View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institutional Knowledge at Singapore Management University

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection Lee Kong Chian School Of Business

Lee Kong Chian School of Business

1-2009

Decision aids for addressing the validity-adverse impact trade-off

Paul R. SACKETT

Wilfried DE CORTE

Filip LIEVENS

Singapore Management University, filiplievens@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/lkcsb_research

Part of the Human Resources Management Commons, and the Organizational Behavior and Theory Commons

Citation

SACKETT, Paul R.; DE CORTE, Wilfried; and LIEVENS, Filip. Decision aids for addressing the validity-adverse impact trade-off. (2009). *Adverse Impact: Implications for Organizational Staffing and High Stakes Selection*. 459-478. Research Collection Lee Kong Chian School Of Business.

Available at: https://ink.library.smu.edu.sg/lkcsb_research/5821

This Book Chapter is brought to you for free and open access by the Lee Kong Chian School of Business at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection Lee Kong Chian School Of Business by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Sackett, P.R., De Corte, W., & Lievens, F. (2009). Decision aids for addressing the validity-adverse impact tradeoff (pp.453-472). In J. Outtz (Ed.) Adverse Impact: Implications for Organizational Staffing and High Stakes Selection. SIOP Frontiers Series, Lawrence Erlbaum.

17

Decision Aids for Addressing the Validity-Adverse Impact Trade-Off

Paul R. Sackett, Wilfried De Corte, and Filip Lievens

Introduction

Typically, adverse impact (AI) is an after-the-fact analysis: Once predictor scores for a pool of applicants are available, AI is evaluated. Sometimes the analysis is made in real time, as predictor scores are obtained on a set of applicants, and AI calculations are done on a "what if" basis as input to decisions about features such as where to set a cutoff score. The focus of this chapter, however, is on attempts to estimate in advance the likely impact of a given selection system. Here, estimates are made based on available information about the features such as the expected magnitude of subgroup differences, expected interpredictor correlations, and expected predictor-criterion correlations. Such information may be local (e.g., group differences observed the last time a predictor was used) or based on a more general research literature (e.g., group differences reported in publisher manuals or in the published literature for a given predictor type and a given job category).

These projections of AI and other outcomes are generally made in one of two ways. The first is via simulation, in which multiple samples of data are generated from populations with specified parameters (e.g., means, standard deviations [*SD*s], interpredictor *r*s, subgroup differences). Indices of interest (e.g., AI ratios [AIRs], proportion of positions filled by minority group members) are computed for each sample, and the distributions of these indices are tallied and examined. The second is via analytic solution, in which the outcomes of interest can be determined precisely via equation. For example, while one can determine the expected value of an

Y100250_C017.indd 459 3/20/09 3:16:19 PM

AIR obtained if a selection device with a *d* of 1.0 *SD* is used with a selection ratio (SR) of 10% by drawing repeated random samples from a normal distribution, one can determine this more directly via the equation for the area under a normal curve. Simulations are more useful in settings for which an analytic solution is not available.

We use the standardized mean difference *d* as the index of group differences. This is the majority mean minus the minority mean divided by the pooled within-group standard deviation. This index expresses the group difference in standard deviation units, with zero indicating no difference, a positive value indicating a higher mean for the majority group, and a negative value indicating a higher mean for the minority group.

In this chapter, we summarize a number of decision aids for AI planning. These design tools address a range of applied questions. They fall into two major categories. The first involves those that focus solely on AI as an outcome. While these are useful for understanding the likely AI in a specific selection setting, they are silent regarding the consequences for other outcomes of attempts to reduce AI. The second involves those that focus on both AI and other outcomes, with the mean criterion performance of those selected as the most common additional outcome. Studies in this second category permit examining trade-offs between AI and mean criterion performance (e.g., documenting the performance consequences of setting a low cutoff score). In the remainder, we examine each category in turn.

Category 1: Approaches That Focus Solely on Adverse Impact

AI as a Function of *d* and SR

A basic starting point for insight into AI is a clear understanding of the major components that contribute to it. If top-down selection on a given score distribution (which may be a single predictor or a composite of multiple predictors) is used, and if normality assumptions are met, the expected value of the AIR is a function of the standardized mean difference *d* between the two groups of interest and the SR. The relationship among *d*, SR, and AI can then be derived from properties of the normal distribution. Tables showing this relationship were presented by Sackett and Wilk (1994) and expanded to a broader range of SRs by Sackett, Schmitt, Ellingson, and Kabin (2001). They presented separate tables for the effects of *d* and majority group SR on two outcomes: the minority group SR and the AIR. Table 17.1 integrates this information into a single table.

TABLE 17.1

Minority Group Selection Ratios and Four-Fifths Ratios When the Majority Group Selection Ratio Is 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, or 99%

