learning process of rewarded and punished behaviors shaped through past experience and observations.[19-21] For instance, youth might learn actual criminal techniques (e.g., how to steal things from others), psychological coping strategies (e.g., how to deal with guilt or shame as a result of criminal activities), and attitudes about crime (e.g., the norms and values related to criminal activities) from direct exposure to antisocial behavior[22-23] or from relationships with a delinquent group.[24-27]

Furthermore, it has been suggested that criminals are at lower stages of moral development than law-abiding citizens.[28-30] This reasoning suggests that people’s perceptions of their environment influence moral development. In fact, Thornberry[26] found that peer influence was a crucial element during mid-adolescence, and having delinquent peers helped form delinquent values. Menard and Elliott[31] also found that antisocial behavior attenuated a sense of social morality.

Considering influences that move youth away from antisocial behavior, in contrast, Hirschi[32] focused on four important prosocial bonds that detach adolescents from delinquency: attachment (i.e., sensitivity to and interest in others); involvement (e.g., participation in social activities); commitment (i.e., investing time, energy, and effort in conventional behaviors); and belief (i.e., respecting social values). According to his social bond theory, if youths have weak bonds of attachment, involvement, commitment, and belief, then they are more likely to engage in delinquent behavior. Extending this theoretical model, social bond theory was transformed into the general theory of crime (GTC), in which impulsive adolescents who receive poor socialization are more likely to be low in self-control and to weaken their social bonds to conventional groups, which, in turn, encourages them to seek criminal opportunity (e.g., joining gangs, using illegal drugs).[33]

Contrary to theoretical predictions, however, it has been reported that some youths who did not actually reject social bonds nevertheless developed associations with delinquents.[24] Thus, it is suggested that a relationship between social bonds and delinquent behavior is moderated by other factors, such as socioeconomic status.[24] Alternatively, path analyses of the National Youth Survey from 1976 to 1978 concluded that prior delinquency and involvement in delinquent peer groups were direct causal influences on delinquency and drug use, and conventional bonds and strain in-directly influenced later delinquency.[24] This research implies that delinquency is recidivistic probably because such youth have been labeled negatively and stigmatized, making it difficult for them to be rehabilitated into conventional society.[34-35]

Thus, previous research has provided rich information explaining sociological and

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

psychological mechanisms underlying delinquency. Our goal in this study is to combine previous theoretical perspectives and research findings to examine delinquency more comprehensively than has been done previously. Most prior research has examined only bivariate or linear relationships with delinquency and has analyzed a limited number of predictors. In this study, we investigated many different potential predictors in a single integrated model and explored how these various predictors interact non-linearly with each other. We hypothesized that both social and personal factors would mutually influence delinquent behaviors. We also considered several personal, social, and family-related variables that are potentially associated with delinquency, such as attitudes toward deviance, social isolation, family isolation, and demographic characteristics. Our dependent variable was youth's delinquency status—delinquency versus non-delinquency—and we used a newly available non-linear multivariable method of classification tree analysis, based on optimal data analysis (ODA), to classify observations into delinquents or nondelinquents.[36]

Advantages of Classification Tree Analysis (CTA)

Traditionally, linear classification methods such as discriminant analysis and logistic regression analysis have been used to solve statistical classification problems. Nevertheless, linear classification methods have several weak points that might produce statistical solutions that are less than optimal. For example, discriminant analysis can produce probabilities beyond the range of 0 to 1 and requires restrictive normality on the independent variables, which is usually not met in practice.[37] Furthermore, both discriminant analysis and logistic regression analysis simplify complex real-world phenomena by using a linear model although real phenomena are typically not linear.[38] In addition, these linear methods assume three conditions that are often unrealistic—namely, that the magnitude of importance, the direction of influence, and the coefficient value for each predictor variable is the same across all observations.[38] *It is not our intention to argue that statistical results found by linear methods are invalid, but rather to note that the level of accuracy of these methods is constrained by the above limitations.*

In contrast to traditional linear classification techniques, the ODA paradigm offers a non-linear multivariable classification method known as hierarchically optimal classification tree analysis (CTA).[38] Independent and dependent variables are referred to respectively as "attributes" and "classes" in CTA. An attribute is defined as: "any variable that can attain two or more levels, and reflects the phenomenon that one hopes will successfully predict the class variable," and a class variable is defined as "any variable that can attain two or more levels, and reflects the phenomenon that one desires to successfully predict."[36]

Note that a class variable must be categorical, although an attribute can be either categorical or continuous. CTA has distinct advantages over linear classification methods. First, CTA can handle non-linear, complicated real-world phenomena. With CTA, the shape or form of a given phenomenon does not matter, whereas linear methods assume that a straight line or a sigmoidal curve characterizes the underlying phenomenon.[38] In addition, a CTA model produces a high level of classification accuracy by adopting optimal decision rules, rather than trying to maximize explained variance or minimize a fit function (see Method for more detail). Moreover, CTA is free from the restrictive assumptions about independent variables. In particular, unlike linear methods, CTA does not assume constant importance, direction of influence, and coefficient value (unstandardized or standardized regression coefficient) for each attribute across all observations.[38]

Another strength of CTA is it provides a hierarchically optimal classification model, which can be very informative. In CTA, the at-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

tribute with the strongest effect size for the total sample, called the first node, enters the top of a hierarchically optimal classification tree model. One level or branch of the first node leads to a second node through a predictive pathway, while another level of the first node leads to another second node through a different predictive pathway. At these second nodes, the attributes with the strongest effect size under each condition are entered to produce, in turn, different pathways to the third nodes. These patterns are repeated until prediction endpoints are reached.

The final CTA model reveals two important pieces of information. First, tracing combinations of nodes in CTA visually identifies crucial interaction effects. For example, imagine the final CTA model indicates a certain subgroup (endpoint) is predicted to engage in delinquency when the first node of the model (e.g., attachment) is at a low value and the second node (e.g., moral belief) is also low. This result indicates that moral belief predicts delinquency, depending on the strength of attachment. Note that in contrast to traditional linear approaches, CTA automatically detects important interactions by examining all attributes in the statistical model. Second, the CTA model allows us to trace multiple stages branching into each level of a class variable and to discover the critical profiles linked to each outcome. In the above example, the CTA model would show attachment (the first stage) and moral belief (the second stage) at which youths move toward delinquency or non-delinquency. This result implies that one profile of delinquency is the combination of weak attachment and moral beliefs.

In contrast, linear methods cannot identify ordinal predictors leading to each outcome. Furthermore, unlike CTA, linear methods have difficulty finding combinations of multiple variables predicting each level of an outcome simultaneously, making it more difficult to use linear methods to identify predictive profiles.

These advantages make CTA a powerful procedure for solving statistical classification problems in comparison with the linear classification methods. CTA models are manually constructed using statistical software which conducts ODA and classifies observations optimally by following “a prediction rule that explicitly achieves the theoretical maximum possible level of classification accuracy”.[36] We used ODA in this study for three reasons in addition to the fact that ODA enables us to capitalize on all the strengths of CTA. First, ODA can analyze all types of attributes measured by ratio, interval, ordinal, and nominal scales.[36,39] Second, as noted in the Method section below, ODA empirically tests the expected cross-sample generalizability of optimal classification models. [36,39] Finally, ODA simultaneously analyzes as many attributes as one wants without the limitations of the ratio of attributes to sample size or problems of multicollinearity.[36] This is because ODA tests the overall effect of each attribute on a class variable individually and selects only the single most influential attribute at each node. This strategy differs from multiple regression analysis, which calculates the partial effect of each variable independent of the effects of other variables when considered simultaneously.

Method

Participants and Materials. Archival data from the National Youth Survey, a 1976-1978 longitudinal design with multiple birth cohorts, were used.[24,40-41] In early 1977, the first wave of the survey gathered a multistage, cluster (area) probability sample of 1,725 American adolescents aged from 11 to 17 in 1976. Thus, by design, the sample included not only delinquents but also non-delinquents. The survey assessed events and behaviors theoretically linked with delinquency during calendar year 1976, and the subsequent wave tracked most of the individuals in 1978. Because the National Youth Survey followed the same individuals over time, we selected theoretically relevant attributes from the 1976 dataset to predict later self-reported delinquency in the 1978 dataset. Partici-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

pants interviewed for the first survey were representative of the youth population aged 11-17 in the U.S. measured by the U.S. Census Bureau, and the attrition rate for the subsequent wave was only 6% (N=99).[24] ODA software[36] was used to manually construct a hierarchically optimal CTA model of juvenile delinquency.

Measures. Our class variable of general delinquency was a composite index consisting of the frequency of the following behaviors reported by youths in 1978: aggravated assault, larceny, burglary, robbery, marijuana use, hallucinogens use, amphetamines use, barbiturates use, cocaine use, vandalism, buying stolen goods, hitting, joyriding, runaway, carrying a hidden weapon, prostitution, and selling drugs. Note that there were no questions about homicide and arson in the survey. Alcohol use, lying about age, hitchhiking, and buying liquor for a minor from were excluded from our measure of delinquency because they were rather common illegal acts.[24,43] Sexual intercourse, panhandling, and disorderly conduct were also excluded from delinquent behaviors. Sexual intercourse is relatively commonplace among youths, and it is also hard to judge whether sexual intercourse is delinquent.[43] For example, a victim of rape has sexual intercourse against his or her will, but voluntary intercourse is not illegal. Thus, it was reasonable to bar sexual intercourse as a component of delinquency. As for panhandling, begging for money does not hurt anyone and is not delinquent. Finally, people often behave in a disorderly manner (e.g., being loud in public) simply because of their exuberantly positive mood, so disorderly conduct is not always a form of delinquency.

Although our decision to consider some illegal acts as non-delinquent due to the trivial nature of these acts may not be universally accepted, the proportion of youths who performed at least one of these "trivial" illegal acts once or more monthly was 69.1%, whereas the proportion of youths who committed delinquent acts once a month or more as we have operationally defined this construct was 32.8%, which seems much more reasonable as an estimate of the underlying rate of delinquency.

The National Youth Survey offered two sets of questions to measure (a) the actual number of times each delinquent act was committed and (b) the frequency of each delinquent behavior using a scale ranging from one (never) to nine (two-three times a day). Cronbach's α for the frequency rates of the general delinquency was 0.713, which was greater than that for the actual number of delinquent behaviors. Hence, only the frequency rate items were used to construct the class variable for CTA. Committing each delinquent behavior once a month or more (score≥4) was recoded as one point, while committing each delinquent behavior less than once a month (score<4) was recoded as zero points. This rule was the most effective in making our sample as representative as possible of American delinquents and non-delinquents (see the above discussion of the proportion of delinquents). Respondents who scored at least one point were defined as delinquents, whereas respondents who scored zero points were defined as non-delinquents: this was the class variable employed in CTA.

Attributes. A total of 166 attributes were examined, including 17 theoretical "broad band" composite variables, the individual "narrow band" items composing these theoretical attributes, and additional background and demographic characteristics used in prior research.[24] The theoretical variables were: (a) *conventional involvement* measured by a sum of scores on the school athletic and activities involvement scales and community involvement scale (α=0.70); (b*)* *attachment to family* measured by a sum of scores on the family involvement and aspiration scales (α=0.72); (c) *conventional commitment* measured by a sum of scores on the school aspirations scale and future occupational and educational goal scales (α=0.71); (d) *moral belief* measured by a sum of scores on the family, school, and peer normlessness scales (α=0.72);

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

(e) *exposure to peer delinquency* measured by a sum of scores on the number of close friends performing each of some bad behaviors (α=0.82); (f) *involvement with delinquent peers* measured by a sum of scores on the peer involvement scale multiplied by the difference between an observed score for exposure to peer delinquency and its mean (because this is a single index, α was not computed[24]); (g) *socialization* measured by a sum of scores on the perceived sanctions in family scale (α=0.84); (h) *attitudes toward deviance* measured by a sum of scores on the attitudes toward deviance scale (α=0.79); (i) *social disorganization* measured by a sum of scores on the neighborhood problems scale and the reversed and standardized family income scale (α=0.75); (j) *prior self-reported delinquency* measured by a sum of scores on the continuous frequency rate scale (α=0.95) and measured by a sum of scores on the dichotomous frequency rate scale (α=0.91); (k) *social isolation* measured by a sum of scores on the family and school social isolation scales (α=0.73); (l) *family isolation* measured by a sum of scores on the family social isolation scale (α=0.72); (m) *social labeling* measured by a sum of scores on the family and school labeling scales (α=0.86); (n) *perceived labeling by parents* measured by a sum of scores on the family labeling scale (α=0.71); (o) *perceived labeling by teachers* measured by a sum of scores on the school labeling scale (α=0.80); and (p) *strain* measured by a sum of scores recoded 0 (no strain) to 3 (high level of strain), after subtracting scores on the achievement of each goal from scores on the importance of the corresponding goal (α=0.62).[24] Note that in measuring prior delinquency based on both continuous and dichotomous scales, we adopted the same operational definition as that of our class variable.

Procedure and Analysis Strategy. The National Youth Survey data sets were obtained through the Inter-University Consortium for Political and Social Research (ICPSR) of the University of Michigan. After all data were accessed and gathered, the class variable and attributes were selected and computed as described above. Finally, the class variable and the attributes were input into the ODA program to construct the CTA model.

To facilitate clarity of exposition we review how optimal data analysis operates in constructing a CTA model. ODA is first used to determine a cutpoint, or decision rule, for each attribute that maximizes the overall percentage of observations that are correctly classified (i.e., the percentage accuracy in classification, or PAC). For each equal interval or ordinal (i.e., continuous) predictor, ODA identifies an optimal classification cut-point (e.g., if age>14, then predict delinquency; if age≤14, then predict non-delinquency) that maximizes overall PAC. For each nominal or binary (i.e., categorical) predictor, ODA identifies an optimal classification rule (e.g., if ethnicity=Anglo, then predict delinquency; if ethnicity≠Anglo, then predict non-delinquency) that maximizes overall PAC. Thus, ODA can accommodate multi-category nominal predictors, such as race, without dummy coding these variables. Unlike other statistical methods for constructing tree models (e.g., regression-based CART or chi-square-based CHAID), ODA uses an exact permutation probability with no distributional assumptions, assesses the expected cross-sample generalizability of classification rules through an automated jackknife validity analysis procedure, and finds main effects and nonlinear interactions that optimally classify admission decisions. PAC is computed as 100% x (number of correctly classified observations)/(total number of observations).[36]

After determining the optimal cutpoint providing the greatest PAC for each attribute, the next step is to decide which attributes to enter into the hierarchically optimal CTA model. The chosen attribute must have the greatest effect strength for sensitivity (ESS), which reflects how much better PAC is compared to chance, using a standardized scale where chance classification accuracy is 0% and perfect classification accuracy is 100%. ESS is calculated using the following equation:

$$\text{ES }(\%) = \left\{ 1 - \frac{100 - (\text{mean PAC across classes})}{100 - \dfrac{100}{C}} \right\} \times 100$$

where C is the number of response categories for the class variable.[36] By rule-of-thumb, ESS values < 0.25 are regarded as weak, values between 0.25 and 0.50 are considered moderate, and values > 0.50 are defined as strong.[36]

After selecting the attribute with the greatest ESS to serve as a node of a tree model, the attribute's expected cross-sample stability in classification performance is assessed using a leave-one-out (LOO), or jackknife, validity analysis. In LOO analysis, classification performance is evaluated after removing an observation, and then the removed observation is classified again according to the classification performance obtained using the remaining subsample. This process is repeated until every observation has been removed and classified. An attribute is included in the CTA model only if its classification accuracy is stable in LOO analysis. LOO analysis helps to construct a tree model whose constituent attributes are most likely to generalize to a new sample.

If a LOO stable attribute with the greatest ESS is statistically significant, then the attribute enters as the first node of a CTA model. The level of statistical significance is determined by Monte Carlo simulation as a permutation probability, and is isomorphic with Fisher's exact *p* test for binary attributes. After the first node is determined, ODA subsequently searches the second node and lower nodes under each level of the highest node of a hierarchical tree model using the above procedures. These procedures are repeated until no more attributes are below the critical $p<0.05$-level.

Note that a given attribute can re-enter a node at a lower level even if it has already entered as a node at a higher level in the CTA model. This is the case when a re-entered attribute still contributes to the best classification performance with a new cutpoint when combining specific levels of higher nodes. Finally, to control the experimentwise Type I error rate at $p<0.05$ per comparison, a sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure is used to prune attributes selected by inflation of Type I error.[36] These adjustments also help maximize statistical power by rejecting lower nodes tested from very small subsample sizes when the total sample becomes divided and reduced.[36]

Results

Univariate Analyses. To describe simple relationships between delinquency and each attribute, we first conducted univariate analyses using ODA (Table 1). Consistent with previous findings, most theoretical attributes were significantly related to delinquency in the predicted direction: delinquency was significantly associated with weak attachment to family, weak conventional commitment, weak moral belief, greater exposure to peer's delinquency, positive attitudes toward deviance, high level of social disorganization, more experiences of prior delinquency, high level of social isolation, high level of family isolation, negative social labeling, negative social labeling by teachers, and high level of strain.

In addition to these theoretical attributes, race and age were also significantly related to delinquency: Anglo adolescents were more likely to commit delinquency than other racial groups; and adolescents aged 14 or older were more likely to commit delinquency than those aged 13 or younger.

Table 1: Univariate Associations of Theoretical and Demographic Attributes with Delinquent (1) Versus Non-Delinquent Behavior (0) for the Total Sample (N=1,606)

Attribute	ODA Model	*n*	% Delinquent	ESS	*p*-value
Conventional involvement	> 20.5, predict 0	70	30.00	17.93	0.413
	≤ 20.5, predict 1	186	36.56		
Attachment with family	> 29.5, predict 0	1024	25.78	19.94	0.118 x 10^{-13}
	≤ 29.5, predict 1	536	45.15		
Conventional commitment	> 30.0, predict 0	875	24.00	21.38	0.906 x 10^{-15}
	≤ 30.0, predict 1	705	42.98		
Moral belief	> 42.5, predict 0	907	25.58	18.95	0.935 x 10^{-12}
	≤ 42.5, predict 1	653	42.73		
Exposure to peer’s delinquency	≤ 16.5, predict 0	809	21.88	30.96	0.102 x 10^{-26}
	> 16.5, predict 1	538	50.56		
Involvement with delinquent peers	≤ 1.26, predict 0	812	21.80	31.19	0.107 x 10^{-25}
	> 1.26, predict 1	532	50.75		
Socialization	> 30.5, predict 0	57	26.32	1.08	0.175
	≤ 30.5, predict 1	1520	33.16		
Attitudes toward deviance	> 25.5, predict 0	878	21.75	27.32	0.524 x 10^{-24}
	≤ 25.5, predict 1	719	46.04		
Social disorganization	≤ 12.15, predict 0	1377	31.30	3.79	0.0112
	> 12.15, predict 1	135	41.48		
Prior self-reported delinquency	≤ 33.5, predict 0	1053	20.42	36.86	0.215 x 10^{-46}
	> 33.5, predict 1	553	56.42		

Social isolation	≤ 20.5, predict 0	662	29.15	6.49	0.0082
	> 20.5, predict 1	917	35.01		
Family isolation	≤ 10.5, predict 0	1018	29.76	8.59	0.000519
	> 10.5, predict 1	577	37.95		
Social labeling	> 81.5, predict 0	1050	26.67	19.20	0.462 x 10^{-13}
	≤ 81.5, predict 1	479	46.35		
Perceived labeling by parents	> 37.5, predict 0	1146	28.88	13.34	0.682
	≤ 37.5, predict 1	403	44.17		
Perceived labeling by teachers	> 43.5, predict 0	1010	25.94	19.97	0.132 x 10^{-13}
	≤ 43.5, predict 1	541	45.29		
Strain	≤ 11.5, predict 0	171	23.98	3.66	0.0479
	> 11.5, predict 1	1095	30.50		
Exposure to peer's alcohol use	≤ 2.5, predict 0	880	22.05	32.13	0.332 x 10^{-30}
	> 2.5, predict 1	501	52.89		
Attitudes toward marijuana use	> 3.5, predict 0	1042	23.61	27.01	0.553 x 10^{-25}
	≤ 3.5, predict 1	556	49.82		
Sex	Male, predict 0	849	40.64	-18.75	0.999
	Female, predict 1	757	24.04		
Race	Black/Chicano/American Indian/Asian/other, predict 0	322	25.47	6.69	0.000902
	Anglo, predict 1	1281	34.66		
Age	≤ 13, predict 0	732	24.45	17.28	0.346 x 10^{-10}
	> 13, predict 1	874	39.82		

Grade at School	8th grade or lower, predict 0	819	26.01	17.28	0.439
	9th grade or higher, not in school, or other, predict 1	787	39.90		
GPA	F, predict 0	10	60.00	-0.78	0.983
	A, B, C, or D, predict 1	1585	32.49		
Family Income	≤ $14,000, predict 0	141	33.33	-0.43	0.646
	> $14,000, predict 1	1375	32.22		
Parent's Marital Status	Single or married, predict 0	1300	31.23	5.11	0.593
	Divorced/separate/other, predict 1	280	38.93		

Note: "ODA Model" indicates the cutpoint or decision rule by which ODA classified (non)delinquents.[36] Total sample sizes varied across attributes due to incomplete data. A sequentially-rejective Bonferroni adjustment procedure was *not* employed for univariate analyses.[36] The total number of respondents who answered the set of questions associated with conventional involvement was 256, so the response rate for this set of items was only 15.94%. ESS values indicated in red were stable in jackknife ("leave-one-out") validity analysis, and are expected to show cross-sample generalizability.

However, contrary to previous theory and research, attributes unrelated to delinquency included conventional involvement, socialization, and perceived labeling by parents. Moreover, LOO analysis concluded that a significant relationship between involvement with delinquent peers and delinquency was not cross-sample generalizable.

Classification Tree Analysis. Our primary interest was not to see simple relationships between each attribute and delinquency, but to see how multiple attributes combine to explain predictive roots and profiles of juvenile delinquency and non-delinquency. Therefore, we used ODA to construct a hierarchically optimal CTA model. Following established procedures for constructing optimal CTA models, 68 nodes were initially identified; but after applying a sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure, only five nodes were retained. These five nodes were prior self-reported delinquency measured by continuous scales as the first node (p<0.001) and as the third node (p<0.001), exposure to peer alcohol use during 1976 (p<0.001), exposure to peer delinquency during 1976 (p<0.001), grade level in school during 1976 (p<0.001), and attitudes toward marijuana use during 1976 (p<0.001). Except for grade level, all attributes were significant in the univariate analyses. Figure 1 shows the final hierarchically optimal CTA model for explaining juvenile delinquency. In the figure, circles represent nodes, arrows indicate branches, and rectangles are prediction endpoints (D=delinquency, ND=non-delinquency). Numbers below each node indicate directional Fisher's exact p value for the node, and numbers in parentheses within each node indicate ESS for the node. Also, numbers next to each arrow indicate the value of the cutpoint for the node.

The strongest predictor of delinquency for the total sample was prior self-reported delinquency (ESS=36.86%): the first node of the CTA model. The cutpoint for this attribute was 33.5 (1.94% on the absolute scale).

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

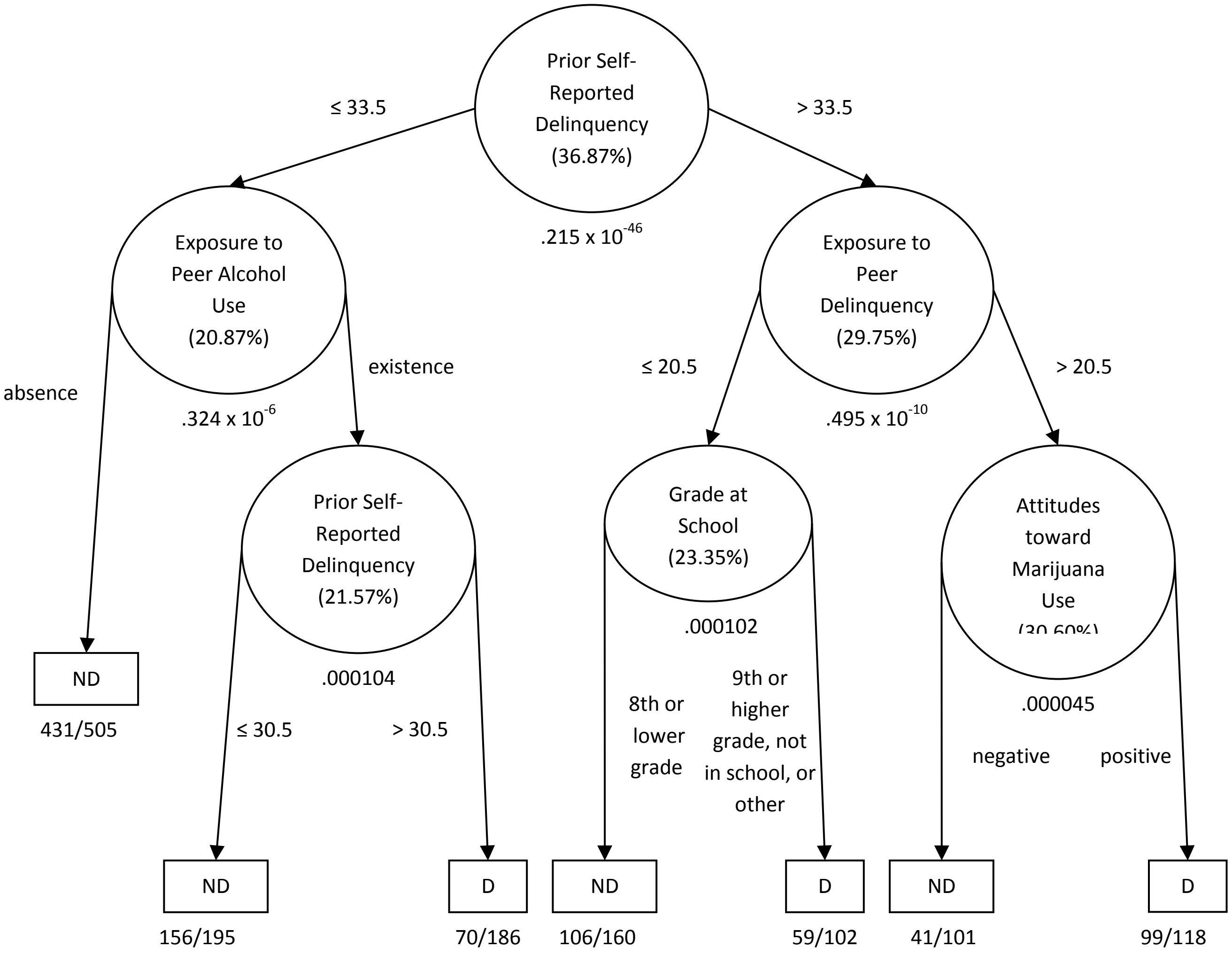

Prior Self-Reported Delinquency (36.87%)
≤ 33.5
> 33.5
.215 x 10-46
Exposure to Peer Alcohol Use (20.87%)
absence
existence
.324 x 10-6
Exposure to Peer Delinquency (29.75%)
≤ 20.5
> 20.5
.495 x 10-10
Prior Self-Reported Delinquency (21.57%)
.000104
≤ 30.5
> 30.5
Grade at School (23.35%)
.000102
8th or lower grade
9th or higher grade, not in school, or other
Attitudes toward Marijuana Use (30.60%)
.000045
negative
positive
ND
431/505
ND
156/195
D
70/186
ND
106/160
D
59/102
ND
41/101
D
99/118

Figure 1: The CTA model for predicting juvenile delinquency versus non-delinquency (*N*=1,367). Ellipses represent nodes, arrows represent branches, and rectangles represent prediction endpoints. Numbers under each node indicate the exact *p* value for each node. Numbers in parentheses within each circle indicate effect strength. Numbers beside arrows indicate the cutpoint for classifying observations into categories (delinquency or non-delinquency) for each node. Fractions below each prediction endpoint indicate the number of correct classifications at the endpoint (numerator) and the total number of observations classified as the endpoint (denominator). Negative attitudes toward marijuana use = Thinking that marijuana use is “very wrong” or “wrong” for a youth or someone his or her age; Positive attitudes toward marijuana use = Thinking that marijuana use is “a little bit wrong” or “not wrong at all” for a youth or someone his or her age; D = delinquency; ND = non-delinquency.

