NISS

Perspectives on Statistics for Educational Research: Proceedings of a Workshop

Valerie S. L. Williams, Lyle V. Jones and Ingram Olkin

Technical Report Number 35
August, 1995

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Perspectives on Statistics for Educational Research: Proceedings of a Workshop

Table of Contents

Table of Contents (continued)

Perspectives on Statistics for Educational Research: Proceedings of a Workshop

Preface

This report presents abbreviated versions of the research papers presented and discussed at the Workshop, *Perspectives on Statistics for Education Research*, in Research Triangle Park, NC, April 7-8, 1995, sponsored by the National Institute of Statistical Sciences (NISS). The workshop focussed on current issues in educational statistics. It was organized by Lyle V. Jones of the University of North Carolina at Chapel Hill, Ingram Olkin of Stanford University, and Jerome Sacks of NISS, and was supported in part by a grant to NISS from the National Science Foundation.

The National Institute of Statistical Sciences was established in 1991 at the initiative of the national statistical community and a consortium of institutions in North Carolina. NISS exists to develop and facilitate collaborative cross-disciplinary statistical research. NISS projects address large-scale, complex problems of a statistical nature on which statisticians and other scientists from diverse disciplines can constructively collaborate.

In January, 1992, Ingram Olkin convened a workshop in Alexandria, VA, to explore needs for methodological and statistical advances in educational research. A primary focus became statistical problems faced by the National Center for Education Statistics (NCES) and how NISS might address those concerns. Following that workshop, a proposal was submitted to and subsequently funded by the National Science Foundation, to study a set of issues pertaining to education statistics. The Principal Investigators are Jerome Sacks, Ingram Olkin, Lyle Jones, and Daniel Horvitz.

This Workshop, *Perspectives on Statistics for Education Research*, provides a forum to report progress on problems that have been addressed, to invite critical comment from discussants who have dealt with the same problems from different perspectives, to promote active discussion among other participants, and to stimulate thought and critical reactions from readers of this report.

Sections 1, 2, and 3 constitute progress reports by NISS researchers on work largely stimulated by problems encountered at NCES. Section 1 addresses the topic of multiplicity and multiple comparisons, and considers adopting the "false discovery rate" as an alternative to

"familywise error rate" when selecting a criterion in multiple hypothesis testing. Section 2 reports methodological and applied research findings related to multilevel modeling. Section 3 describes two projects designed to link the results from regularly administered educational assessments to results from the less frequent and more costly National Assessment of Educational Progress. Section 4 is a discussion of some of the issues raised during the course of the Workshop.

Perspectives on Statistics for Educational Research: Proceedings of a Workshop

List of Participants

David Afshartous *University of California, Los Angeles*

Nancy Allen *Educational Testing Service*

Chris Averett *North Carolina Department of Public Instruction*

Nada Ballator *North Carolina Department of Public Instruction*

Betsy J. Becker *Michigan State University*

Yoav Benjamini *Tel Aviv University*

Kathy Billeaud *University of North Carolina at Chapel Hill*

Bruce Bloxom *Defense Manpower Data Center*

Lloyd Bond *University of North Carolina at Greensboro*

Peggy Carr *National Center for Education Statistics*

John B. Carroll *University of North Carolina at Chapel Hill*

Lee Chen *University of North Carolina at Chapel Hill*

Elliot Cramer *University of North Carolina at Chapel Hill*

Lori A. Davis *University of North Carolina at Chapel Hill*

Jan de Leeuw *University of California, Los Angeles*

John Hattie *University of North Carolina at Greensboro*

Eddie Ip *Educational Testing Service*

Richard M. Jaeger *University of North Carolina at Greensboro*

Valen Johnson *Duke University*

Lyle V. Jones *University of North Carolina at Chapel Hill*

Alan F. Karr *National Institute of Statistical Sciences*

Ita G. G. Kreft *California State University*

Alok Krishen *Glaxo Inc.*

Erich Lehmann *University of California, Berkeley*

Nicholas T. Longford *Educational Testing Service*

Mary McFarlane *Wake Forest University*

Donald McLaughlin *American Institutes for Research*

List of Participants (continued)

Lauren Nelson *University of North Carolina at Chapel Hill*

Agostino Nobile *National Institute of Statistical Sciences*

Ingram Olkin *Stanford University*

Abigail Panter *University of North Carolina at Chapel Hill*

Stephen W. Raudenbush *Michigan State University*

Paul Rosenbaum* *University of Pennsylvania*

Donald B. Rubin *Harvard University*

Jerome Sacks *National Institute of Statistical Sciences*

Eleanor Sanford *North Carolina Department of Public Instruction*

Juliet P. Shaffer *University of California, Berkeley*

Susan E. Stockdale *University of California, Los Angeles*

Larry Suter* *National Science Foundation*

Piyushimita Thakuriah *National Institute of Statistical Sciences*

David Thissen *University of North Carolina at Chapel Hill*

John W. Tukey *Princeton University*

William B. Ware *University of North Carolina at Chapel Hill*

John Wasik *North Carolina State University*

Iris R. Weiss *Horizon Research Inc.*

Peter Westfall *Texas Tech University*

Chris A. Wiesen *National Institute of Statistical Sciences*

Valerie S. L. Williams *National Institute of Statistical Sciences*

Robert Wolpert *Duke University*

Forrest Young* *University of North Carolina at Chapel Hill*

S. Stanley Young *Glaxo Inc.*

Yiu-Fai Yung *University of North Carolina at Chapel Hill*

Michele Zimowski *University of Chicago*

Rebecca Zwick *Educational Testing Service*

* unable to attend

Perspectives on Statistics for Educational Research: Proceedings of a Workshop

Program

Friday April 7

9:00 - 9:15 Jerome Sacks, *The Mission of NISS*

9:15 - 12:00 **Controlling the Proportion of False Discoveries for Multiple Comparisons**

Chair: Ingram Olkin

Lyle V. Jones, *Empirical Findings*

Valerie S. L. Williams, *Simulation Results*

John W. Tukey, *Future Directions*

Discussants Juliet P. Shaffer

S. Stanley Young

General Discussion

1:30 - 5:00 **Multilevel Analysis for Education Research**

Chair: Jan de Leeuw

Jan de Leeuw, *Overview*

Ita G. G. Kreft, *The Effects of Centering in Multilevel Analysis: A Reanalysis of Raudenbush and Bryk (1986) Using NELS:88*

David Afshartous, *Small Sample Properties of Multilevel Model Estimates and Multilevel Model Design*

Susan E. Stockdale, *Gender Differences in Math Achievement: An Empirical Application of the Propensity Score Adjustment*

Nicholas T. Longford, *An Infrastructure for Large-Scale Educational Statistics*

Discussant Stephen W. Raudenbush

General Discussion

Saturday April 8

9:00 - 12:00 **Linking NAEP to Other Assessments**

Chair: David Thissen

Bruce Bloxom, *Linking to a Large-Scale Assessment: An Empirical Evaluation*

David Thissen, *Linking NAEP-TSA to the NC End of Grade Test*

Discussants Chris Averett

Donald B. Rubin

General Discussion

12:00 Adjournment

1 Controlling the Proportion of False Discoveries for Multiple Comparisons

A more extensive report is available; see Williams, Jones, and Tukey (1994).

1.1 Introduction and Empirical Findings *(Lyle V. Jones)*

Multiple comparison procedures are defined as adjustments intended to control the probability of erroneous inferences under conditions of multiplicity, that is, when more than one statistical inference is drawn, or might be drawn, from a given body of data. Data from the National Center for Education Statistics (NCES) typically involve a large number of variables and large number of comparisons. For example, in the Trial State Assessment (TSA), a component of the National Assessment of Educational Progress (NAEP), comparisons are reported on achievement change from one year of administration to another for each state, and states are compared, each with each other.

Traditional multiple comparison procedures control the familywise error rate at a traditional significance level, often at $\alpha = 0.05$, thereby assuring that, in the long run, fewer than 1 in 20 reports will contain even a single "false discovery." A *family* is defined as a set of related inferences or comparisons as, for example, all pairwise comparisons between the states. Family size is the total number of those comparisons. Using traditional procedures, statistical power decreases as family size increases; as the family size becomes indefinitely large the power approaches zero.

Two rather different questions may be asked about comparisons. First, can we be confident about the direction — that is, the sign — of the underlying population comparison? This is analogous to a significance test. Also, for what interval of values can we be confident that the value of the population comparison is contained therein? This is the confidence interval approach.

Assume that we choose to control some error rate for statements of confidence about the direction of a population comparison; $\alpha/2$ for that comparison is a bound on the probability that we decide with confidence that the population comparison has one sign when in reality it has the opposite sign. This is somewhat different from the traditional formulation — it accepts Tukey's (1991, 1993) admonition that for real-world populations and for real-world comparisons the null hypothesis of zero difference is never, in fact, true (if one records enough decimal places). Thus, we replace the unrealistic null hypothesis of a zero population comparison with what may be a more realistic *perinull* hypothesis, namely, that the population comparison is near zero, but not

precisely zero. If the population comparison were truly zero, a declaration in either direction would be erroneous; however, if the population comparison is not precisely zero, as in the perinull situation, the population value has either one sign or the other, and then only one directional statement of confident difference is erroneous. We are right about half the time by chance. Here, α is a bound on twice the probability of being erroneously confident about the sign of the population comparison.

With multiplicity, the probability that a declaration of confident direction for any one or more comparisons will be in error will exceed $\alpha/2$, often very substantially if we make many comparisons. It is therefore necessary to adjust for this increased probability. Methods for accomplishing this may rely on either single-stage or sequential procedures.

The Bonferroni technique is the traditional, single-stage procedure that controls the familywise error rate. There is a single critical *p*-value, $\alpha/2m$, where *m* is the family size. If the observed *p*-value is less than the critical *p*-value, a confident direction is declared. To increase power while still controlling the familywise error rate, Hochberg (1988) proposed a procedure that employs a sequential adaptation of the Bonferroni technique.

