PERSONNEL PSYCHOLOGY
1997, 50

EXAMINING SELECTION UTILITY WHERE COMPETING PREDICTORS DIFFER IN ADVERSE IMPACT

CALVIN C. HOFFMAN
Southern California Gas Company

GEORGE C. THORNTON III
Colorado State University

This paper discusses the roles of validity, cut score choice, and adverse impact on selection system utility using data from two concurrent validation studies. We contrast an assessment center and published aptitude test on several metrics, including validity, testing costs, adverse impact, and utility. The assessment center produced slightly lower validity than the aptitude test while costing roughly 10 times as much per candidate. In spite of these advantages for the aptitude test, the assessment center produced so much less adverse impact its operational utility would be higher given cut scores likely to be chosen in this organization. Potential concerns with applying net utility models to this type of situation are discussed in comparison to gross utility models.

The genesis for this article dates to a critical incident the first author experienced in 1989. While completing a concurrent validation study, he met with the client (a director with a marketing background) to discuss options for combining predictors into a battery and to consider possible battery cut scores. Prior to meeting with the client, he had calculated observed validities for several batteries, adverse impact (AI) ratios for possible cut scores on each battery, and gross utility at each cut score. Battery validity received most of his attention; the proposed battery was the one producing the highest validity. After being provided a brief conceptual overview of utility analysis, the client asked why one of the alternate batteries wasn't being considered—the validity coefficient was almost as high, and adverse impact was reduced so much the cut score could be much higher, providing greater utility. This discussion prompted much self-reflection on the wisdom of the expression "can't see the forest for the trees."

Lisa Holden, Kelly Gale, and Nabil Peter originally conducted the two concurrent validation studies upon which this study is based. The authors gratefully acknowledge the insights of Paul Cardenas for his critical incident, the feedback of four anonymous reviewers, and the advice of editor Richard J. Campbell in preparing this manuscript.

Correspondence and requests for reprints should be addressed to Calvin C. Hoffman, Southern California Gas Company, 555 W. 5th St., ML 15H1, Los Angeles, CA 90013-1011.

This critical incident has a number of implications for practice: (a) industrial/organizational psychologists should not be blinded by the quest to maximize r (or R); (b) cut score choice impacts both adverse impact and selection system utility; (c) at least some clients are willing to "give up" some validity in exchange for increased utility and decreased adverse impact; (d) utility analysis can be very useful in a real world context for evaluating competing selection measures; and (e) managers don't necessarily view utility analysis as futile (Latham & Whyte, 1994).

Several years later, the first author was conducting developmental assessment centers (ACs) for exempt employees. As part of our research on the quality of these ACs, we evaluated their criterion-related validity with a long term view of implementing them into future selection systems. Having already experienced the aforementioned critical incident, we were very interested in mean group differences in AC performance. The much smaller Black/White mean score differences we observed in the AC in contrast to the relatively large mean group differences on aptitude tests, led to the formulation of a question: "Given that the AC costs a lot more, and has similar validity, would the smaller Black/White mean group differences allow setting a high enough cut score to provide positive utility over an aptitude measure?"

This study had two main objectives. First and foremost, we explicitly consider adverse impact resulting from choice of predictor and choice of cut score as relevant to evaluating utility outcomes. In performing these utility analyses, it becomes apparent that the traditional ways of applying and interpreting net utility models can be misleading if researchers wish to evaluate competing selection measures, yet also wish to consider different cut scores for each predictor. The need to evaluate competing predictors is especially critical now that more researchers are recommending the use of biodata (Mitchell, 1995) or integrity-based predictors (Ones, 1995) to augment or replace aptitude-based predictors due to their relative lack of adverse impact. In organizations where clients are concerned about AI, as in the critical incident described here, using traditional net utility models to compare competing predictors can potentially be very misleading. A second objective of this study was to extend the AC utility analysis literature by contrasting an AC with a published cognitive ability test.

Previous research on AC utility has included only limited comparisons with other predictors. Several published studies (Burke & Frederick, 1986; Cascio & Ramos, 1986; Cascio & Silbey, 1979; Goldsmith, 1990; Hogan & Zenke, 1986) have contrasted the utility of ACs with interviews. Given the method's relatively widespread use (Arvey & Faley, 1988; Thornton & Byham, 1982) the lack of comparisons with other predictors is surprising. The AC method has demonstrated reasonably good

validity (ρ = .37) in a meta-analysis conducted by Gaugler, Rosenthal, Thornton, and Bentson (1987) and has usually demonstrated reduced adverse impact relative to aptitude tests. We first discuss utility models, briefly review the AC utility literature, then discuss group differences in the AC literature.