For youths who scored 33.5 or less on the prior delinquency scale based on its frequency rate, the second node was exposure to peer alcohol use (ESS= 20.87%). If a respondent had no friends who used alcohol, then that respondent was predicted to be non-delinquent with 85.35% accuracy. In other words, a few prior experiences with delinquency and no exposure to peer alcohol use jointly led to non-delinquency. For youths who had a few prior experiences of delinquency but who were exposed to peer alcohol use, a third node branched to either delinquency or non-delinquency. This third node was, again, prior self-reported delinquency (ESS=21.57%). That is, prior self-reported delinquency became the strongest attribute again among youths who had committed delinquent behavior less frequently and were exposed to peer alcohol use, but not among youths who fell into the other predictive pathways. At this node the cutpoint was 30.5, representing less than the 1st percentile on an absolute scale. If youths scored 30.5 or lower on the prior delinquency scale, then they were predicted to be non-delinquent with 80% accuracy. Therefore, even if youths had friends who had used alcohol, it was possible that the youths were still non-delinquents when they had been much less likely to perform delinquent behaviors two years earlier. In contrast, under the conditions where youths were exposed to peer alcohol use, if their scores were above 30.5 but 33.5 or less on the prior delinquency scale, then they were predicted to be delinquent with 37.63% accuracy. This was the lowest classification performance at any endpoint predicting delinquency. Overall predictive accuracy for youths who had earlier engaged in delinquent acts less often was 74.15% (657/886).

In comparison, for those who had earlier engaged in delinquent behavior more often, a different hierarchical pattern appeared. Among youths who scored more than 33.5 on the prior delinquency scale, the strongest predictor in the model was exposure to peer's delinquency. The cutpoint for this attribute was 20.5, which represents the 26th percentile on an absolute scale. If youths scored more than 20.5 on the scale of exposure to peer delinquency, then they were classified as being either delinquent or non-delinquent, depending on their attitudes toward marijuana use. In contrast, among youths reporting more frequent prior delinquency and less exposure to peer's delinquency (score≤ 20.5), classification as delinquent or nondelinquent depended on their grade level in school. Specifically, youths were predicted as non-delinquent when (a) they were more exposed to peer delinquency and thought that marijuana use was "very wrong" or "wrong" for them or someone their age (59.41% delinquency rate), or (b) they were less exposed to peer's delinquency and were in the eighth grade or lower (33.75% delinquency rate). In comparison, youths were classified into delinquency when (c) they were more exposed to peer delinquency and thought that marijuana use was "a little bit wrong" or "not wrong at all" (83.90% delinquency rate), or (d) they were less exposed to peer's delinquency and were in ninth grade or higher, did not attend at school, or a trade or business school (57.84% delinquency rate). Overall predictive accuracy for those who reported more frequent delinquent behaviors earlier was 63.41% (305/481).

Table 2 summarizes the overall classification performance of the CTA model, which correctly classified 962 (70.37%) of the total 1,367 youths. The ESS for this model was 30.59%, indicating that the model attained almost one-third of the theoretically possible improvement in classification accuracy versus the performance expected by chance: this is considered to reflect a moderate effect.[36]

Table 2: Confusion Table for CTA DelinquencyModel

		Predicted Class Status		
		Non-Delinquent	Delinquent	
Actual Class Status	Non-Delinquent	860	128	Specificity = 87.0%
	Delinquent	135	70	Sensitivity = 34.1%
		Negative Predictive Value = 86.4%	Positive Predictive Value = 35.4%	

Additional Comments about Cutpoints. Although the cutpoints for prior self-reported delinquency were 33.5 and 30.5, depending on the level of node, what do these values signify? Scores less than 33.5 were located within 1.94% on the absolute possible range, and the scores less than or equal to 30.5 reflects 0.65% of the absolute possible range on the prior delinquency scale. Descriptive statistics showed that the mean of prior delinquency (range=29-261) was 35.02 with *SD*=15.40. Overall, 65.2% of respondents scored 33.5 or less, while 34.8% scored more than 33.5. Conceptually, a respondent who scored 29 (i.e., 1 point x 29 items) had never committed delinquency in 1976, and a respondent who had performed all types of delinquent behaviors once or twice in 1976 should have scored 58 (i.e., 2 points x 29 items). Therefore, respondents who scored 33.5 had performed only a few types of illegal behaviors once or twice in 1976. In addition, because the score of 30 indicates that a respondent committed one kind of delinquent behavior once or twice in 1976, scores less than or equal to 30.5 indicate that respondents were engaged in only one delinquent behavior very few times. Thus, scores below 33.5 on the prior delinquency index were much closer to the score of non-delinquents used to categorize the class variable, and could be considered as reporting very few prior delinquent experiences.

What about exposure to peer delinquency? The cutpoint for exposure to peer delinquency was 20.5. Descriptive statistics revealed that the mean of this attribute (range=10-50) was 16.72 with *SD*=5.87. For exposure to peer delinquency, 77.8% of respondents scored 20.5 or less, and 22.2% scored greater than 20.5. Scores less than 20.5 fell within 26.25% on an absolute scale. A score of 20 (i.e., 2 x 10 items) would indicate that a respondent was exposed to peers who committed all ten types of delinquent behaviors. Therefore, a score of 20.5 or less indicates that a respondent was exposed to relatively few delinquent peers.

Discussion

Implications of the CTA Model of Delinquency. As hypothesized, this study yielded a parsimonious model identifying social (exposure to peer alcohol use, exposure to peer delinquency, and grade level in school) and personal variables (prior delinquency and attitudes toward marijuana use) that together predicted American youths as either delinquent or non-delinquent, supporting the critical influence of these factors on young people’s anti-social behavior. The optimal CTA model achieved about a third of the possible improvement in classification accuracy relative to chance, which represents a moderate effect size. The model identified three profiles of juvenile delinquency: (a) lay delinquency, reflecting infrequent prior delinquency with exposure to peer alcohol use (37.63% accuracy), (b) unexposed chronic delinquency, reflecting youth who had frequent prior delinquency with less exposures to peer delinquency, but being in the ninth grade or higher (57.84% accuracy), and (c) exposed chronic delinquency, reflecting youth who had frequent prior delinquency with exposure to peer delinquency and positive attitudes toward marijuana use (83.90% accuracy). In contrast, the model yielded four profiles of non-delinquency: (a) unexposed non-delinquency, reflecting youth who have infrequent prior delinquency with no exposure to peer alcohol use (85.35% accuracy), (b) exposed non-delinquency, reflecting youth who had extremely infrequent prior delinquency with exposure to peer alcohol use (80.00% accuracy), (c) unexposed reformation, reflecting youth who had frequent prior delinquency with less exposure to peer delinquency, but who were in eighth grade or lower (66.25% accuracy), and (d) exposed reformation, reflecting youth who had frequent prior delinquency with greater exposure to peer delinquency, but who had negative attitudes toward marijuana use (40.59% accuracy).

The CTA model provides additional insights into the prospective predictors of delinquency. Prior delinquency was the strongest predictor of subsequent delinquency—a conclusion that is consistent with previous reports that prior general delinquency directly influences later delinquency and drug use.[24] Our results extend prior findings, by identifying combinations of variables that exert a differential influence for experienced delinquents versus other subgroups of youth. For experienced delinquents, the factors important in maintaining delinquency appear to be exposure to peer delinquency, grade level in school, and attitude toward marijuana use. Youths who maintained their status as delinquents were categorized as unexposed or exposed chronic delinquents with 71.82% accuracy (Table 3). Previous studies showing the effect of exposure to antisocial behavior on criminal actions[22-23] and the effect of peers on the formation of delinquent values[26,31] support the profile of exposed chronic delinquency. Thus, with exposed chronic delinquency, prior delinquent experiences and exposure to delinquent peers might lead youths to form positive attitudes toward marijuana use, and these antisocial attitudes might encourage them to commit delinquent actions later. Note, however, that there is also a predictive profile reflecting exposed reformation, implying that not all youths with frequent prior delinquency and more exposure to delinquent peers automatically adopt positive attitudes toward marijuana.

In contrast, for adolescents who have infrequent prior delinquency, the variables predictive of changing non-delinquency into delinquency were exposure to peer alcohol use and prior delinquency. However, the combination of these factors predicted lay delinquency with only 37.63% accuracy, indicating that other factors not measured in the survey also operate.

Table 3: Summary of Cross-Classification by Year (N=1,367)

Year of 1978	Year of 1976	
	Non-Delinquency	Delinquency
Non-Delinquency	587/700 (83.86%)	147/261 (56.32%)
Delinquency	70/186 (37.63%)	158/220 (71.82%)

Note. The numerator of each fraction indicates the number of observations classified correctly. The denominator of each fraction indicates the number of observations predicted as a given category by the CTA model. Percentages reflect the proportion of correctly classified observations.

Another important implication is that the factors that maintain non-delinquency are different from the factors that terminate delinquency (Figure 1). The CTA model demonstrated that unexposed and exposed non-delinquents maintained their status of non-delinquency with 83.86% accuracy, whereas unexposed and exposed reformers became non-delinquents with only 56.32% accuracy (see Table 3). Future researchers should include measures of the variables composing these profiles, in order to enhance accuracy in predicting and understanding the dynamics of juvenile delinquency.

The CTA model identified protective factors more accurately than risk factors, and classification accuracy for non-delinquency was greater than for delinquency. This is probably because the surveys did not assess some critical risk factors. For instance, impulsivity[33], attention deficit/hyperactivity disorder[44], criminal opportunity[33.45], and historical contexts, such as a change in the level of surplus value[46] have all been identified as important risk factors, but were not directly assessed by the surveys. Another interesting implication concerns the crucial roles of adolescent exposure to peer delinquency and substance use in relation to delinquency. Regardless of prior delinquency, youths are sensitive to influence from peers perhaps because they desire to maintain intimacy and to avoid being rejected by peers. Also, alcohol use seems to be a “gateway” to performing delinquent behaviors by youths with infrequent prior delinquency, while marijuana use may be an obstacle to stopping delinquent behaviors.

Some variables found to be predictive of delinquency in previous research did not appear in the final CTA model. These predictors were socialization[17,24,33], social disorganization and social strain[18,24], involvement with delinquent peers[24-27], any types of social bonds[24,32-33], and any form of labeling.[34-35] It should be noted that in the univariate analyses all of these predictors—except for involvement with delinquent peers, conventional involvement, socialization, and perceived labeling by parents—were significantly predictive of delinquency (Table 1). The reason why these particular predictors failed to enter the final CTA model was that these predictors had smaller ESS than attributes selected for entry in the model, had low generalizability across samples, and/or had weaker effects when combined with variables in higher nodes of the hierarchical tree model. In contrast, grade in school was not significant in the univariate analysis, yet it was a node in the CTA model. This indicates that grade in school is significant among only a certain group, that is, American young people who had more prior delinquent experiences and were more likely to be exposed to peer delinquency, but not among general American young population.

Limitations. Our results are not without limitations. Although the strongest predictor of delinquency was prior self-reported delinquency, this result subsequently raises a follow-up question, “What factors, if any, predict prior delinquent behavior?” In our model, the profile of lay delinquency included not only those who had no prior delinquent experience, but also those who had very few prior delinquent experiences. Future research should explore the additional profile of delinquent youth who have no prior experiences of delinquency whatsoever.

Another limitation of the present research is the time frame of the survey data we analyzed. The National Youth Survey was conducted in 1976 and 1978. Thus, our results might reflect phenomena that are no longer generalizable to the present time period. Future research should address this limitation by constructing CTA using more recent data.

In terms of methodological limitations, our model reflects roughly 60% of the eligible youths originally selected by the multistage cluster sampling method. Although there is no agreed-upon standard for what constitutes an acceptable rate of inclusion, excluding 40% of respondents raises the possibility of potential selection and non-response biases. However, no

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

particular group of the youth population appears to be over- or under-represented in our sample, compared to the original sample who agreed to participate in the National Youth Survey.[24]

Other methodological issues concern the particular measures used in the National Youth Survey. In particular, the self-report items used to assess delinquency and other socially negative behaviors might not accurately reflect the actual levels of these behaviors because of social desirability, memory limitations, and motivation to recall. Moreover, the National Youth Survey did not include some variables that we wanted to examine as potential predictors of delinquency (e.g., impulsivity). Future research needs to include measures of other unanalyzed variables so that the classification accuracy of the hierarchical tree model can be further improved. Finally, although some theoretical composite attributes showed acceptable values of Cronbach's α, other attributes, including exposure to peer alcohol use and attitude toward marijuana use, were each measured by only a single individual question and had unknown reliability. Future research should measure attributes, especially exposure to peer alcohol use and attitude toward marijuana use, using multiple items, obtain acceptable Cronbach's α for these composite subscales, and then re-test them by including them in an ODA model.

Finally, it should be noted that an alternative definition of delinquency might yield different findings concerning the prospective predictors of juvenile delinquency. Although we contend that the classification of delinquency or non-delinquency based on our definition produced representative samples of youths who engage in these two forms of behavior, other theorists or researchers might well adopt an alternative definition of these two constructs. Or, they might suggest examining more specific delinquent actions (e.g., theft) independently rather than a broader, comprehensive category of juvenile delinquency because the factors might vary across different delinquent actions. Nevertheless, while we should avoid over-generalizing the factors found in our study to all delinquent actions, it is also informative to focus on the large-scale pattern of delinquency. This macro-level analysis is important because (1) the society and citizens tend to be more interested in getting a general idea (e.g., how to prevent delinquent crime in general) than a specific idea (e.g., how to prevent each potential delinquent actions specifically), and (2) each specific delinquent action is not exclusive or independent but accompanies another illegal action (e.g., robbery and assault could occur at the same time). Thus, our findings provide an overview of delinquent behavior, and the next goal should be to focus on each specific delinquent action to examine whether our model is applicable to it.

Another limitation concerning our definition of delinquency is the inevitable loss of precision in analyzing delinquency as a dichotomy as opposed to a continuous rate of frequency. In doing so, we have limited ourselves to investigating variables that predict whether or not youths exceed a threshold frequency that we have defined a priori as representing juvenile delinquency versus non-delinquency. These predictive variables may well differ from those that explain variation in the absolute frequency of delinquent behaviors.

Applications of the Present Study. The findings suggest potentially effective strategies for crime prevention. For example, shifting positive attitudes toward marijuana use toward negative attitudes may reduce delinquent behavior among exposed but reformed delinquent youths. Furthermore, our results suggest that an effective approach to protect non-delinquent youths from moving toward delinquency is to keep them away from peers who use alcohol. Future research should test these hypotheses.

References

[1]Federal Bureau of Investigation. (2004). *Crime in the United States 2003: Uniform crime re-*

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

ports. Retrieved November 9, 2004, from http://www.fbi.gov/ucr/03cius.htm

[2]Soler, M. (October, 2001). *Public opinion on youth, crime, and race: A guide for advocates*. Retrieved November 9, 2004, from http://www.buildingblocksforyouth.org/advocacyguide.html#juvcrime.

[3]Farrington, D.P. (1986). Stepping stones to adult criminal careers. In D. Olweus, J. Block, & M. Radke-Yarrow (Eds.), *Development of antisocial and prosocial behavior* (pp. 359-384). New York: Academic Press.

[4]Farrington, D.P. (1989). Early predictors of adolescent aggression and adult violence. *Violence and Victims*, *4*, 79–100.

[5]Farrington, D.P. (1991). Childhood aggression and adult violence: Early precursors and later-life outcomes. In D.J. Pepler & K.H. Rubin (Eds.), *The development and treatment of childhood aggression* (pp. 5-29). Hillsdale, NJ: Erlbaum.

[6]Farrington, D.P. (1995). The development of offending and antisocial behavior from childhood: Key findings from the Cambridge study in delinquent development. *Journal of Child Psychology and Psychiatry*, *36*, 929–964.

[7]Farrington, D.P. (1998). Predictors, causes and correlates of male youth violence. In M. Tonry & M.H. Moore (Eds.), *Youth violence: Vol. 24*. (pp. 421-447). Chicago: University of Chicago Press.

[8]Farrington, D.P., Barnes, G.C., & Lambert, S. (1996). The concentration of offending in families. *Legal and Criminological Psychology*, *1*, 47–63.

[9]Farrington, D.P., & Hawkins, J.D. (1991). Predicting participation, early onset, and later persistence in officially recorded offending. *Criminal Behavior and Mental Health*, *1*, 1–33.

[10]Farrington, D.P., Loeber, R., & Van Kammen, W.B. (1990). Long-term universal outcomes of hyperactivity-impulsivity-attention deficit and conduct problems in childhood. In L.N. Robins & M. Rutter (Eds.), *Straight and devious pathways from childhood to adulthood* (pp. 62-81). Cambridge, England: Cambridge Univ. Press.

[11]Sampson, R., & Laub, J. (1993). *Crime in the making: Pathways and turning points through life*. Cambridge, MA: Harvard University Press.

[12]Siegel, L.J. (1998). *Criminology: Theories, patterns, and typologies* (6th ed.). Belmont, CA: Wadsworth Publishing Company.

[13]Curry, G.D., & Spergel, I. (1988). Gang homicide, delinquency, and community. *Criminology*, *26*, 381-407.

[14]Park, R. (1915). The city: Suggestions for the investigation of behavior in the city environment. *American Journal of Sociology*, *20*, 579-583.

[15]Park, R., Burgess, E., & McKenzie, R. (1925). *The city*. Chicago: University of Chicago Press.

[16]Shaw, C.R., & McKay, H.D. (1972). *Juvenile delinquency and urban areas* (Rev. ed.). Chicago: University of Chicago Press.

[17]Cohen, A. (1955). *Delinquent boys*. New York: Free Press.

[18]Cloward, R., & Ohlin, L. (1960). *Delinquency and opportunity*. New York: Free Press.

[19]Bandura, A. (1973). *Aggression: A social learning analysis*. Englewood Cliffs, NJ: Prentice Hall.

[20]Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.

[21]Bandura, A. (1979). The social learning perspective: Mechanisms of aggression. In H. Toch

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

(Ed.). *Psychology of crime and criminal justice* (pp. 198-236). NY: Holt, Rinehart, and Winston.

[22]Sutherland, E. (1939). *Principles of criminology*. Philadelphia: Lippincott.

[23]Sutherland, E., & Cressey, D. (1970). *Criminology* (8th ed.). Philadelphia: Lippincott.

[24]Elliott, D., Huizinga, D., & Ageton, S. (1985). *Explaining delinquency and drug use*. Newbury Park, CA: Sage Publications.

[25]O'Donnell, J., Hawkins, J. D., & Abbott, R. (1995). Predicting serious delinquency and substance use among aggressive boys. *Journal of Consulting and Clinical Psychology*, *63*, 529-537.

[26]Thornberry, T. (1987). Toward an interactional theory of delinquency. *Criminology*, *25*, 863-891.

[27]Thornberry, T., Lizotte, A., Krohn, M., Farnworth, M., & Jang, S.J. (1994). Delinquent peers, beliefs, and delinquent behavior: A longitudinal test of interactional theory. *Criminology*, *32*, 601-637.

[28]Kohlberg, L. (1969). *Stages in the development of moral thought and action*. New York: Holt, Rinehart, and Winston.

[29]Henggeler, S. (1989). *Delinquency in adolescence*. Newbury Park, CA: Sage.

[30]Kohlberg, L., Kauffman, K., Scharf, P., & Hickey, J. (1973). *The just community approach in corrections: A manual*. Niantic, CT: Connecticut Department of Corrections.

[31]Menard, S., & Elliott, D. (1994). Delinquent bonding, moral beliefs, and illegal behavior: A three wave-panel model. *Justice Quarterly*, *11*, 173-188.

[32]Hirschi, T. (1969). *Causes of delinquency*. Berkeley, CA: University of California Press.

[33]Gottfredson, M., & Hirschi, T. (1990). *A general theory of crime*. Stanford, CA: Stanford University Press.

[34]Erickson, K. (1962). Notes on the sociology of deviance. *Social Problems*, *9*, 397-414.

[35]Schur, E. (1972). *Labeling deviant behavior*. New York: Harper & Row.

[36]Yarnold, P.R., & Soltysik, R.C. (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Book Co.

[37]Howell, D.C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury, Thomson Learning.

[38]Yarnold, P.R., Soltysik, R.C., & Bennett, C.L. (1997). Predicting in-hospital mortality of patients with AIDS-related pneumocystis carinii pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, *16*, 1451-1463.

[39]Bryant, F.B. (2005). How to make the best of your data. [Review of the book *Optimal data analysis: A guidebook with software for Windows*] [Electronic version] *Contemporary Psychology: APA Review of Books*, *50*.

[40]Elliott, D. (1977). *National youth survey [United States]: Wave I, 1976* [Computer file] ICPSR version. Boulder, CO: University of CO, Behavioral Research Institute [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1994.

[41]Elliott, D. (1986). *National youth survey [United States]: Wave III, 1978* [Computer file] ICPSR version. Boulder, CO: University of CO, Behavioral Research Institute [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1994.

[42]Johnston, L.D., O'Malley, P.M., Bachman, J.G., & Schulenberg, J.E. (December 21, 2004). *Overall teen drug use continues gradual de-*

cline; but use of inhalants rises. University of Michigan News and Information Services: Ann Arbor, MI. [On-line]. Available: http://www.monitoringthefuture.org; accessed 03/16/05.

[43]Henry J. Kaiser Family Foundation. (May 19, 2003). *National survey of adolescents and young adults: Sexual health knowledge, attitudes and experiences.* Menlo Park, CA. [On-line]. Available: http://www.kff.org; accessed 03/16/05.

[44] Ostrander, R.,Weinfurt, K.P., Yarnold, P.R., & August, G. (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology*, *66*, 660-672.

[45]Clarke, R. (1995). Situational crime prevention. In M. Tonry (Series Ed.) & D. Farrington (Vol. Ed.), *Building a safer society: Vol. 19. Strategic approaches to crime preventioni (pp. 91-151)*. Chicago: University of Chicago Press.

[46]Lizotte, A., Mercy, J., Monkkonen, E. (1982). Crime and police strength in an urban setting: Chicago, 1947-1970. In J. Hagan (Eds.), *Quantitative criminology*. (pp. 129-148). Beverly Hills, CA: Sage.

Author Notes

Work on this project was completed as a master’s thesis by Hideo Suzuki while he was a graduate student in the Department of Psychology, Loyola University Chicago. He is currently a Postdoctoral Research Associate in the Department of Psychiatry, Washington University. Fred B. Bryant and John D. Edwards are faculty in the Department of Psychology, Loyola University Chicago. We are grateful to Paul R. Yarnold and Robert C. Soltysik for advice and guidance in data analysis, and to Delbert S. Elliott, who conducted the National Youth Survey. eMail: suzukih@psychiatry.wustl.edu .

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Automated CTA Software: Fundamental Concepts and Control Commands

Robert C. Soltysik, M.S. and Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Fundamental methodological concepts are reviewed, and automated CTA software commands are annotated.

A decade in the making, commercially-available software which conducts automated hierarchically optimal *c*lassification *t*ree *a*nalysis[1] (CTA) is now being offered to organizations and individuals. This article reviews motivation underlying use of nonlinear models; shortcomings of suboptimal nonlinear methods; CTA methods, model interpretation and reporting; and use of automated software. Software commands and sample code used for solving (un)weighted classification problems are annotated.