The Benjamini-Hochberg (1995) procedure is a sequential technique which controls the proportion of all discoveries that are false; that is, of the comparisons declared to be confident in direction in the population, a certain proportion will be in error. The proportion expected to be in error is no greater than $\alpha/2$. This is a different criterion from the traditional familywise error control, but it is analogous to $\alpha/2$ as the probability of error in the single comparison case. It assures that the expected value of the false discovery rate is no greater than $\alpha/2$:

False Discovery Rate = False Discoveries/(1+Total Discoveries)

where the denominator of the ratio is 1 plus the total number of declarations of confidence and the numerator is the number of declarations that are in fact erroneous. (The addition of 1 in the denominator serves to avoid the possibility of dividing by zero.)

The Bonferroni and the Hochberg procedures control the familywise error rate at $\alpha/2$, thereby assuring that the probability of one or more erroneous declarations of confidence per family is no greater than $\alpha/2$. The Benjamini-Hochberg method assures that $\alpha/2$ is an approximate bound on the expected value of a false discovery ratio. For example, if $\alpha = 0.05$, we expect no more than 2½ percent of all declarations to be erroneous declarations. NCES has traditionally used the Bonferroni approach for all reporting purposes, and the Educational Testing Service has used the Bonferroni method in reporting NAEP results.

Table 1.1

Decision rules for four alternative criteria

For m ordered p-values, $p_1 \leq \ldots \leq p_m$, m being the number of comparisons in the family:	
Criterion	Rule
Unadjusted	Declare a confident direction for the ith comparison if $p_i \leq \alpha/2$.
Bonferroni	Declare a confident direction for the ith comparison if $p_i \leq \alpha/2m$.
Hochberg (1988)	Declare a confident direction for the ith comparison if, beginning with $i = m$ (largest p-value) and continuing toward $i = 1$ (smallest p-value), $p_i \leq \alpha/2(m-i+1)$; declare confidence in direction for all $j < i$ remaining comparisons.
Benjamini-Hochberg (1995)	Declare a confident direction for the ith comparison if, beginning with $i = m$ and continuing toward $i = 1$, $p_i \leq i\alpha/2m$; declare confidence in direction for all $j < i$ remaining comparisons.

Table 1.1 summarizes the decision rules for these alternative criteria. Table 1.2 provides a display of the change in average eighth-grade mathematics achievement scores for the 34 states that participated in both the 1990 and the 1992 NAEP Trial State Assessments. The 34 states are ordered from the largest p-value, or smallest contrast, at the top, to the smallest observed p-value at the bottom.

For both the Hochberg and Benjamini-Hochberg sequential techniques, the critical p-value varies as a function of the ordered size of the mean difference. Starting at the top of Table 1.2 with the largest p-value, the p-values for both techniques are equivalent to the unadjusted p-value; at the bottom of the table, with the largest mean difference, both techniques have the same critical p-values as the traditional Bonferroni correction. To implement these sequential procedures, we begin with the largest p-value and move down the column until the p-value associated with the test statistic is less than the critical value. For the Benjamini-Hochberg procedure, the first evidence of an increase in mathematics performance is that for the state of Kentucky; every remaining p-value also denotes a confident direction. For the Benjamini-Hochberg technique, there are 11 confident directions as compared with 15 for the unadjusted per-comparison approach; both the Hochberg and the Bonferroni procedures yield only 4 differences that are sufficiently large to be declared confident in direction.

Table 1.2

Mean achievement change for m = 34 states in eighth-grade mathematics, 1990 to 1992, and p_{crit}-values for alternative criteria: Unadjusted (UNA), Benjamini-Hochberg (B-H), Hochberg (HOC), Bonferroni (BON)

State	$\bar{X}_{92}$-$\bar{X}_{90}$ (*se*)	t	p-value†	p_{UNA}	$p_{B\text{-}H}(i)$	$p_{HOC}(i)$	p_{BON}
GA	-0.32 (1.8)	-0.18	.42814	.025	.025000	.025000	.000735
AR	-0.78 (1.5)	-0.52	.30141	.025	.024265	.012500	.000735
AL	-1.57 (2.0)	-0.78	.22004	.025	.023529	.008333	.000735
NJ	1.57 (1.9)	0.81	.20999	.025	.022794	.006250	.000735
NE	1.33 (1.5)	0.87	.19320	.025	.022059	.005000	.000735
ND	1.53 (1.7)	0.91	.18445	.025	.021324	.004167	.000735
DE	1.37 (1.3)	1.02	.15581	.025	.020588	.003571	.000735
MI	2.22 (1.8)	1.20	.11761	.025	.019853	.003125	.000735
LA	2.64 (2.1)	1.27	.10482	.025	.019118	.002778	.000735
IN	2.15 (1.6)	1.31	.09694	.025	.018382	.002500	.000735
WI	2.80 (2.0)	1.43	.07936	.025	.017647	.002273	.000735
VA	2.86 (1.9)	1.48	.07187	.025	.016912	.002083	.000735
WV	2.33 (1.4)	1.67	.05013	.025	.016176	.001923	.000735
MD	3.40 (1.9)	1.77	.04113	.025	.015441	.001786	.000735
CA	3.78 (2.1)	1.79	.03956	.025	.014706	.001667	.000735
OH	3.47 (1.9)	1.87	.03295	.025	.013971	.001563	.000735
NY	4.89 (2.5)	1.93	.02901	.025	.013235	.001471	.000735
PA	4.30 (2.2)	1.95	.02786	.025	.012500	.001389	.000735
FL	3.78 (1.9)	1.96	.02745	.025	.011765	.001316	.000735
WY	2.23 (1.1)	2.03	.02339	.025*	.011029	.001250	.000735
NM	2.33 (1.1)	2.03	.02325	.025*	.010294	.001190	.000735
CT	3.20 (1.5)	2.09	.02052	.025*	.009559	.001136	.000735
OK	4.18 (1.8)	2.38	.01018	.025*	.008824	.001087	.000735
KY	4.33 (1.6)	2.67	.00482	.025*	.008088*	.001042	.000735
AZ	4.99 (1.9)	2.70	.00452	.025*	.007353*	.001000	.000735
ID	2.96 (1.1)	2.77	.00374	.025*	.006618*	.000962	.000735
TX	5.65 (1.9)	2.99	.00202	.025*	.005882*	.000926	.000735
CO	4.33 (1.4)	3.12	.00141	.025*	.005147*	.000893	.000735
IA	4.81 (1.5)	3.23	.00100	.025*	.004412*	.000862	.000735
NH	4.42 (1.4)	3.27	.00090	.025*	.003676*	.000833	.000735
NC	7.27 (1.6)	4.58	.00001	.025*	.002941*	.000806*	.000735*
HI	5.55 (1.2)	4.74	.00001	.025*	.002206*	.000781*	.000735*
MN	6.42 (1.4)	4.75	.00001	.025*	.001471*	.000758*	.000735*
RI	5.10 (0.9)	5.37	.00000	.025*	.000735*	.000735*	.000735*

† The p-values are obtained using Student's t, df = 60.

* Confident direction with error rate < 0.025.

If one is willing to sacrifice control of familywise error in favor of control of the proportion of false discoveries, a great deal of power can be gained. This is demonstrated in another example of all pairwise comparisons among the 41 states participating in the 1992 Trial State Assessment for $m = 41 \times 40/2 = 820$. By the traditional Bonferroni adjustment there are 480 confident directions between states; the Hochberg technique admits an additional 13 declarations, and the use of the Benjamini-Hochberg results in an additional 159, while the unadjusted per-comparison approach increases the number of confident directions beyond the Benjamini-Hochberg by only 6.

Using either the Hochberg or the Bonferroni procedure, the probability is 0.025 that there are one or more erroneous statements in the whole set; using the Benjamini-Hochberg approach, it is expected that about 2.5% of the declarations will be incorrect. A willingness to accept that risk gains considerable additional information.

1.2 Simulation Results *(Valerie S. L. Williams)*

Because our interest in multiple comparisons was largely stimulated by the problems posed by the National Assessment of Educational Progress, we set out to design simulations based upon the structure of these data. True "achievement levels," μ_i, were defined for 48 states as the approximate median values in repeated realizations of each of 48 ordered observations from a normal distribution with mean 0 and variance σ_A^2. In order to generate an observed mean, $\bar{X}_i$, for each state, a random (standard normal) deviate was added to each μ_i. By manipulating the value of σ_A, five conditions of effect size were studied: the *perinull* condition of negligible differences among the μ_i where $\sigma_A = 0.001$, and four non-null conditions of increasingly larger effect sizes, $\sigma_A = \{0.3, 1.0, 3.0, 5.0\}$.

Two types of families of comparisons were investigated: uncorrelated differences ($m = 48$ independent comparisons, $\mu_i - M$), conceptualized as the 48 states compared with a constant national mean (M), and all pairwise differences among the 48 states ($m = 48 \times 47/2 = 1128$ nonindependent comparisons, $\mu_i - \mu_j$). The value of α was set to 0.05, with 10,000 replications.

Each of the three adjustment techniques — the Bonferroni, the Hochberg (1988), and the Benjamini-Hochberg (1995) — do maintain a false discovery rate bounded by $\alpha/2$ with independent hypotheses (or uncorrelated differences) and all pairwise differences, in all effect-size conditions; for all three procedures, the maximum false discovery rate is in the perinull situation, diminishing rapidly as effect size increases.

For our simulations, we considered what is referred to as *all-pairs* power, the probability of claiming significance for all differences among all pairs. For both uncorrelated and pairwise nonindependent families, the Benjamini-Hochberg technique generally results in greater power than that for the Hochberg or Bonferroni procedures, providing a substantial increase in power for the large effect sizes. The increase in power of the sequential Bonferroni technique over the traditional Bonferroni becomes detectable only for very large effect sizes.