Standardized group difference (*d*)	Majority group selection ratio[a]								
	1%	5%	10%	25%	50%	75%	90%	95%	99%
0.0	.010	.050	.100	.250	.500	.750	.900	.950	.990
	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
0.1	.008	.041	.084	.221	.460	.716	.881	.938	.987
	.80	**.82**	**.84**	**.88**	**.92**	**.95**	**.98**	**.99**	**.99**
0.2	.006	.033	.069	.192	.421	.681	.860	.925	.983
	.60	**.66**	**.69**	**.77**	**.84**	**.91**	**.96**	**.97**	**.99**
0.3	.004	.026	.057	.166	.382	.644	.837	.910	.978
	.40	**.52**	**.57**	**.66**	**.76**	**.86**	**.93**	**.96**	**.99**
0.4	.003	.021	.046	.142	.345	.606	.811	.893	.973
	.30	**.42**	**.46**	**.57**	**.69**	**.81**	**.90**	**.94**	**.98**
0.5	.002	.016	.038	.121	.309	.568	.782	.873	.966
	.20	**.32**	**.38**	**.48**	**.62**	**.76**	**.87**	**.92**	**.98**
0.6	.002	.013	.030	.102	.274	.528	.752	.851	.957
	.20	**.26**	**.30**	**.41**	**.55**	**.70**	**.84**	**.90**	**.97**
0.7	.001	.010	.024	.085	.242	.488	.719	.826	.947
	.10	**.20**	**.24**	**.34**	**.48**	**.65**	**.80**	**.87**	**.96**
0.8	.001	.007	.019	.071	.212	.448	.684	.800	.936
	.10	**.14**	**.19**	**.28**	**.42**	**.60**	**.76**	**.84**	**.95**
0.9	.001	.006	.015	.058	.184	.409	.648	.770	.922
	.10	**.12**	**.15**	**.23**	**.37**	**.54**	**.72**	**.81**	**.93**
1.0	.000	.004	.011	.047	.159	.371	.610	.739	.907
	.00	**.08**	**.11**	**.19**	**.32**	**.49**	**.68**	**.78**	**.92**
1.1	.000	.003	.009	.038	.136	.334	.571	.705	.889
	.00	**.06**	**.09**	**.15**	**.27**	**.45**	**.63**	**.74**	**.90**
1.2	.000	.002	.007	.031	.115	.298	.532	.670	.869
	.00	**.04**	**.07**	**.12**	**.23**	**.40**	**.59**	**.71**	**.88**
1.3	.000	.002	.005	.024	.097	.264	.492	.633	.846
	.00	**.04**	**.05**	**.10**	**.19**	**.35**	**.55**	**.67**	**.85**
1.4	.000	.001	.004	.019	.081	.233	.452	.595	.821
	.00	**.02**	**.04**	**.08**	**.16**	**.31**	**.50**	**.63**	**.83**
1.5	.000	.001	.003	.015	.067	.203	.413	.556	.794
	.00	**.02**	**.03**	**.06**	**.13**	**.27**	**.46**	**.59**	**.80**

Add leading zeros in table if appropriate.

[a] Selection ratio = number of applicants hired/number of applicants applied. Per cell, two values are given. The first value refers to the minority group selection ratio. The second value in bold represents the four-fifths ratio (i.e., the minority group selection ratio/majority group selection ratio). Tabled values in bold less than .80 represent scenarios that violate the four-fifths rule.

This table illustrates a variety of useful general principles. First, at a given *d*, the AIR increases as SR increases. This is certainly a well-known result, but the table is useful in making clear the magnitude of this effect. For example, with *d* = .5, the AIR ranges from .20 at a majority SR of 1% to .62 at a majority SR of 50% to .98 at a majority SR of 99%. Of course, as SR approaches 100%, subgroup SRs must converge, and the AIR must approach 1.0. Second, at a given SR, the AIR increases as *d* increases. This is also a well-known result; again, the table is useful in making clear the magnitude of the relationship. Third, the table illustrates the combination of SRs and *d* values that results in a violation of the four-fifths rule. For small *d* values (e.g., .1 to .2), the four-fifths rule is violated only when SR is less than 50%. For *d* values larger than .5, the four-fifths rule will be violated unless SR is very large, typically 90% or higher.

This decision aid can help project the likely effects of using a particular predictor with a particular SR. It permits addressing questions such as, If *d* could be reduced by adding additional valid predictors with lower *d*, how much change from the current *d* would be needed to avoid violating the four-fifths rule? or How large a change from a planned SR would be needed to avoid violating the four-fifths rule? Other similar questions might focus on target levels other than the four-fifths rule, such as, How much of a change from the current *d* would be needed to improve the AIR by a specified magnitude?

The discussion to this point has dealt with expected values of the AIR. However, given the small-to-modest sizes of applicant pools in many settings, it is certainly possible for a given sample to deviate from the population value. The AIR, like any sample statistic, has a sampling distribution, and De Corte and Lievens (2005) extended the work with an explicit treatment of this sampling distribution. They presented the relevant equations and offered illustrative examples. Table 17.2 shows the distribution of AIRs for various *d* values for the situation in which there are 300 applicants, a 10% SR, and 20% of the applicant pool is from the minority group. The table shows each possible AIR value as well as the likelihood of obtaining an AIR value of that magnitude or lower. For example, it shows that even if *d* were .00, such that we would expect no AI, the AIR would drop below 80% for 24.2% of samples. Note that large deviations from the expected value are more likely when a small minority applicant pool is paired with a small SR. Further exploration of the sampling variability in AI can be found in the work of Roth, Bobko, and Switzer (2006).

While the discussion to this point has focused on the AIR as the outcome of interest, AI is sometimes operationalized as a finding that the difference in selection rates for the two groups of interest is statistically significant. De Corte and Lievens also extended prior work by examining the probability with which a planned selection using a predictor with a given effect size *d* will result in a selection outcome that reflects AI according

Provide a reference.