"One Size Fits All" versus "Different Strokes for Different Folks"

Examples of linear models broadly used in applied research include models derived via logistic regression, log-linear, and discriminant analysis.[2,3] Regardless of derivation, all linear models share three important, usually unfulfilled assumptions.

First, linear models assume attributes in the model are important for every observation in the sample. In contrast, with nonlinear models different attribute sets can be used with different partitions of the sample: one set of attributes is used for classifying one partition of the sample; another set of attributes is used for classifying a different sample partition; and so forth.

Second, linear models assume the model attributes have identical direction of influence (positively or negatively predictive) for every observation. In contrast, with nonlinear models an attribute may predict class category 1 for one partition of the sample, versus category 0 for a different sample partition.

Third, linear models assume attributes in the model have the same coefficient value (or decision weight) for all sample observations. In contrast, in nonlinear models the coefficient for an attribute may assume two different values for two different sample partitions: for example, 0.2 and -1.8, respectively.

Traditional Nonlinear Methods

Nonlinear classification methods based on general linear model (GLM) or maximum-likelihood (ML) paradigms maximize variance ratios, or the value of the likelihood function for the sample, respectively. Examples of such suboptimal methods are chi-square automatic interaction detection, classification and regression tree analysis, genetic algorithms and neural networks. A problem for GLM-based methods involves satisfying the multivariate normally distributed (MND) assumption required for *p* to be valid, and a problem for ML-based methods is

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

that model coefficients are biased except in the limit for enormous samples.[2,3] A common issue is that neither GLM nor ML methods explicitly maximize model *accuracy*.[1]

Example of a CTA Model

The first CTA model published involved exploratory research discriminating geriatric (at least 65 years of age) versus nongeriatric adult ambulatory medical patients on the basis of self-reported well-being.[4] Forty geriatric and 85 non-geriatric ambulatory medical patients completed a survey assessing five functional status dimensions (Basic and Intermediate Activities, Mental Health [absence of depression], Social Activity, Quality of Social Interaction), and including five single-item measures assessing health satisfaction, physical limitations, and quantity of social interaction. The CTA model (Figure 1) was constructed manually using ODA software.[1]

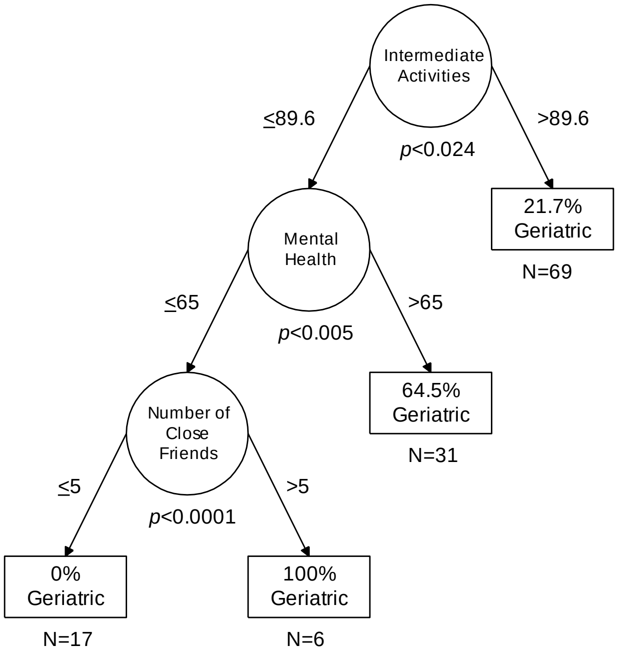


Figure 1: CTA Model Discriminating Geriatric *vs*. Nongeriatric Ambulatory Medical Patients

On first glance a depiction of any classification tree model may appear similar to results obtained by decision analysis (DA), because both methods depict findings using tree-like representations.[4] As seen, CTA models initiate with a *root node*, from which two or more *branches* emanate and lead to other *nodes*: branches indicate pathways through the tree, and all branches ultimately terminate in model *endpoints*. The CTA algorithm determines the attribute subset which predicts the outcome with maximum accuracy, beginning with the attribute which best discriminates the class variable (geriatric status) with maximum accuracy for the total sample. DA *estimates valence and likelihood associated with all possible decision-making strategies and outcomes*. In contrast, CTA *identifies a specific decision-making strategy which maximizes accuracy in predicting a specific outcome*.

Circles represent nodes in this schematic illustration of the CTA model, arrows indicate branches, and rectangles represent model endpoints. Numbers (or words, when attributes are categorical) adjacent to arrows indicate the value of the *cutpoint* (or *category*) for the node. Numbers underneath nodes give the generalized (per-comparison) Type I error rate for the node. The number of observations classified into each endpoint is indicated beneath the endpoint, and the percentage of geriatric observations is given inside the rectangle representing the endpoint.

Using CTA models to classify individual observations is straightforward. Consider a hypothetical person having an Intermediate Activities score=85, a Mental Health score=64, and 7 close friends. Starting with the first node, since the person's Intermediate Activities score is <89.6, the left branch is appropriate. At the second node the left branch is again appropriate because the person's Mental Health score is <65. Finally, at the third node the right branch is appropriate since the person has more than 5 close friends. The person is classified into the corresponding model endpoint: as seen, all six observations classified into this model endpoint were geriatric. Note that endpoints represent sample strata identified by the CTA model. The probability of being geriatric for this endpoint is $p_{geriatric}$=1 for the sample (in light of the small sample size at this endpoint, it may be more meaningful, depending on the application, to report $p_{geriatric}$≥6/7). In this example, had the patient instead reported 5 or fewer close friends, then the left-hand endpoint would be appropriate, with $p_{geriatric}$=0 (i.e., $p_{geriatric}$≤1/18).

Some intuitive aspects of CTA models are immediately obvious. For example, model "coefficients" are cutpoints or category descriptions expressed in their natural measurement units. In addition, sample stratification unfolds in a "flow" process which is easily visualized across model attributes. The manner in which CTA handles observations having missing data is also intuitive: linear models drop observations missing data on any attributes in the model, but CTA only drops observations which are missing data on attributes required in their classification. In the present example, imagine an observation having an Intermediate Activities score of 89.6 or greater, but missing data on number of close friends and/or on Mental Health. Using a linear model the observation would be dropped, but using CTA the observation would be classified.

Staging Tables

Staging tables (see Table 1) represent an alternative intuitive representation of CTA findings, and are useful for assigning "severity" or "propensity" scores (weights) to observations based on the findings of the CTA model. The rows of the staging table are simply model endpoints reorganized in increasing order of percent of class 1 (geriatric) membership. Stage is an *ordinal index* of geriatric propensity, and $p_{geriatric}$ is the corresponding *continuous index*: increasing values on either index indicates increasing propensity. Compared to Stage 1 (with $p_{geriatric}$ set at ≤1/18, or 0.056), $p_{geriatric}$ is approximately 4-times higher in Stage 2, 12-times higher in Stage 3, and 15-times higher in Stage 4 (with $p_{geriatric}$ set at ≥6/7, or 0.857).

To use the table to stage geriatric propensity for a given observation, simply evaluate the fit between the observation's data and each stage descriptor. Begin at Stage 1, and work sequentially through stages until identifying the descriptor which is *exactly true* for the data of the observation undergoing staging. Consider the hypothetical person discussed earlier. Stage 1 does not fit because the person has more than five close friends. Stage 2 does not fit because the person's Intermediate Activities score is ≤89.6. Stage 3 does not fit because the person's Mental Health score is ≤65. The staging table has only one degree of freedom, so through the process of elimination, it is clear that Stage 4 must be appropriate. Because the person has an Intermediate Activities score ≤89.6, a Mental Health score ≤65, and >5 close friends, Stage 4 clearly fits the data of this hypothetical person.

Table 1: Staging Table for Predicting Geriatric Status

Stage	Intermediate Activities	Mental Health	Close Friends	N	$p_{geriatric}$	Odds
1	≤ 89.6	≤ 65	≤ 5	17	0	≤1:17
2	> 89.6	------	-----	69	.217	1:4
3	≤ 89.6	> 65	-----	31	.645	2:1
4	≤ 89.6	≤ 65	> 5	6	1	≥6:1

Note: Increasing scores on Intermediate Activities indicate increasing adaptability, and increasing scores on Mental Health indicate decreasing depression.

Assessing Model Performance

Performance measures for CTA (and for all ODA methods) are also intuitively appealing, and are derived from a *confusion table*, as indicated for the present example in Table 2. Rows of the confusion table indicate the *actual* class category of any given observation in the *training sample* (used for model development), and columns indicate the class category *predicted* for an observation by the CTA model. For predictions involving the class category status of

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

individual observations in the training sample, when the actual and predicted class categories are identical (e.g., a geriatric person is predicted to be geriatric) then the model is correct; otherwise it is incorrect. Row and column marginal totals (the sum of all table entries in the row or column, respectively) are presented in the borders of the confusion table. For example, for actual class=geriatric, the *row marginal* is 15+26=41. For predicted class=geriatric, the *column marginal* is 11+26=37. Finally, the total sample size which is classified by the model is given in the lower right-hand corner of the table: this total is equal to the sum of row marginals, and also to the sum of column marginals.

Table 2: Confusion Table for the Example CTA Model

	Predicted Class		
Actual Class	Nongeriatric	Geriatric	
Nongeriatric	71	11	82
Geriatric	15	26	41
	86	37	123

Assessing the performance of a CTA (or any classification) model begins by computing five standard epidemiological indices.[1] The first pair of indices assess the ability of the model to accurately classify observations in the different class categories. *Sensitivity* is the likelihood of correctly classifying an observation from Class 1, and is defined as the number of correctly classified Class 1 observations divided by the total number of Class 1 observations: here, 26/41=0.634. *Specificity* is the likelihood of correctly classifying an observation from Class 0, and is defined as the number of correctly classified Class 0 observations divided by the total number of Class 0 observations: 71/82=0.866.

The next set of indices address the accuracy of the model when it is used to make classifications. *Positive predictive value* (PPV) is the likelihood that an observation predicted to be a member of Class 1 is accurately classified (i.e., is in reality a member of Class 1): here, 26/37=0.703. *Negative predictive value* (NPV) is the likelihood that an observation predicted to be a member of Class 0 is accurately classified: here, 71/86=0.826.

Finally, overall accuracy, or *percentage accuracy in classification* (PAC), is 100% times the number of correctly classified observations divided by the total number of observations classified by the model: 100% x (71+26)/123=78.9%. In the literature, sensitivity, specificity, PPV and NPV are typically multiplied by 100% in order to report all five indices in a common, familiar metric, and because the focus of CTA (and all statistical models in the optimal data analysis paradigm) is predictive accuracy rather than probabilistic likelihood.[1,5]

Summarizing a confusion table is a methodic, straightforward process, as illustrated for the present example: Using the CTA model, a total of 30.1% [100% x (26+11)/123] of the sample is predicted to be geriatric. These predictions are correct 70.3% [100% x PPV] of the time, and correctly identify 63.4% [100% x sensitivity] of all geriatric observations. Also, 82.6% [100% x NPV] of the model-based predictions that an observation is nongeriatric are correct, and correctly classify 86.6% [100% x specificity] of all the nongeriatric observations. Overall, the model correctly classified 78.9% [PAC] of the observations in the sample.

Foregoing indices are bounded by 0 and 1 (or, equivalently, between 0% and 100%), and reference the *absolute predictive capacity* of a classification model. The ultimate objective is for all of these indices to reach their theoretical upper limit of 100% correct prediction. However, in the likely event that a statistical model fails to achieve perfect prediction, statistical criteria are used to assess the performance of CTA (and other) models, in terms of their *predictive capacity relative to chance*.

Effect Size for Sensitivity (ESS)

None of the five absolute performance indices are normed relative to chance, or have an associated exact *p* value.[1] Accordingly, the performance of all models in the optimal data analysis paradigm, including CTA, is summarized using the *effect strength for sensitivity* (ESS) statistic, a normed index ranging between 0 (representing the level of classification accuracy expected by chance) and 100 (representing errorless classification).[1]

The formula for computing ESS for problems with class variables involving two categories (automated CTA software solves only two-category problems: CTA for more than two class categories has never been reported) is:

ESS=100% x (Mean PAC – 50)/50 (1),

where

Mean PAC=100% x (sensitivity + specificity)/2

(2).

For example, if a CTA model had sensitivity=0.85 and specificity=0.74, then mean PAC= 100% x [(0.85+0.74)/2]=79.5%, and ESS=100% x [(79.5-50)/50]=59.0%.

Using ESS one may directly compare the performance of different models, relative to chance, regardless of structural features of the analyses, such as sample size, number of class categories, number of attributes and attribute metric, sample skew, and so forth. The rule-of-thumb which is used for evaluating *ecological significance* of results achieved by classification models is: ESS<25% (one-quarter of the improvement in classification accuracy theoretically possible to attain beyond the performance achieved by chance) is a *relatively weak* effect; $25\% \leq$ESS<50% is a *moderate* effect; $50\% \leq$ESS <75% a *relatively strong* effect; and ESS$\geq$75% is a *strong* effect.[1] Thus, in order to complete the summary of the confusion table which was presented earlier, append the following conclusion: “The CTA model yielded ESS=50.0%, a relatively strong effect.”

It is noteworthy that linear models may classify all observations in the sample into the dominant class if the sample is highly skewed (e.g., more than 75% of the sample falls into one class category). In this case Mean PAC is 50%, and ESS=0. For expository purposes, Table 3 illustrates how Mean PAC and ESS are related if one class category is classified perfectly, and Table 4 emphasizes that mean PAC=50 is what is anticipated by chance.

Table 3: PAC in Each of Two Groups (PAC= 100% in One Group), Mean PAC, and ESS

Group A	Group B	Mean PAC	ESS
100	0	50	0
100	10	55	10
100	20	60	20
100	30	65	30
100	40	70	40
100	50	75	50
100	60	80	60
100	70	85	70
100	80	90	80
100	90	95	90
100	100	100	100

Table 4: Patterns of PAC in Each of Two Groups that Yield ESS=0

Group A	Group B	Mean PAC	ESS
100	0	50	0
90	10	50	0
80	20	50	0
70	30	50	0
60	40	50	0
50	50	50	0

Ostrander et al.[6] note that, in contrast to sensitivity and specificity, PPV and NPV are influenced by base rate of class category *c* (e.g., 0 or 1) in the population, and by the false posi-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

tive rate—the likelihood that the model will classify an observation into class category *c* when the observation is *not* a member of *c*. A method is given for easily assessing the models *efficiency* over different base rates (an efficient model provides PAC for category *c* which is greater than the category *c* base rate).[6]

Model Interpretation

In addition to its greater accuracy versus logistic regression analysis or Fisher's discriminant analysis, CTA also produced substantively richer findings. In the present example the linear models identified two patient clusters: relatively active, depressed nongeriatric people; and relatively inactive, non-depressed geriatric people.

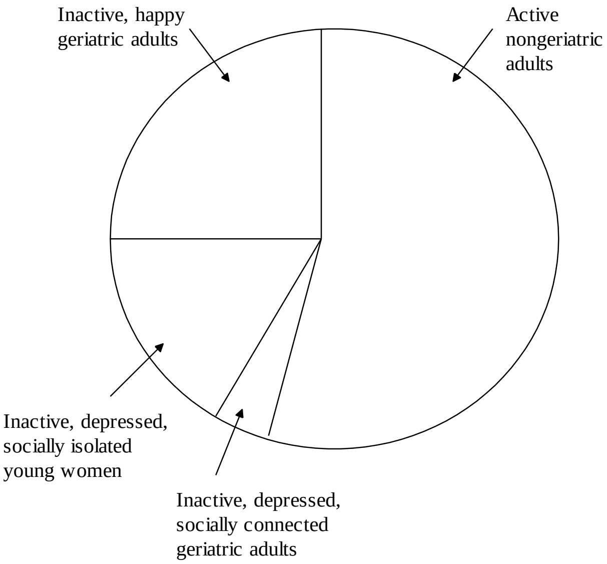


Figure 2: Pie-Chart Illustrating Distribution of Total Sample in Four CTA-Based Strata

In contrast, the CTA model identified four patient strata. Patients scoring >89.6 on Intermediate Activities were primarily (78.3%) relatively active nongeriatric adults (56% of total sample). Patients scoring at lower levels on Intermediate Activities, and at high levels (>65) on Mental Health, were largely (64.5%) relatively inactive, nondepressed geriatric adults (25% of sample). All the patients scoring at lower levels on both Intermediate Activities and Mental Health, and having fewer than six close friends, were inactive, depressed, socially isolated nongeriatric adults (14% of total sample, primarily young depressed women). Finally, all patients scoring at lower levels on both Intermediate Activities and Mental health, but having more than five close friends, were inactive, depressed, socially-connected geriatric adults (5% of sample).

Illustrating the portion of the total sample represented by CTA-identified strata, using a pie-chart, can facilitate understanding and development of policy implications of CTA-based findings: for example, by indicating the percentage of the sample that falls into each strata, the likelihood of attributing undue attention to comparitively rare strata is diminished (see Figure 2).

Table 5: AID Analysis for CTA Example

Attribute	Percent of Sample Evaluated in Part on the Basis of the Attribute	
Intermediate Activities	123/123	100.0%
Mental Health	54/123	43.9%
Number of Close Friends	23/123	18.7%

It is also informative to evaluate the attributes loading in the CTA model in terms of their importance in the prediction-making process. Conceptually related to the R^2 statistic from regression analysis, which indicates the percentage of the variance in the class (independent) variable which is explained by attributes (dependent measures) in the model[2], an *Attribute Importance in Discrimination* (AID) analysis indicates the percentage of the sample of classified observations which were influenced by the attribute (Table 5).

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Only the root attribute is involved in the classification decisions for all observations in the sample. Easily seen in Figure 1, Mental Health was involved in classification decisions for all of the observations except for those classified on the right-hand side of the root attribute: 123–69=54 observations. Mental Health therefore influenced classification decisions for 100% x 54/123, or 43.9% of the total sample. Also easily seen in Figure 1, the Number of Close Friends influenced classification decisions for 100% x 23/123, or 18.7% of the total sample.

Validity Assessment in CTA

Limited by the daunting computational burden associated with manual construction of CTA models, *experimental* research addressing validity issues in CTA has been infeasible in the absence of automated software. Psychometric properties of scores created using optimal data analysis methods has been a major focus of the paradigm since its inception[1], and rigorous investigation in this area is underway.

Nevertheless, some preliminary research in this area has been reported. For example, a Bayesian method was developed for estimating the efficiency of a CTA model versus chance for any class variable base rate.[6] And, the first CTA model published in the field of medicine used a manual *hold-out* methodology to create a CTA model which was optimal for two random *split-halfs* of a single large sample.[7] This study used CTA to create a severity-of-illness score for predicting in-hospital mortality from *Pneumocystis carinii* pneumonia, which cross-generalized to independent random samples with strong ESS.[8]

For all models created in the optimal data analysis paradigm, the upper-bound of expected cross-generalizability of the findings to an independent random sample is estimated via jackknife ("leave-one-out") analysis, whereby each observation in the sample is classified by a model created using a sample omitting the observation's data.[1] In the absence of automated CTA software, only attributes with stable jackknife classification performance (i.e., with ESS that did not vary between training versus jackknife analyses) were used in manually-derived CTA models. However, an estimate of Type I error associated with the jackknife procedure may be determined by computing the ESSj from the confusion table generated by this procedure. The proportion of ESS values greater than ESSj obtained from randomly shuffled classes in the original Monte Carlo procedure estimates the jackknife Type I error, and setting this proportion to the desired value (e.g., 0.05) may be used in a decision rule to admit these attributes into the final model.

Obtaining CTA Models

The mechanics underlying construction of CTA models was described previously.[1,7,9] Recursively-derived CTA models chain together series of models, derived by univariate optimal discriminant analysis (UniODA), on monotonically diminishing sample strata.[1] Because they chain together UniODA models, CTA models may be derived manually[10] via ODA software[1] which conducts UniODA (advantages of using automated software are discussed ahead). Exact statistical distribution theory and Monte Carlo simulation methodology are available for testing one- (confirmatory, *a priori*) and two-tailed (exploratory, *post hoc*) hypotheses.[1]

Researchers are encouraged to construct at least one CTA model manually using ODA software, in order to gain a deeper understanding of the recursive mechanical nature of CTA. Furthermore, ODA and CTA software use identical command syntax, so skill and knowledge acquired by using ODA will generalize to operation of CTA.

Submitting a Program for Analysis

Automatic CTA software can be used to analyze problems with two class categories, 500 attributes, and 65,535 observations (methods to solve problems involving massive samples are undergoing alpha testing), and is available under either commercial or individual license: custom systems are also created for special-purpose applications. The software is available through the ODA webpage.[11] To run an analysis, registered users login to the ODA webpage and upload the associated command and data file. Analyses are executed in the order they were received, and all associated output is returned via eMail.

A quick word seems in order regarding why Optimal Data Analysis, LLC, adopted a "software as service" model for distributing access to the automated CTA software. From the perspective of users there are several advantages of this model: (1) users needn't tie-up their (probably slower) computers, our fast computers will do the work; (2) the most current version of the software is always immediately available; (3) one can work 24/7/365 from any computer, anywhere; and (4) if the system crashes then specialists will be scrambling to fix the problem immediately—and any problems may well be fixed before most users are even aware that an issue had occurred. Another advantage to both user and Optimal Data Analysis, LLC, is savings in money and time, because the software doesn't need to be adjusted to run in the context of many different types of constantly changing computers, operating systems and data-base programs. Users simply send text files to the CTA system, and the CTA system returns a text file output via eMail.

Interpreting Automated Software Output

The module which produces schematic illustrations of CTA models is currently under development, and investigation addressing optimal information display in this context is underway in our laboratory.[12] The present software reports CTA models using an intuitive shorthand notation describing the node constituents of the CTA model. To facilitate clarity, Figure 3 gives a schematic illustration of node structure underlying all CTA models.

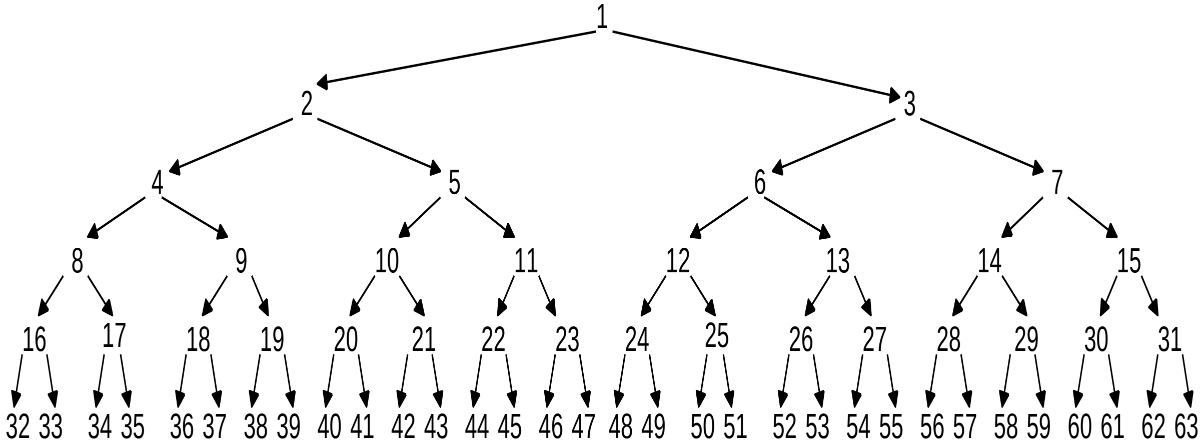


Figure 3: CTA Node Structure

It is a simple matter to determine the "identity number" of a node existing at a deeper depth than is illustrated in this five-level-deep tree (depth level 1 of the tree includes node 1; level 2 includes nodes 2 and 3; level 3 includes nodes 4-7; level 4 includes nodes 8-15; level 5 includes nodes 16-31; and level 6 includes nodes 32-63). From the perspective of node X (for X>1), the identify number of the node emanating from X's left-hand side is 2X, and from

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

X's right-hand side is 2X+1. For example, from node 47, node 94 (2x47) emanates to the left, and node 95 (94+1) emanates to the right. From node 94, node 188 emanates to the left, node 189 to the right, etcetera. Note that after the root attribute (depth 2 and deeper), all even-numbered nodes lie on the left-hand branch, and odd-numbered nodes on the right-hand branch, of the tree.

CTA software produces output employing node identity numbers to describe the CTA model: an example of CTA software output is presented in Figure 4 (hypothetical data). Respectively, the automated CTA software output lists: attribute name (D2, D3 and D4 loaded in the hypothetical CTA model); node identity number; tree depth level; sample size for the analysis indicated; ESS for the attribute; whether jackknife (leave-one-out, or LOO) validity analysis was stable (indicated) or unstable; jackknife ESS; *p* for the jackknife ESS; attribute metric (ORD=ordered, CAT=categorical); and CTA model shorthand.