All three procedures maintain familywise error rates at approximately $\alpha/2$ in the perinull situation. However, whereas the Benjamini-Hochberg technique does not maintain the familywise error rate at $\alpha/2$ with increased effect size, the Hochberg and Bonferroni adjustments become excessively conservative.

Further simulations demonstrated that the Benjamini-Hochberg advantage in power is associated with increasing family size and is little affected by the degree of dependence among the contrasts. It appears to be the larger family size that is driving the Benjamini-Hochberg procedure's increased familywise error rate.

To summarize the simulation results, the Benjamini-Hochberg technique provides only weak control of the familywise error rate, whereas the Hochberg and Bonferroni techniques are excessively conservative. All three procedures maintain a false discovery rate bounded above by $\alpha/2$ under all conditions, although the control of the false discovery rate becomes extremely conservative in the large effect-size conditions for all three adjustment procedures. The Benjamini-Hochberg technique provides substantially greater statistical power than either the Hochberg or Bonferroni procedures for the larger effect sizes and with very large families of comparisons.

1.3 Future Directions *(John W. Tukey)*

> *In politics, a young man*[1] *is often supposed to start on the left wing, but to turn more conservative with age. In conclusions, perhaps some young men will always tend to begin by believing that null hypotheses really can — and even do — happen, but they should learn better as fast as their world view permits.*[2]

****

Pidgin confidence is defined as the condition in which a data analyst is confident that the direction is either negative or positive, or is unsure. Pidgin confidence about direction corresponds much more closely to the real world than an unoriented accept-or-reject decision and

---

[1] young man *or* young woman

[2] A test for the existence of ESP is, perhaps, the sole exception.

it has serious consequences, for example, for asymptotics. So, all should "accept" that:

- Everything is different (at some decimal place), although we may not know the sign of each difference.
- 5% means one-sided 2.5%.
- If we are concerned with direction and not amount — either because we have too little data to go further or because we have no need to go further — pidgin confidence is called for.
- The next step about confidence in direction is full confidence, complete with a full set of confidence intervals, explicit or implicit.
- We must learn how to combine pidgin confidence with full confidence.

A major contribution of Benjamini and Hochberg (1995) is taking the false discovery rate (FDR) seriously as an alternative error rate to control. For the simple false definite rate, or false discovery rate as conceived of by Benjamini and Hochberg, there are some clear benefits:

(1) When very few comparisons are definite, the Benjamini-Hochberg procedure is similar to the Bonferroni adjustment or the Studentized Range, and you worry whether any are definite, particularly in the context of a large number of comparisons.

(2) If there are only a very few that are not definite, then the Benjamini-Hochberg procedure behaves very much like individual comparisons, and this is reasonable, too, because if you are convinced that all but one of the differences is real, then there is no reason for that one to have any multiplicity to it because it is the only thing of importance. (This is very different from increasing α from 5% to 20%.)

(3) The Benjamini-Hochberg procedure satisfies a relatively simple nominal requirement, making it moderately easy to describe — some of the fancier procedures people are inventing these days to sop up the last little drops of the Bonferroni procedure do not satisfy this requirement of simplicity.

(4) Often, the Benjamini-Hochberg procedure greatly reduces dependence on the definition of family size because it is not too far, in many practical cases, from the unadjusted per-comparison approach which is independent of family size.

On the negative side, the disadvantages of the Benjamini-Hochberg procedure are:

(1) Benjamini-Hochberg's weakest definite statements have larger error rates than the nominal familywise error rate. This is inevitable in a technique that does not correspond to techniques which control the familywise error rate in the strong sense, such as the Bonferroni procedure, Hochberg (1988) procedure, and Studentized Range technique.

(2) With Gaussian true values (and Gaussian errors), the Benjamini-Hochberg procedure expends only a fraction of what its nominal error rate provides when differences among effect sizes are substantial.

Any definite statements in a perinull situation have about a 50% chance of being wrong. (A *perinull* situation is close to, but not at, the null case.) Accordingly, there has to be an initial "bump" in what the cutoffs of FDR-controlling procedures provide. The Benjamini-Hochberg procedure seems to be a plausible way to include that initial bump.

A procedure which also controls the false definite rate can be combined with the Studentized Range to create a "compound" FDR procedure. Such a technique would enable claims of confident direction of very unequal strength to be sorted out, and well-measured differences to be recognized. However, a disadvantage of a compound FDR technique is that it will introduce a somewhat larger initial bump and the overall error rate will be increased.

How the two levels in a single compound procedure should be related will require continuing thought. Two pidgin-compatible choices that seem plausible because the FDR-confidence is at $\alpha/2$ are:

- pidgin-confidence at $\alpha/2$ and full confidence at $\alpha/2$ familywise,
- pidgin-confidence at $\alpha/2$ and full confidence at sufficiently less than $\alpha/2$ to average $\alpha/2$ erroneous statements per family.

As the supporting data grow stronger, the appropriate confidence procedures for controlling familywise error rates start with pidgin Studentized Range, continue with pidgin Welsch (1977) technology, then progress through pidgin complex techniques (e.g., the Braun and Tukey (1983) maximum subrange procedure), and eventually reach the full (non-pidgin) Studentized Range. How should this be reflected in procedures that control the FDR? Other alternatives should be considered; it is unclear whether compound methods will eventually be replaced by more integrated methods.

It is important to gain some understanding of whether a modified procedure for controlling the false discovery rate can come closer to the nominal error behavior in a Gaussian-Gaussian situation without pushing error rates for other situations too far above the nominal level — this must be assured for pidgin confidence. It is also necessary to determine the distribution of the number of errors for the Studentized Range at 1% and 10% familywise for assorted family sizes and degrees of freedom, and to study the error rates for comparisons at the margins of the Benjamini-Hochberg procedure which are dependent on the configuration of the hypotheses.

****

Once upon a time, we accepted the F-test as an omnibus procedure — a procedure that too often, in practice, degenerated into three steps:

(1) perform the omnibus F-test,

(2) if positive, believe all apparent differences,

(3) consider, as the only error to be considered, being definite when there are no differences at all.

This degeneration was clearly unacceptable. Today, we — or most of us — look to individual comparisons, adjusting our looking for multiplicity, and begin with confidence about direction of individual differences.

A two-sided t-test can degenerate in a similar way:

(1) perform the omnibus (two-sided) t-test,

(2) if positive, believe the sign with no thought of error,

(3) consider, as the only error, being definite ($\neq$) when there is no difference.

With luck, we will soon regard this strategy as almost equally unacceptable, and we will then begin with confidence about direction, and often proceed to full confidence.

1.4 Discussion *(Juliet P. Shaffer)*

Many have stated in the literature that the null hypothesis is never true, but before now there has not been a thorough reinterpretation of multiple hypothesis testing from this perspective. The authors here should be complimented for fleshing out the implications of their approach.

Given the assumption that the null hypothesis is never true, H_0: $\theta = 0$ can be replaced by the two incompatible hypotheses that H: $\theta > 0$ and H: $\theta < 0$. An error is rejecting the wrong hypothesis; lack of rejection means we don't know which is true. If we use a procedure that assures a low probability of errors, a rejection is then a statement that we are confident that we have a correct rejection.

Another issue is often confused with this one when the results of surveys are involved: It is true that the means of two finite populations are virtually never equal, and it doesn't make sense to test the hypothesis that they are. To quote Cochran (1963, pp. 37-38):

> It is seldom of scientific interest to ask whether $\bar{y}_j = \bar{y}_k$ because these means would not be exactly equal in a finite population, except by a rare chance, even if the data in both domains were drawn at random from the same infinite population. Instead, we test the null hypothesis that the two domains were drawn from infinite populations having the same mean.

Examples are the mean achievement in two schools, or the means of boys and girls within a single school.

It does sometimes make sense to assume that some intervention makes no difference. Another possibility is that the effect is so minute that it would take millions of observations to detect it — we would, therefore, like to treat it as zero. Consider also the cases in which one assumes that means must be ordered in a given way ($\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$) and we would like evidence that at least one of the inequalities is strict. If one accepts that the null hypothesis is never true, it seems there is nothing to test in this case — should these tests be abandoned?

The appropriate formulation of the null hypothesis depends upon the situation:

- In some situations it may be reasonable to assume that the null hypothesis is true; then rejecting it in any direction is an error. Usually, we also want to state a direction, but not necessarily.
- In some situations the only errors we care about are errors in direction. Then we can think of the two hypotheses as H: $\theta \leq 0$ and H: $\theta \geq 0$. (If θ is exactly zero, we can reject either hypothesis without it counting as an error.)

With a single hypothesis or a pair of directional hypotheses referring to a single parameter, it is relatively simple to move from one formulation to another and determine the change in error properties. However, with a stepwise multiple comparison procedure involving a number of parameters, this isn't necessarily so.

For directional inferences, it is unknown whether the FDR-controlling methods are valid. For stepwise methods with multiple hypotheses, it is not clear that the α is maintained when rejection of the null hypothesis of equality is followed by directional decisions. A counterexample in Shaffer (1980) shows that the probability of making a directional error can be even greater than 0.05 under some distributional assumptions. Does the $\alpha/2$ limit always hold under the perinull assumption?

If all parameters are zero (or very close to zero, as in the perinull situation) the familywise error rates (FWE) are:

- FWE under the null hypothesis is equal to α ,
- FWE under the perinull hypothesis is equal to $\alpha/2$,
- FWE with consideration of directional error only is 0 (when the parameters are all exactly 0),
- FWE under the null hypothesis but with directional conclusions : unknown.

If all the parameters are nonzero, then the FWE under the null hypothesis is 0, but the other FWE rates are all equal and we don't know what they are. It is important to investigate different combinations of null and non-null configurations and alternative definitions of error.

Both the Hochberg (1988) and Benjamini-Hochberg (1995) methods are based on the Simes (1986) equality. Simes proved that if all the null hypotheses are true, and the significance probabilities are independent, then for $p_1 \leq p_2 \leq \ldots \leq p_m$,

$$\text{Prob}\{p_i > i\alpha/m \text{ for all } i\} = 1 - \alpha$$

or,

$$\text{Prob}\{p_i \leq i\alpha/m \text{ for any } i\} = \alpha\,. \qquad (1)$$

Therefore, if the tests are independent, we can reject the global hypothesis that all m hypotheses are true at level α if $p_i \leq i\alpha/m$ for any i.

