Academy Panel Joins the Fray Over Job Testing #24

It endorses the controversial practice of reinterpreting scores of blacks and Hispanics on a widely used employment test

A PANEL of the National Academy of Sciences has endorsed a system for reinterpreting the ability test scores of blacks and Hispanics to make them competitive with those of whites.

This controversial recommendation is part of a report attempting to resolve a dispute between the Labor Department and the Justice Department over the use of the General Aptitude Test Battery (GATB), the most widely used civilian employment test in the country.

Job candidates who take the GATB at state-run employment services are referred to employers according to a "race norming" formula that helps employers identify the highest scorers within ethnic categories. The practice, promoted by the Labor Department, has been attacked by the Justice Department as "intentional racial discrimination." However, the academy panel, headed by Yale University statistician John Hartigan, concluded that the practice is justified because of the imprecision of the test.

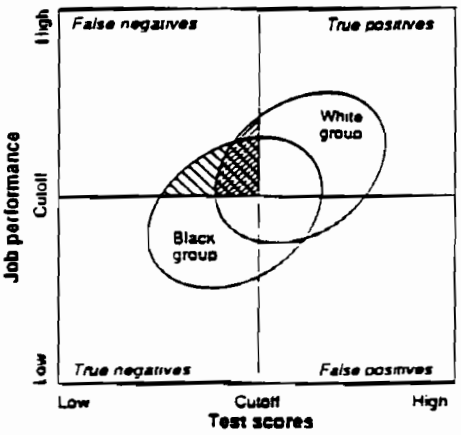
The report,* issued on 22 May, has a direct bearing on two different but often intertwined issues: the value of ability tests in predicting a candidate's future performance on the job, and appropriate strategies for minority applicants, who argue that tests unfairly discriminate against them.

While research has shown that "objective" selection procedures such as the GATB are the best single predictor of employee performance, it has also shown that such tests put blacks and Hispanics at a severe disadvantage. The resolution to this conflict between minority interests and concerns about worker quality will affect millions of job referrals at state employment services and will have implications for how tests are used throughout private industry.

In coming years, the problem is likely to escalate: the country's rapidly changing demographics means that the majority of future work force entrants will be minorities. Meanwhile, test use is on the rise. Since the debate cannot be resolved scientifically, no matter what the government decides, the ultimate outcome will depend on the courts.

The focus of this particular controversy, the GATB, tests three major domains: general cognitive ability, perceptual ability, and psychomotor ability. Since 1981, the Labor Department's Employment Service has been encouraging its use to improve screening at state employment services, which annually process about 20 million job-seekers and refer them to private employers.

To offset the fact that blacks and Hispanics have lower scores than whites on the GATB (as they do on virtually all standardized ability tests), the government has promoted within-group scoring. Applicants label themselves as "black," "Hispanic," or "other," and their test scores are recomputed to reflect their percentile ranking within their racial group. Top scorers from all groups are then referred to employers.



How good a predictor? *Many potentially able workers fail the GATB test (false negatives), and since blacks as a group score lower, they are more likely to be in this category.*

In November 1986, the Justice Department ordered the employment service to "cease and desist" from the practice. The service agreed to put a moratorium on expansion of the program pending the completion of the study. (The stay is continuing as the government studies the findings.)

The Academy's Committee on the GATB was charged by the Labor Department with assessing whether the test is a good one, whether it is valid (a useful predictor of job performance), whether it is fair to minorities, and whether race norming is a good idea. The answer on all counts was yes.

By supporting GATB, the academy panel was endorsing a concept that was formulated in the late 1970s. That was when psychologists Frank Schmidt of the University of Iowa and John Hunter of Michigan State University came up with an answer to the landmark 1971 decision of the Supreme Court that had almost crushed testing. The case, *Griggs* v. *Duke Power Co.* established that any employment practice having "adverse impact" on minorities constituted evidence of discrimination. This put the burden of proof on employers who had to show that their criteria were directly job-related. Many employers abandoned ability testing rather than devote the enormous resources necessary for constructing and validating job-related tests.

Until the late 1970s, psychometricians could see no way around the need for separately validated tests because the results of validation studies were so varied that it did not appear that those from any one study could be generalized from one place to another.