The root attribute (here, D2) is listed first in the report. For each attribute the report *first indicates* the cutpoint and outcome for *the left-hand branch* emanating from the attribute, and *second for the right-hand branch*. *Branches ending in model endpoints are marked by an asterisk.* As seen, the left-hand branch emanating from D2 has a cutpoint of ≤6.2 units: observations having D2 scores ≤6.2 units are predicted to be a member of class 4, and this branch terminates in a model endpoint representing a total of 242 observations, of whom 165 (68.18%) are correctly classified. The remaining 242–165=77 observations having D2 scores ≤6.2 units were members of class 5, and were misclassified by this branch of the CTA model.

```
ATTRIBUTE NODE LEV  OBS    p     ESS      LOO     ESSL   LOOp TYP           MODEL
--------- ---- ---  ---    -     ---      ---     ----   ---- ---  -------------------------
       D2    1   1  704  .000  48.44%  STABLE  48.44% .000 ORD  <=6.2-->4,165/242,68.18%*
                                                                 >6.2-->5,375/462,81.17%

       D3    3   2  292  .000  41.60%  STABLE  41.60% .000 ORD  <=4.5-->4,29/63,46.03%
                                                                 >4.5-->5,206/229,89.96%*

       D4    6   3   62  .039  28.99%  STABLE  28.99% .039 ORD  <=1.9-->4,18/30,60.00%*
                                                                 >1.9-->5,22/32,68.75%*
```

Figure 4: Sample CTA Software Output (Hypothetical Expository Data)

The right-hand branch emanating from D2 has a cutpoint of >6.2 units: observations having D2 scores >6.2 units are predicted to be members of class 5, but this branch does not terminate in a model endpoint. Rather, the model includes attribute D3 at node 3.

As seen, the left-hand branch emanating from D3 has a cutpoint of ≤4.5 units: observations having D3 scores ≤4.5 units are predicted to be members of class 4, but this branch does not terminate in a model endpoint.

The right-hand branch from D3 has a cutpoint of >4.5 units: observations with D3 scores >4.5 units are predicted to be members of class 5, and this branch terminates in a model endpoint representing a total of 229 observations, of whom 206 (89.96%) are correctly classified. The remaining 229–206=23 observations having D3 scores >4.5 units were members of class 4, and were misclassified by this branch of the CTA model.

Both branches emanating from D4 terminate in a model endpoint (this is always true for the last attribute listed in the output). The left-hand branch has a cutpoint of <u>&lt;</u>1.9 units: observations with D4 scores <u>&lt;</u>1.9 units are predicted to be members of class 4; this endpoint represents 30 observations of whom 18 (60.00%) are correctly classified and 30–18=12 (40.00%) are misclassified. And, the right-hand branch has a cutpoint of >1.9 units: observations having D4 scores >1.9 units are predicted to be members of class 5; this endpoint represents 32 observations of whom 22 (68.75%) are correctly classified and 32–22=10 (31.25%) are misclassified.

To construct an illustration of the final CTA model, referring to Figure 3 select nodes 1, 3 and 6 (see Table 3, column 2): these are depicted by circles (Figure 1). Branches are then depicted using arrows emanating from the left-hand side of the root attribute (D2), the right-hand side of D3, and both sides of D4, terminate in model endpoints depicted using rectangles (Figure 1). Add the Type I error rate beneath each attribute, cutpoint values adjacent to arrows, and text indicating the outcome for each endpoint—and the CTA model is complete.

Automated CTA Command Syntax

Table 6 gives an alphabetical roster and description of automated CTA software control commands and keywords (an example of an automated CTA program is provided ahead).

Table 6: Control Commands for
Automated CTA Software

ATTRIBUTE

Syntax ATTRIBUTE *variable list* ;

Alias ATTR

Remarks The ATTRIBUTE command lists the attribute(s) to be used in the analysis. The TO keyword may be used to define multiple attributes in the list. For example, the command

ATTR A1 to A4;

indicates that A1, A2, A3 and A4 will be treated as attributes. Further exposition of the TO keyword is found in the discussion for VARS.

CATEGORICAL

Syntax CATEGORICAL {ON | OFF} ;
CATEGORICAL *variable list* ;

Alias CAT

Remarks The CATEGORICAL command specifies that categorical analysis will be used, and is required when the attribute to be analyzed is categorical. Using the ON keyword indicates that all variables in the variable list are categorical. CAT with no parameters is the same as CAT ON. The TO keyword may be used in the variable list (see the discussion under VARS).

CLASS

Syntax CLASS *variable list* ;

Remarks The mandatory CLASS command specifies the class variable to be used in the analysis. A separate analysis will be run for each class variable named. The TO keyword may be used in the variable list (see discussion under VARS).

DIRECTION

Syntax DIRECTION {< | LT | > | GT | OFF} *value list* ;

Aliases DIR, DIRECTIONAL

Remarks The DIRECTION command defines the presence and nature of a directional (i.e., *a priori*, one-tailed, or confirmatory) hypothesis. The

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

parameter < or LT indicates that the class values in the value list are ordered in the "less than" direction. The parameter > or GT indicates the class values are ordered in the "greater than" direction. The value list must contain every value of the class variable currently defined. The default is OFF.

ENUMERATE

Syntax ENUMERATE {ROOT} {MINOBS *value*} ;

Remarks The ENUMERATE command with no options specifies that all combinations of attributes in the top three nodes will evaluated. ENUMERATE ROOT specifies that only the top node will have all attributes evaluated. ENUMERATE MINOBS *value* allows only solution trees with at least *value* observations in them.

EXCLUDE

Syntax EXCLUDE *variable* {= | <> | < | > | <= | >= | OFF} *value* (,*value2*,…) ;

Aliases EX, EXCL

Remarks This command excludes observations having the indicated *value* of *variable*. For example,

```
EXCLUDE D=4 ;
```

drops all observations with the value of 4 for attribute D. The command

```
EXCLUDE B=2 Z>=113 ;
```

drops all observations with the value of 2 for attribute B or values greater than or equal to 113 for attribute Z. Commas in the exclude string enable the user to exclude multiple values of a variable using a single command:

```
EXCLUDE C=2,4 ;
```

excludes all observations having a value of 2 or 4 for attribute C. Multiple EXCLUDE commands may be entered, up to a maximum of 100 clauses. Observations which satisfy any of the EXCLUDE clauses will be excluded.

FORCENODE

Syntax FORCENODE *node var* ;

Remarks The FORCENODE command forces CTA to insert the attribute *var* at node *node* in the solution tree. If the UniODA solution for this attribute is not significant, or this node is subsequently pruned, an error message will be printed.

GO

Syntax GO ;

Remarks The GO command begins execution of the currently defined analysis.

INCLUDE

Syntax INCLUDE *variable* {= | <> | < | > | <= | >= | OFF} *value* (,*value2*,…) ;

Aliases IN, INCL

Remarks The INCLUDE command functions in the same manner as the EXCLUDE command, except that only those observations with the indicated *value* for *variable* are included. If multiple INCLUDE statements exist, only those observations will be kept which satisfy all these INCLUDE statements.

LOO

Syntax LOO {*p*value | STABLE} ;

Remarks The LOO command indicates that leave-one-out analysis will be performed for every attribute in the tree. LOO STABLE allows only attributes with LOO ESS equal to the ESS for that attribute. LOO *pvalue* allows only those attributes in the solution tree which have an ESS that yields a $p \leq pvalue$.

MCARLO

Syntax MCARLO {ITERATIONS *value* | CUTOFF *pvalue* | STOP *confvalue* } ;

Alias MC

Remarks The MCARLO command controls Monte Carlo analysis for estimating Type I error, or *p*. The keywords specify stopping criteria; if any criterion is met, then the analysis stops. ITERATIONS (ITER) specifies the maximum number of Monte Carlo iterations. STOP xxx indicates the confidence level (in percent), which will stop processing for the current attribute, if the estimated Type I error rate (specified with the CUTOFF keyword) drops below this level. For example, the command

```
MCARLO ITER 70000
CUTOFF .05 STOP 99.9 ;
```

indicates a Monte Carlo analysis will be conducted, and will stop when one of the following occurs: (1) 70,000 iterations have been executed, (2) a confidence level of less than 99.9% that p<.05 has been obtained.

MAXLEVEL

Syntax MAXLEVEL *value* ;

Remarks The MAXLEVEL command specifies the deepest level or depth allowed in the solution tree.

MINDENOM

Syntax MINDENOM *value* ;

Remarks The MINDENOM command specifies that only attributes which yield a denominator of *value* or more will be allowed in the solution tree.

MISSING

Syntax MISSING {*variable list* | ALL} (*value*) ;

Alias MISS

Remarks The MISSING command tells ODA to treat observations with value (*value*) as missing for each variable on the list. For example, the command

```
MISSING X Y Z (-4) ;
```

indicates that observations with attrbutes X, Y, or Z equal to -4 will be dropped if they are present in a CLASS, ATTRIBUTE, WEIGHT, or GROUP variable. ALL specifies that the indicated missing value applies to all variables. The TO keyword may be used in the attribute list (see discussion under VARS).

OPEN

Syntax OPEN {*path\file name* | DATA} ;

Remarks The OPEN command specifies the data file to be processed by ODA. This file must be in ASCII format. DATA indicates that a DATA statement, with inline data following, appears in the command stream.

OUTPUT

Syntax OUTPUT *path\file name* {APPEND} ;

Remarks The OUTPUT command specifies the output file containing the results of the ODA run. The default is ODA.OUT. APPEND indicates that the report is to be appended to the end of an already existing output file.

PRIORS

Syntax PRIORS {ON | OFF} ;

Remarks The PRIORS command indicates whether the ODA criterion will be weighted by the reciprocal of sample class membership. The default is ON. PRIORS with no parameters is the same as PRIORS ON.

PRUNE

Syntax PRUNE *pvalue* {NOPRIORS} ;

Remarks The PRUNE command indicates the p-value with which to optimally prune the classification tree. The NOPRIORS keyword should be used when PRIORS is turned OFF.

SKIPNODE

Syntax SKIPNODE *node* ;

Remarks The SKIPNODE command specifies that the node *node* will be empty of any attribute in the solution tree.

TITLE

Syntax TITLE *title* ;

Remarks The TITLE command specifies the title to be printed in the report. TITLE with no parameters erases the currently defined title.

USEFISHER

Syntax USEFISHER *value* ;

Remarks The USEFISHER command specifies that all probability calculations for categorical variable will be determined by Fisher's exact test, rather than by Monte Carlo.

VARS

Syntax VARS *variable list* ;

Remarks The VARS command specifies a list of attribute names corresponding to fields in the input data set. The TO keyword may be used to define multiple variables in the variable list. For example, the command

```
VARS X Y Z V1 TO V4 ;
```

specifies that the input file contains, in order, variables X, Y, Z, V1, V2, V3, and V4, and that there is at least one blank space separating all adjacent data. Alternatively, the data points may be separated by a single comma (with no spaces).

The TO keyword may be used to input a range of variables which have the same name except for the integer at the end of the name: the integers must be positive and ascending, increasing one unit per variable. Thus, VAR1 TO VAR10 is admissible (defining 10 variables). In contrast, VAR10 TO VAR1, VARA TO VARJ, or A TO X10, are not admissible.

The data for each observation may all exist on a single line of the data set, or may be placed on multiple adjacent lines. It is not recommended that a new observation is included on a line containing data from the previous observation.

WEIGHT

Syntax WEIGHT {*variable* | OFF} ;

Alias RETURN

Remarks The optional WEIGHT command specifies the weight variable for the analysis. The data values for the WEIGHT variable supply the weight the corresponding observation. The default is OFF.

Two Example Automated CTA Programs

Imagine an application in *finance*. In light of the recent calamitous failure of home mortgages, it is decided that a new credit-screening methodology is needed. Toward this objective a bank creates a dataset consisting of records describing all mortgages granted in the past three years (for exposition, imagine N=300 loans were made, of which, 10%, or 30 loans, were in default). The class variable is whether or not the loan went into default (label this class variable "Loan", and use dummy-codes 1=solvent, 0=default). The weight is the value of the loan in dollars (label this variable "Value"). Finally, imagine data are available for twenty attributes (Var1-Var20). Of these, Var1-Var10 are ordered, and the rest categorical.

Imagine that data and program files have been saved, and the output file will be saved, in the "c:cta" directory. As per the automated *CTA system job-naming convention*, *a common name is used for data, program and output files*: the name of the data file is "loan.dat"; the name of the program file is "loan.pgm"; and the name of the output file is "loan.out." The following code defines data and output files, assigns class, weight, and attribute variables, and defines the categorical attributes:

```
open c:\cta\loan.out;
output c:\cta\loan.out;
vars loan value var1 to var20;
class loan;
attr var1 to var20;
cat var11 to var20;
weight value;
```

It is decided *a priori* that, to increase the likelihood of the model cross-generalizing when applied to a validity sample, model endpoints should represent at least 5% of the total sample (5% of N=300 is N=15):

```
mindenom 15;
```

It is also decided *a priori* that to increase the likelihood of the model cross-generalizing, only variables stable in leave-one-out analysis would be allowed as model nodes:

```
loo stable;
```

It is decided *a priori* to use the system default (on) for weighting by prior odds intact, as another means of increasing the likelihood of the model cross-generalizing to an independent random sample, and also to explicitly maximize ESS (*setting priors off explicitly maximizes overall PAC*). The conventional experimentwise Type I error rate (p<0.05) is selected for pruning[13] to maximize ESS (experimentwise p<

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

0.05 is used automatically during model growth to control overfitting[1]):

```
prune .05;
```

Because there are relatively many categorical variables, it is decided to use Fisher's exact test to assess *p* for categorical variables[1] and reduce the number of Monte Carlo simulation experiments conducted:

```
usefisher;
```

Because the sample is modest in size, as is the number of attributes, and in light of the small number of failed loans in conjunction with the minimum denominator specification, it is decided that full enumeration of the first three nodes is feasible and appropriate, using 25,000 Monte Carlo experiments to compute *p* for all ordered attributes:

```
enumerate;
mcarlo iter 25000 cutoff .05 stop 99.9;
title loan default weighted CTA;
go;
```

Imagine an application in *space physics*. A phased array of 16 high-frequency antennas located in Goose Bay (Labrador), with a total transmitted power exceeding 6 kilowatts, was used to target free electrons in the ionosphere.[14] The class variable was labeled "return": "good" returns showed evidence of some type of structure in the ionosphere, and "bad" returns failed to provide evidence of structure (dummy-coded as "1" *vs*. "0", respectively). Received signals were processed using an autocorrelation function with two arguments per signal: time of pulse and pulse number. Because there were 17 pulse numbers for the Goose Bay system, there were thus 34 ordered attributes ("X1-X34"). There was no weight variable, and no categorical attribute. The objective is to maximize overall PAC—the total number of accurately classified good and bad returns.

Imagine that data and program files have been saved, and the output file will be saved, in the "c:cta" directory. As per the automated *CTA system job-naming convention*, *a common name is used for data, program and output files*: the name of the data file is "radar.dat"; the name of the program file is "radar.pgm"; and the name of the output file is "radar.out." The following code defines data and output files, and assigns class and attribute variables:

```
open c:\cta\radar.out;
output c:\cta\radar.out;
vars return x1 to x34;
class return;
attr x1 to x34;
```

It is decided *a priori* that, to maximize overall PAC achieved, the endpoint minimum denominator and model maximum depth would be unconstrained, but rather explicitly optimized by the program (no commands required).

Also, to maximize overall PAC it was decided to let attributes load as nodes even if unstable in LOO analysis, so long as their ESS in LOO analysis exceeded the ESS achieved by any other attribute:

```
loo 0.05;
```

It is decided *a priori* to set priors off in order to *explicitly* maximize overall PAC:

```
priors off;
```

The default setting for optimal pruning is priors on, so the prune command has to be adjusted to indicate that priors is set to off. Also, to maximize overall PAC, a statistical marginal loading will be allowed in the optimally-pruned model:

```
prune .10 nopriors;
```

Because there are no categorical attributes, the usefisher command is omitted. Because the sample is moderate in size, as is the

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

number of attributes, and the attributes are ordered with few ties, analysis will be resource intensive. Also, 100,000 Monte Carlo experiments will be used in order to provide adequate statistical power for the small denominator endpoints that are anticipated:

```
mcarlo iter 100000 cutoff .05 stop 99.9;
```

Because UniODA analysis showed many attributes are loo-unstable, the analysis is judged to be too computationally intensive to attempt full enumeration on the first pass through the data via CTA (omitting the enumeration command results in an *algorithmic* analysis by default). Thus, after specifying enumeration of the root variable only, and providing a title, the program is ready to go:

```
enumerate root;
title RADAR maximum-PAC CTA;
go;
```

Advantages of Automated versus Manually-Derived CTA

Perhaps the most striking advantage of the automated software is that it is able to accomplish the example analyses just described, whereas *neither* of those analyses are *possible* to accomplish using manual derivation. Two specific advantages of the automated software are integrated automated pruning procedures: (a) sequentially-rejective Sidak "Bonferroni-type" multiple comparisons adjustment[1] to prevent model overfitting during the growth phase of the analysis; and (b) optimal pruning to maximize ESS at any specified experimentwise alpha level after growth has ceased.[13] And, those with experience conducting manual CTA using ODA[1] software would likely be amazed to hear that in recent speed trials (N=351, 34 continuous attributes) the automated software was able to solve *enumerated* CTA models averaging 0.7 CPU seconds per model, running 5,000 Monte Carlo experiments on a 3 GHz Intel Pentium D microcomputer. An *algorithmic* CTA derived manually for either type of CTA would typically require one or more man-days.

Initial comparisons of automated versus manual methods clearly reveal that the increased depth of search afforded by the enumeration capabilities of the automated software typically returns stronger, more efficient models.[8] The enumerated models may also be more consistent with original hypotheses than manually-derived counterparts.[15] Preliminary investigations in our laboratory suggest that the advantages of automated software become even more striking in applications which feature numerous, scattered, missing data. We are aware of several studies which compare previously-completed manually-derived CTA models *vs*. models derived using automated software, either planned or in progress. Monte Carlo simulation studies comparing the two methods are obviously warranted.

It is exciting to witness, whether as actor or spectator, the beginning of a new area of inquiry involving a powerful and evolving new methodology. Manually-derived CTA may be likened to an early telescope, focused by moving the body much like a trombone slide. Initial exploration using this early tool was fruitful and informative, and motivated the development of the automated system, which may be likened to a modern telescope. The modern instrument allows for pinpoint placement of the machine in any particular area (forcenode), aspect control including depth of field (maxlevel) and search (mcarlo iter; enumerate), luminosity (minobs; mindenom), fuzzy control (loo stable *vs*. .0x), and a standardized measure of acuity (ESS). It is likely that using these controls in a variety of applications will lead to refinements in the controls themselves, as well as in the methods of their operations, and these developments in turn may result in the creation of additional control features. For these reasons we anticipate surprising findings and major advances in the understanding of absolute and comparative capabilities of automated CTA—soon to come.

2155-0182/10/$3.00

References

[1]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. Washington, DC, APA Books, 2005.

[2]Grimm LG, Yarnold PR. *Reading and understanding multivariate statistics*. Washington, DC, APA Books, 1995.

[3]Grimm LG, Yarnold PR. *Reading and understanding more multivariate statistics*. Washington, DC, APA Books, 2000.

[4]Yarnold PR. Discriminating geriatric and nongeriatric patients using functional status information: an example of classification tree analysis via UniODA. *Educational and Psychological Measurement* 1996, 56:656-667.

[5]Yarnold PR, Soltysik RC. Optimal data analysis: a general statistical analysis paradigm. *Optimal Data Analysis* 2010, 1:10-22.

[6]Ostrander R, Weinfurt KP, Yarnold PR, August G (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology* 1998, 66: 660-672.

[7]Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: an example of hierarchically optimal classification tree analysis. *Statistics in Medicine* 1997, 16: 1451-1463.

[8]Yarnold PR, Soltysik RC. Manual *vs*. automated CTA: optimal preadmission staging for inpatient mortality from *Pneumocystit cariini* pneumonia. *Optimal Data Analysis* 2010, 1:50-54.

[9]Yarnold PR, Soltysik RC. *Hierarchically optimal classification tree analysis: applications in medicine and allied health disciplines*. Submitted.

[10]Yarnold PR. *How to obtain a CTA model manually using ODA software*. In preparation.

[11]Automated CTA software webpage: www.OptimalDataAnalysis.com.

[12]Yarnold PR, Soltysik RC. *Automated CTA: initial standards for exploratory analysis*. In preparation.

[13]Yarnold PR, Soltysik RC. Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis* 2010, 1:23-29.

[14]Sigillito VG, Wing SP, Hutton LV, Baker KB. Classification of RADAR returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest* 1989, 10:262-266.

[15]Coakley RM, Holmbeck GN, Bryant FB, Yarnold PR. Manual *vs*. automated CTA: predicting adolescent psychosocial adaptation. *Optimal Data Analysis* 2010, 1:55-58.

Author Notes

Correspondence: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. eMail: Journal@OptimalDataAnalysis.com.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

How to Save the Binary Class Variable and Predicted Probability of Group Membership from Logistic Regression Analysis to an ASCII Space-Delimited File in *SPSS 17 For Windows*

Fred B. Bryant, Ph.D.
Loyola University, Chicago

This note explains the steps involved and provides the SPSS syntax needed to run two-group logistic regression analysis using SPSS 17 for Windows, and output to an ASCII space-delimited data file the binary class variable and predicted probability of group membership (i.e., “Y-hat”) from an SPSS logistic regression analysis.

1. Obtain an SPSS data set containing a binary class variable (e.g., sex), along with categorical (e.g., city1, city2, city3, colorA, colorB, colorC) and continuous (e.g., age) attributes. Missing data should be indicated with a value (e.g., -9) in the SPSS data set.

2. Open the SPSS data set, and run the following syntax file, which saves predicted probability of group membership as a variable named PRED_1 in the active SPSS data file.

```
LOGISTIC REGRESSION VARIABLES sex
 /METHOD=ENTER age raceA raceH city2 city3
 /CONTRAST (city3)=Indicator
 /CONTRAST (city2)=Indicator
 /CONTRAST (colorA)=Indicator
 /CONTRAST (colorC)=Indicator
 /SAVE=PRED
 /CLASSPLOT
 /PRINT=GOODFIT
 /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20)
   CUT(0.5).
```

3. If desired, in Variable View, edit the SPSS data file to rename PRE_1 as “lryhat,” for example, to reflect “logistic regression y-hat.”

4. From the drop-down SPSS Windows menu, select Transform, Recode into Same Variable, and change the value of “system missing” (blank) to -9 (or value used) for the PRE_1 (lryhat) variable. Then resave the SPSS data set.

5. Run the following SPSS syntax to write a space-delimited ASCII data file which is named “lryhat.dat” and which contains a code for the class variable (e.g., sex) and the predicted probability of group membership (e.g., lryhat):

```
FORMATS sex (f4.0).
FORMATS lryhat (f13.8).
write outfile='c:\lryhat.dat' records=1
   /1 sex lryhat.
execute.
```

6. Locate the file “lryhat.dat” in the root folder for the c:\ drive, and move this file to the ODA directory for analysis.

Author Notes

Correspondence should be sent to Fred B. Bryant at: Department of Psychology, Loyola University Chicago, 6525 North Sheridan Road, Chicago, IL, 60626. eMail: fbryant@luc.edu.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

An Internet-Based Intervention for Fibromyalgia Self-Management: Initial Design and Alpha Test

William Collinge, Ph.D.
Collinge and Associates

Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

and

Paul R.Yarnold, Ph.D.
Optimal Data Analysis, LLC

Self-Monitoring And Review Tool, or SMART, is an interactive, internet-based, self-monitoring and feedback system, which helps people discover and monitor links between their own health-related behaviors, management strategies, and symptom levels over time. SMART involves longitudinal collection and optimal analysis of an individual's self-monitoring data, and delivery of personalized feedback derived from the data. Forty women with fibromyalgia (FM) enrolled in a three-month alpha test of the SMART system. Utilization, satisfaction, and compliance were high across the test period, and higher utilization was predictive of lower anxiety, and improved physical functioning and self-efficacy.

FM is a chronic illness without medical cure and prevalence estimated as high as twelve million Americans—primarily women of child-bearing age, although children, the elderly, and men are also affected.[1] Predominant symptoms include widespread musculoskeletal pain, multiple tender points, fatigue, concentration and memory problems, and gastrointestinal complaints.[2-4] For a sample of 594 FM patients in an HMO, scores on a well-being measure were lower than for patients having advanced cancer, chronic obstructive pulmonary disease, and rheumatoid arthritis.[5] FM is believed to involve a central pain processing dysfunction of the nociceptive system, particularly the central nervous system, and chronic inflammation.[6,7] There is evidence of elevated corticotropin-releasing hormone and substance P in the cerebrospinal fluid of FM patients.[8]

Many conventional and complementary treatment and management options have been tried, with mixed results across patients. The prevailing approach involves a combination of

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

pharmacological treatments for specific symptoms, and recommendations involving education, lifestyle change, exercise, and self-help.[9-13] Meta-analysis of 49 treatment outcome studies concluded that nonpharmacological approaches, specifically cognitive, behavioral and exercise, appear to be more efficacious in improving self-report of FM symptoms and functional well-being than pharmacological treatment alone.[14] There is consensus that success in managing FM depends heavily on the patient's daily efforts in self-management. FM patients who use adaptive problem-focused coping strategies report lower levels of pain and emotional distress, and higher levels of perceived control over symptoms, versus patients who don't use such strategies.[15] A patient's decision to adopt a self-management strategy is influenced by a sense of self-efficacy about pain, a strong internal health locus of control, and a belief that one is not necessarily disabled or damaged by FM. In addition, positive self-management behaviors are related to decreased guarding, increased exercise, seeking support from others, activity pacing, and use of coping self-statements.[16,17]

While habits and patterns of daily living are directly compromised by FM, they also present opportunities for beneficial impact. Those reporting success in managing or recovering from FM commonly report significant—sometimes radical—changes in daily habits, roles, interpersonal relationships, behavior patterns, and life goals.[18] Several self-management strategies which have shown benefit in FM warrant inclusion in an individualized self-management approach. For example, pacing involves learning to regulate one's activity levels and energy expenditure throughout the day and the week, and FM patients have a higher level of pre-morbid "action proneness" in their lifestyles compared to control samples, suggesting failure to appropriately pace oneself may be both a predisposing and perpetuating factor.[19] Studies of mindfulness-based stress reduction and related approaches report moderate to marked improvement in FM symptom levels.[20,21] Performance of aerobic exercise in FM patients is also related to improvement in functioning, depression, pain, range of motion, and general FM impact.[22]

In light of the numerous types of self-management options available, it is important for patients to appraise the impacts of specific approaches and avoid depleting resources on unhelpful strategies. It is also important not to abandon a potentially helpful approach without a fair trial. Yet, the frustrations of carrying out new behaviors without the benefit of systematic tracking of use and effects, may cause the patient to reject a strategy prematurely.[23] Self-monitoring is common practice—and can be life-saving in conditions such as diabetes and hypertension, where methods can be concrete (glucose testing, blood pressure) and the results easily grasped. Self-monitoring is also used in weight loss programs to promote awareness of eating habits, but self-monitoring tools must be user-friendly to enhance the likelihood of their successful use.[24] Many FM patients attempt to monitor their illness with journals or diaries, but problems with memory and concentration make it difficult to track symptoms reliably or process information in an organized or systematic way. However, successful self-monitoring holds great potential of illuminating dynamics of the illness. For example, in a study of 63 FM patients who participated in one week of daily pain, sleep quality, and fatigue assessment, path analysis revealed that poor sleep quality alone fully accounted for the positive relationship between pain and fatigue, thus substantiating the mediational role of sleep quality.[25]

This report describes the design and initial clinical evaluation of a proprietary, inter-active, user-friendly, web-based, systematic approach to self-management of FM.

