Original Article

The Utility of Personnel Selection Decisions

Comparing Compensatory and Multiple-Hurdle Selection Models

Jisoo Ock[1] and Frederick L. Oswald[2]

[1]School of Social Sciences, Division of Psychology, Nanyang Technological University, Singapore
[2]Department of Psychology, Rice University, Houston, TX, USA

Abstract: Compensatory selection is generally more reliable than multiple-hurdle selection. Yet, practitioners may lean toward multiple-hurdle models, because administering an entire predictor battery to every applicant can be time-consuming, labor-intensive, and costly. Using Monte Carlo simulation, we considered some specific cases to illustrate, in terms of selection utility and the cost-reliability tradeoff between compensatory and multiple-hurdle selection models. Results showed that compensatory model selection produced a higher level of expected criterion performance in the selected applicant subgroup, and a higher overall selection utility in most conditions. The simulation provides researchers and practitioners with a practical illustration of the tradeoff between reliable (compensatory) versus cost-efficient (multiple-hurdle) selection models – one that can inspire the exploration of other scenarios and tradeoffs.

Keywords: selection utility, selection model, high-stakes testing, Monte Carlo simulation

Organizations approach and operationalize personnel selection systems in different ways, but two major classes of selection cover much of the approaches. In a *compensatory model* approach, selection occurs on a composite score of multiple predictor measures (e.g., standardizing and adding up the test scores to create a unit-weighted composite). Composite scores are compensatory because higher scores on some predictors will compensate for lower scores in others. In a *multiple-hurdle model* approach, large numbers of applicants are first administered an initial set of tests, perhaps those that are generally less expensive to administer on a per-applicant basis (e.g., cognitive ability tests, personality tests). The subset of applicants who passed the cutoff on this initial hurdle are administered additional tests as hurdles, perhaps those that are more expensive or time-consuming to administer (e.g., assessment centers, structured interviews). Of course, any predictor or subset of predictors in a selection process can be subject to different rules. For example, in a compensatory model, one can use unit weights, expert weights, or regression weights; or in a multiple-hurdle model, hurdles can be individual tests and/or composites of tests. Obviously, the choice of the rule can have meaningful implications for the outcomes of selection (Douglas & Mislevy, 2010).

An important advantage of the multiple-hurdle selection procedure is that it can be designed to save money and time dedicated to administration, testing materials, and labor. However, a single cutoff based on a composite of tests in the compensatory model is more psychometrically reliable than cutoffs applied to each constituent test in a multiple-hurdle selection process (Hambleton & Slater, 1997). In other words, multiple-hurdle selection is less reliable than compensatory selection because applicants who cross only the initial test hurdles are providing less information about the selection decision (Kane & Case, 2004).

Also, the choice of selection method has a meaningful influence on the range restriction of job applicants' test scores, and this needs to be accounted for because it ultimately impacts the estimation of utility or the bottom-line dollar value of the selection system. In particular, multiple-hurdle selection creates both direct and incidental range restriction that reduces the variance in predictor scores as well as the criterion scores to which the predictors are supposed to relate (Guion, 2011; Sackett, Laczo, & Arvey, 2002). For example, in multiple-hurdle selection, selection on the first hurdle creates direct range restriction on that variable and its validity; and it also creates incidental range restriction on predictor variables in the subsequent hurdles

https://doi.org/10.1027/1866-5888/a000205

where direct selection is yet to occur, as well as incidental range restriction on the criterion to which these predictors relate. Selection on a composite also creates incidental range restriction on all variables (Sackett, Lievens, Berry, & Landers, 2007), but this effect is generally less than that for multiple-hurdle selection.

Given all this, there is a real reason and value for practitioners and the organizations they serve, then, to be comparing compensatory and multiple-hurdle models of selection: Compensatory approaches seem to have greater validity benefit in theory, but in practice, multiple-hurdle models might be justified in terms of the cost and time savings relative to any decrease in validity.

