

1994 Language Aptitude Invitational Symposium
25-28 September 1994
THEME: Partnerships in Language Aptitude Research

Title: **Let Computers Use the Past to Predict the Future:**
Using Machine-Based Retrospective Correlation Data for Prospective Aptitude
Assessment

Partnership: Science & Technology — Scientific Statistical Exploration delivered
in Hi-Tech Environment

Co-author:
Dr. Frank L. Borchardt
Professor of German
Box 90256
Duke University
Durham, NC 27708-0256
Phone: (919) 660-3161
FAX: (919) 660-3166
e-mail: frankbo@acpub.duke.edu

Co-author:
Dr. Ellis Batten Page
Professor Educational Psychology and
Research Duke University
213 West Duke Bldg.
Box 90739
Durham, NC 27708-0739
Phone: (919) 684-3924 e-mail:
ebp@acpub.duke.edu

Presentation Format: Paper B (45 minutes with 15 minutes for questions)

Presentation Category: Assessment of Language Aptitude

Objectives of the presentation or research:

The objectives of the research proposed by this presentation is to combine the ramifications of one of the authors' scientific research (Page), dealing with the role of the computer in what are traditionally considered "hard-to-quantify" areas of human activity, with the technological delivery power of the co-author's (Borchardt) program, for the purpose of bringing automation to the aptitude assessment process.

Background/Rational:

It has been demonstrated that machine-based evaluation of "hard-to-quantify" activities (such as the grading of English compositions written by high school students) can produce cross-validation results as high as .87 when compared to a team of human judges. Furthermore, machine-based grading routinely surpasses the cross-validation quotient of two human judges, i. e., machine-based evaluation will accord with a human judge far more probably than that two human judges will accord with one another.

By contrast to human rating, computers employ indirect criteria. In the case of written essay evaluation, "direct" or "intrinsic" variables of interest, called "trins," might include fluency, diction, style, organization, logic. Without computerized measures of these

variables, “trins,” measurable substitutes, have been invoked to approximate intrinsic variables, approximations or “proxes,” such as essay length for fluency, variation in word length for diction, proportion of subordinating conjunctions and relative for complexity of style.

Application of the underlying principles of this research to aptitude testing would seem a highly plausible and, in due course, labor-saving strategy.

Methods/ Design/ Procedures/ Techniques

As distinguished from the English essay experiment, where criteria, both direct and indirect, “intrinsic” and “approximate” were established *a priori* and applied to the testing data, the aptitude testing experiment proposed here would

first, identify a cadre of individuals determined by whatever means or general consensus to have a demonstrably high degree of language aptitude,

second, gather into one data store all available electronic performance, drill-and-practice, quizzes, tests, questionnaires, and

third, seek without prejudice, retrospectively, *all* statistical correlatives, whether later considered “intrinsic” or “approximate,” to which a third, more remote and accidental category, such as “contingent” should be added. (It is possible, for example, that fly-fishing correlates poorly to language aptitude and expert macramé correlates well.)

Results/ Conclusions

Inductive, *a posteriori*, discovery of correlations between electronic testing data and actual language aptitude should, over time, produce exceedingly interesting patterns worth further study and theory. In the meantime, however, results could be employed to identify, prospectively and with high probability, individuals especially well equipped with language aptitude.

Importance/ Implications of the Results

The practical consequences of the above experiment include, above all, the necessity of employing a common electronic data store with a common analysis engine. The U.S. Government presently has proprietary rights to precisely such a mechanism, WinCALIS (for “MS-Windows” and “Computer Assisted Language Instructional System”). WinCALIS, uniquely among authoring environments, observes the UNICODE standard and is therefore actually capable of serving many difficult and unusual languages and potentially capable serving *all* the languages of the world.

WinCALIS is already equipped to deal with the “Paradox Engine,” which means that the proposed retrospective aptitude testing data store would require only appropriate flagging and a suitable user interface. The electronic means for actualizing the statistical relationships or patterns or profiles inheres in the “Paradox Engine.”