Design and Method

Subject Recruitment. The study was conducted with cooperation of the MaineHealth Learning Resource Center (MHLRC), the patient education program of the MaineHealth Network consisting of ten non-profit medical centers, hospitals and outpatient clinics serving the ten counties of southern, central and western Maine. The IRB was the Maine Medical Center Research Institute. Primary publicity was a mailing regarding the "Fibromyalgia Wellness Project" to self-identified FM patients in the MHLRC program mailing list. The mailing guided prospective applicants to the study website to read details of the project, download the consent document, download a required Medical Information Form, and complete a secure on-line application.

The on-line application collected contact information, demographics and date of FM diagnosis. Data were entered into the applicant's personal database on the project's secure server, and automatically forwarded to the PI via email. Upon receipt the PI contacted the applicant for a telephone interview to cover eligibility criteria: (1) a diagnosis of primary FM (i.e., not secondary to lupus, rheumatoid arthritis, or other condition) according the official American College of Rheumatology criteria (documented on Medical Info Form); (2) not under current treatment for another serious medical condition; (3) able to read and speak English and complete the assessment forms; (4) physically able to attend the introductory meeting; and (5) daily access to the internet. Eligible applicants mailed or faxed the Medical Information Form signed by their physician attesting to criteria 1 and 2 above, with the date of diagnosis. After receiving this and the signed consent form, the subject was assigned a user-name and password to access the project web site to complete the first Monthly Survey. Recruitment was terminated once 40 subjects completed this process: using statistical power analysis this sample size was determined to be sufficient to detect a moderate increase in longitudinal survey scores with one-tail $p<0.05$ and 90% power.

Data Collection. All data collection was conducted via the project website using SMART Log software, with data for a given subject stored in the subject's personal database. Each subject completed a Monthly Survey five times, on days 1, 30, 60, 90 and 120 of the study. Compensation was $20 each time. Monthly survey instruments used were the Fibromyalgia Impact Questionnaire[26] or FIQ (uses a one-week recall period to obtain ratings of pain, physical functioning, fatigue, depression, anxiety, stiffness, morning tiredness, job difficulty, days of paid work missed, number of good days in the past week, and total score); the SF12[27] (uses a four-week recall period to obtain ratings of role limitations due to illness or emotional problems, physical functioning, bodily pain, mental health, vitality, social functioning, and general health perceptions); Self-Efficacy for Chronic Disease Scale[28] or SECDS (uses a six-item scale rating self-confidence in managing the challenges of illness); and the Health Locus of Control Scale[29] (HLCS), Form C (condition-specific for FM and obtaining ratings of perceived control in terms of internality, chance externality, and powerful others externality for doctors and other people). The first two administrations of the Monthly Survey were treated as baseline data, and the last three treated as intervention phase data.

Data on utilization of SMART Log consisted of weekly counts of submissions on the project website. Ratings of *satisfaction* with the program, and of the perceived *relevance* of the program to the subject's health, were recorded by subjects at the end of each usage of SMART Log using four-point Likert-type rating scales anchored at the extremes by "not at all" and "completely".

The Intervention. After completing the second baseline Monthly Survey each subject attended a three-hour orientation meeting introducing SMART Log and supporting features on the project website: five meetings each with 3-12 subjects were conducted, one meeting per subject. Subjects were instructed how to use the SMART Log and requested to use it at least three to four times per week. It was explained that the more they used it, the more powerful would be their personal database in its ability to provide meaningful feedback via the SMART Profile. Subjects were compensated $2 each time the used the system up to five times per week, although they were free to use the system for as many days as they wished. Compensation for travel and time was $50.

One of the greatest challenges in any behavioral intervention is compliance. We chose the prescription of "at least three to four times per week" to be a reasonable goal given the illness burden and stress load commonly experienced by people with FM. We did not want compliance to be perceived as an added burden (e.g., "required every day"), yet we needed enough submissions to capture sufficient data within the 13-week use period of the study to perform the desired statistical analyses. As per the instructions, a mean of 3.5 uses per week was considered full compliance in this study.

Immediately after the orientation meeting (day 31 of the project) the SMART Log function was activated on the web site and came available for subject use. SMART Log consists of two sections. First, the Inputs Checklist captures 24-hour reporting of lifestyle behaviors and self-management strategies and stressors in five categories: sleep and rest; meals and snacks; self-care; general activities; and "unique items"—up to five user-defined inputs unique to each individual. Second, the Symptom Rating Scale captures user ratings of ten of the most prominent FM symptoms over the past 24 hours.

After 30 days of using SMART Log (day 60 of the project), a SMART Profile was posted on a weekly basis on the subject's private database until the end of the project (day 120). To accomplish this, each week, each subject's cumulative SMART Log data were analyzed using exact single-subject statistical methods via optimal data analysis (ODA): power analysis simulating a moderate effect revealed univariate optimal discriminant analysis (UniODA) was appropriate for a sample of 22 to 47 days, and hierarchically optimal classification tree analysis (CTA) was appropriate thereafter.[30] Findings for each subject were individually summarized in a narrative ("Profile") comprising statements about statistically significant associations found between the subject's Inputs Checklist and ratings of specific symptoms. After logging in a subject could click the SMART Profile tab and access the latest Profile. Over time, as more data accumulated for the subject, more (and more detailed) statements became possible. Subjects received a total of ten weekly Profiles by the end of the project.

If a subject's data did not yield at least one significant association, then a general Profile statement as follows was received:

> "Your SMART Profile does not yet show significant connections between your inputs and symptom levels. Either you need more submissions, or there's not enough variation in your data so far, or both.
>
> *More submissions*: To date you have 15 SMART Log submissions. As you add more you are more likely to accumulate enough data to show connections in future Profiles. More frequent use of SMART Log may help get you there sooner. For example, if you've been submitting only two or three times per week, increase to four or more.
>
> *More variation*: You can also boost your odds of finding connections by adding more variation to your inputs. If you're doing the same things all the time (the same inputs), your symptom levels are more likely to stay the same. This is good reason to begin changing inputs you think could possibly affect your symptoms—like your bedtime, eating schedule, meal sizes, work hours, self-care practices, stressors, or other inputs. Consider some changes you can try and begin experimenting with them."

If the subject's data yielded at least one significant association, then a Profile in the following format was received:

> "Your SMART Profile shows significant connections between your inputs and symptom levels. Keep in mind the more you use SMART Log, the more likely you are to see other connections in future Profiles. Also remember that if you're doing the same things all the time (same inputs), your symptom levels are more likely to stay the same. This is good reason to try changing inputs you think could possibly affect your symptoms—like your bedtime, eating schedule, meal sizes, self-care practices, work hours, stressors or other inputs. Consider some changes you can try and begin experimenting with them.
>
> *Your Profile for this week*: Your pain level is least when you have lunch by 1:37 PM. Your concentration problems are least when your afternoon nap is no longer than 25 minutes. Your stiffness is least when work or school-related activity is longer than 2 hours 36 minutes AND morning exercise is longer than 5 minutes. Your fatigue level is least when domestic activity is no more than mildly stressful. Your gastrointestinal symptoms are least when your afternoon snack is very light OR childcare stress is 3 or less."

Results and Discussion

For this study usability criteria comprise data on recruitment and retention, utilization and compliance, and satisfaction.

Recruitment and Retention. A total of 40 women having an official diagnosis of FM (ACR criteria) enrolled in the project. The study was originally planned for adults 18 or over, but one 16 year-old appealed for an exception, and with the permission of the NIAMS program officer, the IRB and her parent, she was enrolled in the study. A summary of demographic features of the total recruited alpha test sample are summarized in Table 1.

Table 1: Sample Demographic Characteristics

Age	Mean 46.5 (SD 12.4) Range 16-66
Sex	Female: 40 Male: 0
Ethnicity/ Race	White: 34 African American: 3 Hispanic/Latino: 2 Native American: 1
Years FM Diagnosed	Mean 7.9 (SD 5.7) Range 0.8-27.3
Marital Status	Married: 27 Not married: 13
Education	Some HS: 1 HS grad: 2 Some college: 17 BA: 9 Some grad school: 5 Grad degree: 6
Employment	Full time: 9 Part time: 11 Seeking: 1 Retired: 3 Disabled: 14 Student: 2

Utilization. Aggregate and individual measures of utilization are considered presently: utilization and compliance were high and stable over the course of the use period.

Aggregate Utilization. Because subjects were instructed to utilize the SMART Log at least three to four times per week, perfect compliance is operationalized as a mean of 3.5 times per week. Considered as a whole, the sample of 39 patients completing the study submitted a mean of 4.05 (SD=1.61) SMART Log entries per week over the 13-week study period. This corresponds to a 95% confidence interval (95% CI) of 3.65 to 4.45 mean submissions per week, or mean aggregate compliance between 104.2% and 127.2% over the duration of the study.

Aggregate SMART Log utilization was also examined weekly for the total sample. Seen in Table 2, lowest aggregate mean weekly utilization (indicated in red) occurred in the first week of data collection: this was the only week for which the upper bound of the 95% CI was lower than 100% mean aggregate compliance. Four of the six highest mean aggregate utilization weeks (indicated in blue) occurred within the first six weeks of data collection: for these the lower bound of the 95% CI exceeded 100% mean aggregate compliance. And, four of the six intermediate mean aggregate utilization weeks (indicated in black) occurred within the final six weeks of data collection: for these the 95% CI overlapped 100% mean aggregate compliance. The temporal trend was unreliable: the correlation between study week and mean utilization was insignificant (r=-0.31, p<0.30), indicating mean weekly aggregate utilization was consistent over the data collection period. Examination of 95% CI overlap indicated that lowest mean aggregate utilization occurred in the first week of data collection, and that mean aggregate utilization in weeks 2 and 4 were greater than in weeks 8, 10, 11 and 12.

Individual Utilization was defined using measures of relative quantity, and change over the three-month use period. *Utilization quantity* was defined via a mean-split procedure whereby the mean number of SMART Log entries per week for an individual is compared against the aggregate mean number of SMART Log entries per week: individuals having a mean which is greater than the aggregate mean are considered "higher utilizers" (N=20, 51% of sample), and those having a mean lower than the aggregate are considered "lower utilizers" (N=19, 49% of sample). Absence of skew in weekly utilization data yielded nearly identical samples sizes using this procedure.

Table 2: Weekly SMART Log Mean Aggregate Utilization

Week	Mean (SD)	95% CI	Corresponding % Compliance
1	2.95 (1.70)	2.54 – 3.36	72.6 – 96.0
2	5.00 (1.93)	4.56 – 5.44	130.2 – 145.4
3	4.44 (1.73)	4.02 – 4.86	114.8 – 138.8
4	5.00 (2.70)	4.48 – 5.52	128.0 – 157.8
5	3.97 (1.93)	3.53 – 4.41	100.8 – 126.0
6	4.70 (2.41)	4.20 – 5.18	120.0 – 148.0
7	4.00 (2.41)	3.51 – 4.49	100.2 – 128.2
8	3.94 (2.54)	3.44 – 4.46	98.2 – 127.4
9	4.26 (3.29)	3.68 – 4.84	105.2 – 138.2
10	3.36 (3.08)	2.80 – 3.92	80.0 – 112.0
11	3.21 (2.88)	2.67 – 3.75	76.2 – 107.2
12	3.51 (2.95)	2.96 – 4.06	84.6 – 116.0
13	4.26 (4.18)	3.61 – 4.92	103.2 – 140.6

Note: For % compliance values, lowest values are indicated in red, intermediate values in black, and highest values in blue.

Change in utilization across time was operationalized using lag-1 autocorrelation, or ACF(1): for a single individual all pairs of measurements recorded at times i and i-1 are constructed, the data pairs are combined, and the i and i-1 data are correlated.[31] The result is a Pearson correlation coefficient bounded by 1.0 and -1.0: a negative value of ACF(1) indicates scores recorded recently are lower than scores recorded previously, and thus the individual is making fewer SMART Log entries (decreasing utilization) as the study proceeds. A positive value of ACF(1) indicates scores recorded recently exceed scores recorded previously, and thus the individual is making more SMART Log submissions (increasing utilization) as the study proceeds. Presently, 16 subjects (41% of sam-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

ple) had ACF(1)>0 and thus increasing SMART Log utilization, and 23 (59% of sample) had decreasing utilization. The small number of data pairs available to compute ACF(1) had weak statistical power to identify reliable ACF(1) coefficients, yet two negative and three positive statistically reliable (p<0.05) coefficients materialized. Quantity of utilization and change in utilization over time were negatively correlated (r=-0.52, p<0.0009), indicating moderate (r^2= 0.27) regression to the mean.[31]

Satisfaction. Using consistent methodology and obtaining results consistent with those obtained for utilization, aggregate as well as individual satisfaction measures were considered presently: mean ratings of perceived relevance and overall satisfaction with the intervention were scaled as “mostly satisfied” over the course of the study; data were stable (the 95% CI for both variables overlapped across all use weeks); and an aggregate decline in satisfaction over time was attributable to outlying negative ratings, as the majority of participants reported increasing satisfaction over time.

Aggregate Satisfaction. After every use of the SMART Log its *relevance* (i.e., “How satisfied are you that today's SMART Log addressed matters relevant to your well-being?”) and *satisfaction* (“On the whole, how satisfying has your use of this program been to date?”) was rated by the patient. The sample of 39 patients recorded a mean relevance rating of 3.20 (SD=0.57, 95% CI=2.96–3.44), and a mean satisfaction rating of 3.16 (SD=0.64, 95% CI= 2.90–3.41). Considered in relation to the rating response scale used (1=not at all, 2=somewhat, 3=mostly, 4=completely), both the mean ratings correspond to “mostly satisfied.” Examination of weekly aggregate relevance and overall satisfaction data indicated that for both measures the 95% CI for weekly aggregate means overlapped across all weeks. Mean aggregate relevance (r= -0.63, p<0.02) and satisfaction (r=-0.56, p<0.05) declined over time, but analysis of individual satisfaction shows the decline was attributable to outlying ratings of a minority of patients.

Individual Satisfaction. Paralleling the individual utilization measures, individual program relevance and satisfaction ratings were conceptualized in terms of both quantity and change over time. *Quantity* was operationalized by a mean split procedure: mean relevance and satisfaction scores for an individual were compared against mean aggregate scores. Individuals having a mean relevance score greater than the aggregate mean considered the SMART Log relatively “more relevant” to their well-being (N=14, 35.9% of sample), and individuals having a mean lower than the aggregate mean considered the SMART Log relatively “less relevant” (N=25, 64.1% of sample). Similarly, individuals having a mean satisfaction score greater than the aggregate mean were relatively “more satisfied” (N=16, 41.0% of sample), and individuals having a mean lower than the aggregate mean were “less satisfied” (N=23, 59.0% of sample).

And, also paralleling treatment of utilization data, analysis of temporal effects in individual relevance and satisfaction measures was operationalized by ACF(1) coefficients. Insufficient variance made computation of ACF(1) impossible for the relevance data of 14 patients, and for the satisfaction data of 16 patients. For *relevance*, 20 patients (80% of sample) had ACF(1)>0 and perceived increasing relevance, and five patients (20% of sample) reported decreasing relevance. For *satisfaction*, 19 patients (78.3% of sample) had ACF(1)>0 and reported increasing satisfaction, and four patients (21.7% of sample) reported decreasing satisfaction. Negative trends across time were noted for aggregate mean scores dominated by outlying ratings made by a few dissatisfied patients, but the majority of the sample had positive trends when their data were examined individually: a type of Simpson’s Paradox.[32] In spite of weak statistical power to test the reliability of the ACF(1) coefficients, for relevance 12 ACF(1) coeffi-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

cients had $p<0.05$ (6 exceeded 0.92), and for satisfaction 13 coefficients had $p<0.05$ (9 exceeded 0.89). Quantity and change over time were unrelated for relevance ($p<0.38$) and satisfaction ($p<0.94$) ratings.

Outcome Data. In the present study the outcome data were obtained from instruments administered in the Monthly Survey.

Utilization. First, UniODA and CTA[30] were used to identify relationships involving the *quantity of utilization*. Higher utilization predicted greater reduction over the use period in FIQ Anxiety scores ($p<0.04$), as 13/15 (86.7%) patients reporting decreased Anxiety were higher utilizers, and 6/11 (54.6%) patients with increased Anxiety were low utilizers of SMART Log. The effect was moderate (ESS=41.2%) and stable in jackknife validity analysis.

Higher utilization also predicted greater increase over the use period in SF12 Physical Functioning ($p<0.05$): 5 of 10 (50.0%) patients with increased Physical Functioning over the use period were higher utilizers of the SMART Log, while 12 of 13 (92.3%) patients reporting decreased Physical Functioning over the use period were lower utilizers. The effect was moderate (ESS=42.3%) and jackknife-stable.

Higher utilization was marginally predictive of improvement over the use period in Health Locus of Control (HLC) versus lower utilization ($p<0.09$): 10 of 18 (55.6%) patients reporting decreased Chance locus of control over the use period were higher utilizers, while 14 of 18 (77.8%) patients reporting increased Chance locus of control were lower utilizers. The effect was moderate (ESS=33.3%) and stable in jackknife analysis.

Higher utilization was also marginally predictive of older age ($p<0.09$): 17 of 20 (77.8%) patients who were at least 43 years of age were higher utilizers, and 10 of 19 (52.6%) patients who were 42 years old or younger were lower utilizers. The effect was moderate (ESS= 37.6%) and stable in jackknife analysis.

Finally, CTA was used to compare high- versus low-utilizers, and a two-attribute model emerged. Five of six (83.3%) patients reporting increased SF12 Physical Functioning over the use period were higher utilizers, versus one of six (16.7%) who were lower utilizers ($p<0.05$). Five of six (83.3%) patients reporting reduced Physical Functioning *and* decreased external (Doctor) locus of control over the use period ($p<0.005$) were higher utilizers, versus 7 of 7 (100%) lower utilizers who instead reported increased external (Doctor) locus of control. The effect was strong (ESS=77.8%) and stable in jackknife validity analysis.

UniODA and CTA were next employed to identify relationships involving *change in utilization over time*, and only one statistically marginal association emerged involving Internal locus of control ($p<0.07$). Of patients having positive ACF(1) coefficients (indicating increasing utilization of the SMART Log over the use period), 13 of 15 (86.7%) reported increased Internal locus of control over time; of patients having negative ACF(1) coefficients (indicating decreasing use over time), 10 of 21 (47.6%) patients reported a decreased Internal locus of control over time. This effect was moderate (ESS= 34.3%), and stable in jackknife validity analysis.

In summary, higher utilization predicted significantly greater improvement over the use period in anxiety and physical functioning, and marginally greater improvement over the use period in internal health locus of control. The strength of these associations was moderate, and the findings are likely to cross-generalize to an independent random patient sample. There was also a significant positive association between *change* in utilization and *change* in health locus of control over the use period.

Satisfaction. Next, UniODA and CTA were used to identify relationships involving the *quantity of satisfaction*. Ratings of satisfaction and relevance were nearly perfectly associated: all 16 patients reporting higher satisfaction had a mean relevance score greater than 3.12, and 22 of 23 (95.6%) patients with lower satisfaction had a mean relevance score of 3.12 or lower: p< 0.0001, ES=95.6%. Accordingly, only ratings of satisfaction are considered further. The only statistically significant association identified involved change in Stiffness over time (p<0.04): 8/14 patients reporting increased satisfaction reported decreased Stiffness over the use period, and 14/17 (82.4%) patients reporting decreasing satisfaction over the period reported increasing Stiffness: this moderate effect (ESS = 39.5%) was stable in jackknife validity analysis.

Profiles Delivered. As a direct measure of the performance of the SMART system in fulfilling the intended function of delivering person-specific SMART Profile reports, we assessed system velocity—the quantity of statistically reliable feedback reports produced by the SMART system for patients as a function of time, over the use period. The first four weeks of the 13-week SMART Log use period was used to build the subject's personal database. At the end of the fourth week the SMART Profile function was activated and subjects began receiving weekly reports. As described earlier, Profiles were of two types: if the analysis yielded no significant associations between inputs and symptom levels, the patient received a general SMART Profile statement explaining how the Profile is produced and encouraging further submissions and variation of inputs. If the analysis yielded one or more significant associations ("significant profile") the patient received a general SMART Profile statement appended with the associations found.

As expected, the number of subjects receiving significant profiles increased over time as databases accumulated more SMART Log data. As seen in Table 3, in the first week of SMART Profile production 20% of subjects had significant associations, and this percentage increased until reaching asymptote at a mean of 88% over the final five weeks.

Qualitative Data. All 39 subjects who completed the study provided qualitative data at follow-up via focus groups. Response content received from participants was analyzed via QSR-N6 software.[33]

Table 3: Weekly Number of Statistically Significant SMART Profiles

Week	Number of Significant Profiles Divided by Number of Patients	Percent Significant Profiles
1	8/40	20.0
2	16/40	40.0
3	28/40	70.0
4	30/39	76.9
5	31/39	79.5
6	34/39	87.2
7	34/39	87.2
8	35/39	89.7
9	35/39	89.7
10	34/39	87.2

Concerning overall impression of the SMART Log instrument, 54% of responses were unqualified positive, and an additional 41% were positive with qualifications. Issues identified included time pressures due to family, travel, work, tiredness, and some confusion as to how to apply some SMART Profile statements in one's daily life. Overall, 95% of the responses expressed were favorable to the tool. A modal comment was that the use of SMART Log raised consciousness regarding day-to-day activities and their effects on the individual independently of the content of the Profiles.

Concerning overall impression of the SMART Profile reports, 74% of responses were coded as "useful." Dissenting respondents indicated they failed to submit SMART Log entries often enough over the study period, and thus didn't receive as many significant SMART Profile statements as they would have liked.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Concerning relevance of the reports to one's health, 79% of responses reflected a positive evaluation. Components singled out as especially relevant were eating patterns, exercise, spirituality, meditation, social activity, getting out, behavior awareness, use of routines, and sleep habits.

All subjects identified a "most important" personal discovery achieved as a consequence of using the SMART system. Most frequently mentioned discoveries involved consciousness-raising regarding the impact of daily behavior choices (48%); the importance of self-care activities (27%); the need for more information on the disease and its effects, both personally as well as on others (14%); the value of socialization as a form of support (7%); and the role of daily exercise in well-being (2%).

Finally, during follow-up focus group discussions, obstacles or difficulties in using the SMART system were framed in terms of life getting in the way, time commitments, work, not being much of a computer user, family commitments, vacation, time of year, depression reducing motivation, and limited computer access.

Eighteen-Month Follow-Up. A total of 22 subjects in the Phase I study were located 18 months after the study terminated and asked to complete a nine-item self-report survey, and 19 subjects returned responses that could be analyzed. Subjects were asked to answer three questions about each of three different time periods—just before the Phase I project began, right after the Phase I project terminated, and now. The first question was: "about how many hours per day *were you able* to be productive?" The second question was: "about how many hours per day *did you want* to be productive?" The third and final question was: "on a scale from one to ten ("not at all" to "completely", respectively), how satisfied were you *with your productivity*?" Productivity was defined to subjects as: "what *you* would consider being productive in your life. This could include work, school, homemaking, or taking care of others, whatever it means to be productive." Table 4 gives the mean (and standard deviation) for survey items.

Table 4: Mean Productivity Self-Ratings

Survey Item	Baseline	Study End	18-Months Post-Test
Hours ABLE to be productive	6.4 (4.1)	7.5 (3.9)	7.6 (4.0)
Hours WANT to be productive	10.5 (3.4)	10.4 (3.5)	11.0 (3.0)
SATISFACTION with productivity	4.1 (2.3)	6.2 (2.5)	5.7 (2.7)

As seen, patients reported a 17% increase in the mean number hours per day they were able to be productive by the end of the study (*p*<0.008), and this gain was maintained after the study terminated. Desired mean daily productivity did not vary across rated time periods (*p*>0.17). Mean satisfaction increased 51% by the end of the study (*p*<0.0004) and diminished slightly but not reliably (*p*<0.55) to an increase in mean satisfaction of 39%, 18 months after the study terminated (p<0.02). Thus, exposure to the SMART Log intervention significantly increased mean self-reported productivity and satisfaction ratings, and these gains were maintained in follow-up.

Consistent with the thesis of the disconfirmation paradigm[34], diminishing difference between actual and desired hours of productivity was significantly correlated with increasing satisfaction for ratings made before the study (*p*<0.02) and at 18-month follow-up (*p*<0.0005), and marginally for ratings obtained when the study ended (*p*<0.12).