TABLE 17.2

Sampling Distribution Function of the AI Ratio When Selecting 30 Candidates From a Total of 300 Applicants (60 Minority and 240 Majority Candidates) Using a Selection Test With Population Mean Difference Equal to 0, 0.2, 0.5, and 1.0

			Population mean difference			
J	**K**	**AI ratio**	**δ = 0.0**	**δ = 0.2**	**δ = 0.5**	**δ = 1.0**
0	30	0.000	0.001	0.007	0.058	0.394
1	29	0.138	0.008	0.044	0.237	0.770
2	28	0.286	0.037	0.146	0.495	0.940
3	27	0.444	0.110	0.321	0.732	0.988
4	26	0.615	0.242	0.532	0.885	0.998
5	25	0.800	0.420	0.725	0.960	1.000
6	24	1.000	0.609	0.862	0.988	1.000
7	23	1.217	0.770	0.942	0.997	1.000
8	22	1.455	0.883	0.979	0.999	1.000
9	21	1.714	0.949	0.994	1.000	1.000
10	20	2.000	0.981	0.998	1.000	1.000
11	19	2.316	0.994	1.000	1.000	1.000
12	18	2.667	0.998	1.000	1.000	1.000
13	17	3.059	1.000	1.000	1.000	1.000

Note: *J* indicates the number of selected minority applicants. *K* indicates the number of selected majority applicants

Leading zeros correct throughout table?

to Fisher's exact test. They referred to this probability as the risk of AI. For both extensions (examining the sampling distribution of the AIR and assessing the risk of AI), De Corte and Lievens offered the needed equations and illustrative examples as well as a flexible computer program permitting the user to input values of specific interest. The program can be downloaded from the Internet at http://users.ugent.be/~wdecorte/software.html. This site also offers access to most of the other programs that are mentioned in this chapter.

Prospects for Reducing *d* by Adding Additional Low-*d* Predictors

One potential strategy for reducing AI is to supplement a high-*d* predictor with one or more additional predictors with lower *d*. Sackett and Ellingson (1997) offered a set of formulas that permit an estimation of the expected effect of supplementing an existing predictor with additional predictors. They offered the following formula for determining the degree of group differences present when two or more predictors are combined to form an equally weighted composite:

Indicate mathematical notation for the formula.

$$d = \frac{\sum_{i=1}^{k} d_i}{\sqrt{k + k(k-1)r_{ii}}}$$

where d_i indicates the d value for each predictor included in the composite, k indicates the number of predictors combined to form the composite, and r_{ii} indicates the average correlation between the predictors included in the composite. The equation for d reduces to the following when only two predictors are combined to form a composite:

$$d = \frac{d_1 + d_2}{\sqrt{2 + 2r_{12}}}$$

where d_1 indicates the d value of the first predictor, d_2 indicates the d value of the second predictor, and r_{12} indicates the correlation between the two predictors. Table 17.3 presents the d values that would be observed when two predictors are combined to form a composite. The two factors that influence composite d (i.e., the summation of standardized difference

TABLE 17.3

Standardized Group Differences (d) for Two Predictors Combined to Form a Composite

Sum of *ds*	Correlation between the two predictors										
	.00	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**.90**	**1.0**
0.0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
0.2	.14	.13	.13	.12	.12	.12	.11	.11	.11	.10	.10
0.4	.28	.27	.26	.25	.24	.23	.22	.22	.21	.21	.20
0.6	.42	.40	.39	.37	.36	.35	.34	.33	.32	.31	.30
0.8	.57	.54	.52	.50	.48	.46	.45	.43	.42	.41	.40
1.0	.71	.67	.65	.62	.60	.58	.56	.54	.53	.51	.50
1.2	.85	.81	.77	.74	.72	.69	.67	.65	.63	.62	.60
1.4	.99	.94	.90	.87	.84	.81	.78	.76	.74	.72	.70
1.6	1.13	1.08	1.03	.99	.96	.92	.89	.87	.84	.82	.80
1.8	1.27	1.21	1.16	1.12	1.08	1.04	1.01	.98	.95	.92	.90
2.0	1.41	1.35	1.29	1.24	1.20	1.15	1.12	1.08	1.05	1.03	1.00
2.2	1.56	1.48	1.42	1.36	1.31	1.27	1.23	1.19	1.16	1.13	1.10
2.4	1.70	1.62	1.55	1.49	1.43	1.39	1.34	1.30	1.26	1.23	1.20
2.6	1.84	1.75	1.68	1.61	1.55	1.50	1.45	1.41	1.37	1.33	1.30
2.8	1.98	1.89	1.81	1.74	1.67	1.62	1.57	1.52	1.48	1.44	1.40
3.0	2.12	2.02	1.94	1.86	1.79	1.73	1.68	1.63	1.58	1.54	1.50

Note: Sum of d = the d value for one predictor + the d value for the second predictor.

Use leading zeros in table if appropriate., as they seem to be for all columns after the first.

scores for each predictor and the correlation between the two predictors) are systematically varied.

A review of Table 17.3 reveals a number of trends. First, holding sum of d constant, as the correlation between the two predictors increases, the level of composite d decreases. When two predictors become more highly correlated, they share increasing amounts of common variance. Combining two such predictors in a composite creates additional common variance, which produces decreased group differences. Second, Table 17.3 demonstrates that, in certain contexts, supplementing a predictor with a large d with another predictor with a smaller d actually produces a composite with a larger d than either of the individual predictors. Third, in discussions about this issue we find that the intuition of many of our colleagues is that the d for a composite of two predictors will be approximated by "splitting the difference" between the d values for the two predictors (e.g., a composite of a predictor with a d of 1.0 and another with a d of 0.0 will have a d of .5). Particularly when the correlation between the predictors is low, this intuition will severely underestimate the composite d (e.g., in the example, with two uncorrelated predictors, the composite d will actually be .71). Thus, the degree to which group differences, and subsequently AI, can be reduced by supplementing a predictor with a large d with a second predictor with a small d may be commonly overestimated.

If incorrect, provide reference.