Using Monte Carlo simulation, we demonstrate the distinct difference in expected selection utility between two specific compensatory and multiple-hurdle selection methods defined by a realistic set of key parameters. Specifically, we integrated different aspects of previous research on utility analysis to illustrate the effect of selection method on expected criterion performance of selected applicants (Naylor & Shine, 1965), and on overall utility (Brogden, 1946; Cronbach & Gleser, 1965). Simulations results will provide a useful basis for comparing a compensatory model to a multiple-hurdle model in exploring the tradeoff between validity and cost savings using the metric of utility to assess that tradeoff.

A Brief Review of the Traditional Utility Models in Personnel Selection

Traditional utility models in personnel selection are used to obtain and compare the value of different selection outcomes, in terms of dollars or job performance gains, on the basis of different selection tools and their validities (Brogden, 1946; Cronbach & Gleser, 1965; Naylor & Shine, 1965; Taylor & Russell, 1939). For these models, a selection test or a composite of selection tests is considered within a compensatory selection model; multiple-hurdle selection is typically not considered (though see Roomsburg, 1989, for a notable exception).

All of these utility models operationalize selection utility in different ways: The Taylor-Russell model estimates percent gains in successful performance in those who were selected, where "success" is a dichotomous outcome; the Naylor-Shine model also estimates gains but in terms of mean predicted performance, where performance is a continuous outcome; and the Brogden-Cronbach-Gleser model considers estimated monetary gains from conducting selection.

Utility estimates in all three of these traditional selection utility models consider specific conditions relevant to the selection context, such as the base rate of success, the validity coefficient, and the fixed and variable costs of implementing the selection procedure. One might think that the loss in utility due to these factors can be estimated directly, such as through utility formulas and psychometric corrections. Although this is true in simple cases, such estimates are not always straightforward or even possible sometimes, due to a multitude of complex factors found in personnel selection settings, especially in multiple-hurdle selection (Sackett & Roth, 1996). Psychometric corrections might be available for some selection factors when they are taken in isolation (e.g., criterion-related validity and the standard error adjusted for incidental selection; Allen & Dunbar, 1990; Mendoza, Bard, Mumford, & Ang, 2004) but when factors such as these are considered simultaneously, the standard error associated with the psychometric corrections is often not easily estimated by formula (however, for reliability and range restriction taken together, see Raju & Brand, 2003). Most importantly, psychometric corrections and associated standard errors alone do not directly communicate practical outcomes for personnel selection practices, such as utility and variation in utility across samples, which is the central goal of this paper. Simulations help ensure that the variability (selection successes and errors) for a personnel selection situation is summarized appropriately and tied to utility in a more straightforward and practical manner.

Utility Comparison Between Compensatory Selection Versus Multiple-Hurdle Selection

First, we examined the difference in mean true criterion scores between applicants selected through a compensatory selection method and applicants selected through a multiple-hurdle selection method. Second, we calculated the difference in estimated monetary gains from conducting selection with compensatory versus multiple-hurdle selection methods. Specifically, we varied the dollar value associated with the criterion score of the selected applicants and the cost of the selection process on per-applicant basis; then we compared the difference in net monetary gains across study conditions, with special consideration toward comparing compensatory and multiple-hurdle selection methods.

Method

The current study focused on specific selection scenarios that are based on a realistic set of correlations derived from the Roth, Switzer, Van Iddekinge, and Oh's (2011) meta-analysis of employment research, containing predictors

relevant to employment (i.e., cognitive ability, structured interview, conscientiousness, and biodata), and an overall job performance criterion: Table 1 shows the correlation matrix that was used to generate data with a multivariate normal distribution.

First, to estimate the population correlation matrix that generated the simulation data, we corrected the meta-analytic predictor intercorrelations and validity coefficients from Roth et al. (2011) for measurement error variance using realistic reliability coefficients found in the organizational psychology literature: .83 for cognitive ability (Salgado, Anderson, Moscoso, Bertua, & De Fruyt, 2003), .68 for the structured interview (Huffcutt, Culbertson, & Weyhrauch, 2013), .78 for conscientiousness (Viswesvaran & Ones, 2000), .77 for biodata (Shaffer, Saunders, & Owens, 1986), and .52 for overall job performance (Viswesvaran, Ones, & Schmidt, 1996). Even though these served as reasonable estimates, we acknowledge there is heterogeneity associated with each reliability estimate. As an obvious case of this, we know that the structured interview and biodata predictors are methods with reliability coefficients that surely vary as a function of the construct being measured (Arthur & Villado, 2008).