Conclusion

For this four-month pilot study with 39 patients retention was very high and compliance with use of the intervention—a significant

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

challenge in most behavioral medicine research, and especially with FM—was high and stable across the study period. Retention, compliance, and satisfaction ratings indicate that the intervention has high usability.

Analyses revealed moderate to strong clinical benefits predicted by the subject's frequency of utilization of the SMART system. Specifically, higher utilization was associated with significantly greater positive change over the use period in anxiety, physical functioning and health locus of control. These effects were stable in jackknife validity analysis, indicating they are likely to cross-generalize to independent random patient samples. Discussed earlier, impaired physical functioning and emotional well-being are central aspects of the disease burden of FM, and increased health locus of control is essential to patients taking an active role in self-management.

An important qualitative finding was that subjects reported the systematic and regular use of SMART Log raised their awareness of the impacts of daily behaviors independently from the feedback they received from SMART Profile. Focus groups revealed instances where subjects became aware of impacts and changed behavior even though their SMART Profile had not yet provided feedback about that issue. Thus, one mechanism of benefit of using the SMART system may be a heightened sense of self-awareness that it engenders. This is important because subjects are not guaranteed that a given SMART Profile will present significant associations, as there may not always be enough submissions, sufficient variability, or actual underlying association within their submitted data to show reliable associations. It thus appears that the success of the SMART Log program does not derive exclusively from contents of SMART Profile.

Preliminary findings reveal the SMART system is a user-friendly, structured and systematic approach to self-management of a chronic illness affecting 6 to 12 million Americans. The success of the alpha project motivated a beta test using a substantially larger sample of FM patients, which is currently underway in our laboratory.

References

[1] http://www.rheumatology.org/public/factsheets/fibromya_new.asp? Accessed 12/1/07.

[2] Brown MM, Jason LA. Functioning in individuals with chronic fatigue syndrome: increased impairment with co-occurring multiple chemical sensitivity and fibromyalgia. *Dynamic Medicine* 2007, 31:6:6doi:10.1186/1476-5918-6-6.

[3] Sephton SE, Studts JL, Hoover K, Weissbecker I, Lynch G, Ho I, McGuffin S, Salmon P. Biological and psychological factors associated with memory function in fibromyalgia syndrome. *Health Psychology* 2003, 22:592-597.

[4] Adler GK, Manfredsdottir VF, Creskoff KW. Neuroendocrine abnormalities in fibromyalgia. *Current Pain and Heada3che Reports* 2002, 6: 289-298.

[5] Kaplan RM, Schmidt SM, Cronan TA. Quality of well being in patients with fibromyalgia. *Journal of Rheumatology* 2007, 27:785-789.

[6] Wood PB, Patterson JC 2nd, Sunderland JJ, Tainter KH, Glabus MF, Lilien DL. Reduced presynaptic dopamine activity in fibromyalgia syndrome demonstrated with positron emission tomography: a pilot study. *Journal of Pain* 2007, 8:51-58.

[7] Bazzichi L, Rossi A, Massimetti G, Giannaccini G, Giuliano T, De Feo F, Ciapparelli A, Dell'Osso L, Bombardieri S. Cytokine patterns in fibromyalgia and their correlation with clinical manifestations. *Clinical and Experimental Rheumatology* 2007, 25:225-230.

[8] Lucas HJ, Brauch CM, Settas L, Theoharides TC. Fibromyalgia—new concepts of pathogen-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

esis and treatment. *International Journal of Immunopathology and Pharmacology* 2006, 19: 5-10.

[9]American College of Rheumatology. http://www.rheumatology.org/public/factsheets/fibromya.asp, 2004.

[10]Jones KD, Burckhardt CS, Clark S, Bennett RM, Potempa K. A randomized controlled trial of muscle strengthening versus flexibility training in FM. *Journal of Rheumatology* 2002, 29: 1041-1048.

[11]Jones KD, Lipton G. Exercise interventions in fibromyalgia: clinical applications from the evidence. *Rheumatic Disease Clinics of North America* 2009, 35:373-391.

[12]Friedberg F. *Fibromyalgia and chronic fatigue syndrome: seven proven steps to less pain and more energy*. New Harbinger, Oakland, CA, 2006

[13]Friedberg F. Chronic fatigue syndrome, fibromyalgia, and related illnesses: a clinical model of assessment and Intervention. *Journal of Clinical Psychology* In press.

[14]Rossy LA, Buckelew SP, Dorr N, Hagglund KJ, Thayer JF, McIntosh MJ, Hewett JE, Johnson JC. A meta-analysis of fibromyalgia treatment interventions. *Annals of Behavioral Medicine* 1999, 21:180-191.

[15]Turk DC. Suffering and dysfunction in fibromyalgia syndrome. *Journal of Musculoskeletal Pain* 2002, 10:85-96.

[16]Cronan TA, Serber ER, Walen HR. Psychosocial predictors of health status and health care costs among people with fibromyalgia. *Anxiety, Stress & Coping: An International Journal* 2002, 15:261-274.

[17]Nielson WR, Jensen MP. Relationship between changes in coping and treatment outcome in patients with Fibromyalgia Syndrome. *Pain* 2004, 109:233-241

[18]Mengshoel AM, Heggen K. Recovery from fibromyalgia—previous patients' own experiences. *Disability and Rehabilitation* 2004, 7:46-53.

[19]Houdenhove B, Neerinckx E, Onghena P, Lysens R, Vertommen H. Premorbid “overactive” lifestyle in chronic fatigue syndrome and fibromyalgia: an etiological factor or proof of good citizenship? *Journal of Psychosomatic Research* 2001 51:571-576.

[20]Grossman P, Tiefenthaler-Gilmer U, Raysz A, Kesper U. Mindfulness training as an intervention for fibromyalgia: evidence of postintervention and 3-year follow-up benefits in well-being. *Psychotherapy and Psychosomatics* 2007, 76: 226-233.

[21]Collinge W, Yarnold PR, Raskin E. Use of mind/body self-healing practice predicts positive health transition in chronic fatigue syndrome: a controlled study. *Subtle Energies & Energy Medicine* 1998, 9:171-190.

[22]Valim V, Oliveira L, Suda A, Silva L, de Assis M, Barros Neto T, Feldman D, Natour J. Aerobic fitness effects in fibromyalgia. *Journal of Rheumatology* 2003, 30:1060-1069.

[23]Burckhardt CS. Nonpharmacologic management strategies in fibromyalgia. *Rheumatic Disease Clinics of North America* 2002, 28:291-304.

[24]Mossavar-Rahmani Y, Henry H, Rodabough R, Bragg C, Brewer A, Freed T, Kinzel L, Pedersen M, Soule CO, Vosburg S. Additional self-monitoring tools in the dietary modification component of The Women's Health Initiative. *Journal of the American Dietetic Association* 2004, 104:76-85.

[25]Nicassio PM, Moxham EG, Schuman CE, Gevirtz RN. The contribution of pain, reported sleep quality, and depressive symptoms to fa-

tigue in fibromyalgia. *Pain* 2002, 100:271-279.

[26]Burckhardt CS, Clark SR, Bennett RM. The fibromyalgia impact questionnaire (FIQ): development and validation. *Journal of Rheumatology* 1991, 18:728-733.

[27]Ware J, Kosinski M, Dewey B. *SF-36v2 health survey user's manual*. Quality Metric, Inc, Lincoln, RI, 2002.

[28]Lorig K, Stewart A, Ritter P, Gonzalez V, Laurent D, Lynch J (1996). *Outcome measures for health education and other health care interventions*. Sage, Thousand Oaks, CA, 1996, pp. 24-25, 41-45.

[29]Wallston KA, Stein MJ, Smith CA. Form C of the MHLC scales: a condition-specific measure of locus of control. *Journal of Personality Assessment* 1994, 63:534-553.

[30]Yarnold PR, Soltysik RC. *Optimal data analysis: guidebook with software for Windows*. APA Books, Washington, DC, 2005.

[31]Yarnold PR (1992). Statistical analysis for single-case designs. In: FB Bryant, L Heath, E Posavac, J Edwards, E Henderson, Y Suarez-Balcazar, and S Tindale (Eds.), *Social psychological applications to social issues, volume 2: methodological issues in applied social research*. Plenum, New York NY, 1992, pp. 177-197.

[32]Yarnold PR. Characterizing and circumventting Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement* 1996, 56:430-442.

[33]QSR International: http://www.qsrinternational.com.

[34]Thompson DA, Yarnold PR. Relating patient satisfaction to waiting time perceptions and expectations: the disconfirmation paradigm. *Academic Emergency Medicine* 1995, 2:1057-1062.

Author Notes

Mail correspondence to Dr. Collinge at: Collinge and Associates, PO Box 309, Kittery Point, ME 03904. Send Email to Dr. Collinge at: William@Collinge.org.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Junk Science, Test Validity, and the Uniform Guidelines for Personnel Selection Procedures: The Case of *Melendez v. Illinois Bell*

Fred B. Bryant, Ph.D. and Elaine K.B. Siegel
Loyola University Chicago Hager & Siegel, P.C.

This paper stems from a recent federal court case in which a standardized test of cognitive ability developed by AT&T, the Basic Scholastic Aptitude Test (BSAT), was ruled invalid and discriminatory for use in hiring Latinos. Within the context of the BSAT, we discuss spurious statistical arguments advanced by the defense, exploiting certain language in the current Uniform Guidelines for evaluating the fairness and validity of personnel selection tests. These issues include: (a) how to avoid capitalizing on chance; (b) what constitutes "a measure" of job performance; (c) how to judge the meaningfulness of group differences in performance measures; and (d) how to combine data from different sex, race, or ethnic subgroups when computing validity coefficients for the pooled, total sample. Pursuant to the Uniform Guidelines' standard for unfairness, when one ethnic group scores higher on an employment test, the test is deemed "unfair" if this difference is not reflected in a measure of job performance. Although studies validating selection instruments often survive the unfairness test, such data are vulnerable to bias and manipulation, if appropriate statistical procedures are not used. We consider both the benefits (greater clarity and precision) and the potential costs (loss of legal precedent) of revising the Uniform Guidelines to address these issues. We further discuss legal procedures to limit "junk science" in the courtroom, and the need to reevaluate validity generalization in light of Simpson's "false correlation" paradox.

The purpose of this paper is to share our insights from a recent federal court case, which we refer to as *Melendez*, involving a claim of employment discrimination in personnel selection, *Melendez v. Illinois Bell Telephone Company*, No. 90 C 5020 (N.D. Ill. Sept. 16, 1994), aff'd, 79 F.3d 661 (7th Cir. 1996).[1] These insights arise from certain defenses advanced by the employer, in which dubious statistical procedures were justified by language from current federal guidelines for validating personnel selection tests, the Uniform Guidelines for Employee Selection Procedures, promulgated jointly by the United States Equal Employment Opportunity Commission and the United States Departments of Labor, Justice, and the Treasury [43 Fed. Reg. 38,290 (August 25, 1978); EEOC, 29 CFR Part 1607]. We refer to these as the Uniform Guidelines.

After providing some background to the particular legal case involved, we describe the original validation studies that formed the heart of the litigation, and present research evidence which was the main point of contention at trial. After summarizing the evidence against the validity of the personnel selection test in question—the Basic Scholastic Aptitude Test (BSAT)—we highlight some apparent ambiguities in the Uniform Guidelines. Comparable ambiguities exist in both the Standards for Educational and Psychological Testing[2] and in the Society for Industrial and Organizational Psychology's Principles for the Validation and Use of Personnel Selection Procedures.[3] Ironically, although the Uniform Guidelines are intended to promote equality of employment opportunity regardless of race, religion, and gender, they do not expressly prohibit the use of certain research practices that produce spurious artifacts, and which actually perpetuate discrimination in the workplace.

In this paper we share our observations with professionals within the psychological testing, statistical analysis, human resources and legal communities; discuss the application of Uniform Guidelines in maintaining consistency vis-à-vis professional standards; and conclude by recommending a reevaluation of the procedure of validity generalization in light of Simpson's "false correlation" paradox (i.e., paradoxical confounding).

Historical Context

What was this trial all about? Plaintiff Carmelo Melendez claimed he was denied equal employment opportunity in applying for a job with defendant Illinois Bell Telephone Company. Mr. Melendez was born and raised in Puerto Rico, and moved to East Chicago in the middle of his grade school years. Though he spoke no English, Mr. Melendez was placed in a monolingual English classroom. A straight-A student in Puerto Rico, in the United States he got F's. By struggling hard, he learned English, taught himself the skills he needed to advance, and raised his grades until, by the time he graduated from high school, he was earning B's.

It was then, however, that Mr. Melendez first encountered an obstacle that he could not overcome, and that he would confront throughout his adult life: standardized ability tests. He performed miserably on the SAT, and could not attend college. He decided to apply for an entry-level position in metallurgy at the local steel mill. He failed the standardized entry examination, however. Yet another standardized test kept him out of the military.

Mr. Melendez persevered, and eventually got his college degree. He also became a certified x-ray technician, and he eventually worked for the federal Civil Rights Commission. He went on to become the host of a Chicago-area television talkshow. Then, in 1988, he applied for a job as Assistant Manager of Urban Affairs for Illinois Bell.

The job description called for a person who could interface with the local Latino community, to assess emerging urban trends for use in marketing telecommunications services. The successful applicant should be able to interact with community leaders and residents, and to communicate effectively in a bilingual setting, orally and in writing.

Illinois Bell required all external applicants for its first-level management jobs to surmount three separate pass-fail hurdles. Applicants had to have a college diploma, graduating

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

in the top half of the class. Applicants had to pass a structured, standardized interview, demonstrating a sufficient level of leadership. Finally, applicants had to take the standardized Basic Scholastic Aptitude Test (BSAT), scoring at or above a raw pass-fail cutoff score of 196. This cognitive ability test was the central focus of the court case.

The BSAT is a standardized paper-and-pencil test, purporting to assess verbal and quantitative ability, much like the SAT. It also includes questions designed to tap the ability to follow directions, in which one must indicate answers while listening to a tape-recording which contains complex, conflicting instructions. Each subsection of the test is timed, or "speeded," and the entire test takes about one hour.

Despite his college degree and his success on the leadership interview, Mr. Melendez failed the BSAT. He grew depressed and despondent, and became estranged from his family for more than a year. Not long after his rejection by Illinois Bell, however, Melendez won a position with the federal government. He has performed successfully there ever since, and has risen to a position of authority.

Based on his experience, Mr. Melendez believed that the BSAT was unfair because it was not job-related. He saw no connection between the skills required to do well on the job of Assistant Urban Affairs Manager, and the skills required to pass the BSAT. To right the wrong, he filed suit against Illinois Bell for employment discrimination.

Adverse Impact of the BSAT

Before turning to the evidence concerning test validity, we first consider the BSAT's impact on applicants of different ethnicity (i.e., the BSAT pass-fail rates for different racial or ethnic groups). Table 1 presents pass-fail rates for whites, African-Americans and Latinos on the BSAT separately for two time periods: 1979 and 1987-88. The 1979 statistics are for 591 managerial applicants, and are taken directly from the original AT&T validation report: in 1979, about 3 in 4 whites passed the test, versus 1 in 5 African-Americans, and 1 in 2 Latinos.[4]

Table 1: Rates of Success and Failure on the BSAT for Different Racial Groups

		Racial Group					
		White		Black		Latino	
Time Period		P	F	P	F	P	F
1979	*n*	265	79	42	151	25	29
	%	77	23	22	78	47	53
1987-88	*n*	344	51	83	62	50	44
	%	87	13	57	43	53	47

Between-Group Pairwise Comparisons via Fisher's Exact Test

	W79	W87	B79	B87	L79
W87	.000459				
B79	.000001	.000001			
B87	.000018	.000001	.000001		
L79	.000010	.000001	.000827	.21	
L87	.000014	.000001	.000001	.60	.50

Note: Pairwise comparisons were performed using two-tailed Fisher's exact test computed using ODA software.[5] Row and column headings indicate both ethnic class (W= white, B=Black, L=Latino) and time period (79=1979, 87=1987-88). Tabled for each unique combination of row and column is the *p*-value (six significant digits) for the exact test comparing pass/fail rates of the corresponding samples. *P*-values indicated in red are statistically significant at experimentwise $p<0.05$ based on an appropriate Bonferroni criterion (see discussion in paper: $p<.05/1115$, or $p<0.000046$); *p*-values indicated in blue are statistically significant at the generalized criterion (per-comparison $p<0.05$); *p*-values indicated in black are not significant.[5]

The 1987-88 pass-fail statistics are from Illinois Bell's records, from a sample of 634 applicants for first-level management positions. During the 1987-88 period, most whites—nearly 9 in 10—passed the test, versus 6 in 10 African-Americans and 5 in 10 Latinos.

To evaluate these pass-fail rates, there is a guideline for judging the impact of an employment test on different ethnic groups. This rule-of-thumb is known as the "four-fifths rule." According to this guideline, a test has an *adverse impact* on an ethnic group whose pass rate is less than four-fifths the rate of the group with the highest test pass-rate: "A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact" (Uniform Guidelines, §1607.4.D). The Uniform Guidelines define "adverse impact" as: "A substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group" (Uniform Guidelines, §1607.16.B).

In 1979, for example, whites had the highest pass-rate on the BSAT, at 77% (see Table 1). The BSAT, then, had an adverse impact on any group in 1979 whose BSAT pass-rate falls below four-fifths of 77% (or below 61.6%). The 1979 pass-rates for African-Americans (22%) and Latinos (47%) are clearly lower than the four-fifths mark of 61.6%.

For the 1987-88 period, under the Uniform Guidelines' four-fifths rule, the BSAT had an adverse impact on any group whose pass-rate falls below four-fifths of the white pass rate of 87% (or below 69.6%). Because pass rates for African-Americans (57%) and Latinos (53%) are below this four-fifths mark of 69.6%, the BSAT had an adverse impact on both of these groups during 1987-88, according to the Uniform Guidelines' standard.

This evidence of strong and consistent adverse impact makes test validity even more vital. Rejecting such a large number of minority applicants might be defensible, if the test accurately predicted important on-the-job performance. For example, imagine using a valid test of visual acuity to select fighter-pilots; if minority applicants have worse eyesight than majority applicants, then so be it. It is an entirely different matter, however, if the test has nothing to do with on-the-job performance. If minorities do not actually have worse eyesight, then the test unfairly denies them equal employment opportunity. In the case of the BSAT, the evidence for test validity is particularly critical, given the unequivocal adverse impact on minorities. In the words of the Uniform Guidelines: "Reliance upon a selection procedure which is significantly related to a criterion measure, but which is based upon a study involving a large number of subjects and has a low correlation coefficient will be subject to close review if it has a large adverse impact..." (Uniform Guidelines, §1607. 14(B)(6).

BSAT Validation Studies

Two validation studies of the BSAT formed the heart of the litigation, and the trial gravitated around certain research evidence from these studies. In the late 1970s, AT&T industrial/organizational psychologists developed the BSAT, using test components originally written by the Educational Testing Service (ETS), which also developed the SAT, LSAT, GRE, and other cognitive ability tests. One of the AT&T psychologists drafted the final research report containing two validation studies, which assessed the relationship between BSAT scores and job performance. These studies purported to evaluate the BSAT's predictive validity, i.e., its ability to predict subsequent on-the-job performance. Illinois Bell relied on these validation studies in using the BSAT to screen its job applicants.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

The first of the two validation studies, referred to as the *Preliminary Study*, focused on entry-level managers already hired at 8 different company locations throughout the country. This Preliminary Study included 229 managers who had earlier taken a large battery of standardized tests, including the School and College Ability Test (SCAT) and the predecessor of the BSAT, the Bell System Qualification Test (BSQT). One year after these applicants were hired their job performance was evaluated by their supervisors, who rated each applicant's job performance using a set of 13 criterion measures, developed through a job analysis of management positions, including ratings of skills in planning, decision making, oral and written communications, leadership, resistance to stress, interpersonal awareness, and a global rating of overall job performance. The test developers then selected a subset of verbal and math items based on correlations with supervisor ratings, and these items became the BSAT. Researchers then examined the relationship between test score and rating of overall job performance to establish a pass-fail cut-score for the test, which was implemented throughout AT&T companies.

The second validation study, referred to as the *Followup Study*, focused on 286 job applicants who were applying for entry-level management positions in 11 different AT&T company locations. Applicants selected for participation were given the BSAT (using the pass-fail cut-score determined in the Preliminary Study), and then one year later, their supervisors were asked to rate each employee on a set of 15 performance criteria. As in the Preliminary Study, researchers examined the correlation between test scores and performance ratings, trying to cross-validate the findings from the Preliminary Study. Thus, both validation studies concern the predictive validity of the test, that is, whether the test accurately predicts job performance and is therefore job-related.

Validity Evidence for the BSAT

What evidence is there concerning the predictive validity of the BSAT? The primary validity evidence in the validation studies consists of Pearson product-moment correlation coefficients relating applicants' test scores to supervisors' performance ratings.

Preliminary Study. Turning first to Table 2, note that the Preliminary Study reports no figures for Latinos. Instead, for African-Americans and whites separately and for the pooled data set, it reports correlations between BSAT scores and each of the 13 performance ratings. Note that the BSAT shows a statistically significant correlation with ratings of overall job performance for the total sample, $r(151)=0.38$, $p<0.00001$. For whites, however, only 4 of the 13 criterion measures show a statistically significant ($p<0.05$) relationship with BSAT score. Indeed, BSAT scores had no significant relationship with ratings of overall job performance for whites. Averaging across all correlations for whites (mean $r=0.128$, $p<0.08$), the BSAT predicts about 2% of the variance in whites' performance ratings. This represents a Hedges corrected effect-size of 0.26, equivalent to an experimental effect in which the treatment group scores about one-quarter of a standard deviation above the control group.

Also note that, for African-Americans, 7 of the 13 performance ratings (including overall job performance) show a statistically significant relationship with BSAT score. Averaging across all correlations (mean $r=0.314$, $p<0.006$), the BSAT explains about 10% of the variance in African-Americans' performance ratings (Hedges corrected $\underline{g}=0.65$). Considered together, this evidence from the Preliminary Study suggests that the BSAT is largely invalid for use with whites, but has marginal validity for use with African-Americans. We return later to the first column of Table 2, giving validity coefficients for the total group.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Table 2: Preliminary Study Correlations Between BSAT Score and Job Performance Ratings for Different Groups

	Groups		
	Total	White	Black
Job Skills	*n*=153	*n*=94	*n*=39
Organizing and Planning	.28*	.09	.34*
Decision Making	.30*	.20*	.27
Decisiveness	.39*	.25*	.36*
Oral Communications	.23*	.08	.43*
Written Communications	.28*	.21*	.26
Leadership	.36*	.02	.54*
Interpersonal Awareness	.25*	.09	.30*
Behavior Flexibility	.20*	.04	.20
Fact Finding	.38*	.29*	.24
Resistance to Stress	.21*	.11	.18
Energy Management	.15	.04	.08
Potential	.42*	.11	.42*
Overall Job Performance	.38*	.13	.46*

Note: Adapted from Tables 4 and 8 of the original validation report.[4] An asterisk (*) indicates $p<0.05$ at the generalized (per-comparison) criterion.[5] *N* for the total sample is greater than the sum of the *n*s for the white and black groups because the Preliminary Study included 16 Hispanics and 4 "other minorities" whose data were pooled in the analysis of the total sample. Discussed further ahead in the paper, the "false correlation paradox" (paradoxical confounding) is present when an index for pooled samples lies outside the range of index values for individual samples considered separately (indicated in red).

Followup Study. Table 3 gives validity coefficients for the Followup Study. Again the BSAT shows a significant correlation with ratings of overall job performance for total sample, $r(284)=0.21$, $p<0.001$. For whites, 4 of 15 performance ratings show a significant relationship with BSAT score: averaging coefficients (mean $r=0.077$, $p>0.19$), the BSAT predicts about 2% of the variance in whites' performance ratings (corrected $g=0.19$). For African-Americans, 8 of 15 validity coefficients are significant: averaging coefficients (mean $r=0.215$, $p<0.01$), the BSAT predicts about 6% of the variance in African-Americans' performance ratings (corrected $g=0.44$). BSAT score was significantly related to ratings of overall job performance for both whites and African-Americans, though these effect sizes again were relatively small.

The fourth column in Table 3 reports the only direct empirical evidence available concerning the validity of the BSAT for use in hiring Latinos. Only one of the 15 validity coefficients was significantly different from zero for Latinos ($r=0.24$, $p<0.05$, one-tailed) for Latinos. The sole significant coefficient (for coordination) was reported as nonsignificant in the original validation study. Essentially, this means that the BSAT does no better than chance in predicting how Latinos will perform on the job (mean $r=0.093$, $p>0.32$, corrected $g=0.21$).

In relation to the present case, this is the single most relevant piece of validity evidence in the entire report. *Plainly, these data do not support the validity of using the BSAT to hire Latinos.*

Inflation of Apparent Validity Vis-à-Vis Extensive Analysis: The "Trolling" Problem

It would be one matter if the coefficients were the only analyses in the validation studies. If this were the case, then there would be 49 tests of statistical hypotheses in the Preliminary Study (Table 3) and 60 tests in the Followup Study (Table 4), for a total of 109 tests.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Table 3: Followup Study Correlations Between BSAT Score and Job Performance Ratings for Different Groups

		Groups		
Job Skills	Total (*n*=286)	White (*n*=147)	Black (*n*=76)	Latino (*n*=57)
Organizing and Planning	0.17*	0.08	0.19	0.15
Decision Making	0.18*	-0.12	0.21*	-0.08
Oral Communications	0.17*	0.10	0.26*	0.01
Written Communications	0.28*	0.18*	0.44*	0.10
General Administration	0.11*	0.09	0.22*	0.07
Supervision	0.01	0.02	0.10	0.09
Coordination	0.19*	0.01	0.30*	0.24*
Behavior Flexibility	0.10*	0.03	0.20	0.08
Fact Finding	0.25*	0.10	0.33*	0.18
Problem Solving	0.22*	0.17*	0.25*	0.08
Resistance to Stress	0.05	0.06	0.05	0.05
Ability to Learn and Develop	0.16*	0.05	0.17	0.10
Tolerance of Ambiguity	0.12*	0.08	0.17	0.07
Management Potential	0.16*	0.16*	0.08	0.12
Overall Job Performance	0.21*	0.14*	0.26*	0.14

Note: Adapted from Table 18 of the original validation report .[4] *N* for the total sample is greater than the sum of the *n*s for the three subgroups because the Followup Study included six Asians whose data were pooled for total sample analysis. An asterisk (*) indicates p<0.05 at the generalized (per-comparison) criterion. The coefficient indicated in red was reported as being nonsignificant in the original validation report, but is actually statistically significant at the generalized criterion (p<0.05, one-tailed).