But then "validity generalization" emerged on the scene. This concept means, simply put, that a general measure of cognitive ability that is valid for some jobs is valid for all jobs. The theory is based primarily on the work of Schmidt (formerly at the U.S. Office of Personnel Management) and Hunter, who applied new analytical techniques to 500 validity studies. They found that when the results of the studies were corrected for various distortions—primarily those imposed by small study samples—they yielded substantial correlations with a wide range of jobs. Their conclusion: "professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all jobs . . . in all settings."

The employment service was excited by these findings, particularly in light of Hunter and Schmidt's calculations that widespread adoption of the GATB would result in an $80-billion-a-year savings to the economy through increased productivity. The services implemented a pilot program to test out the concept.

Because blacks and Hispanics get lower scores, they added within-group scoring to achieve parity in referrals. For example, because blacks as a group score a standard deviation below whites, the raw scores of those who fall in the 50th percentile are assigned to the 84th percentile. Hispanics with the same score are in the 66th percentile. As of 1986, when the Justice Department stepped in, the system had met with

*"Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery" (National Academy Press, Washington, D.C., 1989).

considerable favor from employers and was being used in 40 states on about 8% of applicants.

The NAS panel members generally affirmed the whole program. They confirmed that the GATB is a valid predictor of job performance, although they concluded that Hunter and Schmidt's estimated correlation of 0.5 between test scores and job performance was too high. Citing additional, more recent, studies, they adjusted it down to the neighborhood of 0.3. Although many critics of ability testing believe a 0.3 correlation represents an improvement over random choice so small as to be meaningless, panel leader Hartigan said at a Labor Department briefing that the critics have misinterpreted the effectiveness data, and that use of the tests actually improves the ability to predict a worker's productivity by 30%. And he reaffirmed the panel's agreement with Hunter and Schmidt that the GATB is more reliable than any other single selection criterion, including interviews, educational background, skills, and job experience. Said Hartigan, "we probably cannot afford not to use" the GATB.

The most crucial aspects of the report have to do with questions of whether blacks and Hispanics are unfairly dealt with, and what to do about the fact that their scores are significantly lower than those of whites. With regard to whether the GATB is racially biased, the NAS endorsed findings from a vast body of research on the subject showing that it is not—that is, the tests predict equally well for blacks and whites. The NAS confirmed that, if anything, the test slightly favors blacks by "overpredicting" their job performance.

Many test critics, including Richard T. Seymour of the Lawyers Committee for Civil Rights, have claimed that the test is racially biased because more potentially able black workers are rejected by the test, and more poor white workers get passing scores. In scientific terms, more blacks than whites fall in the category of "false negative," and more whites are "false positives."

The NAS panel said, however, that this disproportion has nothing to do with race per se but arises from the fact that it is the marginal scorers who are most likely to fall in the false-negative category (see graph). This can be demonstrated by performing the same analysis using one racial group.

As the academy panel pointed out, the problem of false negatives is an inevitable result of the limited predictive capability of the test. But the panel has put itself in a somewhat awkward position. Study director Alexandra Wigdor emphasized that "this correction is not for racial underprediction, it is underprediction for low scorers." But the report presents race norming as a way to "ensure that able black and white workers have the same chances of referral"—thus implying that the test is biased.

What the academy has done is to take a remedy adopted by the employment service on purely pragmatic grounds and present it as one that is scientifically justified—even though, according to James Sharf, an industrial psychologist at the Office of Personnel Management, the vast bulk of research shows that pure rank-ordering of scores "is the only scientifically justified position." Sharf, a member of the committee's liaison group, quotes Hartigan as saying, at a meeting 2 years ago, that "this committee is not about to put a scientific fig leaf on a naked political argument." Sharf says the feeling at OPM is that the committee has done just that. Hartigan could not be reached for comment.

The widespread adoption of race norming could open up a Pandora's Box of new questions and litigation. Nonetheless, overt and systematic policies of racial preferment may be better than informal arrangements that are often neither efficient nor fair. As Wigdor observed, employers are in a "tremendous bind" because they risk adverse impact suits when they use objective selection procedures, and reverse discrimination suits when they set up programs favoring minorities. As a result, "a lot have turned to quiet, unobtrusive quota systems that can't be recognized in court."