Simulation results suggest that the probability (1) is less than α for pairwise comparisons. But it is known that the probability (1) is greater than α for some patterns of nonindependent test statistics. In fact, the upper limit of the probability (1) is $\min\{1, \alpha\Sigma 1/i\}$, and this bound is sharp. For what types of test statistics is the probability (1) greater than α? Table 1.3 gives the upper limit of the probability (1) for selected family sizes.

Table 1.3

Maximum probability (1) for selected values of m

m	Maximum Probability (1)
2	0.075
3	0.09
10	0.15
20	0.18
100	0.26
1000	0.37

Another issue to consider is robustness under asymmetry. The t-statistic is symmetric around zero when the parent distribution is symmetric — if the nondirectional level is approximately $\alpha = 0.05$, the test can be reinterpreted as a directional test at the 0.025 level. However, if the parent distribution is asymmetric, the directional test is much less robust than the

nondirectional test, and the two-sided level may still be approximately 0.05 whereas the one-sided level is considerably different from 0.025. Although it may be possible to use robust procedures, the computational burden in large surveys makes this difficult in practice.

Appropriateness of the FDR criterion. In small studies in which the results must be interpreted as a whole, we would sometimes like the assurance that all our assertions are correct with a high probability. In that case, the familywise control of Type I error is the appropriate criterion. But in large surveys, with large families of comparisons, it seems unnecessary to have such an assurance — control of the false discovery rate may be more appropriate in such cases. Note that the FDR-controlling method of Benjamini and Hochberg (1995) is conservative. Benjamini and Hochberg have proposed FDR-controlling methods with greater power and higher familywise error.

The interpretation of the FDR criterion is often ambiguous. When there are many hypotheses, and many of these are false and easily rejected, the interpretation of an FDR-controlling technique is approximately valid: Not more than 5% of rejected hypotheses are erroneous. However, in the complete null case, if there are any rejections, they are all erroneous.

1.5 Discussion *(S. Stanley Young)*

The different multiple comparison methods — the Bonferroni, the Benjamini-Hochberg, and so on — protect against different error rates and have different characteristics. The relative cost of a Type I error and the value of an extra discovery should dictate the method. It is useful to study alternative methods of multiplicity adjustment, but if we are not careful, by including all these tests, we will only add to multiplicity.

The Benjamini-Hochberg procedure controls the familywise error under the complete null case, but what about the partial null case? The gains in statistical power are in places where the error control is weakest, when the partial null hypotheses are true, not when the complete nulls are true. There needs to be more study of partial null configurations.

Will one significant result determine the interpretation? In toxicology, in long-term rodent carcinogenicity testing, if there is *any* statistically significant increase in the incidence of *any* tumor in either sex of *any* species, then the test compound is labeled a "carcinogen." In clinical trials for drug safety, if *any* side effect is statistically significant, then the product must be safety-labeled and can then be at a competitive disadvantage or may not be approved. If there is shown to be *any* statistically significant benefit, then a drug is deemed effective. Committees try to interpret hundreds of hypothesis tests and these studies are seldom replicated.

What are the goals of the study? The goals should be clearly stated in the protocol:

- exploratory analysis,
- exploratory analysis with claims,
- specific analysis with estimated effects,
- specific analysis with estimated effects and claims,

with increasingly rigorous attention to Type I error. The best way to improve power in multiple testing situations is by limiting the testing to questions that matter, on the basis of recommendations by subject matter experts. This means limiting or eliminating data dredging by protocol — the number of tests and the possibility of additional testing must be set in protocol, as should be the error rate, methods, etc. It should also include reaching some consensus on the magnitude of effects that are to be considered important and meaningful. Once the data are on the computer, multiple testing is too easy; moreover, investigators become psychologically committed to the observed results. Does the purpose of the study require that all $m(m-1)/2$ comparisons be tested? The largest gains in power will come from carefully limiting hypothesis testing to the important questions, and allocating error accordingly. It is critical to establish, and follow, a protocol.

1.6 Discussion *(Yoav Benjamini)*

A true null hypothesis can sometimes be encountered, but it should be considered an extreme case. Yet the role of the null condition is useful as can be seen in the case of estimating multivariate parameters: If mean square error is important, it pays to set the smaller coefficients to zero rather than to use their estimates. Smaller coefficients are those smaller than the standard deviation of their estimators. A nearly optimal procedure for deciding which are these "small" parameters is equivalent to testing whether each parameter is actually 0.

There seems to be a disagreement as to the definition of the FDR criterion, but there is probably very little difference between the definitions FDR = FD/TD and FDR = FD/(1+TD), unless all hypotheses are nearly null. I prefer the former definition because I think there is usually an action related to the rejection of an hypothesis, and therefore a cost is involved. However, if no hypothesis is rejected, there is no cost; in that case, I like to set FD/TD = 0/0 to be 0.

A property of the FDR-controlling approach which is evident in the presented study — and we have had similar experience — is that, practically, the inference is quite insensitive to the size of the family. This is unlike the familywise Type I error rate which is very sensitive to family size. This property is extremely important, as I believe that there are actually no small studies! There are only studies of which small parts are published or quoted. In fact, the

familywise error approach can be manipulated to increase power by dropping parts of the study, and it is impossible for a reader to know it. The FDR approach also can be manipulated for that purpose, by including in the study many hypotheses which are obviously not true. But, here, it is done by over-reporting and, therefore, can be detected by a skeptical reader.

Finally, we should distinguish between the approach of controlling the FDR, and the specific FDR-controlling method used here. Just as there are many procedures for controlling familywise Type I error rate, there can be other FDR-controlling methods. The Benjamini-Hochberg (1995) procedure used here is simple and appealing, but is also known to be conservative — more so as the number of true hypotheses becomes smaller. An adaptive, more powerful procedure is available, and it, in fact, controls the FDR in the pairwise comparisons problem discussed here. Other procedures also may be designed.

1.7 References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289-300.

Braun, H. I., & Tukey, J. W. (1983). Multiple comparisons through orderly partitions: The maximum subrange procedure. In H. Wainer and S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 55-65). Hillsdale, NJ: Erlbaum.

Cochran, W. G. (1963). *Sampling techniques* (2nd edition). New York: Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-803.

Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *Annals of Statistics, 8*, 1342-1348.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73*, 751-754.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Tukey, J. W. (1993). Where should multiple comparisons go next? In Fred M. Hoppe (Ed.), *Multiple comparisons, selection, and applications in biometry* (pp. 187-207). New York: Marcel Dekker, Inc.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association, 72*, 566-575.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1994). *Controlling error in multiple comparisons, with special attention to the* National Assessment of Educational Progress. Technical Report #33. Research Triangle Park, NC: National Institute of Statistical Sciences.

2 Multilevel Analysis for Education Research

See also the Summer 1995 Special Issue, *Hierarchical Linear Models: Problems and Prospects*, of the *Journal of Educational and Behavioral Statistics*, *20*, 109-240, as well as Kreft, de Leeuw, and Aiken (1995) and de Leeuw and Kreft (1995).

2.1 Aspects of the Analysis of Large Educational Databases *(Jan de Leeuw)*

Large educational databases (LEDs) such as the National Longitudinal Survey of 1972, High School and Beyond, the National Educational Longitudinal Survey of 1988, and the National Assessment of Educational Progress (NAEP), have an incredible richness of data. They can be used to answer a large number of questions about relationships between variables, and even about mechanisms in the school-attainment system (comparisons of schools, of states, of race and gender, about discrimination, tracking, and so on).

In fact, the number of questions that can be asked is very, very large, and some questions are framed in such general terms (or in such explicit causal terminology) that they are impossible to answer. We are only interested here in questions which are framed in terms of relationships between variables, either about the existence of such relationships, or about their strength. Answers to such questions can be used by educators and policy makers, they can be translated into causal terminology, and they can be related to choices that have to be made in either small-scale or large-scale educational policy making.

Making such translations and interpretations is not the primary task of the statistician. The statistician's job is to describe the relationship in a clear and convincing way, and to give an indication about the stability of the description.

We shall concentrate on some of the LED projects which have been and which are still being carried out at the Division of Statistics, University of California, Los Angeles, as part of the NISS education research project.

2.1.1 Paradigm

The paradigm for analysis of LEDs is still linear regression analysis, the workhorse of applied statistics, and this is as it should be, because many of the questions LEDs suggest are formulated in terms of the relationship between an outcome variable and a number of predictors. Nevertheless, the linearity, homoscedasticity, and independence assumptions could conceal more than they reveal, and with increasing sample size and increasing computer power, we can perhaps become a little bit more daring.

2.1.2 Complications

Hierarchical structure. Most LEDs contain information about students, teachers, and schools, i.e., there are variables describing units at various levels. For some questions, it is necessary to combine units from various levels in a single analysis. This creates a multilevel problem, in which more complicated error structures are needed in the regression analysis to take care of the correlation within teachers, classes, or schools.

Various design and implementation questions related to multilevel models have been studied in considerable detail in the last few years, but the answers are still rather tentative.

Sampling design. Typically, LEDs are not simple random samples (certainly not with replacement). They are stratified and/or clustered, with oversampling of certain groups of students or certain geographical regions. It is unclear, so far, what the consequences of the sampling design are for the standard errors typically computed in LEDs. But the problem is certainly important, and one that we are working on.

Temporal dependencies. Although LEDs are sometimes cross-sections, they often have longitudinal aspects as well. This is especially true for cohort studies in which a number of students are followed for a comparatively long period. The longitudinal aspects of LEDs have not been explored much, and this is rather unfortunate. A lot of structure can be derived from the temporal aspect of a study, and process variables can be used to enter the black box and get a better idea of mechanisms (think of tracking or choice of classes). The few attempts to analyze longitudinally have been mostly along LISREL lines. We should explore explicitly the state-space and event-history approaches (the latter, especially, seem quite promising).