Utility

In this study, the Brogden-Cronbach-Gleser (BCG) gross utility model was used to evaluate two competing systems, allowing for different cut scores. This was important because we believed the AC would produce much less adverse impact than the aptitude test, allowing for higher AC cut scores before finding adverse impact. The net utility model allows comparisons between two competing predictors, but requires that the same cut score be used for both predictors being considered.

Although the BCG utility model has existed for many years, it was little used until more convenient methods for estimating the standard deviation of dollar-valued performance (SD_y) were developed, beginning with the research reported by Schmidt, Hunter, McKenzie, and Muldrow (1979). (See Cascio, 1987, or Raju & Burke, 1986, for more detail on SD_y estimation methods). Two variations of the BCG model allow comparing a selection system either to random selection (gross utility) or to a competing selection system (net utility; Cascio, 1987). Both versions allow researchers to include expected tenure and number of candidates selected when making utility estimates.

Assessment Center Utility Research

Cascio and Silbey (1979) examined assessment center utility, comparing hypothetical assessment center (AC) and interview-based selection processes. They systematically varied six parameters: AC validity, validity of competing procedure (interview), selection ratio, SD_y, costs, and number of centers. They found positive utility for all cases where AC validities were greater than assumed interview validity (.25). Of the variables examined, size of SD_y had the greatest impact on utility. Significantly, they found AC costs played a relatively small role in utility outcome. They concluded that although the AC was much more expensive than the interview, its higher validity provided greater utility. Cascio (1987) emphasized the interplay of several variables in the utility equation. These included validity, standard deviation of dollar-valued performance, predictor cutoff, selection ratio, and testing costs, but did not

consider the effects of different levels of adverse impact resulting from competing assessment methods.

Burke and Frederick (1986) contrasted an interview and AC for selecting sales managers using Boudreau's (1983) utility model. AC validity data came from a single criterion-related validation study while interview validity was estimated. Their paper focused mainly on several alternative methods for calculating SD_y, including Schmidt et al.'s (1979) point estimate, two versions of Burke and Frederick's (1984) modification of the Schmidt et al. procedure, and Schmidt and Hunter's (1982) 40% and 70% of mean salary estimate. Reported SD_y estimates varied widely, ranging from $12,789 to $38,333, with the 40% and 70% estimates being the most conservative. The estimated AC true validity was .59 and the selection ratio was .22. Net utility per selectee ranged from $3,991 to $21,222; these values incorporated an assumed 4-year tenure for selectees. Putting these estimates on a 1-year term, net per selectee utility ranged from $998 to $5306.

Cascio and Ramos (1986) contrasted an operational AC and interview on validity, testing costs, and utility. The AC replaced an earlier interview-based selection procedure for screening first-level supervisors. AC observed validity was .19; corrected for attenuation and range restriction, this value increased to .39. Their estimated interview validity was .13 and costs were estimated at $300 per applicant. SD_y was estimated at $10,250 using the Cascio-Ramos estimate of performance in dollars (CREPID; Cascio & Ramos, 1986). Their utility analysis applied the BCG net utility model, assuming a selection ratio of .32 for both interview and AC, and 4.4 year tenure for selectees. The AC net utility over the interview was $11,776 (over a 4.4-year tenure) or $2,676 per selectee per year.

Previous AC utility studies have several limitations. First, ACs were compared only with interviews, not with cognitive ability tests. Second, interview validity has been estimated, and those estimates were set relatively low (i.e., .13) in contrast to the higher interview validity estimates reported in recent meta-analyses (Harris, 1989; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Weisner & Cronshaw, 1988). Third, the relative difference in validity for ACs versus interviews is probably an over-estimate (ρ_{AC} = .37, Gaugler et al., 1987; $\rho_{interview}$ = .47, Weisner & Cronshaw, 1988). AC versus interview utility comparisons have not considered the extent that interview validity seems to be moderated by interview structure. Fourth, the studies did not consider differences in adverse impact between the selection methods. There is reason to believe ACs may produce less adverse impact than do aptitude tests. It is not clear at this time how ACs and structured interviews might compare on relative adverse impact.