Let Computers Use the Past to Predict the Future:

Using Machine-Based Retrospective Correlation Data for Prospective Aptitude Assessment

Status of the Research (Borchardt)

Reviewing the pertinent literature in advance of this presentation, an author new to the field (Borchardt) was impressed by the magnitude of the literature on testing in general (Green 1991), by the enormous controversy attending aptitude testing (Owen 1985; Crouse and Trusheim 1988; Brown 1991), and by the relatively peaceful tone and far more manageable scope of the literature on computerized testing (Wainer 1990). The literature on computerized testing grasped correlations (Farr 1991) but was, with the exception of the contributions of co-author Ellis Page, strangely silent on the concept and application of *indirection* as a testing strategy. Correlations, as it seems, had to demonstrate self-evident and direct connections to the object of the testing. The notion that remote and indirect correlations might also provide useful information on a testing object seems, with few exceptions, to have been either too eccentric or, potentially, too emotionally charged to pursue safely in staid academic educational journals.

One of the contributions to the study of “Computer use in the classroom” observes with refreshing and sobering candor: “The myth that computers have revolutionized the educational system is alive and well” (Green 1991 p. 245). This observation refers to the condition of computing in the classroom environment altogether, including but not limited to the area of testing. After rehearsing the practical failures of technological policy, these same researchers provide twelve good reasons for employing “computerized testing,” a term employed here to embrace both computer based and computer adaptive testing:

A computerized testing system:

- automates the process of creating tests
- automates the process of scoring tests
- facilitates the creation of equivalent versions of the same exam
- provides access to existing test banks
- can be used to train preservice or inservice teachers to administer tests
- standardizes test administration procedures
- provides more detailed feedback to the student
- enables placement testing
- minimizes data entry problems
- enables teachers to become classroom researchers
- provides students with information about strengths and weaknesses, not just a summary test score
- provides analyses of the errors in strategies students are using.

Educational technologies lack, so far, a “killer app,” that is, an application so obvious, so time-saving, so directly usable as to be both irresistible and enough reason for a died-in-the-wool yellow-pad person to go over to computing. It would seem that testing and evaluation represent the most obvious candidates for the “killer app” designation.

Imagine:

entry and exit testing
placement and proficiency testing
achievement and aptitude testing

put in the hands of the classroom teacher in such a way as to

save some of the countless hours spent in

- a. generating
- b. correcting, and
- c. grading

quiz and test materials, and

providing

- a. data and
- b. analysis

in such a way that the classroom teacher could use the results

educationally, to improve testing over time,
institutionally, to help support the achievement of local goals with hard data,
professionally, to support and to document publishable research.

If that would not be the “killer app” for educational technology, then it is hard to imagine what would be.

On Computing “hard-to-quantify” Objects (Borchardt)

On the face of it, there should be large areas of human experience in which the rigorous binary logic of digital computing has no role to play whatsoever or ever will. All of those areas that have to do with the subjective, with instinct, intuition, nuance, hunch, taste, connoisseurship, and “judgment” would seem to be closed forever to the hard, discriminating logic of the digital computer.

Not everyone has always thought so.

One of the claims made for the binary computer from the very outset was that it was a “universal calculator,” that is to say, if anything could be calculated at all could be calculated by a binary computer. Alan Turing, who gave his name to the mathematic concept of the binary computer, the “Turing Machine,” recognized further, that binary arithmetic was formally identical with certain kinds of logic, that a “Turing Machine” did not care if the

binary mode was designated as zeroes and ones, or ons and offs, or opened and closed, or true and false. This suggested to Turing that any process that could be conducted to a conclusion of “true” or “false” could be emulated on a binary computer, that if a judgment were “logical,” it could be represented by a binary computer. His thesis had long-term consequences of which he was clearly aware. His “machine” qualified as a “universal calculator.” It could calculate everything that could be calculated. If calculation includes logic, then, one day, his “machine” should be able to act as a human being might act, insofar as such a human being was acting logically. The “Turing Test” posits an observer--denied visual access to the source of an utterance--unable to tell for sure whether the originator was human or mechanical.