Spatial dependencies. LEDs include spatial information. This could be in the crude form of state, but also could involve the actual location of the school in some smaller scale studies (such as the California Learning Assessment System). Spatial-statistics techniques have not been used much, until now, to map school achievement and related variables. An increasing number of techniques and software are available now to explore these spatial dependencies (variograms, kriging, ALEX).

Mixed measurement level. LEDs have many, many variables, and inevitably some of these variables will be numerical, some will be ordinal, some will be nominal. In the same way that the mixing of levels from the hierarchy can be problematic, mixing levels of measurement can be problematic, too. Standard techniques of multivariate analysis tend to treat all variables as nominal (log-linear), or all variables as numerical (normal). Mixed-level analysis is comparatively rare, and often is only available for a small number of variables in the analysis.

Nonlinearities. There is no reason to suppose that the relationships between variables in a LED (regressions, for instance) will necessarily be adequately approximated by linear structures. By now, various computerized forms of nonlinear and nonparametric regression are available, using kernels or local linear fits, and they should at least be tried out on LED regression problems (comparatively small ones, preferably).

Censoring of school careers. Among the missing-data problems that are common in LEDs, censoring is an important one. In data from Delaware, we have information about all school children in the state between 1981 and 1993 (tests, school achievement, family background, schools, courses, discipline, etc.). This means that we have one full cohort of students from kindergarten to twelfth grade; the other cohorts in the data are either censored on the left or censored on the right. It makes sense in the analyses (planned for the next couple of years) to take this censoring into account. In the same way, it seems interesting to analyze data on dropouts, using survival analysis techniques.

Selection. There are various missing-data problems in LEDs. We have already mentioned dropouts and censoring, but attrition in cohorts is another example. There may be other less systematic forms of missing data (schools in the Delaware cohort do not take all tests in all years, some NAEP information may be collected in some states and not in others, and there may be various forms of "undercount"). In general, missing information may not be missing and random, and we have to think of ways of describing the process generating the missing data.

2.2 The Effects of Centering in Multilevel Analysis: Is the Public School the Loser or the Winner? A New Analysis of an Old Question *(Ita G. G. Kreft)*

Multilevel models raise new problems that need solutions. One of them is centering, with two centering choices, centering predictors within context, or grand mean centering. Both are statistically sound ways to improve the estimation of the parameters in the model. Users of the software package, HLM (Bryk, Raudenbush, Seltzer, & Congdon, 1989), commonly center within context. The literature shows that this type of centering is applied in two ways, with or without adding the mean back to the model. It is not discussed in this literature that centering — especially when means are not reintroduced in the model — produces different results from raw-score models, especially for second-level estimates.

The effects of centering are examined using a large national data set, NELS:88. The research question is an old one, regarding the success of the private school over the public school,

but now analyzed in a new way based on the model used by Raudenbush and Bryk (1986).

The results obtained with the NELS:88 data are in agreement with Raudenbush and Bryk when centering is applied to predictors. Among the results is that public/private sector has no significant main effect on mathematics achievement. Using raw scores produces another result, among them a significant effect of the private sector on achievement. Adding more omitted means to the model reverses this effect and the public sector shows a significant effect on mathematics achievement. Based on the disagreement among models, obtained only by treating the data differently, I conclude that an explicit choice has to be made between centering or not centering, and between reintroducing the means back or not. This choice requires the support of a theory, or research question, and will also depend on the goal of the analysis, either as policy research or for theory development. The examples illustrate that the new multilevel methodology is valuable as a tool for theory development, but caution should be taken when used to answer policy questions.

In the early days, before random coefficient models became available, Coleman, Hoffer, and Kilgore (1982, p. 196) concluded that "there is a tendency to converge over time among students from different backgrounds in Catholic schools, and a tendency toward divergence in the public school." Similar findings are reported by Raudenbush and Bryk (1986). My analyses do not support these findings or any conclusion that the private sector produces better results than the public sector. A voucher system, as proposed by then-president Ronald Reagan, to solve the "Nation at Risk" problem, is a political statement not founded in research. School-effectiveness research is still too much plagued by methodological problems and lack of theory to support the idea that private schools do better than public schools. In my analyses, it seems that the reverse is true: If the between-school part is corrected for the pre-existing student population differences, the public sector does a better job. The conclusion reached at this moment is that centering predictors has consequences, and the higher private school achievements may be artifacts of student-body characteristics. A similar conclusion was reached in my analysis with Dutch data, where no public/private sector difference was found, and results supported selectivity of schools as a more likely cause of higher achievements.

The past tense in Raudenbush and Bryk's concluding comments in their 1986 article can be read as optimistic. They wrote: "Research on school effects has been plagued by both methodological and conceptual problems" (p. 15). I think it still is plagued by problems, but multilevel modeling has made a breakthrough possible in various ways. It offers more ways to analyze the same data, it forces researchers to conceptually rethink their models, and it enhances

new developments. I agree with the part that follows the above citation in Raudenbush and Bryk's 1986 paper, where they mention that multilevel models are a promising development that greatly expands the range of methods for investigating schools, thereby expanding conceptualization.

2.3 Determination of Sample Size for Multilevel Model Design *(David Afshartous)*

The major purpose of this paper is to investigate the small-sample properties of multilevel model estimates, thereby providing information with which to guide sample-size considerations. If one desires to gather multilevel data on a large scale, the cost savings incurred by having a firm understanding of sample-size determination could be quite significant. Since the total sample size is merely the sum of the level-1 units (e.g., students), this problem is similar to the problem of examining the behavior of parameter estimates under various specifications of level-1 (n_j, students) and level-2 (J, schools) sample size. For example, if one were planning to gather educational data on a national scale, one would need to determine (among other things) two things:

(1) How many schools to sample.

(2) How many students to sample from each school.

The differential monetary costs of these two processes make sample-size determination an important issue. For instance, although it may be relatively inexpensive to obtain information from an additional student within an already sampled school, the sampling of an additional school may be prohibitively costly. Thus, sample-size determination may be viewed as a constrained optimization problem. Although the sampling literature (Cochran, 1977) provides some simple results for cluster sampling, there exists little discussion of this issue within the multilevel-model literature where most papers deal with estimation issues.

I investigate the effects of sample size on multilevel-model estimates from a subsampling perspective. Moreover, my method fits nicely into the "sample reuse" or "resampling" methods that are currently popular in various statistical literatures. I investigate an actual hierarchical data set (NELS:88) as follows. First, care is taken to specify a reasonable two-level model with respect to the entire data. Next, repeated subsamples of various sizes are drawn from the population of level-2 units (schools). (For a general discussion of subsampling methods, see Hartigan, 1969.) Given the already small number of level-1 units (students) within each level-2 unit (school), only level-2 units are subsampled. Under this resampling scheme, both fixed effects and covariance component estimates are observed.

Finally, the results are presented in a graphical manner such that the researcher may obtain

useful information, given his/her specific needs. This subsampling scheme may be viewed from a variety of perspectives:

- Model Bootstrap: Given an estimated model, the distribution of the subsampled estimates is an empirical measure of the stability of the original estimates. In addition, the location of the original estimate within this distribution provides an indicator of the "extremeness" of the original estimate.
- Data Compression: Using multilevel models to analyze large data sets is often tediously slow. It is often useful to perform exploratory analysis with a smaller portion of the data, formulating a number of hypotheses that may be examined with respect to the entire data. How large should this subsample be in relation to the entire data?
- Sampling Design: If one desires to collect multilevel data, how should resources be allocated to collecting data at various levels of the design, e.g., among schools and students? (If one's sample consists of the entire population, there exists little difference between the issues involved in the previous item.)

The subsampling routines were carried out under the following design conditions. Random samples of size 40, 80, 160 and 320 schools were drawn from the sample population of schools, and the above model was fit to each of these subsamples. This procedure was repeated 100 times, thereby providing data with which to assess the sampling variability of estimates both within and across the given design conditions. Thus, for each design condition, e.g., a sample of 40 schools, we have 100 values for each parameter in our given model. Student mathematics score is modeled as a function of the sex, race, and socioeconomic status (SES) of the student, while schools are differentiated according to urbanicity and the extent to which school lunches are subsidized, an indicator of the poverty level of the students within a given school.

Unbiased estimates of fixed effects are readily obtainable from subsamples of relatively small size, e.g., 40 schools. With regard to the variability of these estimates, there exists substantial improvement each time the subsample size is doubled. Furthermore, there exists preliminary evidence that the rate of this improvement is dependent upon the type of fixed effect being considered. Specifically, intercept fixed effects evince a proportionally greater reduction in spread the first time the subsample size is doubled, while the corresponding reduction-for-slope fixed effects remain relatively constant each time the subsample is doubled.

With regard to the variance components, I concentrate on the T matrix of level-2 variance components. Recall that this is a 2×2 matrix for our given model, containing elements for the estimated variance of level-1 intercept, level-1 SES effect, and the covariance between them.

With these three estimates, one may estimate the correlation between level-1 intercept and level-1 SES effect, i.e., the ratio of their covariance to the square root of the product of their respective variances. Thus, for each design condition, we obtain 100 values of the estimated correlation between level-1 intercept and level-1 SES slope. The estimate based on the entire data is 0.60, which suggests that schools with a high average mathematics score are likely to exhibit a high SES effect, i.e., the impact of student SES on student mathematics is likely to be more pronounced in such schools. My analyses demonstrate that, unlike the results for the fixed effects, a relatively unbiased estimate is unlikely to be obtained from a small sample of schools; indeed, even for samples as large as 160 schools, an unbiased estimate is unlikely. Matters improve greatly once the subsample size is increased to 320 schools, but this is eight times as large as the corresponding size that was necessary to produce unbiasedness for the fixed effects. With regard to the spread of our estimates, the situation is also worse than that for the fixed effects. The initial doubling of the subsample design has little effect; indeed, the interquartile range actually increases. Moreover, given that the correlation statistic lies in the [-1, 1] interval, the repeated subsamples of 40 and 80 schools do not provide much guidance for narrowing the original parameter space.