Adverse Impact of ACs

AC performance of Blacks and Whites has been compared in numerous studies. Typically, only mean overall assessment ratings are reported. Results are fairly evenly split between studies showing no significant differences in average performance of Blacks and Whites (Byham, 1983; Firefighters, 1980; Jaffee, Cohen, & Cherry, 1972; Marquardt, 1976; G. Russell, 1975; Schaab & Byham, 1983; Wentworth, 1983) and studies showing Blacks scoring lower on average than Whites (Clingingpeel, 1979; Friedman, 1980; Huck & Bray, 1976; Moses, 1973; P. Russell & Byham, 1980), usually by less than one standard deviation. These latter studies must be interpreted cautiously because the samples of each ethnic group may not be equally representative of their respective populations. In many cases, broader samples of Blacks were selected for AC attendance in hopes of identifying a larger pool of qualified minority group candidates.

This Study

This study offers several advantages over previous research. Previous AC utility research has only contrasted ACs with unstructured interviews. The Cascio and Silbey (1979) paper relied on purely hypothetical data for both the AC and the interview, while both the Burke and Frederick (1986) and Cascio and Ramos (1986) papers calculated AC validity while estimating interview validity. Previous AC versus interview utility analyses could be viewed as an unfair comparison because unstructured interviews are relatively less valid than ACs, whereas structured interviews may be more valid than ACs. Empirical validity estimates used in this study are slightly lower for the AC than the aptitude test. Also, the AC is much more expensive than the aptitude test it is contrasted against.

Cascio and Silbey (1979) noted that the net per selectee payoff for the AC was always negative when AC validity was lower than the alternate selection procedure because AC costs were greater. The net utility model compares competing selection systems on validity and testing costs, but requires using the same cut score for both systems. Because cut score choice impacts both adverse impact and utility, using the net utility model with identical cut scores (and ignoring adverse impact) would always make the AC (or any other predictor) "lose" in comparison with a less expensive measure having the same or higher validity.

This paper explicitly considers decreased adverse impact as a reasonable goal of applied selection research, which is fully consistent with the Uniform Guidelines (1978) admonition to search for alternative predictors with similar validity and less adverse impact. We hypothesized the

AC would produce so much less adverse impact than the aptitude test that it would allow setting sufficiently higher cut scores to yield utility greater than that produced by the test. The utility analyses reported below use the BCG gross utility equation to compare both AC and aptitude test utility with random selection. This allowed us to apply a series of possible cut scores and determine the utility at the point where adverse impact appeared for each system.

Method

Problem and Setting

This study was conducted in a large utility company. Data came from two concurrent validity studies with nonoverlapping samples. One study utilized results from a developmental AC while the other used results from an aptitude test to be used in selecting nonexempt employees into exempt supervisory and nonsupervisory job classifications. AC validation was done purely for research purposes. Criterion contamination was ruled out because criterion data (file appraisals) already existed when the assessment was conducted, but were not available to assessors. Aptitude test data came from a concurrent validation study conducted prior to its use in a promotional selection program.

Aptitude Test

The aptitude test was the Professional Employment Test (PET; Psychological Services, Inc., 1987) short form, a 20-item multiple choice test. The PET was administered for research purposes to a sample of 104 employees (69 Whites, 5 Asian Americans, 9 Blacks, 1 American Indian and 20 Hispanics). Of 104 subjects completing the test, 60 had previously been promoted from nonexempt job classifications. The remaining 44 subjects were still in nonexempt classifications. Adverse impact calculations reported below were based on the full sample of 104.

Assessment Center

AC data were from 118 second-line supervisors attending a 1-day "development-only" program. Most assessees had previously been promoted from nonexempt classifications through an earlier version of the promotional program described above (i.e., different cognitive ability measures), but prior to the company's implementation of a formal pass/fail evaluation of test performance. The AC sample included 85 Whites, 10 Asian Americans, 6 Blacks and 17 Hispanics. The AC included four

exercises: in-basket, budget analysis, employee counseling and leaderless group discussion. For this study an average AC score was calculated across all dimension ratings in all exercises. No overall assessment rating was provided by assessors.

Criterion

Both studies used each individual's current administrative performance appraisal on file as a validation criterion. Appraisal data were slightly different because the AC and PET projects were conducted about 2 years apart. In the interim the performance appraisal instrument was changed. The performance appraisal for the AC was the current system (5-point global scale). The performance appraisal the PET was validated against was an earlier instrument (7-point global scale). In both cases, performance appraisals were completed by the immediate supervisor to gauge accomplishment of performance objectives and to guide salary treatment. Time lag between collection of predictor and criterion data was similar in both samples (1 year or less).