Another of the sainted geniuses responsible for the modern computer, John von Neumann wrote in 1951 (McCurdock 1979):

It has often been claimed that the activities and functions of the human nervous system are so complicated that no ordinary mechanism could possibly perform them. It has also been attempted to name specific functions which by their nature exhibit this limitation. It has been attempted to show that such specific functions, logically, completely described, are per se unable of mechanical . . . realization. [A logical circuit proposed by the team of McCulloch and Pitts in 1943] puts an end to this. It proves that anything that can be completely and unambiguously put into words is ipso facto realizable by a suitable finite [mechanical] . . . network.

Von Neumann’s famous pronouncement goes beyond Turing’s claim to suggest that anything that can be symbolically represented in words, completely and without ambiguity, can be mechanically represented and manipulated. The trap, of course, is in the phrase “completely and unambiguously.” Furthermore, the McCulloch-Pitts logical circuit is not, by itself, a universal calculator, lacking the ability, without redundancy, to solve a class of logical problems. It is also a good deal more complicated than a Turing Machine, even though a Turing Machine, being a universal calculator, can imitate it.

The strategies suggested by these mathematical propositions anticipated decades of research in two areas of machine intelligence, Natural Language Processing (NLP) and Artificial Neural Networks (ANNs). In the first, Natural Language Processing, the machine is taught as much as it needs to know, in this case, about human or “natural” language, specifically content (the “lexicon”) and structure (the parser). Human utterances are then compared to those data and “comprehended” in so far as they conform; and potential human utterances can be produced, also insofar as they conform, given the risk of producing sentences like Chomsky’s famous “Colorless green ideas sleep furiously”(1957). In the second area of machine intelligence, Artificial Neural Networks (ANNs), the machine is given as much data as is deemed reasonable, and the mechanism then learns from the data, seeks patterns and correspondences, and produces its results, not in True/False certainties but rather in series of relative probabilities. Both strategies have had their successes and failures in dealing with “hard-to-quantify” objects.

Experiments with Grading Student Prose (Ellis Batten Page)

One such object, subject to a long history of exploration, has been the economic and educational problem of grading student prose. In the late 1960s, enjoying the influence of the Natural Language Processing of that time (Kuno 1964), a team of researchers (Stone, Dunphey, Smith, & Ogilvie 1966) contemplated the role computers might take in the task of evaluation student writing. In the years from 1966 to 1973 “Project Essay Grade” (PEG-1) investigated that role. The Project sought out student essays that had already been through the process of grading by English teachers. Those essays then made their way onto main-frame computers (by way of punch cards) and underwent processing to count various events in each essay. These computer scores then ran through multiple regression to predict human judgments (McColly & Remstadt 1963). The results surprised everyone. Hidden in a cross-validation with four human graders, the computer correlated at .50, as the human graders did with one another. The right to ask the question “Which judge is the computer?” suggests that the program was close to passing the “Turing Test,” even at this early date.

In the next set of experiments (PEG-2) the Project subjected the computer to competition with eight human judges and produced even more astonishing results. The computer appeared to achieve greater accuracy (in comparison with human graders) in scoring creativity and style than in scoring mechanics! (Daigon 1966; Page 1966; Page & Paulus 1968). A third set of experiments (PEG-3) resulted in even closer approximation of single human grader performance (Page 1967a; Page 1967b; Page, Fisher, & Fisher 1968; Paulus, McManus, & Page).

The best results so far have been obtained in an analysis of the National Assessment of Progress (NAEP) writing samples by seniors from the years 1988 and 1990. There the multiple regression was .87 in comparison to machine grading with a pool of six human graders (Ajay, Tillett & Page 1973; Page 1985; Tillett & Ajay 1989).

The approach by which the computer was made amenable to the measurement of a “hard-to-quantify” object, like a student essay, was to evaluate first the distinction between the features the Project wished to measure and those the computer actually could measure. To identify the distinctions the Project coined the following terms: Trins were those features of intrinsic interest, such as fluency, diction, grammar, punctuation; failing direct measurements of these, Proxes were approximations or possible correlates of the trins. Examples of these correspondences would be:

Trins	Proxes
fluency	length of essay
diction	variety of word length
sentence structure	count of relative pronouns, subordinating conjunctions, prepositions

Application to Language Aptitude Testing (Borchardt)

The quality of this research over time has all but proven that the measurement of secondary, indirect features of a “hard-to-quantify” object can emulate with a remarkably high degree of success the personal, human judgment of primary, direct but unmeasurable or “hard-to-measure” features.