The results of this paper provide some guidelines with regard to sample-size consideration. For instance, the fixed effects and variance components behave quite differently under small-sample-size situations. Thus, if one's research interests are concerned mainly with obtaining accurate and reliable estimates of variance components, a relatively large number of level-2 units are necessary. On the other hand, if one is interested solely in the estimates of fixed effects, the number of level-2 units that are necessary decreases substantially. In either case, additional level-2 units improve the accuracy and reliability of the estimates. Moreover, the reliability of the fixed-effects estimates may be related to the type of fixed effect, e.g., intercept or slope, being studied.

2.4 Gender Differences in High School Mathematics Achievement: An Empirical Application of the Propensity Score Adjustment *(Susan E. Stockdale)*

Research over the past 25 years indicates that gender differences in mathematics achievement favoring males are not typically found prior to high school. In high school, differences favoring males are common, particularly in the areas of problem solving and applications. The gender-related differences in mathematics achievement have been attributed to

a number of variables, most notably, differential course-taking patterns and exposure to math, different learning styles, teacher behavior and learning environment, parental attitudes and expectations, and socioeconomic status as well as other background characteristics of students. One problem that is not addressed in this literature is the inability to isolate the effects of gender socialization from observed biological sex when observational data are used.

One method by which the effects of socialization may be partialled out involves the use of a propensity score, as developed by Rosenbaum and Rubin (1983, 1984), for gender. The propensity score is a quantification of the gender/mathematics socialization effect and is here used as an independent variable in an ordinary least squares regression model of mathematics achievement on biological sex. One important goal of this research is to examine gender differences in mathematics achievement over time while controlling for factors that contribute to the gender differences and are associated with gender, such as mathematics course-taking patterns. It was anticipated that the gender differences in mathematics achievement would be greatly diminished.

This research uses panel data from the student component of the National Educational Longitudinal Survey of 1988 (NELS:88). The three outcome variables include the item response theory (IRT) number correct on the mathematics achievement tests for the eighth, tenth, and twelfth grades. The analysis reveals the imposition of gender-stereotypical expectations upon students and gender-stereotyped behaviors. Males receive more encouragement than females from parents and teachers to take mathematics classes, and the expectations for achievement and educational aspirations for males are higher than for females. Females are more likely to rely on the opinions of peers and siblings when making decisions about taking mathematics courses. Neither observed sex nor gender socialization, indicated by the gender-propensity score, have a large effect on mathematics achievement. The greatest effects on mathematics achievement were produced by previous mathematics achievement. Because the differences between males and females in mathematics performance, as measured by the IRT scores, are very small, it is unclear from this analysis if the gender-propensity score, either as a theoretical measure of socialization or as a statistical balancing adjustment, improves the prediction of mathematics achievement.

2.5 An Infrastructure for Large-Scale Educational Statistics *(Nicholas T. Longford)*

Multilevel analysis has played an important role in the methodological developments of educational statistics over the last ten years or so. This paper reflects on some of the statistical principles that have encouraged or brought about these developments and outlines a more

comprehensive statistical infrastructure for educational research. The following is an incomplete list of topics for such an infrastructure:

- Variances as effective summaries
- Small-area-like estimation of large numbers of quantities
- Combining sources information
- Informative missingness
- Integral use of prior information
- Formulation of substantive problems
- Tests of statistical significance that reflect the research interest
- Communicating uncertainty
- Statistics as a profession in educational research

The first five points are principally of technical nature, while the last four have minimal technical context. Although some of the topics are familiar to the educational-research community, their potential is far from exhausted.

The central theme of the paper is that multilevel analysis, however useful in numerous applications, is far from sufficient when used as the sole statistical equipment of an analyst. Often it is not the method itself, but merely its implementation in a software package, that is used to analyze data. This invariably leads to the software dictating the research issues that will be addressed. A more effective application of statistics is achieved by carefully selecting statistical tools to fit the research problems. With few tools, good fit is unlikely, unless the tools are universal.

The greatest potential for the application of statistics in educational research is in dismantling the barriers between the so-called substantive research and statistics. To the extent that statistics is looked upon as a system of delivery of estimates, their standard errors, p-values, and confidence intervals, there is good reason not to take statistics seriously. Statistics is a profession in which thorough understanding of the subject matter is an absolute must, and so contractual arrangements to deliver numbers represent a poor application of the profession.

2.6 Discussion *(Stephen W. Raudenbush)*

In organizing these remarks, I have decided to discuss the papers out of order, considering first a pair of papers (by Afshartous and by Longford) that deal with general methodological issues and then two (by Kreft and by Stockdale) that describe applications.

2.6.1 General Methodological Issues

The audience posed several fundamental questions during de Leeuw's overview, and a

response to these provides a framework for discussing issues raised in the papers. The questions concerned the relationships between full maximum likelihood, restricted maximum likelihood, and fully Bayes inference for two-level models. Using de Leeuw's notation, we have the linear model

$$Y_j = X_j b_j + \delta_j \,, \tag{1}$$

where $\delta_j \sim N(0, \sigma^2)$, and $b_j \sim N(Z_j\gamma, \Omega)$.

This can be viewed as a Bayesian linear model with a normal exchangeable prior for b_j. Inferences about the b_j are based on their conditional distribution given the data, Y, and the "hyperparameters" γ, σ^2, and Ω. Widely available software provides estimates of the hyperparameters via "full maximum likelihood" (ML) based on all units $j = 1,...,J$, and bases inferences about b_j on point estimates of the hyperparameters.

A second approach adds a "flat prior" for γ, e.g.,

$$\gamma \sim N(0, \Gamma) \tag{2}$$

where Γ is an arbitrarily large matrix, and bases inferences about b_j and γ on their joint conditional distribution given the data and the hyperparameters, where the hyperparameters are now conceived as σ^2, Ω. Often, these hyperparameters are estimated via restricted maximum likelihood (RML) and inferences about b_j and γ are based on point estimates of these parameters.

A third approach adds a flat prior for the variance-covariance parameters σ^2, Ω as well. Inferences about all unknowns are based on their joint conditional distributions given only the data. This is the fully Bayesian approach.

The results of the three approaches converge as J increases. For small J, many statisticians would regard the fully Bayes approach as the "best" with the RML approach ranking a distant second, with ML close behind. The reason is that reliance on point estimates is most problematic when J is small, especially when the data are highly unbalanced. (When the data are fully balanced, the three approaches give quite similar results.) However, in computational efficiency the ranking is reversed, with ML being first, RML a close second, and fully Bayes a distant third. The implication is that for large-J problems, one would tend to use ML or RML, retaining fully Bayes for small-J, unbalanced problems. When J is small, the computational burden of going fully Bayes is comparatively small and the inferential gains comparatively large.

This all seems neat and clean except that little is known about what values of J are sufficiently large to justify the simpler approaches. The answer must be context-specific, depending on the degree of imbalance in the data, the dimension of b_j, the size of the variances to be estimated, and whether interest focuses primarily on b_j, γ, or the variance-covariance elements themselves. Here is where studies like that of Afshartous are helpful. He takes a large,

widely analyzed data set, NELS:88, a nationally representative longitudinal study of secondary school students. He then takes subsamples of various sizes from this data set and applies ML, watching to see how stable the results are across subsamples, depending upon the sample size, J. Estimation theory would suggest that, when J is sufficiently large, the likelihood for a given variance parameter in Ω will be quite peaked and symmetric about the mode, so that each of the three methods will give similar results. However, when J is small, that likelihood may be substantially positively skewed with a mode (near zero) that poorly represents plausible values of the parameter. If the data are balanced, the impact on inferences about γ will be negligible because those inferences do not depend on the unknown variance component. However, inferences about b_j will be sensitive, because its conditional variance does depend on the unknown variance components. However, Afshartous does not consider inferences about b_j . The greater the imbalance in the data, the more dependent are inferences about γ to point estimates of the variance component.

Afshartous' results follow this script as expected, but they give us details about sample size in the context of an important and representative data base. His results and method also have implications for how future multilevel research might be designed for optimal inference.

Longford discusses a variety of important questions, non-technical and technical. Given the increased flexibility and complexity of modeling and estimation, he suggests that the relationship between substantive researcher and statistician ought to be recast, and I agree. Incorporating the statistician from the start in the conceptualization and design may be essential if the analysis and interpretation are to fill the bill.

Longford discusses the problem of estimating many parameters and, especially, small-area estimation as topics for future application of these methods. Actually, my whirlwind tour of models at the beginning of these remarks is based on a long history which might be viewed as starting with Lindley and Smith and Novick in the early 1970s and continued with Dempster, Rubin, Tsutakawa and others in the early 1980s. Educational applications contributed heavily to this history, which focused primarily on inferences about b_j, but yielded the basic model of Equation (1) and all of the estimation alternatives. The Bayesian framework appears most helpful in conceptualizing the estimation of many parameters; no sampling mechanism is required to construe the unknowns as random. Thus, if we have states j = 1,...,50 within the United States as key units, Equation (1) still makes complete sense within the Bayesian framework even though a sampling theory approach would view the collection of states as an enumeration of the population rather than a sample.

2.6.2 Applications

After seeing Kreft's paper, I reviewed my 1986 article with Bryk and noticed the following quotation (p. 15):

> We expect researchers to encounter difficulty in interpreting the results of an HLM analysis, which are considerably more complex than results from an ordinary linear model. The ensuing technical discussions, however, should not deflect us from primary concerns. Ultimately, Cooley's (1981) recommendation that we engage in more substantive discussions about the causal models we assume in conducting research on schools remains paramount. By facilitating the explicit modeling of processes that occur both within and between various levels of school organization, HLM analyses can enrich such discussions and advance research on school effects.