Analyses

Statistical analyses included correlating AC or PET scores with the criterion, calculating gross utility for each selection system over a range of possible cut scores, and calculating adverse impact (AI) for the same cut scores. Adverse impact was calculated using the four-fifths rule of thumb (Uniform Guidelines, 1978; i.e., we determined the point in the predictor distribution where the non-White selection rate would be less than 80% of the white selection rate). Because both samples were relatively small (AC $n = 118$; PET $n = 104$), minority group representation was insufficient to calculate meaningful adverse impact ratios separately for Asian Americans, Blacks, or Hispanics versus Whites. Therefore, AI was calculated based on White/non-White pass rates for each assumed cut score. This probably underestimated AI for certain groups, especially on the PET, because Blacks typically perform about one standard deviation lower than Whites on cognitive ability tests, Hispanics score about one half standard deviation lower than Whites, while Asian Americans typically perform about as well as Whites (Gottfredson, 1988).

The Brogden-Cronbach-Gleser gross utility model was applied to each predictor, assuming Ø (selection ratio) values varied from .9 to .1 in .1 increments. The formula used was:

$$\Delta_u = r_{xy} SD_y \lambda / \text{Ø} - C/\text{Ø}$$

where Δ_u is gross utility, r_{xy} is validity, SD_y is the dollar-valued criterion, λ is the ordinate of the normal curve associated with the cut score, C is testing costs, and Ø is the selection ratio. The value $\lambda/Ø$ is tabled and available in Cascio (1987).

Several assumptions were made: PET testing costs (C) were $50 per candidate, two AC testing cost estimates (C) were evaluated ($500 and $1,000), and SD_y (standard deviation of dollar-valued performance) was $20,000 using the Schmidt and Hunter (1982) 40% rule (i.e., 40% of a $50,000 salary). This is a conservative estimate because employees in these grades are now paid $60,000 to $65,000 per year. Less conservative SD_y estimates would range between $24,000 and $26,000 using the 40% rule.

Results

Table 1 provides means and standard deviations for AC dimensions, overall AC score, PET score, and respective criteria. AC observed validity was .26 ($df = 116; p < .01$ one-tailed), while PET observed validity was .30 ($df = 58; p < .05$ one-tailed). Correcting these values for attenuation due to unreliability in the criterion (.60; Pearlman, Schmidt & Hunter, 1980) provided estimated corrected validities of .34 (AC) and .39 (PET). This corrected AC validity is similar to the .37 true value reported by Gaugler et al. (1987) in their meta-analysis of AC validity. The validity estimate for the PET is lower than the corrected validity for general cognitive ability found in one meta-analytic study (Pearlman, Schmidt, & Hunter, 1980). Although the SIOP *Principles* (1987) endorse correcting observed validity coefficients for both attenuation and range restriction, the range restriction correction was not applied here for several reasons. First, the AC was a developmental center, so there was no direct selection impacting scores of attendees; second, because the AC was custom developed, there are no published norms available to guide range restriction corrections (Hoffman, 1995; Sackett & Ostgaard, 1994); third, this correction was not applied to the PET because there was no range restriction in this sample; the *U* value commonly used to evaluate range restriction (Thorndike, 1949) was 1.01. The net result of not applying this correction is that the AC and PET validities applied in the utility analyses are probably both slightly conservative.

Utility calculations were conducted assuming both $500 and $1,000 AC testing costs. Results for $500 costs are hereafter reported as $AC_{\$500}$, an analogous format is used for results assuming $1,000 costs. Gross utility for $AC_{\$500}$ ranged from $770 (10th percentile cut score) to $6,934 (90th percentile cut score). $AC_{\$1,000}$ gross utility ranged from $215 (10th percentile cut score) to $1,934 (90th percentile cut score). Under

TABLE 1

Predictor and Criterion Descriptive Statistics

Variable	*M*	*SD*
Assessment center (n = 118)		
Adaptability	3.5	0.6
Analysis	3.3	0.6
Presentation	3.6	1.0
Decision making	3.4	0.6
Delegation	3.1	0.7
Interpersonal relations	3.6	0.5
Leadership	3.5	0.8
Oral communications	3.6	0.6
Planning	3.0	0.7
Writing	3.6	0.7
AC dimension average	3.4	0.5
AC average (Whites)	3.4	0.5
AC average (non-Whites)	3.3	0.5
Criterion[a]	3.4	0.7
Whites	3.5	0.8
Non-Whites	3.2	0.6
Aptitude test (n = 60)		
Prof. employment test	13.0	4.3
PET (Whites)	14.3	3.9
PET (non-Whites)	11.3	3.9
Criterion[b]	4.7	0.7