■ CONSTANCE HOLDEN

Consorting on Superconductors

They may be the most powerful corporate rivals in U.S. research, but IBM and AT&T have decided to join forces—along with the Massachusetts Institute of Technology and Lincoln Labs—to ensure American primacy in superconductivity in the 21st century. The venture, to be known as the Consortium for Superconducting Electronics, will attempt to transform what has been largely an interesting laboratory phenomenon into real-world applications. If it appears to be working, it could become something of a model for corporate rivals in other fields to work together with universities on long-term applied research programs.

The initial focus of the consortium will be applications in the world of microelectronics, such as high-speed signals processing circuits and junctions between electronic devices, that are expected to constitute the first uses of the new superconductors.

This may prove particularly wise because superconducting electronic devices are expected to be less affected than many other putative applications by the recently reported (*Science*, 26 May, p. 914) phenomenon known as "flux creep" that can destroy the superconducting properties of the new materials when they are exposed to magnetic fields. Still, the most promising electronics applications are, as yet, uncertain. Says William Brinkman, director of physics research at AT&T's Bell Laboratories, the consortium should "find an answer to the question of whether there are technical opportunities open to us."

Indeed, the fact that the big players in high-temperature superconductivity have decided to join forces is being viewed by some as an indication that they are looking for a way to share some of the costs while they explore the formidable barriers that lie before them. Says Dean Eastman, a vice president of IBM's research division: "We believe that it's going to take considerable time to achieve applications, so we need to look at this over the long haul, not just when it's in vogue among scientists."

A novel feature of the consortium that sets it apart from other university-industry research arrangements is that it is built around a detailed plan, complete with technical milestones, and it will be managed by a central group to be located at MIT. "It is not a consortium in which IBM, AT&T, Lincoln Labs, and MIT are each following their own programs and sharing results; they will be following a single technical plan," says MIT provost John Deutch. Adds Eastman of IBM, "the consortium will act like a small company."

Not so small, though, when ranked against other superconductivity start-ups. Indeed, the new entity will command an annual budget of $12 million to $15 million a year. A grant of $4 million to $6 million is being sought from the Defense Advanced Research Projects Agency to finance work at MIT; the rest will be kicked in by each industrial partner. Each institution will have the equivalent of five or six full-time researchers working for the consortium.

Deutch says he will be spending some time over the next year seeking additional members for the consortium from industry, the national laboratories, and other universities. Similar consortia could follow, Deutch predicts. "We have it in mind as being a model for how universities, industry, and the national labs can work together on things that are in the national interest."

■ COLIN NORMAN

Letters

Fairness in Employment Testing

Constance Holden (News & Comment, 2 June, p. 1036) has written a useful summary of the issues addressed in the National Research Council's recent report on the General Aptitude Test Battery (GATB). However, she has seriously misrepresented the committee's position on adjusting the scores of black and Hispanic test-takers so that *able* nonwhite workers have the same chance of referral to jobs as *able* white workers.

As the figure reproduced in Holden's article illustrates, the direct use of test scores, without adjustments, will result in the false rejection of a larger proportion of able black and Hispanic workers than of able white workers (although some able workers in all groups will be erroneously rejected). This is not because the test is biased, as Holden says the report implies. The "false negative" effect is not a function of race or ethnicity. Rather, the disproportionate rejection of able minority workers is due to the interplay of two factors: the modest predictive accuracy of the test and the lower average test scores of these applicants. The combination of the two means that proportionately more black and Hispanic test-takers who could perform well on the job will be falsely predicted to be unsatisfactory.

Moreover, the committee does not believe, as a reader of the article might surmise, that the within-group percentile scoring system currently being used by the Labor Department's Employment Service is the only way—or in all circumstances an appropriate way—to ensure equal referral chances for *able* minority and white applicants. Our endorsement of the within-group percentile method is clearly linked to the current predictive power of the GATB. As long as the GATB predicts job performance with only modest accuracy (correlation, 0.3), scores based on group norms will achieve approximately equal referral rates for able white, black, and Hispanic workers.

A crucial point is that the size of the adjustment needed to effect the recommended outcome will necessarily depend on the accuracy with which job performance is predicted by the test. The attraction of the second scoring strategy specifically endorsed in the report (a so-called performance-based method in which test scores are adjusted by group so that the distribution of test scores at a given level of job performance is the same for all groups) is that it is responsive to changes in the predictive accuracy of a test. Highly accurate prediction would mean small score adjustments; at current levels of accuracy the adjustments would be just about the same as those produced by the within-group percentile system.