Sackett and Ellingson (1997) also showed that adding additional supplemental measures has diminishing returns. For example, when $d_1 = 1.0$ and each additional measure is uncorrelated with the original measure and has $d = 0.0$, the composite d values when adding a second, third, fourth, and fifth measure are .71, .58, .50, and .45, respectively. Finally, they also offered an expanded equation for composite d when differing weights are applied to the predictors.

While the approaches discussed thus far shed light on the features driving AI, they are silent regarding the effects of modifying a selection system to reduce AI on mean criterion performance. We turn now to a set of decision aids that do attend to both AI and performance.

Category 2: Focus on Both AI and Criterion Performance as Outcome

Estimating AI and Other Selection Outcomes for Single-Stage and Multistage Selection

De Corte and Lievens (2003) and De Corte, Lievens, and Sackett (2006) described analytic procedures that enable selection researchers and

practitioners to explore the consequences in terms of several key outcomes of single- and multistage selection decisions. These procedures extend earlier related work by Cronbach and Gleser (1965) to the case for which applicants belong to several subpopulations with different mean predictor and performance structures. The procedures build on and generalize from earlier work by Tallis (1961) and Muthen (1990). They focused on the prototypic scenario that the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) labeled as "fixed applicant pool." In this scenario, the organization has information on the size and makeup of the applicant pool and considers using several predictors with known effect sizes, validities, and intercorrelation values to select the required number of applicants. Because single-stage selection is a special case of the more general multistage selection decisions, only the latter type of decisions are henceforth considered.

When planning a fixed-pool multistage selection system in which the applicants belong to different populations, a variety of decisions are to be made, each of which affects the selection cost, the mean performance of those selected, and the minority hiring rate. The first is determining which predictors to administer at an initial stage and which to administer at subsequent stages. The second relates to the proportion of the pool that will advance to subsequent stages in the selection procedure. The third is determining how final selection decisions should be made. Here, the key decision is whether the predictors used in initial screening also play a part in the final selection decision (i.e., if A is administered at Stage 1 and B at Stage 2, is the final selection done on the basis of B only or on A + B?).

The analytical procedure described by De Corte et al. (2006) is designed to assist the selection practitioner in making these decisions. Compared to the simulation approach proposed by Doverspike, Winter, Healy, and Barrett (1996), which may serve the same purpose, the analytical procedure is more flexible and permits dealing with the common situation in which only approximate values for some or most of the decision parameters (e.g., the predictor validities, effect sizes, and intercorrelations) are available. Also, whereas the results of the simulation method vary over repeated applications on the same input data, the analytical method always results in the same point estimate.

To illustrate the potential of the analytical procedure, we consider a situation in which the applicant population is a mixture of white and black candidate populations (with mixture proportions of .80 and .20, respectively) and four predictors are available (i.e., biodata [BI], a cognitive ability test [CA], a measure of conscientiousness [CO], and a structured interview [SI]). Table 17.4 displays the input parameter data for the predictor and criterion (i.e., task performance) mean subgroup differences, the predictor

TABLE 17.4

Standardized Mean Differences, Validities, and Intercorrelations for a Planned Selection System

Predictors	*d*	Validity	Intercorrelation matrix		
1. Cognitive ability (CA)	0.72	0.51			
2. Structured interview (SI)	0.31	0.48	0.31		
3. Conscientiousness (CO)	0.06	0.22	0.03	0.26	
4. Biodata (BI)	0.57	0.32	0.37	0.17	0.31

Eliminate numbers if unnecessary.

Eliminate leading zeros if appropriate.

validities, and the predictor intercorrelation values. The reported data correspond to the meta-analytic values provided by Potosky, Bobko, and Roth (2005) and to estimates obtained from Cortina, Goldstein, Payne, Davison, and Gilliland (2000); Ployhart, Weekley, Holtz, and Kemp (2003); and Dalessio and Silverhart (1994).

Is 2000 correct as in references? If not, provide 2001 reference and mention 2000 reference in text.

With these input parameter data in hand, the selection practitioner may now explore the likely consequences of alternative courses of action. For example, the practitioner may contrast, for a planned two-stage selection with equal selection rates of .5 in the stages, (a) the usage of the unit-weighted composite of the low-impact predictors (i.e., SI and CO) in the first stage followed by the unit-weighted composite of the high-impact predictors (i.e., CA and BI) in the second stage (Scenario 1) with (b) the reverse approach in which the initial selection is based on the unit weighted high-impact predictor composite, and the unit-weighted low-impact composite is used in the second stage (Scenario 2). Other possibilities, such as giving zero weight to one or more predictors, can also be explored. The expected effect of using regression-weighted composites instead of unit-weighted composites in Scenarios 1 and 2, leading to the Scenarios 3 and 4, respectively, as well as the expected merits of a single-stage approach in which either the unit-weighted or the regression-based composite of all four predictors is used with a selection rate of .25 (i.e., Scenarios 5 and 6), may also be of interest.

Is "high" correct?