[a] Global rating on 5-point scale; n = 118.
[b] Global rating on 7-point scale; n = 60.

either cost scenario, AC utility peaked and then decreased at the 90th percentile, although the drop for $AC_{\$500}$ was relatively small ($120 or 1.7%) compared to the drop for $AC_{\$1,000}$ ($2,620 or 57.5%). PET gross utility ranged from $1,465 (10th percentile cut score) to $13,189 (90th percentile cut score). All utility estimates are for 1 year; tenure was not included in the utility equation to provide conservative estimates.

The AC produced AI at the 60th percentile while the PET produced AI at the 20th percentile. Gross utility per selectee for these respective cut scores is $5,312 ($AC_{\$500}$), $4,062 ($AC_{\$1,000}$), and $2,668 (PET). Table 2 presents these findings in more detail. Because aptitude test validity was somewhat lower than the validity generalization literature might lead one to expect, we also calculated the validity needed for PET utility to match AC utility, assuming the PETs 20th percentile cut score was maintained. The PET would have to provide a validity of .5982 with a 20th percentile cut score to match $AC_{\$1,000}$ utility with a 60th percentile cut score, and a validity of .7678 to match $AC_{\$500}$ utility at the same cut score.

TABLE 2

Gross Per Selectee Utility in Dollars (One Year) for Assessment Center and Aptitude Test

Selection ratio	Cut score percentile	Assessment center cost $500	Assessment center cost $1,000	Aptitude test
.1	.9	6,934	1,934	13,189
.2	.8	7,054	4,554	10,707
.3	.7	6,221	4,555	8,881
.4	.6	5,312[a]	4,062[a]	7,402
.5	.5	4,440	3,440	6,140
.6	.4	3,353	2,719	4,948
.7	.3	2,652	1,937	3,790
.8	.2	1,755	1,130	2,668[b]
.9	.1	770	215	1,465

[a] Point where AC produced adverse impact.
[b] Point where PET produced adverse impact.

Discussion

The AC produces higher utility than the aptitude test when cut scores on each are set so as to eliminate adverse impact, even though the AC has slightly lower validity and costs considerably more. Gross utilities at the highest hypothetical PET or AC cut scores evaluated here are unattainable in this organization due to unacceptable levels of adverse impact produced by either predictor. It is likely that some organizations would be unwilling to use either selection system with cut scores passing only 10 or 20% of all candidates. At the 60th percentile, where the AC just started to produce AI, no Blacks in this admittedly small sample would pass the PET. This outcome might be unacceptable in some organizations. Had AC and PET utility been examined via the net utility equation, most psychologists would not consider using the AC for selection—it has lower validity, costs more, and always produces lower utility at comparable cut scores. The real issue is that cut scores need not be comparable because the AC produced so much less adverse impact.

Burke and Frederick (1986) examined AC utility resulting from top-down selection versus use of a cut score and documented a utility loss for using a cut score; they assumed that any cut score chosen would provide a lower mean AC score than that achieved under a top-down strategy. Because Burke and Frederick applied a much higher empirical AC validity estimate (.59) than used here, their utility estimates do not asymptote as in this study. A pure top-down strategy could conceivably produce lower utility than through use of a cut score when selection system costs are very high (as in the $AC_{\$1,000}$ scenario examined here). On a related note, organizations using an AC with a cut score might simply stop processing candidates when a sufficient number of candidates

have passed the assessment with the required score. If a pure top-down approach were used, more candidates might be processed than are really needed, driving selection system costs up, because there is no way to know whether enough candidates attaining an acceptable score have been processed to meet organizational needs.

Burke and Frederick (1986) had this to say regarding the potential benefits of using a cut score rather than a pure top-down approach "These utility estimates, however, do not reflect the gains in social utility resulting from current hiring/promotion practices that attempt to increase minority and female representation in an organization's workforce" (p. 338). The present study addresses this issue, and demonstrates methods for evaluating the utility and adverse impact resulting from varying cut scores.