Finally, Holden quotes an official from the Office of Personnel Management as saying that the vast bulk of research shows that pure rank-ordering of scores "is the only scientifically justified position." We disagree. It is indeed true that selection on "pure rank-ordering" will generate a work force with the highest expected productivity. But it is also true that able black and Hispanic workers will be rejected far more frequently by such a referral policy than whites at the same level of job performance. This is a scientific fact, demonstrated theoretically and empirically in the report.

JOHN HARTIGAN
Committee on the General Aptitude Test Battery, National Research Council, Washington, DC 20418
and *Department of Statistics, Yale University, New Haven, CT 06520*
ALEXANDRA WIGDOR
Committee on the General Aptitude Test Battery, National Research Council

Response: Wigdor and Hartigan basically raise two objections about my article.

1) They say that I say that they imply the GATB is racially biased. In fact, I made it clear that the committee did not find racial bias in the test. However, some statements could lead a rational person to infer that the test is unfair to minorities—for example, the authors assert that able blacks and Hispanics will be rejected by test scores "far more frequently" than whites "at the same level of job performance."

2) The authors disagree with an Office of Personnel Management official that pure rank-ordering of scores is the only "scientifically justified" position. But, as I indicated in my article, while within-group scoring may well be justified socially and politically, the scientific basis is questionable, for at least two reasons.

Since the purpose of the test is to maximize productivity by predicting worker performance, and since, as the authors acknowledge, pure rank-ordering produces a workforce of the "highest expected productivity," it is arguable that there is no scientific justification for tinkering with test scores that does not improve the validity of prediction.

Moreover, the committee has adopted a race-based solution for a problem that it says is not race-based. It might be argued that a more "scientific" solution to the fact that marginal scorers get more false negatives would be to adjust the scores of all low scorers as a group.

My article certainly may be construed as being critical of the committee's reasoning, but I do not see where I have misrepresented their position.—CONSTANCE HOLDEN

Holden writes that the National Research Council committee concerned with job test scores and job performance states that it has "scientifically justified" ways of reporting scores to replace "pure rank-ordering." The "system for reinterpreting the ability test scores of blacks and Hispanics" depends in large part on the committee's distinction between "predictive fairness" and "performance fairness." The former entails predicting performance from test scores. The latter entails predicting test scores from performance, but it is used by the committee to support affirmative action hiring.

"Performance fairness"—which implies group equality in outcome of the selection procedure—does not represent a scientific basis for that purpose because it is "internally contradictory" (*1*). It lacks consistency in applications because there is a reversal in its effect when it is applied to a remedial program for low scorers as opposed to job referrals for high scorers. For example, if within-group scoring were used in determining eligibility for a Head Start program, "performance fairness" would favor whites.

Affirmative action programs for certain minorities rest on value judgments, not on educational and psychological data or on statistical finagling with test scores. Value judgments should be made explicitly and openly, not camouflaged by rhetoric or statistical legerdemain.

The very name "performance fairness" is rhetorical camouflage: The name suggests that tinkering with scores will result in equal performance. But in fact it will not. When the decision to select is made, the only information available on performance of either individuals or groups is from the imperfect selection instrument or instruments.

A related issue is that the committee's rationale can be extended without any empirical or technical qualification to tests and grades used in the selection of undergraduate, graduate and professional school students, and the hiring of professionals. A qualification that procedures suitable for working class occupations are not suitable for the learned professions is not acceptable in a democratic society.

LLOYD HUMPHREYS
Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, IL 61820

REFERENCES

1. N. D. Peterson and M. R. Novick, *J. Educ. Meas.* 10, 3 (1976).

#25

SIOP Symposium: Affirmative Action in the 1990's

April 21, 1990, Miami Beach, Florida

Frank L. Schmidt

Linda Gottfredson's paper

I believe that Linda's comments on recent Supreme Court decisions are correct. But I would go even further, especially with respect to the Wards Cove decision. From a purely legal point of view, Wards Cove puts adverse impact cases on the same basis as general civil law; that is, it requires the plaintiff to prove the charges that he brings, as opposed to requiring the defendant to prove his innocence. However, from a psychological point of view, something more interesting is going on: The Supreme Court is recognizing the fact that adverse impact has been discredited as a trigger for presumptive discrimination. In this sense, the Supreme Court is ahead of many I/O psychologists.