From this population matrix, we generated sample realizations of true scores as well as their corresponding observed scores, given the specified reliability coefficients and intercorrelations for these five variables (see Kaiser & Dickman, 1962, for the singular value decomposition method employed). Predictor composite scores and criterion scores were standardized within the true score data and the observed score data to ensure their comparability (because when unstandardized, observed score variance equals true score variance plus error variance). Note that the validity estimates found in Roth et al. (2011) were corrected for range restriction and criterion measurement error variance. Thus, for us to calculate the observed validity coefficients, we attenuated the validity estimates by introducing the effect of criterion unreliability (i.e., we multiplied the validity estimates by square root of .52).

After correcting for range restriction as we did, then generating observed scores from true scores is a relatively simple matter, given the classical test theory assumption that only measurement error variance affects true scores (Lord & Novick, 1968). Given that the observed score is X_O, the true score is X_T (with a mean of $\bar{X}_T = 0$ and standard deviation of $s_T = 1$), and the population reliability is r_{xx}, then the formula for generating observed scores in the sample is:

$$X_o = X_T\sqrt{r_{xx}} + \dot{s}_T\sqrt{1 - r_{xx}},$$

where $\bar{X}_T$ is the mean of the true scores in the sample, and $\dot{s}_T$ is a score drawn randomly from a standard normal distribution with a mean of zero and a standard deviation of one (the dot notation indicates the score $\dot{s}_T$, not standard deviation itself, $s_T = 1$). This equation shows how the standard error of measurement adjusts this latter standard deviation: when reliability is higher, the standard deviation of the error scores is lower.

Table 1. Meta-analytic correlations between four predictors and job performance

Measure	1	2	3	4	5
1. Cognitive ability	*.83*[a]	.41	.04	.46	.56
2. Structured interview	.31	*.68*[b]	.18	.22	.59
3. Conscientiousness	.03	.13	*.78*[c]	.66	.25
4. Biodata	.37	.16	.51	*.77*[d]	.36
5. Job performance	.37	.35	.16	.23	*.52*[e]

Notes. Observed correlations from Roth et al. (2011) are below the main diagonal. These correlations reflect job incumbent correlations that were corrected for range restriction, with the exception of the biodata – structured interview and biodata – conscientiousness correlations, where the authors could not find appropriately corrected correlations. Note that criterion-related validities in Roth et al. (2011) reflect operational validities corrected for range restriction and criterion unreliability; we substituted them with observed validities by multiplying them by the square root of overall job performance reliability coefficient. True correlations are above the main diagonal; these are corrected for measurement error variance in both the predictor and criterion. Reliability coefficients come from the following sources: [a]for cognitive ability, see Salgado et al. (2003); [b]for the structured interview, see Huffcutt et al. (2013); [c]for conscientiousness, see Viswesvaran and Ones (2000); [d]for biodata, see Shaffer et al. (1986); [e]for overall job performance, see Viswesvaran et al. (1996).

Simulation Conditions

In the real world, selection data and findings are often generalizable, yet the specific conditions that give rise to these data may be only vaguely known. Our selection simulations serve as a useful and informative complement to real-world selection by manipulating a set of realistic parameters, and generating data under reasonable assumptions. Together, this leads to well-understood reasons for the estimates of expected utility that are produced across conditions, as well as the associated variance in utility estimates within conditions across samples.

Selection Method

After all observed scores were standardized, the compensatory selection method made top-down selection decisions on the unit-weighted composite of the observed predictor test scores for each applicant. By contrast, the multiple-hurdle selection method initially selected applicants top-down on the unit-weighted composite of observed scores on cognitive ability, conscientiousness, and biodata; then this subset of screened applicants was selected based on their observed structured interview score. This multiple-hurdle selection sequence screens a large number of applicants first on tests that are relatively cheaper and less time-consuming to administer; then the smaller subset of screened applicants receives further evaluation on tests that are more expensive and/or time-consuming (see Table 2 for more details about the multiple-hurdle model selection procedure).