Table 17.5 summarizes the results in terms of AI and average criterion performance of the six previously described scenarios. As expected, these results reveal that scenarios in which regression-based composites are used result in a higher average quality of the selected candidates and in a somewhat less-favorable AIR than comparable scenarios with unit-weighted composites (cf. Scenario 1 vs. 3 and Scenario 2 vs. 4). Also, comparing the results of Scenarios 1 and 2 to those of Scenario 5 and the results of Scenarios 3 and 4 to those of Scenario 6, it is again quite natural to find that the single-stage Scenarios 5 and 6, which use all the available predictor information at once, show a higher expected criterion score for the selected applicants than their two-stage counterparts. Alternatively, the comparison of Scenario 1 with Scenario 2 and the comparison of Scenario

TABLE 17.5

Projected Selection Quality (i.e., Average Criterion Score of the Selected Applicants) and AI Ratio for Several Planned Selection Scenarios

	Selection rate		Predictor composite		Average criterion score	AI ratio
Scenario	Stage 1	Stage 2	Stage 1	Stage 2		
1	.50	.50	1.00 SI + 1.00 CO	1.00 CA + 1.00 BI	0.70	0.39
2	.50	.50	1.00 CA + 1.00 BI	1.00 SI + 1.00 CO	0.69	0.45
3	.50	.50	0.45 SI + 0.10 CO	0.45 CA + 0.15 BI	0.75	0.38
4	.50	.50	0.45 CA + 0.15 BI	0.45 SI + 0.10 CO	0.74	0.42
5	.25	/	1.00 CA + 1.00 SI + 1.00 CO + 1.00 BI		0.75	0.40
6	.25	/	0.37 CA + 0.32 SI + 0.09 CO + 0.10 BI		0.80	0.37
7	.25	/	0.00 CA + 0.00 SI + 1.00 CO + 0.00 BI		0.28	0.93

Leading zeros correct in last two columns?

Use leading zeros for second and third columns, if appropriate.

3 with 4 suggest the less-intuitive finding that it may be better, in terms of AI, to sequence the high-impact predictors (i.e., CA and BI) before the low-impact predictors (i.e., SI and CO), without incurring any substantial loss of selection quality. However, Sackett and Roth (1996) and De Corte et al. (2006) obtained a similar result, and we refer to the latter authors for a tentative explanation of the phenomenon.

On the basis of the Table 17.5 results and those presented by De Corte et al. (2006), one might be tempted to pursue the quest for a set of rules or guidelines for the design of multistage selection scenarios that optimize the AI and the average quality of the selection. However, both De Corte et al. and Sackett and Roth (1996) warned against such a quest by observing that "there are no simple rules that can be offered about which approach to hurdle based selection is preferred" (Sackett & Roth, 1996, p. 569). Instead, these authors recognized that informative design principles are typically contingent on both generic and specific characteristics of the situation (such as, for example, the set of available predictors and the makeup of the applicant pool).

Provide end of quotation if this is incorrect.

So, although the analytic approach can be used to investigate the expected consequences of different selection designs, its merit as a decision aid remains limited to the exploration of alternative what if approaches. Within such an exploratory perspective, and provided that the boundary conditions for its application are reasonably fulfilled, the procedure is quite versatile. So, provided that the joint distribution of the predictor and criterion variables is approximately multivariate normal in the different subpopulations and that reasonably accurate data on the effect sizes, validities, and intercorrelations of the predictors are available, the procedure is applicable and produces fairly accurate results for a broad class of planned selection designs. As discussed by De Corte et al. (2006), the method can under these boundary conditions be applied to study general

Y100250_C017.indd 468 3/20/09 3:16:25 PM

single- and multistage selection schemes either with or without a probationary period and involving an arbitrary number of protected applicant groups besides the majority group. Selection systems with multidimensional job performance criteria are handled within the same framework. Doing so requires the correlation between performance dimensions and the specification of the relative weights to be assigned to each dimension in creating an overall performance measure.

De Corte et al. (2006) also provided a computer program to apply their procedure. The program output provides detailed information on the expected applicant flow through the stages by calculating for each selection stage the proportion retained and the stage-specific AIR of each applicant group. In addition, the program computes how the initial group differences on the predictors evolve through the subsequent selection stages. Finally, the program enables integrating the analytic procedure within a Monte Carlo approach to handle uncertainty in the selection parameters related to the predictor effect sizes, validities, and intercorrelations as well as to the makeup of the applicant pool.

As emphasized by De Corte et al. (2006), their analytical procedure has, compared to using simulation, the major advantage that it can be integrated within a straightforward approach to the design of selection scenarios that aim to achieve a given set of goals in terms of workforce quality and desired levels of workforce diversity. To highlight the importance of such an integration, we return to Table 17.5 and, in particular, to the results reported there for Scenario 7. This scenario, in which candidates are selected in a single stage on the basis of only the CO predictor scores, corresponds to the best-possible design when only the goal of reducing the expected AI of the selection is of importance, whereas Scenario 6 is the optimal design when only the average criterion score of the selected applicants is valued. The expected outcomes of these two scenarios show a wide range of possible values for the AIR (i.e., between .37 and .93) and the average criterion score of the selected applicants (i.e., between .28 and .80).

Such substantial ranges of possible values for the AI ratio and selection quality are common, and if both workforce quality and workforce diversity are valued, only scenarios that offer an optimal trade-off between these often-conflicting goals will be of interest. To identify these optimal trade-off scenarios, the computational procedure of De Corte et al. (2006) could be used many times, each time inserting different values for the predictor weights and the stage-specific retention weights, but it is obvious that this "trial-and-error" approach is far from practical. Instead, a more direct procedure is to be preferred. Such a procedure, which integrates the De Corte et al. computational method within a multicriteria optimization approach, is discussed next.

Y100250_C017.indd 469 3/20/09 3:16:25 PM

Pareto-Optimal Trade-Offs

To assist selection practitioners in planning future selection systems to optimize both AI and selection quality, De Corte, Lievens, and Sackett (2007) presented a decision tool. This decision tool focuses on the common scenario in which employers have to make decisions on forming a composite of a set of predictors (e.g., cognitive tests, personality tests, interviews, work samples; Sackett & Ellingson, 1997; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). In this scenario, how to maximize the mean criterion score of selected applicants is well known, namely, by inputting all predictors into a regression equation and using the resulting weights. However, employers often ask whether there exists an alternative way of using the predictors that comes close to this optimal solution in terms of the level of criterion performance achieved but does so with less AI.