The adverse impact analyses discussed here represent a relatively simplified scenario where Whites were contrasted against a combined minority sample, and where cut scores were chosen to eliminate adverse impact. In an applied setting, the practitioner would have to consider issues like how many applicants or candidates were available in each ethnic group, whether those numbers reasonably reflect the representation of that group in the geographical area from which applicants are drawn, the degree of AI the hiring organization finds acceptable, the type of job being filled, and the degree of lost productivity the organization is willing to accept. The assumption in this study that AI should be eliminated is unrealistic and unnecessary if the employer has successfully marshaled evidence supporting inferences about a predictor's validity. If a predictor produces no AI, there is probably no legal reason to validate it in the first place (Uniform Guidelines, 1978), although there are still good business reasons to examine validity (and utility).

In many applied settings, the researcher is likely to have larger samples than were available in this study, making sub-group analyses technically feasible. Table 3 presents hypothetical data that contrasts percentiles of possible cut scores, selection ratios, adverse impact ratios (AIR) for Blacks, Hispanics, and Asians versus Whites, and gross per-selectee utility for each cut score. The first author has used this format on several occasions to inform clients regarding the tradeoffs involved in cut score choice, AI, and utility. Examining this table reveals several things: (a) AIRs vary widely between protected groups; (b) a cut score chosen to eliminate AI for Asians (i.e., 70th percentile) would still produce AI against Blacks and Hispanics; (c) eliminating AI against all groups would require setting a 10th percentile cut score, producing an unacceptable loss of utility. Schmidt, Mack, and Hunter (1984) found that setting minimum qualification levels near the 15th percentile re-

TABLE 3

Contrasting Selection Ratio, Cut Score Percentile, Adverse Impact and Gross Utility[a] (Hypothetical Data)

Selection ratio	Cut score percentile	Adverse impact ratio[b] B/W	H/W	A/W	Gross utility
.1	.9	14	28	67	13,189
.2	.8	19	48	69	10,707
.3	.7	22	51	85	8,881
.4	.6	28	61	86	7,402
.5	.5	41	73	89	6,140
.6	.4	50	80	90	4,948
.7	.3	61	86	92	3,790
.8	.2	73	90	93	2,668[b]
.9	.1	83	96	97	1,465

[a] Utility calculations assume validity of .37, SD_y of $20,000, and testing costs of $500;

[b] Adverse impact ratios calculated using percentage White passing as denominator and percentage Black, Hispanic, or Asian passing as numerator.

sulted in productivity losses of about 80 to 90% as large as if valid selection measures were abandoned.

The choice of which cut score to apply remains a qualitative judgment as always, and rests on issues like "how valid" the test is, how extreme the level of AI, the type of organization involved, labor market conditions, and the expected productivity gains or losses associated with different cut scores. Factoring AI into the cut score decision is a fairly practical consideration given that at least one plaintiff's attorney has indicated that level of validity versus level of AI factors heavily into his decision of whether or not to pursue a case (Seymour, 1988). Setting a cut score based on utility estimates as suggested here does not involve treating subgroups differently, such as applying different cut scores by ethnic group or developing race-based norms. Whereas the suggested practice might be inappropriate if routinely applied after inspecting applicant data, it is more defensible after analyzing validation evidence but prior to making actual selection decisions (i.e., before implementing the selection system).

Schmidt (1988) discussed group differences in selection, examining research on differential validity, fairness models, validity generalization, and efforts to remove adverse impact from selection instruments. He noted that efforts to develop valid predictors without adverse impact have not been successful. Although it may not be feasible to eliminate adverse impact, Schmidt (1988) held out a more attainable goal:

> [I]t should be possible to identify non-test procedures that will add to the validity of general ability and at the same time reduce adverse impact. Although research along these lines cannot realistically be expected to eliminate adverse impact, it may have the potential for producing selection

> systems with *reduced* adverse impact and *enhanced* validity [emphasis in original] (p. 284.)

Future selection research should examine Schmidt's (1988) proposition regarding combining aptitude-based measures with alternative predictors, such as structured interviews, ACs, biodata or personality inventories as a means of increasing validity, decreasing AI, and increasing utility. The findings of this study suggest that even if no gain in validity is realized, use of higher cut scores, which are probably more acceptable with decreased adverse impact, could result in greatly enhanced utility.