Table 2. Multiple-hurdle selection procedure

	First stage SR (unit-weighted composite on cognitive ability, conscientiousness, biodata)	Second stage SR (structured interview)
Net SR = .10	.15	.67
Net SR = .20	.25	.80
Net SR = .40	.50	.80

Notes. The product of selection ratio at each hurdle equals the net selection ratio. SR = selection ratio.

Selection Ratio

All other things being equal, the selection ratio affects the accuracy of selection decisions. Specifically, true scores and measurement error become negatively correlated within range-restricted samples (Mendoza & Mumford, 1987). As a result, when selection is either very extreme or very liberal (i.e., has a very low or very high selection ratio, respectively), there are asymmetries in false positives and false negatives whose effects are worth examining; although, of course, when selection is very liberal, any gains in utility are likely to be small. Taking this into account, and because selection ratio vary depending on the purposes of the organization at a given time, the job market and available job applicants, and so on, our simulation likewise varied the selection ratio across a range of plausible values (SR = .10, .20, and .40).

Sample Size

We varied the sample size across three levels (N = 50, 100, 250). In many organizations, the total number of applicants at any one point in time might be much higher than the numbers we considered, especially with the increased prevalence of online recruitment practices that enables organizations to reach job seekers easily and at a relatively low cost (Howardson & Behrend, 2014). However, a constraint of online recruitment systems is that it also invites a high volume of unqualified application traffic (Lievens & Harris, 2003). We assume that organizations have an initial screening system in place to screen out irrelevant responses to job postings (e.g., applicants who do not meet the job description) and very low base rate issues that do not affect our material findings (e.g., screen out those with criminal histories), and so the sample sizes we considered reflect a manageable number of applicants that organizations might seriously consider in a given selection situation, at any given point in time.

Procedures

Simulations were programmed using R Code (R Core Team, 2017). When the true and observed scores for the five study variables were generated, the simulation performed top-down selection on the observed predictor scores two times: once based on the unit-weighted predictor composite, and once using multiple-hurdle selection, both in the manner described earlier. We then calculated the mean and interquartile range of the true criterion performance score for applicants, who were selected on their observed predictor scores. With each combination of the simulation parameters, we replicated the aforementioned process 1,000 times to model sampling error variance within each condition.

Selection Utility Calculation

We calculated selection utility based on the classic utility analysis formula (Brogden, 1946; Cronbach & Gleser, 1965):

$$U = \left(\bar{z}_y \times SD_y \times N_s\right) - \left(N \times C_A\right),$$

where U denotes the overall monetary utility of the selection procedure; $\bar{z}_y$ denotes the mean expected criterion performance score of selected applicants in standard deviation units; SD_y denotes the monetary value associated with a one standard deviation difference in criterion performance; N_s denotes the number of selected applicants; N denotes the total sample size; and C_A denotes the cost per applicant.

We used the $\bar{z}_y$ values from our simulation as input values to the utility equation above, and varied the values of SD_y and the cost per applicant according to the realistic range of values from Sturman (2000). Specifically, we adjusted the values in Sturman for inflation using the online calculator from the US Bureau of Labor Statistics (https://data.bls.gov/cgi-bin/cpicalc.pl), and rounded them to the nearest hundred. The range we used for SD_y was \$6,000–\$60,000 in \$1,000 increments; and the range we used for cost per applicant was \$100–\$1,600 in \$100 increments. We also varied the cost of multiple-hurdle selection at 50%, 30%, and 10% of the cost of compensatory selection. After cost estimates were set, we then examined the difference in utility between compensatory model selection and multiple-hurdle model selection across study conditions (see Table 3 for summary of the study conditions).