Tallying across the entire validation report, however, reveals that more than a thousand statistical tests were performed—all using the p<0.05 level of statistical significance. Of those 1000 tests, 50 would be expected simply by chance alone to be statistically significant at per-comparison p< 0.05, although exactly which effects are attributable to chance cannot be known. The validity evidence is thus inflated, as the excessive statistical testing adds a substantial number of chance correlations to the true correlations. Accordingly, well-known procedures for controlling the experimentwise Type I error-rate should be used.[5] For example, among the most commonly employed methods for reducing the number of “false-positive” results when conducting numerous statistical tests is the so-called “Bonferroni adjustment, in which an adjusted p-value is obtained by dividing the desired alpha-level by the number of p-values examined. For the BSAT validation report, a Bonferroni-adjusted p-value would be roughly .05/1100, or p<0.00005. This is the cost for undertaking vast numbers of analyses indiscriminately, when analyses can and should be more clearly focused.[5,6]

Table 4: Followup Study Means and Standard Deviations for the 15 Job Performance Ratings, and for BSAT Score for Whites (n=147) and Latinos (n=57)

	Whites		Latinos	
Job Skills	Mean	sd	Mean	sd
Organizing and Planning	5.22	1.13	4.99	1.04
Decision Making	5.15	0.93	4.93	0.84
Oral Communications	5.31	1.15	4.88	1.18
Written Communications	5.24	1.16	4.82	1.23
General Administration	5.12	1.06	4.68	0.87
Supervision	4.98	1.23	4.92	1.32
Coordination	5.39	1.00	4.85	0.90
Behavior Flexibility	5.25	1.17	4.83	1.08
Fact Finding	5.38	1.11	4.88	1.06
Problem Solving	5.18	1.03	4.86	1.10
Resistance to Stress	5.22	1.11	5.25	0.97
Ability to Learn and Develop	5.71	1.01	5.41	1.20
Tolerance of Ambiguity	5.08	1.13	4.81	0.86
Management Potential	6.02	1.93	6.65	2.08
Overall Job Performance	5.35	1.06	5.11	1.08
BSAT score	218.62	13.89	209.78	15.49

Note: Adapted from Tables 14 and 17 of the original validation report.[4] Scores on the 7-point rating scales have been reversed so that high scores reflect better ratings. Means indicated in red differ from the mean for whites with $p<0.05$ by Tukey's Honest Significant Difference multiple range test. These statistically significant group differences were found when following up significant F-values from initial one-way analyses of variance with white, Latino, and African-American groups.

In the *Melendez* case, we took the "middle-ground" approach of adjusting the criterion to $p<0.05$ in the validation studies. This reduces spurious effects (Type I errors), without unduly increasing false no-difference conclusions (Type II errors) due to low statistical power. Evaluated at this criterion, *there are no significant validity coefficients in the Followup Study*.

Illinois Bell defended its inflationary statistical procedures with a statement in the Uniform Guidelines that one should usually use the $p<0.05$ level in establishing statistical significance: "...Generally, a selection procedure is considered related to the criterion, for the purposes of these guidelines, when the relationship between performance on the procedure and per-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

formance on the criterion measure is statistically significant at the $p<0.05$ level of significance" (Uniform Guidelines, §1607.14.B(5)). The Uniform Guidelines nonetheless require the use of "professionally acceptable statistical procedures" in computing validity coefficients (Uniform Guidelines, §1607.14.B(5)), and also caution users to avoid using procedures that capitalize on chance: "*Overstatement of validity findings*. Users should avoid reliance upon techniques which tend to overestimate validity findings as a result of capitalization on chance unless an appropriate safeguard is taken. Reliance upon a few selection procedures or criteria of successful job performance when many selection procedures or criteria of performance have been studied, or the use of optimal statistical weights for selection procedures computed in one sample, are techniques which tend to inflate validity estimates as a result of chance. Use of a large sample is one safeguard; cross-validation another." (Uniform Guidelines, §1607.14.B.(7).

Clearly, performing 1100 statistical tests at the $p<0.05$ level is a procedure that capitalizes on chance. Under the Guidelines, an adjustment to the alpha-level is in order, minimally one such as using $p<0.01$. To reduce jury confusion over these technical issues, the Uniform Guidelines should include specific recommenddations (e.g., Bonferroni adjustments) for reducing Type I error when a large number of statistical tests have been conducted.

Filling the Validity Gap with Junk Science: Reinventing Statistics

Through the above evidence, plaintiff demonstrated that the BSAT had, at most, negligible validity for white applicants, and no validity for Latino applicants. And how did Illinois Bell respond to plaintiff's showing? Illinois Bell's expert witness, an organizational psychologist, asserted that if the BSAT truly had a nonsignificant (i.e., zero) statistical relationship with job performance for Latinos, then half of the validity coefficients for Latinos should have been positive, and half negative. In other words, if the true value of the correlation in the population is zero, then there should be just as many positive validity coefficients as negative. He noted, however, that 14 of the 15 coefficients for Latinos in the Followup Study were positive (if not statistically significant). He then calculated the binomial probability of obtaining 14 positive coefficients and 1 negative, given a 0.50 probability for obtaining either sign (i.e., $z=3.30$, $p<0.0005$). From this scenario, he deduced that, despite the complete lack of any correlation in the AT&T validation study, the BSAT was nonetheless valid for Latinos—and at a highly significant *p*-value!

By pitting one expert's statistical analysis against the other's, this form of "junk science" has great potential to confuse the jury. To clarify the issue for the layperson, what is needed is a logical, easy-to-follow explanation of the difference between the two opposing views of the same data. However, this is not always easily developed.

In the *Melendez* case, we explained the statistical issue in commonsense terms by using an archery analogy. Testing the validity of the BSAT is like an archery contest. An archer fires 15 arrows at a target; to determine his proficiency, we count how many arrows hit the target. Using the BSAT to predict the 15 performance criteria for Latinos, we count how many times it shows a statistically significant relationship between test score and job performance. Table 3 shows that for Latinos, all 15 arrows missed the mark. By the rules of the game, the archer does not score, and the BSAT is off target (and invalid).

By Illinois Bell's logic, however, 14 of the 15 arrows flew in the target's general direction (i.e., 14 of the 15 validity coefficients were positive) and only 1 arrow flew in the opposite direction (i.e., there was only one negative validity coefficient), and so therefore the archer was a success (and the BSAT is valid for Latinos because only one of its validity coefficients was negative). This is fallacious. At issue is the *magnitude* of the validity coefficients in the positive direction, not just whether the signs of these coefficients are positive or negative. For the BSAT, the magnitudes were insufficient to establish a statistically significant relationship. As the Seventh Circuit ruled on appeal, there was "strong evidence of the BSAT's inability to predict job performance," which supported the trial court's finding that "the BSAT's discriminatory impact was unjustified by Illinois Bell's legitimate business needs" (79 F.3d at 669). That is, the BSAT explains too little variance in performance ratings to be considered valid for use in hiring Latinos. If the BSAT does not provide useful, job-related information, then its use cannot be justified, given the strong evidence of its adverse impact.

The Admissibility of "Junk Science" in the Courtroom

Illinois Bell's spurious defense, that its test is "valid" because of its positive (though not statistically significant) correlations with performance ratings, exemplifies the dangers of "junk science" in the courtroom. As the U.S. Supreme Court has cautioned: "Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it." (*Daubert*, 509 U.S. at 595 (quoting Weinstein, 1992)). Due to defendant's discovery abuse, Melendez was able to bar, altogether, the testimony of the company's expert witness. More typically, dubious science is precluded through a ruling by the trial court that the information is inadmissible under the Federal Rules of Evidence.

Expert testimony is specifically governed by Federal Rule of Evidence 702, which establishes ground rules for admitting expert testimony: "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise" (Fed. R. Evid. 702). As interpreted in the landmark *Daubert* decision, Rule 702 allows expert testimony when it is both relevant and scientifically reliable. In *Daubert* the Court appointed the trial judge as the "gatekeeper" of expert testimony, asserting: "[t]his entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue." (*Daubert*, 509 U.S. at 592-593). The Court went on to explain: "The inquiry envisioned by Rule 702 is, we emphasize, a flexible one. Its overarching subject is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission. The focus, of course, must be solely on principles and methodology, not on the conclusions that they generate" (*Daubert*, 509 U.S. at 594-595).

More recently, the U.S. Supreme Court held unanimously that a trial court's decision to admit or exclude expert evidence should be accorded great deference (*Joiner*, 118 S.Ct. 512). Noting that trial judges typically are not scientists, Supreme Court Justice Stephen Breyer encouraged judges to take the initiative to clarify scientific issues (*Joiner*, 118 S.Ct. 512, 520-521 (Breyer, J., concurring)). They may, for example, utilize their authority to appoint their own experts, or use pretrial hearings to explore the issues. The *Daubert* Court explains that the goal is a middle ground, between "a 'free-for-all' in which befuddled juries are confounded by absurd and irrational pseudoscientific assertions", and "a stifling and repressive scientific orthodoxy" (*Daubert*, 509 U.S. at 595-596). The Court recalled the differences between scientific inquiry and the law, emphasizing that Federal Rules of Evidence are "designed not for the exhaustive search for cosmic understanding but for the particularized resolution of legal disputes" (*Daubert*, 509 U.S. at 597).

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

The Concept of Test "Fairness"

Besides adverse impact and validity, another critical concept in judging whether or not a test in discriminatory is test "fairness." Although researchers have suggested numerous definitional frameworks and statistical models of test fairness[7-12], two approaches are often used in litigation to define "unfairness," and to determine whether a test is "unfair."

Anne Cleary[13] pioneered one of these definitions at the Educational Testing Service. According to Cleary's model, a test is considered "unfair" when it predicts performance differently for different ethnic groups. This differential prediction is detected in the form of statistically significant differences between groups in the slopes and in the intercepts of the regression lines relating test scores to performance. Thus, a test is considered "fair" when there are no significant differences in errors of prediction between groups, using a common regression line. Ironically, by a strict application of Cleary's definition, an invalid test could be deemed "fair." It would not be unfair, for example, to use a coin-flip to hire job applicants, because this selection procedure does not predict performance better for one ethnic group than for another. It is equally invalid for both groups.

Another definition of "unfairness" prominent in the courts is that used in the Uniform Guidelines, under which a test is "unfair" when: "...members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance..." (Uniform Guidelines, §1607.14.B(8)(a)).

In practice, one determines whether a test is "unfair" by comparing group means on the test, then looking for comparable mean-differences in group performance ratings. If one group scores higher on the test, it must also do better on the job. Stated differently, a test is "unfair" if it denies job opportunities to a group whose actual job performance is up to par.

Applying the Uniform Guidelines' definition of "unfairness" to the BSAT Followup Study, Latinos had significantly lower BSAT scores than whites, and passed the test at a significantly lower rate (77% *vs*. 47% in 1979; 87% *vs*. 53% in 1987-88; Table 1). In contrast, on 12 of the 15 performance criteria, Latino and white performance ratings did not differ significantly (Table 4). In other words, 80% of the performance measures (including overall job performance) failed to show lower scores for Latinos than whites. Considered together, this evidence shows that the BSAT is "unfair" to Latinos within the meaning of the Uniform Guidelines.

Twisting the Uniform Guidelines to Establish Test "Fairness"

In a spurious defense of the BSAT, Illinois Bell purported to rely on the Uniform Guidelines' definition of test unfairness. At trial the defense argued that the company adhered to the letter of the Uniform Guidelines, and advanced two lines of defense based on the Guidelines. Neither the law nor professional standards support these arguments.

What constitutes "a measure of job performance"? On cross-examination, the defense read to the jury the Uniform Guidelines' definition of test unfairness in Section 14.B(8)(a), and then asked:

Q: "Am I correct, Doctor, that this says that the differences in scores are not reflected in differences in a measure of job performance? Do you see that, Doctor?"

A: "Yes, I do."

Q: "And you have just testified that here there are three measures of job performance at which Whites score statistically higher than Hispanics, is that correct Doctor?"

A: "That's correct."

Q: "So according to this definition which you have been relying on, there is not unfairness in this test, isn't that right, Doctor?"

The trial court struck this line of questioning. Illinois Bell's interpretation of the Uniform Guidelines' definition of "test unfairness" lacks any scientific or legal basis. While the term "measure" may signify either a single item, or a set of items measuring a single latent construct, this is no mere semantic quibble. What constitutes a "measure," in a given context, must be determined through appropriate legal and statistical analysis.

As a legal matter, Illinois Bell's interpretation of Section 14.B(8)(a) ignores its precise language. Through the use of the phrase "differences in *a* measure," the Uniform Guidelines plainly contemplate "a measure" as comprising more than one item. This conclusion is reinforced by the language of the definition of "unfairness" in the "Definitions" section of the Uniform Guidelines: "*Unfairness of selection procedure*. A condition in which members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences are not reflected in differences in measures of job performance. See section 14.B.(7)" (Uniform Guidelines, §1607.16.V) [emphasis added]. The two definitions of "unfairness" must be read together, and thus do not support reliance on an isolated difference in measurement ("Definitions" section of the Uniform Guidelines mandates "[t]he following definitions *shall apply* throughout these guidelines" (Uniform Guidelines, §1607.16) [emphasis added]).

Illinois Bell's argument, moreover, would permit an employer to ignore the vast weight of unfavorable evidence, so long as any favorable evidence existed at all. Defendant's interpretation would render the unfairness standard meaningless. The term "measure" cannot be applied arbitrarily, but requires a fact-sensitive analysis.

In the *Melendez* case, we reanalyzed the correlations among the 15 performance ratings using both exploratory and confirmatory factor analysis.[14] We found that the 15 criteria are most accurately represented as a single, global measure of job performance. Statistically, the 15 ratings are sufficiently interrelated so that they comprise not 15 independent measures, but rather only one underlying measure. The separate performance ratings cannot properly be considered individually.

Factor analysis should be used routinely in deciding whether to employ single items or composite scales to measure job performance. This would preclude test developers from treating sets of unidimensional criterion measures as multiple single-item indicators, and then selecting and highlighting, as evidence of test "fairness," any criteria on which the majority group has a higher mean. Confirmatory factor analysis, not subjective preference, should answer the question: "what is a measure?"

Factor analytic methodology adheres to the Uniform Guidelines, which proscribe "...reliance upon techniques which tend to overestimate validity findings as a result of capitalization on chance.... Reliance upon a few... criteria of successful job performance when many... criteria of performance have been studied... tend[s] to inflate validity estimates as a result of chance." (Uniform Guidelines, §1607.14.B(7)).

By what criterion should one judge differences in group means? On cross-examination, the defense inquired where the unfairness standard in the Uniform Guidelines requires that group differences be *statistically significant*. The Uniform Guidelines do not authorize excursions into chance associations, but the unfairness standard does not explicitly require statistical significance as a decision criterion. It should be noted, however, that the "Documentation" requirements of the Uniform Guidelines mandate the *reporting* of methods of data analysis, as well as the magnitude, direction, and statistical significance of results. It expressly re-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

quires that "[s]tatements regarding the statistical significance of results should be made (essential)." (Uniform Guidelines, §1607.15.B(8)). This section of the Guidelines specifically refers to measures of central tendency (e.g., means) and studies of test fairness. Illinois Bell argued, in essence, that professional statistical standards may somehow be suspended in evaluating employment test data.

Abandoning professional standards is scientifically and legally untenable. The Uniform Guidelines are themselves founded on the standards of the psychological profession. The Uniform Guidelines, §1607.1.C, states: "These guidelines have been built upon court decisions, the previously issued guidelines of the agencies, and the practical experience of the agencies, as well as the standards of the psychological profession."

Test developers should always adhere to professional standards for drawing inferences from data. The Guidelines do not require researchers to clear the memory of their calculator between computations, but researchers typically do so as a matter of course. Nor can employers ignore the Guidelines' prohibition against reliance on chance (Uniform Guidelines, §1607.14. B(7)). And yet, that is precisely the result if one relies on apparent group differences that lack statistical significance.

Illusory "Fairness" and Artifactual "Validity"

Under the Uniform Guidelines' "unfairness" standard, if one ethnic group scores higher than another on an employment test, and this difference is not reflected in a measure of job performance, the test is deemed "unfair." The BSAT failed this standard. Despite great disparities in test scores, whites and Latinos performed on the job with substantially similar success.

Importantly, under the Uniform Guidelines, the mere fact that majorities outscore minorities on an examination, while securing more favorable performance evaluations, does not affirmatively establish that the test is "fair." It does not prove the positive, that the test is "fair" and "job related," but it does disprove one possible negative. The standard, that is, should not be understood as establishing an affirmative defense for employers. Evidence that a test is not "unfair" merely forestalls the inference of discrimination that arises in cases when the group that excels on the test, garnering the greater share of job opportunities, does not actually do the job appreciably better. To prove or disprove "fairness," the parties may introduce other evidence.

Ironically, the pattern of data contemplated by the Uniform Guidelines' unfairness standard may result in a serious distortion of the validity evidence. If the data from different ethnic groups are simply (and improperly) combined in a pooled analysis, the distribution of the data will typically create the illusion of a correlation between test scores and performance ratings. Scatterplotting the data, the group with higher test scores and performance ratings will tend to fall in the upper right quadrant of the scatterplot. The group with lower test scores and performance ratings will tend to fall in the lower left quadrant of the scatterplot. This pattern will create an apparent correlation between test scores and performance ratings, despite the lack of any true relationship, and it will inflate obtained validity coefficients for the total sample. This problem is a variation of a phenomenon known as *Simpson's paradox*.[15,16]

The following hypothetical example demonstrates how the "false correlation" paradox can occur. Imagine that you are in the middle of a job interview. The interview is going well, so you broach the topic of salary. "How much would I be paid?" "Well," replies the interviewer, "take off your shoes, and let's find out." Requesting an explanation, you are told that the company has found that shoe size is a valid predictor of a person's worth. The company routinely measures the size of job applicants' feet, and then uses the results of that

2155-0182/10/$3.00

measurement to determine salary. Still skeptical, you ask to see the validity evidence, and the interviewer hands you a copy of a table from a research document (see Table 5).

TABLE 5: Validating Shoe Size as a Predictor of Salary: Hypothetical Raw Data for Women and Men

Women	Occupation	Shoe Size	Annual Salary
Ann	secretary	3	$ 22,000
Beatrice	actress	4	$ 14,000
Carol	teacher	4	$ 30,000
Diane	librarian	5	$ 20,000
Edna	lab technician	5	$ 40,000
Florence	baby sitter	5	$ 10,000
Gwen	journalist	6	$ 28,000
Harriet	bank teller	6	$ 18,000
Iris	nurse	7	$ 32,000
Jacqueline	waitress	7	$ 16,000
	Mean :	5.2	$ 23,000
Men	**Occupation**	**Shoe Size**	**Annual Salary**
Al	salesman	8	$ 48,000
Bob	airline pilot	8	$ 62,000
Carl	chef	9	$ 50,000
Don	chemist	10	$ 55,000
Ed	executive	10	$ 70,000
Frank	mechanic	10	$ 40,000
Greg	plumber	11	$ 52,000
Harold	electrician	11	$ 59,000
Ian	detective	12	$ 45,000
John	architect	12	$ 65,000
	Mean	10.1	$ 54,600
Exact Test of Gender Difference:		*p*<0.000001	*p*<0.000547

This table presents raw (hypothetical) data for a sample of 10 men and 10 women, listing their first name, occupation, shoe size, and salary. Reported at the bottom of the data table are the results of exact nonparametric statistical analyses[5] comparing men's and women's mean shoe-size (predictor) and salary (criterion). Women have smaller feet than men, and have comparably smaller salaries. Therefore, by the Uniform Guidelines' unfairness standard, it is not "unfair" to men or to women to use shoe size to determine salary. Validity coefficients relating shoe size to salary, and scatterplots of shoe size and salary, are presented in Figure 1.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

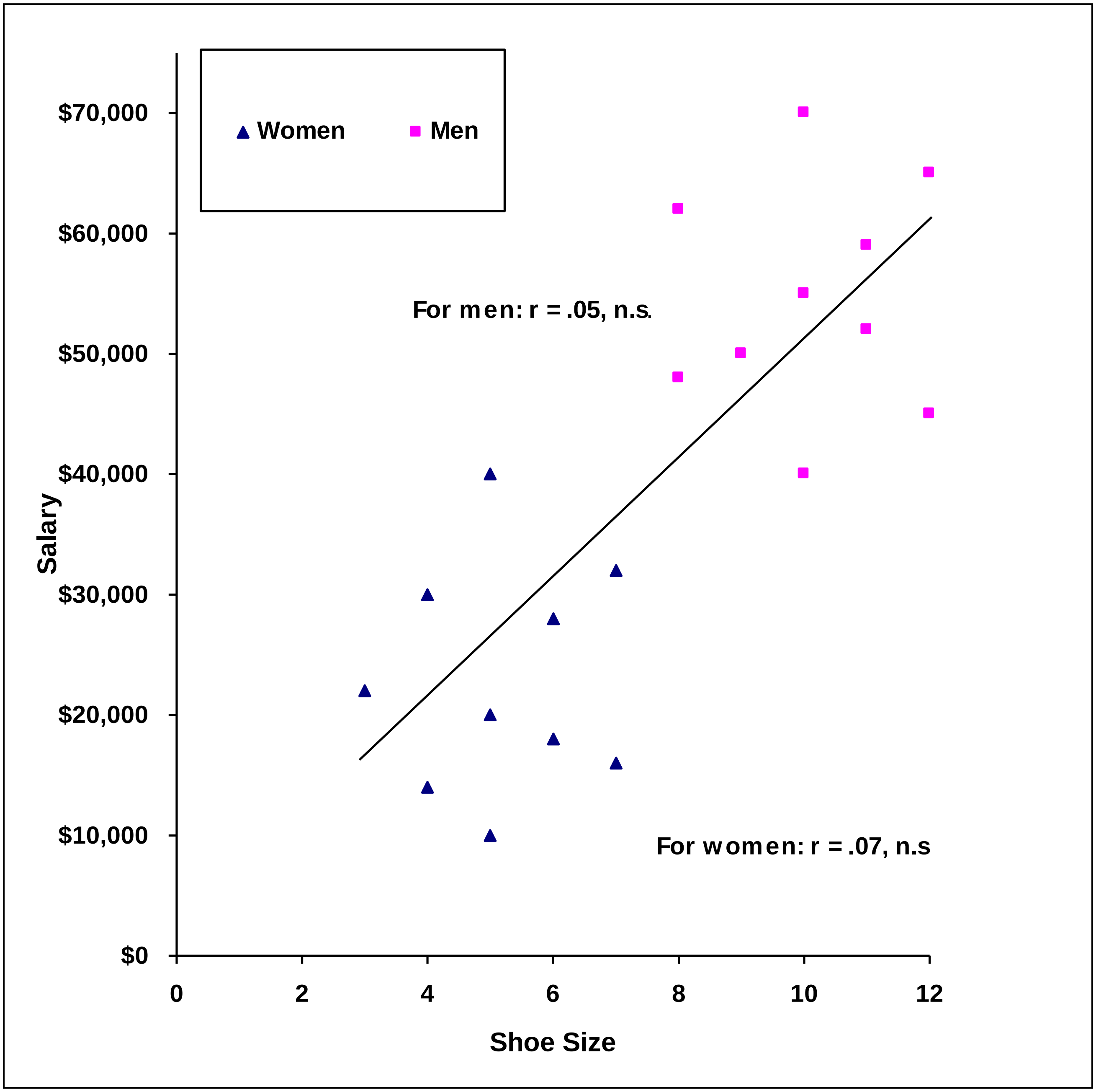


Figure 1: Correlating Shoe Size and Salary using Pooled Hypothetical Raw Data for Women and Men

Examination of validity coefficients for men and women reveals there is no linear relationship between shoe size and salary for either group: *r*=0.05 for men, *r*=0.07 for women, *p*s> 0.05. But, if men's and women's raw data are pooled, the men's data fall into the upper right-hand quadrant of the scatterplot, and the women's data fall into the lower left-hand quadrant (men score higher than women on predictor and criterion measures). When the correlation between shoe size and salary is computed for the total group of 20 subjects, *r*=0.78, *p*<0.001)! Based on this evidence and in accordance with the Uniform Guidelines, it is concluded that it is *both* fair and valid to use shoe size to determine salary.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

This hypothetical scenario is no more absurd than the BSAT validation work. In the Preliminary Study, for example, African-Americans had lower BSAT scores than whites, and they also had comparably lower performance ratings (thus the test does not meet the definition of unfairness, under the Uniform Guidelines' definition).

Figure 2 displays scatterplots of the group means on the BSAT and on overall job performance from the two validity studies. Clearly, these mean differences will inflate the apparent linearity of the relationship between BSAT and performance.

This inflation of correlations strikingly appears in the table of validity coefficients from the Preliminary Study (Table 2). Comparing the correlations of white, African-American, and total groups on the various performance measures, we find an anomalous pattern.