In my opinion, that conclusion still stands. I am afraid, however that the Kreft paper does deflect us from primary concerns. She assesses whether and how to center from the standpoint of maximizing model fit, using deviance statistics to indicate the superiority of one model over another. However, in the 1986 paper, we carefully laid out our choice of models at each level based on the substantive goals of the research. The level-1 model identified for each school its mean and its slope. The level-2 model viewed these as depending upon school mean SES and sector (public versus Catholic). The results of this simple analysis were quite striking: a strong disordinal interaction between SES and sector such that the within-school SES slope was much weaker in the Catholic than in the public sector. This aspect of the results supported prior research by Coleman and others in that they also had found a smaller SES slope in the Catholic sector. However, their analysis, which failed to control the contextual effect of mean SES and which failed to disentangle the SES-outcome relationship into its between- and within-school components, had shown an *ordinal* interaction. This implied that the Catholic minus public mean difference is never negative regardless of SES. Our analysis showed that it is indeed negative for high-SES students.

To examine the sensitivity of this result to model specification, we added homework at level-1, reasoning that differences within schools in homework time would reflect, in part, selection differences. This put the SES-sector interaction to a fairly stiff test, but the interaction was sustained. *It was never our intention to control mean homework at level 2!* We viewed mean homework as endogenous to sector: If a public/private sector effect existed, it would be achieved in part by establishing high expectations and requiring high effort of all students. Thus, in assessing the *existence* of the sector effect, it would be counter-productive to control mean

homework. The fact that adding mean homework at level 2 would increase the fit of the model was not an argument in favor of including it. Thus, the group-mean centering at level 1 (which removes mean homework) fit our purposes. We reasoned that mean homework represented, in part, effects of school policy and practice, while within-school variation in homework reflected individual differences in approaching school work. Our research aims implored us to control the latter but not the former.

Kreft claims that "no software manual or other publications report the difference in macro-parameter estimates obtained by centering around the context mean as compared to raw score solutions." This is false. These issues are discussed in detail in my 1992 book with Bryk (Bryk and Raudenbush, 1992; see tables 5.9 and 5.10 and the surrounding text).

Centering is a tricky topic that needs more consideration in the context of a variety of applications. However, key modeling goals must be kept in mind, and the meaning of variables such as "homework" or "SES" at each level of aggregation must be thoughtfully evaluated if these centering issues are to admit to a solution.

Stockdale's paper provides a number of intriguing findings about gender differences in outlook and treatment that appear to reflect important gender differences in socialization. These are based on a large, nationally representative sample of schools and students, and ought to encourage reflection about equal opportunity in mathematics education.

However, I am not at all persuaded that the Holland-Rosenbaum-Rubin machinery for causal inference in non-experimental studies applies to sex differences in mathematics achievement. First, that machinery is designed to approximate the randomized experiment we wish we could implement but can't. It is not clear what randomized experiment we are attempting to approximate here; thus, the notion of sex or gender as a "treatment" remains problematic. Second, all of the covariates used to "predict" propensity are endogenous to sex. It is not clear how the direct effect of sex could be interpreted in this case. If it disappeared, we could not rule out a biological effect of sex. There is an alternative sociobiological interpretation to which I would personally never adhere, but which cannot be challenged by this analysis.

2.7 References

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., Seltzer, M., & Congdon, R. T., Jr. (1989). *An introduction to HLM: Computer program and users' guide*. Chicago: University of Chicago, Department of Education.

Cochran, W. G. (1977). *Sampling techniques* (3rd edition). New York: Wiley.

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). Cognitive outcomes in public and private schools. *Sociology of Education, 55*, 162-182.

Cooley, W. W., Bond, L., & Mao, B. (1981). Analyzing multi-level data. In R. A. Berk (Ed.), *Educational evaluation methodology* (pp. 64-83). Baltimore: Johns Hopkins University Press.

de Leeuw, J., & Kreft, I. G. G. (1995). *Questioning multilevel models.* Technical Report #31. Research Triangle Park, NC: National Institute of Statistical Sciences.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association, 64*, 1303-1317.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). *The effect of different forms of centering in hierarchical linear models.* Technical Report #30. Research Triangle Park, NC: National Institute of Statistical Sciences.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1-17.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassifications on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

3 Linking Other Assessments to NAEP

The presentations by Bloxom and Thissen are based on material presented more extensively elsewhere: The linkage for the Armed Services Vocational Aptitude Battery has been described by Bloxom, Pashley, Nicewander, & Yan, (1995), and that for the North Carolina End-of-Grade tests by Williams, Billeaud, Davis, Thissen, & Sanford (1995).

3.1 Linking to a Large-Scale Assessment: An Empirical Evaluation

(Bruce Bloxom)

This study develops and evaluates the linkage of a routinely administered measure, the Armed Services Vocational Aptitude Battery (ASVAB), to the proficiency scale of a large-scale assessment, the National Assessment of Educational Process (NAEP). The ASVAB is used in determining eligibility for enlistment in the military. Also, ASVAB results are used in Department of Defense reports to Congress on the aptitudes of cohorts entering the military. Because these reports have depended on the use of aging norms from 1980, it is desirable to link the ASVAB to NAEP to obtain results with reference to more current norms.

The NAEP mathematics scale is composed of the five subscales used in the 1990 NAEP assessment of twelfth-graders, a population that is similar in age and education to the population to which the ASVAB is normally administered. The ASVAB scores used in the linkage included scores on the Arithmetic Reasoning and Mathematics Knowledge subtests, which have substantial content overlap with four of the five NAEP subscales. Although the overlap of content between the NAEP and ASVAB scales is deficient in one of the five areas, regressing proficiency on alternative tests within a projection framework does not require measurement-equivalent or parallel tests (Mislevy, 1992). The extent of overlap that does exist between the assessments supports using them in an illustrative study evaluating this type of linkage.

Our interest is in comparing populations measured with tests other than NAEP; here, the other test is the ASVAB. Scores on the ASVAB are used to estimate the distribution of mathematics proficiency on the NAEP scale. This is a density estimation problem that is defined in terms of a regression model. The focus is not on the precision of the regression coefficients or on the interpretation of them. Rather, the focus is on the accuracy — in terms of both bias and sampling variability — of the estimated proficiency distribution, where the estimate is a projection obtained from the distribution of scores on the measure being linked and the regression-modeled joint distribution of that measure and the assessment to which it is being linked. A specific concern is with biases that can result from shifts in population characteristics.

The real-data illustration of the projection used background information and the subtests of the ASVAB to estimate the NAEP mathematics proficiency distribution in a sample of applicants for military enlistment. Three models were investigated. The first was the NAEP assessment model involving the regression of NAEP subscales on NAEP background characteristics. A second model used an ASVAB-based linkage model regressing NAEP subscales on ASVAB subscale scores and a more limited collection of ASVAB background characteristics. There was also a combined NAEP and ASVAB model, involving the regression of NAEP subscales on NAEP background characteristics, ASVAB subscale scores, and ASVAB background characteristics. For comparative purposes, simulated data were used to illustrate the accuracy of the projection where the true proficiency distribution was known and all assumptions of the linkage model were correct.

There were only very slight differences in sampling variability across these models; no differences in patterns of systematic variability were evident. The linkage-based proficiency distribution estimated by regression on the ASVAB subscales was close to the true distribution in the simulations, and close to the combined-model distribution in the real data analyses.

However, while the mean on the ASVAB mathematics tests was about average for 17-year-olds, the mean for the NAEP tests was 0.5 standard deviation below the NAEP average for 17-year-olds. Proficiency on the NAEP scale may have been systematically underestimated because of motivational factors in the administration of the NAEP measures to examinees in this study — examinees were aware that the NAEP test results, in contrast to the ASVAB results, were of no personal consequence.

3.2 Linking the North Carolina End-of-Grade Mathematics Test to the NAEP Scale *(David Thissen)*

The establishment of linkages between state testing programs and the National Assessment of Educational Progress (NAEP) would reduce the reliance on a national testing program, such as the Trial State Assessment (TSA), for the purpose of tracking student achievement, and would facilitate the comparability of student outcomes across assessment instruments, across different education programs, and across states. State-NAEP linkages would also allow units smaller than states, i.e., school districts, to assess the effects of education reform on student performance, and monitor progress with respect to consensual national achievement standards. The present investigation reports the procedures and results of one successful attempt at linking a statewide assessment program to the NAEP scale using projection methodology, providing statistics useful for description and policy making.

3.2.1 Development of the NC-NAEP linkage

Projection makes use of the empirical relation between scores on tests that do not measure exactly the same thing to predict the distribution of scores on one test (e.g., NAEP) from the distribution of another test (e.g., a state assessment). Mathematics proficiency, as measured by the NAEP exercises, is not identical to mathematics proficiency as measured by the North Carolina End-of-Grade (EOG) tests. Nevertheless, there is considerable overlap in the content frameworks, despite the fact that the two tests were built to different specifications.

Data. The NC-NAEP linkage test administered in February 1994 contained 78 items, including a short form of the EOG mathematics test for grade 8 (40 multiple-choice items) and two blocks of released 1992 NAEP mathematics items (38 items: 29 multiple-choice and 9 free-response). The Educational Testing Service (ETS) provided estimates of *a*, *b*, and *c* parameters for each NAEP item for a unidimensional three-parameter logistic model.

Eighth-grade examinees were selected in a two-stage sampling design, where the primary sampling unit is the school: 103 schools were drawn, and 99 of these participated. A target sample of 30 students was randomly selected in each of the schools; actual counts ranged from 21 to 33 participants. A total of 2824 students were tested. Two ethnic categories reflecting relative educational advantagement were created for the projection analyses: BHN ("Black," "Hispanic," and "Native American" examinees) and WA ("White," "Asian/Pacific Islander," and "Other" examinees).