Results

Impact of Selection Method on Criterion Performance

Simulation results supported the general expectation that compensatory model selection would consistently produce higher levels of mean criterion performance compared with those from multiple-hurdle model selection ($\bar{z}_y$ = 1.72, 1.25, and 0.63 for compensatory model selection versus $\bar{z}_y$ = 0.67, 0.55, and 0.35 for multiple-hurdle model selection when SR = .10, .20, and .40, respectively; see Table 4). Moreover, the difference in mean criterion performance

Table 3. Simulation characteristics and parameters

Characteristics and parameters	Values
Constant	
Predictors	Cognitive ability, structured interview, conscientiousness, biodata
Criterion	Overall job performance
Predictor reliability	Realistic reliability: .83 for cognitive ability; .68 for structured interview; .78 for conscientiousness; .77 for biodata
Criterion reliability	True scores: Perfect reliability (r_{yy} = 1.0)
	Realistic reliability: .52
Variable	
Selection method	Compensatory, multiple-hurdle (two stages)
Selection ratio	.10, .20, .40
Number of applicants	50, 100, 250
Cost per applicant	$100–$1,600
SD_y	$6,000–$60,000
Cost of multiple-hurdle selection relative to compensatory selection	50%, 30%, 10%

Notes. SD_y = Monetary value associated with one standard deviation change in overall job performance. For each condition, simulation was replicated 1,000 times. Reliability estimates are derived from the existing literature as noted in Table 1.

Table 4. Mean and interquartile range of true criterion performance score for selected applicants

	Performance		
	M [IQR]		
Selection method	*N* = 50	*N* = 100	*N* = 250
SR = .10			
Compensatory	1.72 [1.22, 2.22]	1.72 [1.38, 2.07]	1.72 [1.48, 1.95]
Multiple-hurdle	0.65 [0.48, 0.85]	0.68 [0.55, 0.82]	0.68 [0.60, 0.77]
SR = .20			
Compensatory	1.25 [0.95, 1.62]	1.25 [1.00, 1.49]	1.23 [1.09, 1.38]
Multiple-hurdle	0.55 [0.41, 0.70]	0.55 [0.45, 0.65]	0.56 [0.50, 0.62]
SR = .40			
Compensatory	0.62 [0.39, 0.87]	0.66 [0.41, 0.80]	0.62 [0.50, 0.74]
Multiple-hurdle	0.35 [0.27, 0.43]	0.35 [0.30, 0.40]	0.35 [0.32, 0.39]

Notes. *M* = mean; IQR = interquartile range; *N* = sample size; SR = selection ratio.

between the two selection models increased as the selection ratio decreased (across sample sizes, the difference increased to d = 1.05 when SR = .10 from d = 0.28 when SR = .40), indicating that in these scenarios, greater selectivity (a lower selection ratio) is expected to produce greater return in criterion outcome for compensatory model selection than for multiple-hurdle model selection (see Figure 1).

Also, variability in results was greater in both compensatory and multiple-hurdle models when the selection ratio and sample size decreased. This type of variability is not provided in standard utility tables but is important to appreciate. For example, in the most selective and the smallest sample size simulation condition (SR = .10; N = 50), the interquartile range in compensatory model selection was d = 1.22 to 2.22, and the interquartile range in multiple-hurdle model selection was d = 0.48 to 0.85. Relative to the magnitude of the effect, these ranges are roughly similar. Thus, in specific selection situations, the resulting level of criterion performance in the selected subgroup may meaningfully vary from the expected value. What this means for organizations is that utility estimates based on small samples should come with some indication of precision (e.g., a confidence interval) – yet typically, it does not.

For the least selective and the smallest sample size condition (SR = .40; N = 50), we found overlap in the results from the two selection models. Thus, although compensatory model selection is generally expected to produce higher criterion performance in the selected subgroup than multiple-hurdle model selection, this is less true when the overall benefit of selection is smaller, given higher selection ratios and smaller sample sizes.