Interpretation of this result would suggest that, in general, subjective, intuitive objects (as, for example, an opinion) do not exist in a vacuum but consistently reside in a context and against a background. We employ two terms, “context” and “background,” in order to suggest that the environment of a subjective, intuitive, “hard-to-quantify” object may be multi-dimensional and at the very least, three-dimensional. If the object is “hard-to-quantify” because its primary characteristics are subjective and intuitive, perhaps some of its secondary characteristics, as for example, its context and background, may be less “hard-to-quantify.” It is possible that certain aspects of contexts and backgrounds, while not identical to the subjective, intuitive object, relate to it closely, that is, “approximately,” and, if measurable, can provide secondary, “circumstantial” evidence pointing toward a desired outcome or conclusion. It would seem possible, furthermore, that features considerably more remote to the object might hold useful information. What if the remote feature coincides with the object in some degree well beyond that of mere chance? Might that fact also provide information about the otherwise “hard-to-quantify” object? To the established categories of “intrinsic” (primary features) and “approximations” (secondary features), perhaps then a third category of “contingencies” (tertiary or less obviously related correlations) should be added. It is in this category that the possibility resides for the discovery of surprising, hard-to-anticipate correlations. These may provide directions to some interesting research avenues, which might otherwise not leap to mind.

It made good sense in an era when compositions had to be entered into electronic form, manually, one at a time, and with punch cards, to propose correlations and test them against the data. Human preprocessing of the problem much reduces the number of possibilities, focusses the later processing on a finite and comprehensible set of variables, and much enhances the probability of achieving the desired results. The deductive, a priori approach led to such good results in the computerized grading of student essays surely because of human preprocessing. Human preprocessing recognized the most probable correlations and understood the dynamics which connected them (Page 1994): “The trin of diction was approximated by the variance of word length (less common words are often longer).”

However, in an era when compositions are written in word processors to begin with, when many human activities have been and are being computerized, and when huge quantities of raw electronic data are being produced, alternative strategies may be possible and practical. Certain “hard-to-quantify” objects would probably benefit from human neglect, at least to begin with. Inductive, a posteriori approaches would have no difficulty in detecting the great preponderance of correlations between any given “hard-to-quantify” problem and related features in context and background. It would be the business of human postprocessing to look to the classification of the correlations, after the fact, as secondary (approximate) or tertiary (contingent).

This principle should apply to all “hard-to-quantify” problems, including the prediction of performance in language training, otherwise known as “language aptitude.” The process is quite self-evident. One begins with individuals determined by whatever means or general consensus to have a demonstrably high degree of language aptitude, perhaps because they have also demonstrated a high degree of language achievement. Then one gathers into one data store all available electronic performance, drill-and-practice, quizzes, tests, and questionnaires. The more data the better. This, of course, depends on what is legally available and susceptible of “regularization.” By “regularization” is meant the presentation of all data in comparable structures (oranges with oranges, apples with apples). And finally one seeks without prejudice, retrospectively, all statistical correlatives, whether later considered “intrinsic” or “approximate,” to which the third, more remote and accidental category, “contingent” should be added. It is here that the less-than-obvious correlatives emerge, for example, that fly-fishing may correlate poorly to language aptitude and expert macramé may correlate well.

Inductive, a posteriori, discovery of correlations between electronic testing data and actual language aptitude should, over time, produce exceedingly interesting patterns worth further study and theory. In the meantime, however, results could be employed to identify, prospectively and with high probability, individuals especially well equipped with language aptitude.

Practical Execution (Borchardt)

The practical consequences of the above experiment include, above all, the necessity of employing a common electronic data store with a common analysis engine. The U.S. Government presently has proprietary rights to precisely such a mechanism, WinCALIS (for “MS-Windows” and “Computer Assisted Language Instructional System”). WinCALIS is designed to be equipped with the “Paradox Engine,” which means that the proposed retrospective aptitude testing data-store would require only appropriate flagging, a suitable user interface, and a set of statistical formulae provided by the experienced users of automated testing. The electronic means for actualizing the statistical relationships or patterns or profiles inheres in the “Paradox Engine.”