Selection of a model for NAEP means and standard deviations. For each student, a NAEP posterior distribution is obtained based on the individual response pattern, the population distribution, and the IRT parameter estimates provided by ETS. (The prior, also provided by ETS, is a non-Gaussian histogram for the 1992 national NAEP.) The sum of these distributions, weighted by the sampling weights, is the estimate of the 1994 distribution. Students are then categorized into groups based on ethnic classification and EOG scaled score. By summing the weighted posteriors for each ethnic classification × EOG score combination, two distributions of NAEP scores for each EOG scaled score category are created.

The projection equations fit the posterior mean of each ethnic classification × EOG score category as the dependent variable; this is the mean of the posterior distribution created by summing all the individual posteriors for each examinee in an ethnic classification × EOG score category. The predictors are ethnic classification (dummy-coded BHN = 0 and WA = 1) and EOG score category. The standard deviations of the ethnic classification × EOG score category posteriors are predicted from EOG score category only. The projection was accomplished using

weighted least squares regression analysis in which the ethnic classification × EOG score category subgroupings are weighted by the number of students in each subgrouping.

Bootstrap computation of standard errors. Standard errors for the regression coefficients were computed using a bootstrap procedure described by Sitter (1992a, 1992b). The bootstrap plan included finite population corrections at the first and second sampling stages, for school and for student-within-school. The standard error of each of the statistics is the standard deviation computed from the bootstrap distribution.

Projection of February NAEP results from the May EOG administration. A second analysis projected the February NAEP results from the regular May administration of the EOG test. A total of 2313 students from the NC-NAEP linkage sample were matched with their May EOG scores; the average EOG increased about five points for this sample. The regression coefficients and bootstrap standard errors for predicting February NAEP results from the regular May EOG test administration differed very little from those predicting NAEP from the February NC-NAEP linkage test administration.

3.2.2 State results

The 1994 NAEP TSA results for North Carolina were obtained directly from the linkage sample. When the data from the statewide census administration of the EOG test became available, the second set of projection equations were developed, and the data from all 82,657 eighth-grade students were used to project (or, in this case, postdict) the February NAEP results. Comparison of the estimated proficiency distribution from the projection with that obtained directly from the NAEP administration showed that the two distributions correspond closely.

NAEP TSA results are most commonly reported as a set of quantiles. Table 3.1 shows the values of the percentiles typically reported, as observed in the 1994 special administration of NC-NAEP to the linkage sample of 2824 students, and as projected from the (near) population of 82,657. Six of the seven percentiles from the projection are within one standard error of the original sample values, and the seventh is well within two standard errors. The standard errors from the projection are smaller than the standard errors computed from the original sample administered the NAEP test (Observed), because measuring the *population* with the wrong test (EOG) results in less sampling variation than making an inference from a *sample* with data from the right test (NAEP).

The average NAEP scores for North Carolina's 119 school districts were projected. When the data from the 1995 administration of the EOG eighth-grade mathematics test become available, it will be possible to project the state's (hypothetical) 1995 NAEP TSA results. Moreover, with

Table 3.1

Observed and projected percentiles for the distribution of mathematics proficiency for the NC eighth-grade students (bootstrap standard errors are shown in parentheses).

	5th	10th	25th	50th	75th	90th	95th
Observed	206(2.0)	220(2.0)	244(1.7)	267(1.7)	291(1.3)	308(1.4)	319(1.3)
Predicted	208(1.4)	221(1.1)	243(0.8)	268(0.6)	291(0.5)	310(0.6)	320(0.6)

the next administration of an eighth-grade mathematics TSA, the results of the present NC-NAEP projection will be evaluated for accuracy.

3.3 Discussion *(Chris Averett)*

North Carolina began census testing in 1978 with "off-the-shelf," norm-referenced tests, the California Achievement Tests. In the late 1980s, because of concern over inflated national norms and a surging interest in accountability, state education policy makers became interested in developing tests better matched to the state's curriculum. About the time of the first Trial State Assessment in 1990, the State Board of Education adopted a statewide mathematics curriculum patterned after the standards established by the National Council of Teachers of Mathematics, and the State Department of Public Instruction was in the process of developing new statewide achievement tests.

In 1992, North Carolina administered its last norm-referenced test, ending the State's national trend line. That same year, the second Trial State Assessment was administered, providing a second data point in North Carolina's TSA trend line. We began administering our new End-of-Grade Tests in 1993, but in 1994, there was no Trial State Assessment for comparing North Carolina students to the nation.

While we believe it is necessary to have a state assessment that measures what is taught in our schools, it is also very important to make national comparisons which can provide useful information about the progress of North Carolina students as compared with children around the country. By linking our tests to the TSA, we can also inform school districts about how well their students are performing with respect to national data and standards.

3.4 Discussion *(Donald B. Rubin)*

A very useful way to think about linking is as a missing data problem: You wish everyone in the whole population had been given all the relevant tests, but they had not, and instead you have a giant data matrix that is 99% empty. The more complex the pattern of missing data, in

the sense of nonmonotone missingness, non-overlapping variables and so on, the more machinery you need in terms of models, assumptions, and computational procedures.

With regard to the accuracy of the estimated population distributions based on plausible values, which are multiple imputations (see Rubin, 1987), a very important point that was made is that the accuracy of the estimate of the population distribution is not the same as the accuracy of an individual prediction. In a regression model, I can *know* the regression coefficients, but I can never know exactly how an individual will perform; however, I can know exactly the correct average prediction of performance for a large group of people with specific characteristics.

Multiply-imputed estimates of population distributions have more variability than the true population distributions — they are posterior distributions for population distributions and so reflect the uncertainty of parameter estimation. As the sample size for estimation approaches infinity, the variability of multiple imputations will approach that of the population distribution; however, as long as the observed data are not infinite, there will be uncertainty in the estimated population parameters.

A related point about multiple imputations (or plausible values) is that when judging the quality of multiple imputations, do not make use of mean squared errors of data predictions — the focus is on making the correct inference, not on the best prediction of a specific observation.

If you are only asking one question or doing one analysis, an analytical solution is best if it is available. Multiple imputation attempts to answer all questions at the same time in a noisy way, but obviously cannot do as good a job solving one problem or answering a particular question analytically.

If you don't want to live with assumptions, you can try to make the sample size large enough so you can live in *Asymptotia*. Then you can use standard large-sample methods and be relatively confident about your inferences. However, when there are budget constraints, this becomes impossible, so you must trade off dollars versus assumptions. The issue is not whether you are making assumptions but whether the assumptions are reasonable trade-offs versus the dollars you could spend to get rid of the assumptions.

The bootstrap is an *ad hoc* procedure, and as such, has to keep being patched for specific applications (e.g., the inclusion of the finite population correction is obtained by the Bayesian version of the bootstrap). But, any *ad hoc* patch may not work later when combined with other such patches. Thus, I have reservations about obtaining the sampling variability of the posterior distribution with the bootstrap — this is a curious blend of Bayesian methodology and frequentist bootstrapping. Whether this works needs to be validated theoretically or established through

asymptotic correspondence with multiple imputation results. The method of multiple imputation has the virtue of having been derived from a Bayesian perspective and shown to have the right sampling properties.

3.5 References

Bloxom, B., Pashley, P., Nicewander, A., & Yan, D. (1995). Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics, 20,* 1-26.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Sitter, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association, 87,* 755-765.

Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics, 20,* 135-154.

Williams, V. S. L., Billeaud, K., Davis, L. A., Thissen, D., & Sanford, E. E. (1995). *Projecting to the* NAEP *scale: Results from the North Carolina End-of-Grade testing program.* Technical Report #34. Research Triangle Park, NC: National Institute of Statistical Sciences.

4 Further Issues

Near the end of the formal program, John Tukey commented on several issues that had been suggested to him by earlier discussions.

4.1 Final Remarks *(John W. Tukey)*

Intercepts save writing but they confuse interpretation. What you see from the intercept is mainly just a collection of regression coefficients pasted together — the centercept information is totally concealed.

A carrier in multiple regression is not really what it seems to be, but rather it is the result of regressing all the other carriers out of that carrier. And what that means, who knows? It depends on the situation. Suppose there is an X_{ij} carrier, or there is an X_{ij} carrier and $\bar{X}_i$, or there is a centered X_{ij} and $\bar{X}_i$. The essential issue here is that the last two are the same — they have the same coefficients of X_{ij} or centered X_{ij}, and you get the same fit. What matters is not whether you center, but whether you put the $\bar{X}_i$ into the model.

It is careless to include carriers in whatever form of expression they were recorded. One example is "hours of homework" because what is suggested quite strongly from the analysis of the effectiveness of SAT preparation is that some sort of down-curving expression of time is likely to be better than raw time. If I were going to do something with hours of homework, I would examine the square root of the number of hours rather than the raw number of hours of homework. This is not the sort of thing that always can be obtained by analyzing the data; this is the sort of thing often obtained from past experience.

Hybrid expression is something that we should be doing much more frequently.

Make rootograms, not histograms, where the heights of the blocks are proportional to the square root of the count, or more generally, the square root of the count per unit base. This has the following consequences: (a) It stabilizes the variability; (b) it increases the attention to the tails which is usually a good idea (the tails are more important than histograms tend to show); and (c) the square root of a Gaussian density is just a multiple of a Gaussian density, so if things would have been Gaussian in the histogram, they will be Gaussian in the rootogram as well.

(A point about empirical variance and bias inversely proportional to size: There is a "multi-halver jackknife" which uses both halves of each of a collection of intersecting and nearly orthogonal split halves, and which can allow for any reasonable form of blocking, but need not do so.)

Bootstraps are now arbitrarily complicated — almost as complicated as the EM algorithm. With the so-called "percentile bootstrap" (which I call the *seductive bootstrap*), you draw the sample and look where the percent points are. This has a potentially serious difficulty if you have a skewed situation, because it is skewed in the wrong direction. The typical jackknife has only half the trouble because it is not skewed, and being not skewed is only halfway from the right skewness to the wrong skewness. It is time we did something with these skewness problems.

Like the bootstrap, the jackknife has to be considered an asymptotic device in realistic situations.