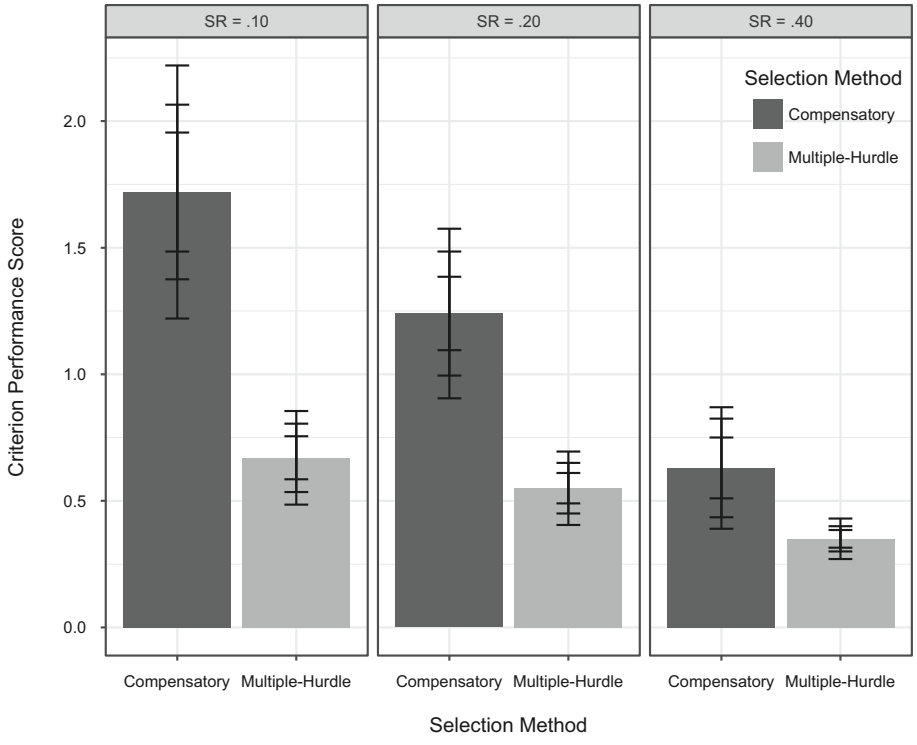

Figure 1. Mean and interquartile range of true criterion performance score for applicants selected using compensatory model selection versus multiple-hurdle model selection on observed predictor test score. Outermost limits of the error bars reflect the interquartile range for N = 50, middle limits of the error bars reflect the interquartile range for N = 100, and innermost limits of the error bars reflect the interquartile range for N = 250.

Impact of Selection Method on Overall Selection Utility

We conducted a hierarchical multiple regression analysis with the utility values as the dependent variable and the main study variables as the predictors, with interaction terms entered in each block. We compared the standardized beta coefficients to examine the relative impact of the main study variables on utility.

Selection utility values were regressed on the following predictors: the selection model used (coded as compensatory = 1 and multiple-hurdle = 0), selection ratio, sample size, SD_y, and cost per applicant. Then, we entered the two-, three-, four-, and five-way interaction terms into the regression equation, each within its own block (total number of conditions was 31). For simplicity, we only included the 10% cost condition to represent the multiple-hurdle model selection condition in the analysis. Thus, the results illustrate the effects of study variables (linear or categorical as appropriate) on utility in two extreme conditions: the most reliable condition (compensatory) versus the most cost-efficient condition (multiple-hurdle at 10% cost of compensatory selection).

Table 5 contains the regression analysis results. The main effects of the predictors (R^2 = .77) and their two-way interactions (R^2 = .96) explained most of the variance in utility values. However, when we included all of the interaction terms into the regression, a few interaction terms accounted for most of the explained variance. Namely, the Sample Size × SD_y interaction term had the strongest impact on utility among the variables we examined (β = .944), followed by Sample Size × SD_y × Selection Model interaction (β = −.588), and Selection Ratio × Sample Size × SD_y interaction (β = .351). These results suggest that the rate of increase in utility was most strongly affected by the combined effect between increase in the sample size (and thus the number of selected applicants who will contribute to the organization), and increase in the monetary value associated with criterion performance. Also, the rate of increase in utility associated with the combined effect of increase in the sample size and the SD_y was greater under compensatory model selection than under multiple-hurdle model selection. Cost had some effect on utility when combined with sample size (β = −.151), but the impact was smaller relative to the other variables. This step for exploring interactions was planned *a priori*; however, most of the variance across simulations was explained prior to the higher-order interaction terms.

Below, we focused on some notable patterns that we observed in comparing the overall selection utility between compensatory model selection and multiple-hurdle model selection. The following website contains an application that allows the users to view a graphical representation of the utility comparison results for all study conditions (https://goo.gl/ZtBZgP). The full range of data (along with the simulation codes) can also be downloaded online (https://osf.io/9cp4w/).