WinCALIS, uniquely among authoring environments, observes the UNICODE standard and is therefore actually capable of serving many difficult and unusual languages and potentially capable of serving all the languages of the world. The chief requirement for successful electronic measurement of aptitude, regardless of the strategy, is quantity and quality of the data. The measurement of language aptitude will require the possibility of reading data in many different languages, not all of them represented by the English alphabet. It is a matter of utmost importance that the language teaching and learning community grasp the fundamental reality that global computing will not be satisfied with the English alphabet. WinCALIS is among very few programs which has made this reality a foundation block of its functioning. WinCALIS is ready to receive data about language aptitude with accent marks and umlauts, in Arabic, Cyrillic, Greek, and Hebrew, why even in Chinese and Japanese, and we’re working on Korean.

The need for advanced record-keeping, expressed by classroom teachers, is bringing the Paradox Engine to WinCALIS. There are three basic reasons for record-keeping in WinCALIS. One is to have a record of student performance for purposes of grading the student or determining educational options. Computer-savvy teachers require of courseware that it keep the record of student accomplishment. The second use of record-keeping is to control branching and the course of the student through the CALIS lesson. If a record of student performance is kept, the student can be branched to alternative parts of the program based on an analysis of performance. The third reason is the one that is the most important to CALIS authors — that is, record-keeping provides a basis for the revision of CALIS lessons, for testing the quality of the tests, of the questions, of the right answers, anticipated wrong answers, and distractors. For this purpose, it is important to have detailed information regarding how student users responded to which items.

Record-keeping is an important capability of the computer which is usually not exploited fully. Most CAI authoring systems only keep track of a cumulative score of right and wrong answers. Typically, there is not even any provision for keeping track of what the wrong answers were. However, in response to the precious input of several classroom teachers, WinCALIS is being readied to interface with the rich data collection capacities of the Paradox Engine and thus to provide useful results for a variety of evaluation processes

Record-keeping in WinCALIS comprises four important components: House-keeping, Questions, Answers and Performance.

House-keeping involves keeping track of basic student information like ID, password. It could also include section, age, and more. House-keeping in WinCALIS is done with the help of two Paradox database tables.

The Questions component of record-keeping includes basic quantitative information about questions assigned by teacher such as learning object, weight, and level.

Answers component records information (student-computer interaction) to a file based on the template specified by the teacher. This template tells WinCALIS about the format, the type, and the amount of data that should be saved.

The Questions and Answers components both draw information from a databank that has been tested previously and can provide raw information for any project the author has in mind.

Finally, performance analysis is carried out, based on the data that is saved to the disk, to compute student's level, strengths and weaknesses, and accomplishment, as determined by parameters laid out by teacher/authors.

A set of formulae (queries) can be applied to the results table (a Paradox table) within the Paradox Engine and produce Reports, that is to say, from the information that has been created by WinCALIS the Paradox Engine can generate statistical patterns or profiles.

In this respect, the in-process implementation of the WinCALIS Paradox Engine resembles Ellis Page's deductive, *a priori* approach, where human preprocessing makes all the difference and anticipates, from experience, the most meaningful correlations.

In due course, the data assembly powers of WinCALIS's in-process implementation of the Paradox Engine will be able to be exploited, as well, by inductive, *a posteriori* approaches which will not require (or even allow) human preprocessing. The correlations will emerge from the data by statistical manipulation or analogous strategies, such as Artificial Neural Networks. Artificial Neural Networks have been employed to predict the reliability of mortgage applicants. There should be no reason that computers could not, given quantitatively and qualitatively adequate data, predict language aptitude as well.



Dr. Frank L. Borchardt took his doctorate at the Johns Hopkins University in 1965 concentrating in late medieval German language and literature. In the early 1980's he stumbled over a computer and has since been professionally involved in Computer Assisted Language Learning, specifically by directing the project at Duke University, which produced CALIS (Computer Assisted Language Instructional System) and its successor, WinCALIS. He has taken an interest in Artificial Neural Networks and their application to the computerized presentation of Natural Language. This interest has led back to a variety of statistical strategies for the solution of Natural Language problems.