Using both cognitive and noncognitive predictors in a selection process raises a number of questions regarding the proper methods for combining such measures to guide decision making. Can the argument be made that high scores on a personality inventory or biodata measure compensate for low scores on a measure of general cognitive ability? If not, then such measures should probably be applied in a multiple hurdle approach rather than in a compensatory model. On a related note, if predictors are used in a multiple hurdle approach, this might not result in any decrement in AI unless the practitioner chose to use a lower cut score on the measure with higher adverse impact (presumably a cognitive ability measure).

Additional research is needed to guide practitioners regarding which combinations of predictors can be reasonably combined into a compensatory battery, and which predictors should be treated as noncompensatory. Sackett and Roth (1996) recently investigated the effects of using predictors in either single-stage or multi-stage selection systems on predicted performance and minority representation. They found that the preferred selection strategy depends on the relative value an organization places on performance and minority representation.

This study has several limitations which should not be overlooked. First, AC and aptitude test data came from independent, nonoverlapping samples. Because we cannot directly calculate the correlation between the AC and the PET, the multiple correlation of both predictors with the criterion cannot be directly calculated, nor can we estimate the adverse impact either selection procedure might generate on a sample taking both measures. Second, both samples in this study were relatively small, although this is somewhat mitigated by the fact that the AC produced a validity coefficient similar to that reported in meta-analytic research, while the PET produced an observed validity similar to those reported in the publisher's manual. A third limitation is that the AC, although valid, was an operating developmental center. It is not clear if this center would produce similar outcomes (validity and adverse impact) if it were conducted in a selection context. Even if the AC pro-

duced more adverse impact in a selection context, its advantage over the PET is so great it is unlikely our utility conclusions would change.

Utility analyses as conducted here expand our understanding of how selection systems work as systems. Reviewing validity evidence without considering adverse impact would lead to conclusions other than those of this study. In particular, the finding that a selection procedure with lower validity and higher costs can produce higher utility when considering real world constraints like cut score choice is probably counterintuitive to many practitioners and academicians.

This study is not meant to provide a definitive examination of the relative amount of AI produced by "the assessment center"—obviously, ACs vary greatly in terms of their length, complexity, speededness, and so forth. Rather, our intent is to illustrate how utility analysis can help evaluate "what if" scenarios, and to encourage other practitioners to consider the validity, cut score, adverse impact, and utility tradeoff we describe. Although utility research is potentially useful for choosing between competing selection system options (Guion, 1991), Guion and Gibson (1988) suggested that as usually conducted, most utility comparisons simply favor the most valid predictor. Guion (1991) discussed the need to expand utility research to consider multiple outcomes or tradeoffs. Our considering adverse impact resulting from cut score choice is a step in that direction.

REFERENCES

Arvey RD, Faley RH. (1988). *Fairness in selecting employees,* (2nd ed.). Addison-Wesley.

Boudreau JW. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. PERSONNEL PSYCHOLOGY, *36,* 551–576.

Burke MJ, Frederick JT. (1984). Two modified procedures for estimating standard deviations in utility analysis. *Journal of Applied Psychology, 69,* 482–489.

Burke MJ, Frederick JT. (1986). A comparison of economic utility estimates for alternative SD_y estimation procedures. *Journal of Applied Psychology, 71,* 334–339.

Byham W. (1983). *Impact of sex and race variables on a middle management assessment center.* Pittsburgh, PA: Development Dimensions, Inc.

Cascio WF. (1987). *Costing human resources: The financial impact of behavior in organizations.* Boston: Kent.

Cascio WF, Ramos RA. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. *Journal of Applied Psychology, 71,* 20–28.

Cascio WF, Silbey V. (1979). Utility of the assessment center as a selection device. *Journal of Applied Psychology, 64,* 107–118.

Clingingpeel R. (1979). *Validity and dynamics of a foreman selection process.* Paper presented at the 7th International Congress on the Assessment Center Method, New Orleans, LA.

Firefighters Institute for Racial Equality v. City of St. Louis, U.S. Court of Appeals, Eighth Circuit (nos. 79-1435 and 79-1461), January 21, 1980.

Friedman M. (1980). *Differences in assessment center ratings as a function of the race and geographic location of the assessee.* Internal technical report: Tennessee Valley Authority.

Gaugler BB, Rosenthal DB, Thornton GC III, Bentson C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Goldsmith RF. (1990). Utility analysis and its application to the study of the cost-effectiveness of the assessment center method. In Murphy KR, Saal FE (Eds.), *Psychology in organizations: Integrating science and practice* (pp. 95–110). Hillsdale, NJ: Lawrence Erlbaum.