The old theory was that if there is adverse impact, that fact indicates a good probability that there is discrimination. The employer then had to prove there was not by demonstrating job relatedness and predictive fairness. But research in I/O psychology indicates that adverse impact is not a plausible trigger for discrimination. Adverse impact almost always exists, and it almost never indicates discrimination. The clearest case is aptitude and

ability tests: they almost always have adverse impact, but we know from 20 years of research that they are virtually always valid and predictively fair. The score differences are not due to bias or any other problem in the tests, and hence the tests are not discriminatory. Life is discriminating but the tests are not. In Wards Cove, the Supreme Court in effect recognized this fact. An anecdote will illustrate this change. In the Berkman case in New York City in the early 80's, plaintiffs alleged that a physical abilities test showed a much larger male-female difference than other such tests. The defense moved to introduce physical abilities testing data from extensive military studies to show this was not so. The judge would not allow this because, in his words, "Title VII assumes all groups to be equal a priori." This is the idea that is now dead.

Some I/O psychologists, in my experience, have not yet recognized this fact. Some I/O psychologists maintain that Wards Cove went too far, that there should be stronger government regulation of employee selection than that decision allows for. Specifically, they say that if a selection procedure has adverse impact, then the employer should be required to show job relatedness. They do not realize that the theory of adverse impact has been discredited--by our own research.

It is hard to escape the suspicion that such attitudes are economically motivated, at least in part. For almost 20 years, I/O psychologists have used government regulation and legal pressures to market their selection and validation services. They have become dependent on this artificial marketing support, that is, employer fears of expensive litigation. Many seem to have lost the ability to market their services based on intrinsic value--increases in efficiency and productivity. This is what we must go back to. In an age of increasing competition, both nationally and globally, this should not be too difficult.

The role of VG is not precisely as stated by Linda. VG showed that validities were generalizable, and that finding laid to rest to spector of the invalid test. Single group validity and differential validity studies showed validities generalized to blacks and Hispanics also. Virtually all ability and aptitude have adverse impact, and they were once believed to be often invalid. The combination of adverse impact and invalidity was one definition--the most frequent definition --of discrimination. But selection procedures could be invalid but still predictively unfair. It was the research on predictive fairness of tests --not VG research--that laid this fear to rest.

Linda's analysis of the Civil Rights Bill now in Congress is excellent. I would go further and say this: If that Bill becomes law in its present form, the damage to the U.S. economy will be so great that Congress will be forced in 2-3 years to repeal it or radically modify it. The productivity losses will be so large that U.S. industry will not be able to function in the increasingly competitive international economy.

This Bill could even become the subject of international trade negotiations. In recent trade talks with Japan, the U.S. agreed, in return for Japanese trade concessions, to take steps to make the U.S. more competitive--improve the schools, reduce the federal debt, increase the U.S. savings rate. I can foresee a day when Japan demands that we repeal legislation like this Act--because such legislation reduces the productivity and efficiency of our economy. We have already successfully made such demands on Japan in the anti-trust area.

Linda's analysis of the NAS panel's use of the discredited Cole-Darlington definition of test fairness to create a false "scientific" justification for test score adjustments is on target. What the committee did in that area is a serious offense against a cherished scientific and scholarly value: intellectual honesty. A panel of experts, under aegis of the NAS, used their expertise and specialized

training to deceive the general public about the nature of a serious social problem. Attempts to use complex statistical sleight of hand techniques to obsfucate and disguise this problem will not succeed in the long run. Further, this sort of approach feeds the cynicism about experts and leaders that is already widespread in society. It is socially corrosive. The report would have been far more honest had it recommended, as its own value judgment, that race norming be used as a way of opening jobs to minorities (while still retaining most productivity gains). What was intellectually dishonest was the deceptive attempt to provide a bogus statistical, psychometric, and scientific justification for score adjustments.

Unlike Linda, I do not believe that it is crystal clear what social policy should be in this case. There are arguments on both sides as Gerry Barrett has noted in his presentation. However, I believe it is clear that social policies--especially policies this important--should not be decided on an intellectually dishonest basis.