Dr. Ellis Batten Page was founding editor of the *Educational Psychologist* for the American Psychological Association, 1963-66, and President of the APA's Division of Educational Psychology, 1976-77. He has also been chair of the Conference on Applied Research in Education for the U.S. Office of Education (1971), and President of the American Educational Research Association (1979-80). He has twice been recipient of the Distinguished Research Award (1981, 1991) of the North Carolina Association for Research in Education (NCARE) and served as its President (1984-85). His publications come to over 250 authored or edited books, monographs, articles, technical reports, invited addresses, contributed papers, and colloquia.



Bibliography of works cited:

Ajay, H.B., Tillett, P. I., & Page, E.B. (1973). *Analysis of essays by computer*, AEC-II. Final Report to the National Center for Educational Research and Development for Project No. 8-0101. Washington, DC: Department of Health, Education, and Welfare.

Brown, Susan, ed. (1991). *Gender bias in testing: current debates, future priorities, a public policy dialogue*. New York: Ford Foundation Office of Communications.

Chomsky, Naom (1957). *Syntactic Structures*. The Hague: Mouton.

- Crouse, James, and Dale Trusheim (1988). *The case against the SAT*. Chicago: University of Chicago Press.
- Daigon, A. (1966). Computer grading of English composition. *English Journal*, 55, pp. 46-52;
- Farr, Charlotte Webb (1991). "Cognitive Psychology and Testing," in Green (1991), p. 282.
- Green, Kathy E., ed. (1991) *Educational Testing: Issues and Applications*. New York & London: Garland Publishing.
- Kuno, S. (1964). *Some characteristics of the Multiple-Path Syntactic Analyzer* (Language Data Processing Series. Report C6). Cambridge, MA: Harvard Computation Laboratory.
- McCollly, W., & Remstad, R. (1963). *Comparative effectiveness of composition skills teaming activities in the secondary school* (Cooperative Research Project No. 1528). Washington, DC: U.S. Office of Education).
- McCorduck, Pamela (1979). *Machines Who Think*. San Francisco: W. H. Freeman. p. 65.
- Owen, David (1985). *None of the above: behind the myth of scholastic aptitude*. Houghton Mifflin, 1985.
- Page, E.B. (1966). The imminence of grading essays by computes. *Phi Delta Kappan*, 48, 238—243.
- Page, E.B. (1967a). *Grading essays by computer Progress report*. Proceedings of the 1966 Invitational Conference on Testing. Princeton, NJ: Educational Testing Service.
- Page, E. B. (1967b). *Statistical and linguistic strategies in the computer grading of essays*. Proceedings of the Second International Conference on Computational Linguistics. 34.
- Page, E. B., Fisher, G. A., & Fisher, M.A. (1968). Project Essay Grade: A FORTRAN program for statistical analysis of prose. *British Journal of Mathematical and Statistical Psychology*, 21, 139.
- Page, E.B., & Paulus, D.H. (1968), *The analysis of essays by computer* (Final Report to the Bureau of Research at the U.S. Office of Education for Project 6-1318). Washington, DC: U.S. Department of Health, Education, and Welfare.
- Page, E.B. (1985). Computer grading of student essays. In T. Husen & T. N. Postlethwaite (Eds.), *International Encyclopedia of Educational Research* (pp. 944-946). Oxford, England: Pergamon;
- Page, E. B. Tillett, P.I., & Ajay, H.B. (1989). Computer measurement of subject-matter essay tests: Past research and future promise. *Proceedings of the Annual Meeting of the American Psychological Society*, 1, 39.

Page, E.B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software, *Journal of Experimental Education*, 62, 2, 127-142; pp. 134-135.

Paulus, D.H., McManus, J.E. & Page. E.B. (1969). Some applications of natural language computing to computer assisted instruction. *Contemporary Education*, 40, 280-285.

Stone. P.J., Dunphey, D.C., Smith, M. S. & Ogilvie, D.M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Wainer, Howard, ed. (1990). *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