Gottfredson LS. (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior, 33,* 293–319.

Guion RF. (1991). Personnel assessment, selection and placement. In Dunnette MD, Hough LM (Eds.) *Handbook of industrial & organizational psychology,* (2nd ed.) (pp. 329–397). Palo Alto, CA: Consulting Psychologists Press.

Guion RF, Gibson WM. (1988). Personnel selection and placement. *Annual Review of Psychology, 39,* 349–374.

Harris MM. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. PERSONNEL PSYCHOLOGY, *42,* 691–726.

Hoffman CC. (1995). Applying range restriction corrections using published norms: Three case studies. PERSONNEL PSYCHOLOGY, *48,* 913–923.

Hogan J, Zenke LL. (1986). Dollar-value utility of alternative procedures for selecting school principals. *Educational and Psychological Measurement, 46,* 935–945.

Huck JR, Bray D. (1976). Management assessment center evaluations and subsequent job performance of black and white females. PERSONNEL PSYCHOLOGY, *29,* 13–30.

Jaffee C, Cohen S, Cherry R. (1972). Supervisory selection program for disadvantaged or minority employees. *Training and Development Journal, 26,* 22–28.

Latham GP, Whyte G. (1994). The futility of utility analysis. PERSONNEL PSYCHOLOGY, *47,* 31–46.

Marquardt L. (1976). *Follow-up evaluation of the second-look approach to the selection of management trainees.* Chicago: Psychological Research and Services, National Personnel Department, Sears, Roebuck and Company.

McDaniel MA, Whetzel DL, Schmidt FL, Maurer SD. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

Mitchell T. (1995, October). *Biodata: Innovations and social issues for a non-traditional selection technique.* Presented at the Fall Conference of the Personnel Testing Council of Southern California, Newport Beach.

Moses J. (1973). The development of an assessment center for the early identification of supervisory potential. PERSONNEL PSYCHOLOGY, *26,* 569–580.

Ones DS. (1995, November). *Using pre-employment tests for personnel selection.* Paper presented to the Personnel Testing Council of Southern California, Los Angeles.

Pearlman K, Schmidt FL, Hunter JE. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Psychological Services, Inc. (1987). *Professional employment test.* Glendale, CA: Test Publications Divisions, Author.

Raju NS, Burke MJ. (1986). Utility analysis. In Berk RA (Ed.), *Performance assessment methods and applications* (pp. 186–202). Baltimore: Johns Hopkins University Press.

Russell G. (1975). Differences in minority/nonminority assessment center ratings. *Assessment and Development, 3,* 3–7.

Russell P, Byham W. (1980). *Reliability and validity of assessment in a small manufacturing company.* Pittsburgh: Development Dimension, Inc.

Sackett PR, Ostgaard DI. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79,* 680–684.

Sackett PR, Roth L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. PERSONNEL PSYCHOLOGY, *49,* 549–572.

Schaab N, Byham W. (1983). *Evaluating management schools by results achieved.* Pittsburgh: Development Dimensions, Inc.

Schmidt FL. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33,* 272–292.

Schmidt FL, Hunter JE. (1982). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68,* 407–414.

Schmidt FL, Mack MJ, Hunter JE. (1984). Selection utility in the occupation of U.S. Park Ranger for three modes of test use. *Journal of Applied Psychology, 69,* 490–497.

Schmidt FL, Hunter JE, McKenzie RC, Muldrow TW. (1979). Impact of valid selection procedures on workforce productivity. *Journal of Applied Psychology, 35,* 333–347.

Seymour RT. (1988). Why plaintiff's counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." *Journal of Vocational Behavior, 33,* 331–364.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures (3rd ed.).* College Park, MD: Author.

Thorndike RL. (1949). *Personnel selection: Test and measurement techniques.* New York: Wiley.

Thornton GC III, Byham WC. (1982). *Assessment centers in managerial performance.* New York: Academic Press.

Uniform Guidelines on Employee Selection Procedures (1978). *Federal Register, 43,* 38290–38315.

Weisner WH, Cronshaw SF. (1988). A meta-analytical investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Applied Psychology, 61,* 275–290.

Wentworth P. (1983). *An analysis of behavioral ratings within a decision-theoretic framework.* Paper presented at the 11th International Congress on the Assessment Center Method, Williamsburg, VA.