Gerry Barrett's presentation

Gerry presented a tremendous amount of material. He presented:

1. a good overview of the current legal status of affirmative action;
2. a good analysis of common misinterpretations of affirmative action;
3. some convincing evidence of the negative consequences that affirmative action can have.

To my knowledge, none of this is controversial. So I will limit my comments to statements later in his presentation that I believe are problematic.

Methods of "Thumb on the Scale" Affirmative Action

One method Gerry described was differential weighting of tests in a battery so that the total "has approximately equal means" for minorities and nonminorities. In my experience, this would not work: any weighted composite will have substantial adverse impact, if the tests are aptitude and ability tests and if they have decent reliabilities. This conclusion is also supported by the research literature.

Gerry stated that these "Thumb on the Scale" approaches have not been adequately discussed or critiqued in the professional literature--except for the recent NAS report. Two points here:

1. As described by Linda (and by Mary Tenopyr in a recent article), the NAS discussion was anything but "adequate."
2. Gerry's statement ignores the many articles in the 1970s examining and critiquing the various test fairness models and the quota model. These appeared in such journals as JAP and Psych. Bulletin. In particular, the Cole-Darlington model adopted by the NAS panel was extensively critiqued. Our professional literature has carefully analyzed all these options.

Gerry stated that "Thumb of the Scale" approaches to AA "are based on unstated assumptions and at present have no scientific status." I have heard this argument frequently but do not accept it. Consider for example race-norming or percentile equating as used by USES. Of course it is true that this method has no scientific status. It was never intended to have scientific status or scientific justification. Its only purpose is to open up job opportunities for minorities by eliminating adverse impact. The score adjustments are made despite the fact that we know the test scores are predictively fair and unbiased. This is done for social and not scientific reasons, and it is done on a policy and not a scientific basis. There are no unstated but false scientific assumptions because there are no

scientific assumptions at all. That is precisely why race norming is honest and the NAS committee rationale is intellectually dishonest. The NAS report purported to present a scientific justification for race norming.

Gerry recommends what he calls "the engineering approach to affirmative action." This approach involves numerous small steps to try to reduce adverse impact at every point in the selection process. It is hard to disagree with this proposition in the abstract. It is clear that it is possible in some cases to find combinations of selection procedures with equal validity and somewhat different levels of adverse impact. However, it would be easy to get the impression from Gerry's comments that this approach is more effective than it is. I would make the following points:

1. We have tried this approach, and it has not been very successful. We have looked for alternatives with equal validity but less adverse impact without much success for about 20 years.
2. Gerry states that "given a choice between two tests of equal validity," one should choose the test with the least adverse impact. There are two problems here: (1) if they truly have equal validity, they will tend to have very similar adverse impact; and (2) maximization of validity and utility requires that both be

used (not one alone), and the composite will tend to have higher adverse impact than the average of the two.

3. Gerry states that in a certain clerical battery, the racial difference ranges from zero to .75 SD's, so adverse impact can be minimized by careful choice of tests. The problems are these. Combining tests into a battery total increases the reliability, which increases the racial difference. If the maximum racial difference among these 10 tests is only .75 SD's, this indicates that the individual tests are not very reliable to begin with. But any combination of several tests will be more reliable and hence will have more adverse impact.

Time does not permit analysis of all the other assertions and recommendation's Gerry makes for "engineering" reduced adverse impact. The key difficulty is this. On the one hand, no one can argue with the idea we should be on the lookout for ways to simultaneously increase validity and decrease adverse impact. For example, a combination of a biodata scale and a mental ability test for selecting supervisors may have higher validity and somewhat lower adverse impact than the mental ability test alone. On the other hand, Gerry's presentation seems to imply that major reductions in adverse impact are possible through this approach (without validity

losses). I reviewed the literature on this question for an article published in 1988. I do not believe that large reductions in adverse impact are possible (without large validity losses).

The essential problem is that there are real mean differences between groups in job performance capabilities. Improved measurement techniques on the predictor end simply cannot eliminate these differences on the criterion end. Gerry is well aware of this, of course. But some of his statements might imply something different to some listeners. It would be wonderful if adverse impact could be "engineered" away. But it can't.

Jim Outy's paper

As of Wednesday noon, Jim's paper had not arrived. Therefore, I was unable to prepare any comments on it.