Consider the performance criterion of Decision Making. Its validity coefficient is r=0.20 for the group of 94 whites, and r=0.27 for the group of 39 African-Americans. For the Total Group, however, the r=0.30 correlation is higher than that for either subgroup. Similarly, the validity coefficients for Written Communications are r=0.21 for whites, r=0.26 for African-Americans, and r=0.28 for the Total Group; for Resistance to Stress, r=0.11 for whites, r= 0.18 for African-Americans, and r=0.21 for the Total Group; and for Energy, r=0.04 for whites, r=0.08 for African-Americans, and r=0.15 for the total group. Cases such as these, in which the correlations for the pooled group actually exceed the correlations found in each constituent subgroup, are a tell-tale sign of the "false correlation paradox," where in fact the "whole" is deceptively greater than the sum (or weighted average) of its parts.[16]

This technical problem is particularly critical because Illinois Bell rested its claim that the test was valid largely based on one number—one validity coefficient: the correlation between BSAT score and the rating of overall job performance, for the *Total Group* in the Preliminary Study. That coefficient is r=0.38, significant for the total sample of 153 subjects at p< 0.00001 (see Table 2).

A *possible* methodology for circumventing such *paradoxical confounding* (the technical terminology for the "false-correlation problem") is to remove mean differences on the x- and y-variables before combining the data: for example, standardizing the x- and y-scores separately for each group using a z-score transformation maps the data into the same metric.[16] How does this work in the shoe size example? After transforming subjects' raw data to z-scores separately within the male and female samples, and subjecting these standardized data to correlation analysis, yields results given in Figure 3. When properly analyzed, the correlation between shoe-size and salary is r=0.05 for men, r=0.07 for women, and r=0.06 for the total group.

This cure for Simpson's paradox (normatively standardizing separately by sample) only works if the true relationship between x and y is consistent across the multiple samples.[16] For example, if x and y are perfectly *positively* correlated in sample A and perfectly *negatively* correlated in sample B, normatively standardizing the data separately by sample and then combining them will yield a correlation coefficient of zero. Thus, it is necessary to verify homogeneity of covariance between x and y across samples before standardizing and pooling the data.[16-18]

Fortunately, instances of reverse validity rarely appear in the personnel selection literature.[19] Indeed, some proponents of validity generalization have even argued against the notion of differential validity altogether, though the BSAT data clearly show stronger evidence of validity for African-Americans than for Latinos or whites.[10] Thus, when analyzing the total sample, it should be routine practice before pooling data to normatively standardize separately within groups (after first verifying between-group equivalence of covariance matrices).

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

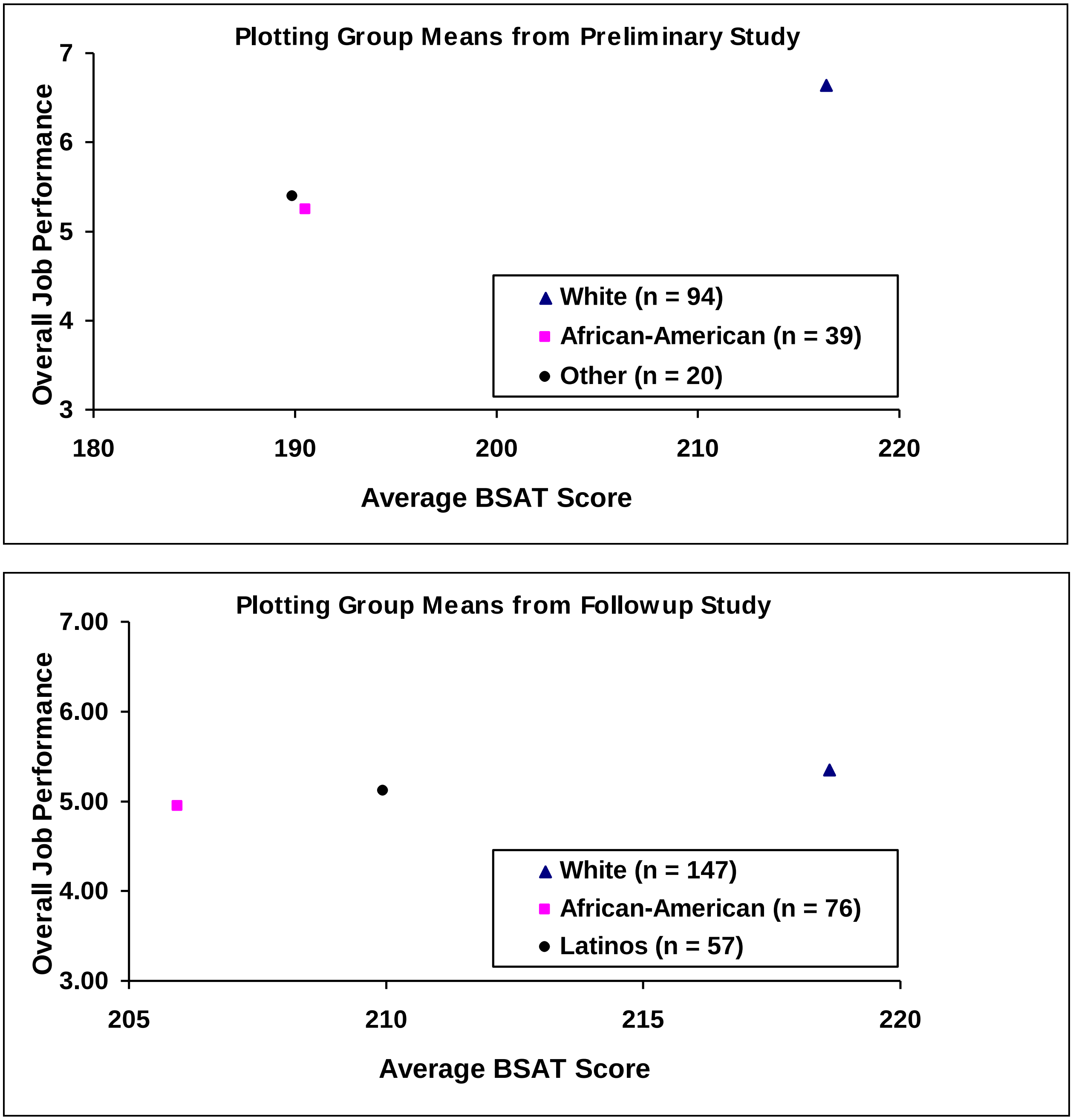


Figure 2: Scatterplotting BSAT score and overall job performance for the Preliminary and Followup Studies. Supervisors rated overall performance using a 9-point Likert-type scale in the Preliminary Study (1,2=exceptionally high; 3,4=very high; 5,6=moderately high; 7,8=moderately low; 9=unsatisfactory) and 7-point Likert-type scale in the Followup Study (1=exceptionable; 2=very high; 3=high; 4=average; 5=below average; 6=passable; 7=unacceptable). Scores on these rating scales have been reversed for ease of presentation.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

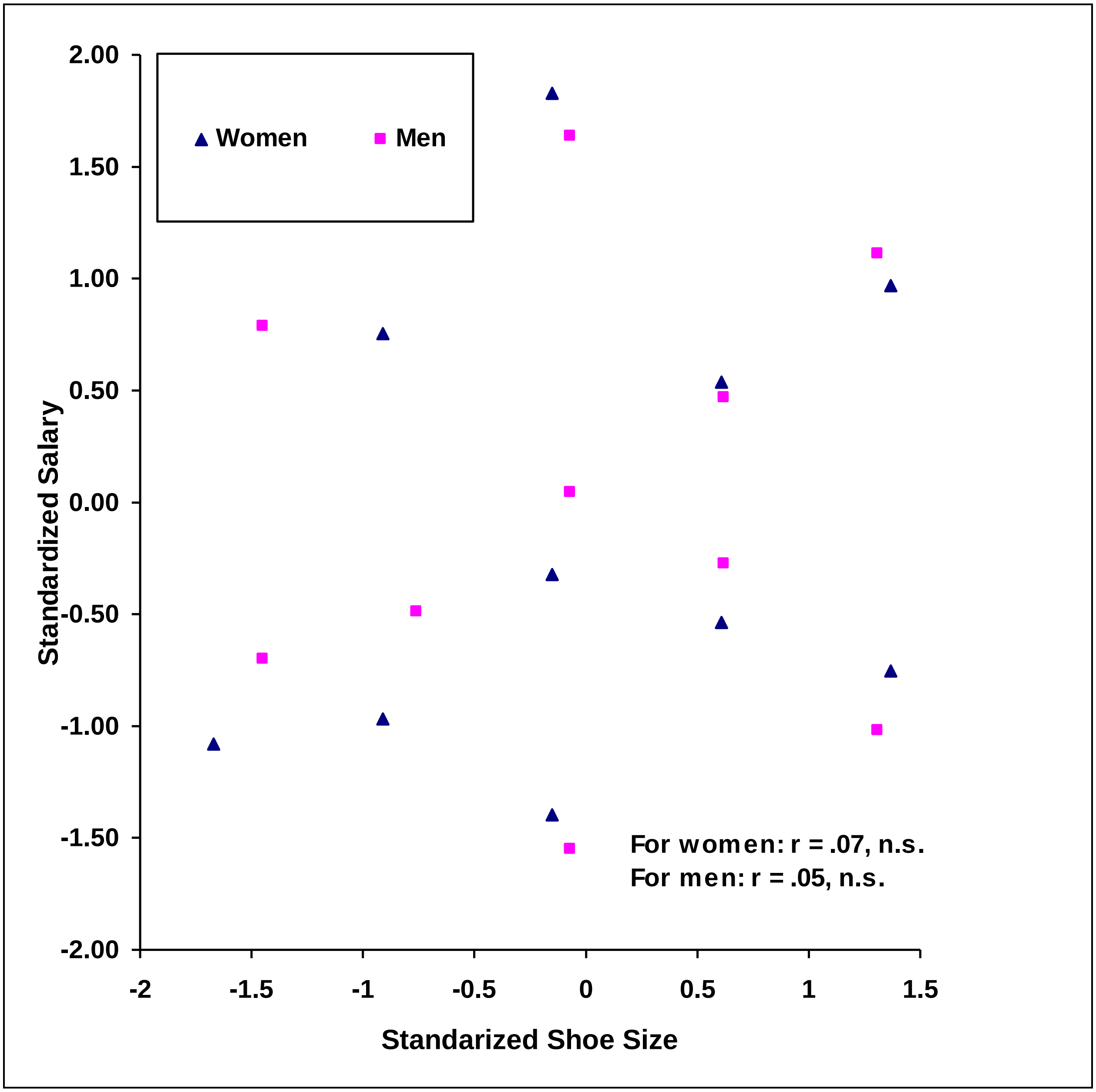


Figure 3: Correlating shoe size and salary using hypothetical data normatively standardized separately for women and men

Yet, typically researchers simply pool data across subgroups in total-sample analyses. This practice inflates total sample validities throughout the testing industry. Among the most robust findings in the literature on cognitive ability testing is that minorities score significantly lower on cognitive ability tests than do whites.[20] And in validation studies, minorities often receive significantly lower performance ratings.[21] Ironically, if test scores are lower for minorities than for whites, to meet the Uniform Guidelines' unfairness standard, minority performance ratings must also be lower. Although it is not unfair within the meaning of the Uniform Guidelines, this very situation will typically make tests appear more valid than they

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

really are, if data are simplistically pooled and correlated. *Test developers should avoid indiscriminately pooling subgroup data, particularly when these subgroups have different means on the test and on the criterion*.

The Uniform Guidelines provide a basis for addressing the distortions arising from the improper pooling of data. Section 1607.14.B (4), entitled "Representativeness of the sample," relevantly provides: *Where samples are combined or compared*, attention should be given to see that such samples are comparable in terms of the actual job they perform, the length of time on the job where time on the job is likely to affect performance, and *other relevant factors likely to affect validity differences*; or that these factors are included in the design of the study and their effects identified (emphasis added).

Hardly restricted to industrial/organizational psychology, this false-correlation problem pervades the life sciences: indeed it has been stated that the problem of paradoxical confounding is the most significant and pervasive challenge to the validity of empirical quantitative analysis in all areas of inquiry.[22] The practice of simply pooling data across subgroups inflates correlation coefficients whenever one group has higher mean scores than the other on both x and y. For example, studies of naturalistic animal behavior often pool data across intact groups to examine relationships among social and behavioral variables, without regard to possible mean differences.[23] Similarly, personality psychologists often pool the data of males and females, examine the correlations among numerous measures of, for example, anxiety, neuroticism, and general maladjustment, and find a single, stable pervasive trait that they label negative affectivity.[24] Given that women tend to report higher levels of negative experience in general than do men[25], pooling male and female data without standardization will inflate the observed intercorrelations for the total group, exaggerateing structural unidimensionality.

The problem of when and how to combine the data of multiple groups remains largely ignored in the social sciences.[16] Haphazardly pooling data across different groups (or time periods[16]) can produce unexpected, counterintuitive relationships, which researchers inevitably scramble to explain *a posteriori*. If one group scores lower than the other on x but higher on y, for example, then simply pooling the data across groups can produce a negative correlation for the total sample, even if the x-y relationship is actually positive in each group (the group with lower x scores and higher y scores will fall in the upper-left quadrant of the scatterplot, whereas the group with higher x scores and lower y scores will fall in the lower-right quadrant, yielding a false negative correlation). As a case in point, when studying psychosocial adjustment to head injury, researchers often combine the data of patients who are aware of functional deficits with the data of patients who are unaware of functional deficits. The correlation between severity of injury and emotional distress is then computed. An unexpected negative correlation often emerges, with greater severity of injury predictive of less distress.[26-28] It seems likely that the correlation between severity of injury and distress is actually positive within both the deficit-aware and deficit-unaware groups (i.e., greater severity linked to greater emotional distress), but that patients aware of their impairment have less severe head trauma (lower x-scores) and report higher levels of emotional distress (higher y-scores) than do patients who are unaware of their impairment, creating a false negative correlation for the pooled sample.

At first blush, the procedure of standardizing data separately within groups before computing pooled validity coefficients may seem similar to so-called *race norming*.[29] This latter practice seeks to ameliorate a test's adverse impact in personnel selection, by expressing individual test scores in terms of their standing relative to the mean of their particular racial group. However, the two approaches have entirely different objectives. Race-norming uses standardization in deciding which job applicants to hire. Standardizing raw data separately within groups before computing pooled validity coefficients, on the other hand, is done simply to avoid bias in estimating test validity, and is not used to select job applicants. Whereas race norming disaggregates data to avoid comparison between groups when selecting applicants, standardizing before computing pooled validity coefficients allows data from different groups to be meaningfully aggregated when evaluating test validity if their covariance is homogeneous.

Implications for Validity Generalization

Besides highlighting ambiguities in the Uniform Guidelines, the *Melendez* case also has implications for meta-analytic research on validity generalization.[10] This area of research entails synthesizing validity coefficients from studies attempting to validate personnel selection tests, in order to draw conclusions about the relationship between cognitive ability and job performance. Typically, these meta-analyses have concluded that cognitive ability tests are generally valid in the workplace across a full range of different racial subgroups, different jobs, different tests, and different settings.[10] Although conclusions about validity generalization have been criticized on a variety of statistical and conceptual grounds[30], the problem of paradoxical counfounding has been overlooked.

Validity coefficients based on pooled unstandardized data will be biased whenever the data contain subsamples that reliably differ on both the predictor and the criterion (e.g., racial subgroups, gender, types of jobs, different sites of data collection). Synthesizing validity coefficients will yield biased conclusions when the coefficients share a common bias (e.g., whites had higher test scores and higher performance ratings than other racial subgroups, and the data of racial subgroups were simply combined). This suggests that previous meta-analyses of test validation studies using total sample correlations have *overestimated* overall effect strength.

Although most statistical adjustments in meta-analysis serve to increase the strength of observed relationships by correcting for sources of unreliability[10], a comparable adjustment is needed to remove the inflation in correlations due to paradoxical confounding. If means and standard deviations are available for racial subgroups from the primary studies, for example, then group differences can be examined on the predictor (x) and the criterion (y). When one group scores higher than others on x or y, a better estimate of the pooled correlation coefficient is a weighted composite of the correlations for the separate subgroups, using *r*-to-*z* methodology.[18] Paradoxical confounding exists whenever the coefficient based on pooled data differs from the weighted mean coefficient across subgroups.

In the name of validity generalization, extravagant claims have been made for the efficacy of cognitive ability tests as personnel selection devices. For example, it has been argued: “[R]eliable measures of the standard aptitudes (e.g., verbal, quantitative, and spatial abilities) are valid predictors of... performance on the job for all jobs in the occupational spectrum... [T]hese findings can be generalized to all jobs in the economy for which tests are used in selection... [T]here are no jobs or job families for which reliable measures of cognitive ability do not have validity”.[31] Couching claims in cosmic hyperbole, validity generalization is likened to “the powerful telescopes used in astronomy,” and it is suggested that the theory is as well-established as the measurement of the speed of light.[31]

Ironically, persistent disparities between test scores and performance evaluations of majority and minority employees is also what one would expect from a pervasive pattern of discrimination. Consistent use of discriminatory employment tests, coupled with racially-biased supervisory evaluations, would produce com-

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

parable statistical outcomes. For this result to obtain, overt and conscious racial discrimination need not exist. For example, unconscious, subjective perceptions favoring majority employees would tend to inflate the mean criterion measure for this subgroup; similarly, the impact of broad societal discrimination would tend to depress the mean test performance of a minority group. Where the data for such racial and ethnic groups are pooled without correcting for differences in means on predictor and criterion, the likely result is a distribution yielding false positive correlations. The resulting evidence of "validity" would be illusory.

The implications for the theory of validity generalization are clear. Meta-analysis is based in a vast pool of results from combined samples, drawn primarily from reported validity studies of employment tests. A systematic bias throughout this data base would correspondingly bias the meta-analysis. Further empirical research is needed to isolate and assess the statistical impact of artifactual validity arising from paradoxical confounding.

Conclusion

The case of *Melendez v. Illinois Bell Telephone Company* highlights ambiguities in the Uniform Guidelines for validating personnel selection tests. Although the Guidelines could be revised to clarify these ambiguities, there is a potential drawback to this approach: namely, the possibility that hard-won legal precedents, gained over the years in the courts, might be lost if the Guidelines were substantially modified.[30] There is an inevitable trade-off here between more specificity in the Uniform Guidelines, and less applicability of previous court rulings.

Although the judgment in the *Melendez* case strengthens the legal means for removing invalid, discriminatory tests from the workplace, it does not immediately reduce the likelihood of such tests being developed in the first place, as might revisions in the Uniform Guidelines. Ultimately, however, the demise of invalid discriminatory tests in the workplace may depend more on their perceived liability costs for the user than on the specificity of the guidelines for test development.

References

[1]*Melendez v. Illinois Bell Telephone Co*., No. 90 C 5020 (N.D. Ill. 1994); *aff'd*, 79 F.3d 661 (7th Cir. 1996).

[2]*Standards for educational and psychological testing*. Washington, DC, APA Books, 1985.

[3]*Principles for the validation and use of personnel selection procedures*. College Park, MD, Society for Industrial and Organizational Psychology, 1987.

[4]Adams E. Development of the BSAT qualification score for general management hires. Trenton, NJ, AT&T, 1982.

[5]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. Washington, DC, APA Books, 2005.

[6]Rosenthal R, Rosnow RL. *Contrast analysis: focused comparisons in the analysis of variance*. London: Cambridge University Press, 1985.

[7]Cascio WF, Outtz J, Zedeck S, Goldstein IL. Statistical implications of six methods of test score use in personnel selection. *Human Performance* 1991, 4:233-264.

[8]Gottfredson LS. Reconsidering fairness: a matter of social and ethical priorities. *Journal of Vocational Behavior* 1988, 33:293-319.

[9]Hartigan JA, Wigdor AK. *Fairness in employment testing: validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC, National Academy Press, 1989.

[10]Hunter JL, Schmidt FL. Methods of meta-analysis. London: Sage, 1990.

[11]Peterson NS, Novick MR. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement* 1976, 13:3-29.

[12]Tenopyr M. Fairness in employment testing. *Society* 1990, 27:17-20.

[13]Cleary TA. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement* 1968, 5: 115-124.

[14]Bryant FB, Yarnold PR. Principal-components analysis and exploratory and confirmatory factor analysis. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association, 1995, pp. 99-136.

[15]Simpson RH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society* 1951B, 13:238-241.

[16]Yarnold PR. Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurment* 1996, 56:430-442.

[17]McClish DK. Combining and comparing area estimates across studies or strata. *Medical Decision Making* 1992, 12:274-279.

[18]Rosenthal, R. *Meta-analytic procedures for social research*. Beverly Hills, CA, Sage, 1984.

[19]Ghiselli, E.E. *The validity of occupational aptitude tests*. New York, Wiley, 1966.

[20]Wigdor, A.K., & Garner, W.R. *Ability testing: uses, consequences, and controversies: part 1. Report to the committee.* Washington, DC, National Academy Press, 1982.

[21]Miner MG, Miner JB. *Employee selection within the law*. Washington, DC, The Bureau of National Affairs, 1978.

[22]Soltysik RC, Yarnold PR. The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, 1:67-100.

[23]Martin P, Bateson P. *Measuring behavior: an introductory guide*. Cambridge, England, Cambridge University Press, 1993.

[24]Watson D, Clark LA. Negative affectivity: the disposition to experience aversive emotional states. *Psychological Bulletin* 1984, 96:465-490.

[25]Gove WR, Tudor JF. Adult sex roles and mental illness. *American Journal of Sociology* 1973, 78:812-835.

[26]Landy PR. The post-traumatic syndrome in closed head injuries and accident neuroses. *Proceedings of the Australian Association of Neurology* 1968, 5:463-466.

[27]McClean A, Dikmen S, Temkin N, Wyler A, Gale JL. Psychosocial functioning at 1 month after head injury. *Neurosurgery* 1984, 14:393-399.

[28]Miller H. Accident neuroses: lectures I and II. *British Medical Journal* 1961, 1:919-952, 992-998.

[29]Gottfredson LS. The science and politics of race-norming. *American Psychologist* 1994, 49: 955-963.

[30]Seymour RT. Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of “validity generalization.” *Journal of Vocational Behavior* 1988, 33:331-364.

Author Notes

The first author served as the plaintiff's expert witness and the second author as attorney for the plaintiff (Melendez) in the case of *Melendez v. Illinois Bell Telephone Company*. We wish to thank Robert Perloff, Richard Seymour, Charles Spielberger, and Scott Tindale for helpful comments on an earlier draft of this manuscript.

Correspondence should be sent to Fred B. Bryant at: Department of Psychology, Loyola University Chicago, 6525 North Sheridan Road, Chicago, IL, 60626. Send e-mail to: fbryant@luc.edu.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Other

Author Instructions. Article categories are described on the inside front cover of *Optimal Data Analysis* (*ODA*). For instructions on how to prepare manuscripts for submission, please visit:

> *ODA* Journal Page → About → Author Instructions.

Special Calls.

1. International Academic Ambassadors. *ODA* seeks bilingual academic ambassadors having research experience using the ODA paradigm, to join the Editorial Board and represent authors publishing in a (their) non-English language (Japanese is already covered). A Vita is required. Please contact the Editor (see Contact Us on Webpage).

2. Reviewers experienced with optimal (maximum-accuracy) and/or traditional statistical methods. *ODA* seeks volunteer reviewers who are experienced in use and reporting of optimal and traditional statistical methodologies (some will be asked to join the Editorial Board). A Vita is required. Please contact the Editor (see Contact Us on Webpage).

3. Papers using optimal (maximum-accuracy) statistical methods. *ODA* seeks manuscripts which involve maximum-accuracy methods. See the inside front cover for a description of article categories, and Author Instructions (above) for a description of how to prepare and submit manuscripts for consideration of publication in *ODA*.

4. Papers investigating theoretical or applied aspects of Simpson's Paradox. *ODA* seeks become a "central hub" of research investigating Simpson's Paradox. Please contact the Editor (see Contact Us on Webpage).

5. Citations to articles involving optimal methods published in journals *other than ODA*. ODA seeks to maintain a database of all manuscripts concerning optimal methods, which is available on the webpage under Resources. Please send complete citations of any articles involving optimal methods which are not in the database to the Editor (see Contact Us on Webpage).

6. Replication attempts of any published application involving optimal methods. *ODA* seeks to publish attempted replications (parallel or conceptually similar) of published works that involving optimal methods. Little has been written concerning confirmatory optimal analyses, but research in this area is on-going in our laboratory. Please contact the Editor (see Contact Us on the Webpage).

7. Data which are consistent with the Finance example given in, *Automated CTA Software: Fundamental Concepts and Control Commands* (*Optimal Data Analysis*, Volume 1, Release 1). *ODA* seeks actual data to conduct the weighted CTA described as the initial example analysis in the cited article, and will work with the individual contributing data to publish the article in *ODA*, and elsewhere if desired.

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Reprint Requests and Bound Copies. For information concerning bound copies of Issues or Releases, annual compilations, individual reprints, and/or customized renditions of *ODA*, please visit:

ODA Journal Page → Products.

Optimal Statistical Software. For information concerning Univariate Optimal Data Analysis (UniODA) software—which is the only software which may be used to manually-conduct hierarchically-optimal classification tree analysis (CTA), and also which may be used to easily perform detailed, pin-point sensitivity analysis to optimize models which were derived using automated CTA, please visit:

Company Home Page → Software → Optimal Data Analysis (ODA);

and for information concerning automated hierarchically optimal classification tree analysis (CTA):

Company Home Page → Software → Classification Tree Analysis (ODA).

Advertiser Instructions. For information concerning advertising on the company or journal webpages, and/or in *ODA*, please contact the Editor (see Contact Us on Webpage).