Prior approaches tried to answer this question by using a trial-and-error strategy for determining various predictor weights to find a composite alternative that comes closest to meeting the two objectives (Hattrup, Rock, & Scalia, 1997; Pulakos & Schmitt, 1996; Sackett & Ellingson, 1997; Schmitt et al., 1997). Such ad hoc trial-and-error strategies are also exemplified by technical reports that typically present a series of alternative models that use varying combinations of available predictors, weighted in differing ways.

De Corte et al. developed an analytical and formal procedure to determine in advance whether there is an alternative way of using the predictors that comes close to the regression-based weighting in terms of predictive efficiency but does so with less AI. Thus, this procedure enables the determination of values for the predictor weights such that the resulting predictor composites provide an optimal balance or trade-off between productivity (i.e., high-validity) and diversity (i.e., low-AI) objectives. To this end, the notion of Pareto-optimal trade-offs between the two outcomes was presented. Given a set of predictors, there are an infinite number of possible weighting schemes that could be applied in forming predictor composites. A Pareto-optimal trade-off corresponds to a weighting scheme for which one outcome cannot be improved without harm to the other outcome given the details of the intended selection scenario (e.g., SR) and the available selection predictors. For example, there may be multiple weighting schemes that would result in a given correlation between the predictor composite and the criterion; of these schemes, the Pareto-optimal one is the set of weights that result in the highest AIR. Similarly, there may be multiple weighting schemes that would result in a given AIR; the Pareto-optimal one is the set of weights that result in the highest level of validity. So, Pareto-optimal composites offer optimal trade-offs between the AI and the validity objective, and the entire collection of these Pareto-optimal trade-offs is usually referred

Provide a reference.

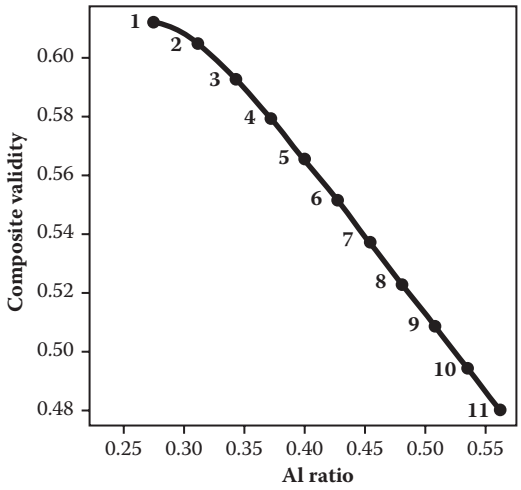

FIGURE 17.1
Pareto-optimal validity-adverse impact ratio trade-off curve for a selection with selection rate of .10 using a composite of cognitive ability and a structured interview as based on values from Potosky et al. (2005; cf. Table 17.1 of Potosky et al.).

Change correct? Or Table 17.1 of this chapter?

to as the Pareto-optimal trade-off curve or function (Keeney & Raiffa, 1993; Pareto, 1906).

De Corte et al. (2007) wrote a computer program to implement the multicriteria optimization procedure for identifying the set of Pareto-optimal composites. As input for the program, a set of predictors with given validity, intercorrelations, and subgroup differences and the specification of an SR are needed. Both probationary and nonprobationary selection as well as situations in which the applicants come from several different minority populations can be addressed.

The results of the procedure are expressed in tabular or graphical form. Figure 17.1 illustrates the graphical outcome of the technique; it presents the Pareto-optimal trade-off curve for a composite of cognitive ability and a structured interview, based on values from Table 17.1 of Potosky et al. (2005) (cf. the present Table 17.4 values). The figure shows the optimal levels of AIR achievable at each level of validity or, equivalently, the optimal level of validity achievable at each level of AIR. Table 17.6 shows the tabular presentation as it further details a selected number of optimal trade-offs. For each selected trade-off (the numbered trade-off points on Figure 17.1), the table summarizes the validity and AIR value as well as the weighting (with weights scaled to have unit sum) of the predictors that characterize the corresponding optimal composite.

The definition of the set of Pareto-optimal composites implies that the regression-based composite is one particular element of the set. As

TABLE 17.6

Selected Pareto-Optimal Trade-Off Composites of Cognitive Ability (CA) and Structured Interview (SI)

			Predictor weights	
Point	**Validity**	**AI ratio**	**CA**	**SI**
1	0.61	0.27	0.53	0.47
2	0.61	0.31	0.42	0.58
3	0.59	0.34	0.35	0.65
4	0.58	0.37	0.30	0.70
5	0.57	0.40	0.25	0.75
6	0.55	0.43	0.20	0.80
7	0.54	0.45	0.16	0.84
8	0.52	0.48	0.12	0.88
9	0.51	0.51	0.08	0.92
10	0.49	0.53	0.04	0.96
11	0.48	0.56	0.00	1.00

Eliminate leading zeros if appropriate throughout.

regression-based weights maximize the validity of the resulting composite, no other weighing of the predictors can outperform this composite in terms of the validity criterion. In the figure, the regression-based composite refers to Point 1, with a mean quality of 0.61 and an AIR of 0.27. The minimal impact composite, defined as the composite with the highest possible AIR value (0.56), is another example of the set (see Point 11 in Figure 17.1). Under the common condition of all positive predictor effect sizes, the regression-based and the minimum impact composite are the boundary points of the Pareto-optimal set, with all the other Pareto-optimal composites showing more balanced trade-offs between validity and AI. More specifically, these intermediate composites are all characterized by a smaller validity than the regression-based composite, and they all show a smaller value for the AIR than the minimum impact composite. Table 17.6 also illustrates how this technique can be used to answer whether there exists a different weighting of predictors that will come close (i.e., within a specified distance) to the maximum mean quality attainable, but with less adverse impact. To address this, the definition of *close* must be specified; once a given decision maker defines it (e.g., anything within 95% of the maximum mean quality attainable), then Figure 17.1 permits this question to be answered. As noted, the maximum mean quality attainable with these predictors at this SR is 0.61. Thus, we can move down the optimal trade-off curve to the point at which mean quality (as gauged by the validity coefficient) is 0.58 (i.e., 95% of 0.61); we find that the Pareto-optimal weighting of predictors at this point produces an AIR of 0.37 compared to the value of 0.27 for the weighting that maximizes

Zero correct?

Zero correct?

quality. The table also presents the predictor weights that would be used to obtain this result. Finally, it can be determined that the gain in AI from 0.27 to 0.37 corresponds to a 32% improvement of minority hiring.

Alternately, suppose another decision maker is willing to accept 10% reduction in validity (rather than the 5% in the example). Here, we can move down the optimal trade-off curve to the point at which mean quality is 0.55 (i.e., 90% of 0.61) and find that the Pareto-optimal weighting of predictors at this point produces an AIR of 0.43 compared to the value of 0.27 for the weighting that maximizes quality.

Zero correct?

Possible reactions to the Pareto-optimal approach might include questions about whether it is permissible to deviate from a validity maximization strategy and whether the Civil Rights Act of 1991 precludes any selection strategy that takes AI into account when weighting predictors. Regarding the first issue, there is no general requirement to maximize validity; in fact, the use of methods that depart from validity maximization is routine. Unit weights are often used for administrative ease; score bands (e.g., "green-yellow-red" or "pass-fail") are commonly used to simplify decision making; shorter forms of tests are commonly used to reduce costs and testing time. What is restricted by the U.S. Civil Rights Act of 1991 is treating scores differently by subgroup. The key point is that the Pareto-optimal approach does not involve such differential treatment. All candidates are treated the same: Any decision about the predictor weights applies to all of the candidates. The procedure simply includes workforce diversity as an additional objective to be met by the selection system. Note that the Pareto-optimal approach does not tell the selection system designer which weights should be used. Instead, it serves essentially as a method of choosing among differing weighting schemes given a set of predictors, providing information regarding relative gains and losses in terms of validity and AI if differing weights are chosen. It is a matter of values about whether an employer is willing to accept a given reduction in validity (i.e., 1% or 5%) for a given reduction in AI. The phrase "willing to accept" is important because the approach does not specify a particular trade-off that one should accept. Finally, it is important to emphasize that investigating weighting schemes a priori may be legally more defensible than waiting until after predictor data have been gathered. In some settings, organizations are even required by statute or policy to reveal the weights given to the components of a selection system to applicants prior to testing.

Aguinis and Smith (2007)

Aguinis and Smith (2007) offered a decision aid that is quite different in nature from those discussed. They presented an approach that integrates four variables: (a) magnitude of the predictor-criterion relationship, (b) AI,

(c) selection errors (false positives and false negatives), and (d) test bias. Their approach incorporates the specification of a desired level of criterion performance; that specification permits the determination of false positives and false negatives at a given SR. It also incorporates the Cleary model of test bias, in which a test is viewed as unbiased if the regression line relating predictor and criterion scores is identical for the subgroups compared. If the regression lines are not identical (i.e., they differ in slopes, intercepts, or both), the test is viewed as biased, and the use of a common regression line would result in systematic errors of prediction being made. The Aguinis and Smith approach distinguishes between prediction errors due to imperfect validity and error made due to treating a biased test as if it were unbiased (e.g., using a common regression line when, in fact, the regression lines for the groups under consideration differ).

Aguinis and Smith (2007) developed an analytical approach that integrates all four of these features and offered a computer program that permits users to enter values for the predictor and criterion of interest to them and to examine the resulting AI, mean criterion performance, and false-positive and false-negative rates by subgroup. One important way in which their approach differs from others discussed is in the information needed as input to the program. While the other approaches focus on correlations, the Aguinis and Smith formulation focuses on regression analysis. It requires as input means and standard deviations for predictors and criteria for each subgroup as well as predictor-criterion correlations. As such, it requires more concrete and detailed information than the approaches described. For example, the other approaches permit addressing a question such as, What would we expect to happen if we added a conscientiousness measure to a cognitive ability measure? The approaches discussed would require an estimate of subgroup differences on both predictors, predictor-criterion correlation estimates for both predictors, and the correlation between the two predictors. The Aguinis and Smith approach requires predictor and criterion means and standard deviations as well and thus seems to focus on specific measures in specific situations rather than on general planning strategy prior to selecting specific measures. Nonetheless, in settings in which these specific details are available, the approach does incorporate issues of rates of false positives and false negatives as well as information about test bias.

Discussion

One crucial point is that all of the approaches are descriptive: They outline the consequences of various courses of action (e.g., What would we

expect to happen if we lower a cutoff score? What would we expect to happen if we add a structured interview to our selection system?). These decision aids do not tell the user what they *should* do as that is a matter of values. This is perhaps made most explicit in the work on Pareto-optimal selection by De Corte, Lievens, and Sackett (2007). That approach specifies the amount of improvement in AI that would be expected to result from any given reduction in the mean criterion performance of those selected (i.e., a reduction in validity). Whether a given validity loss for a given AI gain is seen as acceptable is a value judgment, not a technical issue. A trade-off that seems reasonable to some will be seen as inappropriate by others. We anticipate that some readers will take the stance that validity is the only outcome of interest, and that it is inappropriate to even consider AI-validity trade-offs. Our response is that it is our experience that many organizations do value both diversity and performance and are willing to consider trade-offs between the outcomes. Our stance is that one is best served by as clear an understanding as possible of the implications of any choices made regarding trade-offs between these outcomes, and thus we have pursued the series of investigations and developed the series of decision aids described in this chapter.

A second issue worthy of discussion is the fact that some of the values required as input for the approaches described in this chapter may not be known with certainty. For example, what does one do if one is considering adding a new predictor to a selection system that already includes a predictor with known validity and known d, but the correlation between the new predictor and the existing predictor is unknown? Here, we advocate a sensitivity analysis, in which a range of possible values are input into the decision aid. In some cases, the emergent finding is that variation on the unknown parameter has little effect on the outcomes of interest, in which case one can proceed without concern. In other cases, the finding may be that the outcomes of interest do indeed hinge on this parameter. Here, one option is to work harder to locate an estimate of the parameter, perhaps conducting a local study to obtain the needed value. Another option is to "prepare for the worst" by identifying the worst-case scenario and estimating its effect. Yet another is to note that one truly is uncertain about the expected outcome and thus shy away from offering a priori statements about the likely degree of AI. In short, in some cases one may conclude that one does have a pretty good idea about likely outcomes prior to actual data collection; in other cases, one is best off admitting to a high degree of uncertainty.

A third issue concerns the limitations of the present decision aids. Some of these limitations are tied to the assumption underlying these methods, whereas others point to aspects of the decision situation that still need to be addressed. Thus, several of the presented methods are based on the assumption that the predictor-criterion space is multivariate normal, or

that the within-group regressions of the criterion on the predictors is linear. Features of the decision situation that require further elaboration include job refusal and a focus on classification rather than selection decisions. The extension to multiple-hurdle selection situations of the multicriteria optimization approach to uncover Pareto-optimal trade-offs is another example. All these extensions should provide the selection practitioner with a set of more realistic and generally applicable tools when planning selection decisions to achieve given valuable goals in terms of workforce quality and diversity.

Fourth, we note that work on trade-offs has focused on AI and mean performance among those selected as outcomes. A broader range of outcomes are certainly of interest to organizations. These range from narrow outcomes, such as costs of implementing the selection system (e.g., De Corte et al., 2006) or administrative ease in administering a selection system, to much broader outcomes, such as organizational effectiveness and firm reputation. These broader outcomes are more difficult to measure and model. Nonetheless, we do note that there are additional trade-offs of potential interest that are worthy of investigation.

Fifth, we acknowledge that adding low-impact predictors and predictor weighting are only some routes to workforce diversity (Ployhart & Holtz, 2008). Apart from these routes, there exist other routes to workforce diversity, such as banding and the development of innovative test presentation (e.g., video; see Chan & Schmitt, 1997) and response (e.g., constructed responses; see Edwards & Arthur, 2007). Clearly, these strategies also have important merits. While prior research has typically used these strategies in isolation, we need studies that examine the combination of various strategies for reducing AI.

References

Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology, 60*, 165–199.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143–159.

Civil Rights Act of 1991, Pub. L. No. 102-166, 7 U.S.C. §701, 702, 703, 705, 706, 717 (1991).

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53,* 325–351.

Cronbach, L. R., and Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Dalessio, A., & Silverhart, T. (1994). Combining biodata test and interview information: Predicting decisions and performance criteria. *Personnel Psychology, 47,* 303–315.

De Corte, W., & Lievens, F. (2003). A practical procedure to estimate the quality and the adverse impact of single-stage selection decisions. *International Journal of Selection and Assessment, 11*, 89–97.

De Corte, W., & Lievens, F. (2005). The risk of adverse impact in selections based on a test with known effect size. *Educational and Psychological Measurement, 65,* 737–758.

De Corte, W., Lievens, F., and Sackett, P. R. (2006). Predicting adverse impact and mean criterion performance in multi-stage selection. *Journal of Applied Psychology, 91,* 523–537.

De Corte, W., Lievens, F., and Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92,* 1380–1393.

Doverspike, D., Winter, J., Healy, M., & Barrett, G. (1996). Simulation as a method of illustrating the impact of differential weights on personnel selection outcomes. *Human Performance, 9,* 259–273.

Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*, 794–801.

Hattrup, K., Rock, J., & Scalia C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology, 82*, 656–664.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value tradeoffs.* Cambridge: Cambridge University Press.

Muthen, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology, 43,* 131–143.

Pareto, V. (1906). *Manuale di economica polittica* [Manual of political economy]. Milan, Italy: Societa Editrice Libraia.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56,* 733–752.

Ployhart, R. E., & Holtz, B.C. (2008). The diversity-validity dilemma: strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153–172

Potosky, D., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Assessment and Selection, 13,* 304–315.

Y100250_C017.indd 477 3/20/09 3:16:28 PM

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

Roth, P. L., Bobko, P., & Switzer, F. S. (2006). Modeling behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology, 91,* 507–522.

Sackett, P. R., & Ellingson, J. E. (1997). On the effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 708–721.

Sackett, P. R., & Roth, L. (1996). Multistage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49*, 549–572.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318.

Sackett, P. R., & Wilk, S. L. (1994) Within-group norming and other forms of score adjustment in pre-employment testing. *American Psychologist, 49*, 929–954.

Schmitt, N., Rogers W., Chan, D., Sheppard L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82,* 719–730.

Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society, Series B 23,* 223–229.

Y100250_C017.indd 478 3/20/09 3:16:28 PM