Table 5. Regression analysis results with overall selection utility as the dependent variable

Model	Ordered predictors	β	R^2	ΔR^2
Model 1	SR	.000	.77	.77*
	N	.001		
	Cost	.001		
	SD_y	−.005		
	Model	.001		
Model 2	SR × *N*	−.001	.96	.19*
	SR × Cost	−.001		
	SR × SD_y	.024		
	SR × Model	.000		
	N × Cost	−.151*		
	N × SD_y	.944*		
	N × Model	−.001		
	Cost × SD_y	.000		
	Cost × Model	−.001		
	SD_y × Model	.000		
Model 3	SR × *N* × Cost	.001	.98	.02*
	SR × *N* × SD_y	.351*		
	SR × *N* × Model	.000		
	SR × Cost × SD_y	.000		
	SR × Cost × Model	.001		
	SR × SD_y × Model	−.019		
	N × Cost × SD_y	.000		
	N × Cost × Model	.122*		
	N × SD_y × Model	−.588*		
	Cost × SD_y × Model	.000		
Model 4	SR × *N* × Cost × SD_y	.000	.98	.00
	SR × *N* × Cost × Model	.000		
	SR × *N* × SD_y × Model	.018		
	SR × Cost × SD_y × Model	.000		
	N × Cost × SD_y × Model	.000		
Model 5	SR × *N* × Cost × SD_y × Model	.000	.98	.00

Notes. The entire set of weights is from Model 5. Total number of conditions in the regression was 31. SR = selection ratio; *N* = sample size; Cost = cost per applicant; SD_y = monetary value associated with one standard deviation change in overall job performance, Model = selection model (coded as compensatory = 1 and multiple-hurdle = 0). *$p < .01$.

Consistent with the simulation and results, compensatory model selection generally outperformed multiple-hurdle model selection in terms of overall utility, despite the higher cost associated with the selection procedure (see graphs at https://goo.gl/MDd4vz; for simplicity, we only included the 10% cost condition for multiple-hurdle model selection). However, even in the low selection ratio conditions (i.e., SR = .10, .20), where we found a medium-to-large effect size difference in mean criterion performance between compensatory model selection and multiple-hurdle model selection, multiple-hurdle model selection produced higher utility as the cost per applicant increased, and the monetary return associated with performance decreased (i.e., lower SD_y). For example, across selection ratios and sample sizes, multiple-hurdle model selection at 10% of the cost of compensatory model selection resulted in higher utility when SD_y was between \$6,000 and \$13,000, and multiple-hurdle model selection at 30% of the cost of compensatory model selection resulted in higher utility when SD_y was between \$6,000 and \$10,000 when the cost per applicant reached \$1,600 (although a few exceptions were observed when $N = 100$).

Moreover, because there was greater variability in mean criterion performance scores in compensatory model selection than in multiple-hurdle model selection, the variability in overall selection utility was also greater in compensatory model selection. As a result, when the selection ratio was at .40 (where there was only a small effect size difference in criterion performance between compensatory and multiple-hurdle model selection) and the sample size was at $N = 100$, the interquartile range of selection utility values for compensatory model selection and multiple-hurdle model selection (at 10% of the cost of compensatory model-selection) overlapped even in the highest SD_y condition (i.e., SD_y = \$60,000) across all cost per applicant conditions (see Figure 2 for a sample illustration of the interquartile ranges of selection utility values at SR = .40, $N = 100$, and cost per applicant = \$1,600). In the small sample size condition ($N = 50$), where the variability was greater, the interquartile ranges overlapped across SD_y conditions even when the cost of multiple-hurdle model selection increased to 50% of the cost of compensatory model selection (except when cost per applicant = \$100). Thus, although compensatory model selection is generally expected to produce higher level of selection utility than multiple-hurdle model selection, especially when the SD_y is relatively high, multiple-hurdle model selection may result in higher utility in specific situations.

Additional Exploratory Analyses

Of course, the increased effectiveness of predictor test scores associated with compensatory selection should be limited to the extent that predictor measures with low criterion-related validity or unreliable scores are used. Unfortunately, the use of such poor selection tools (e.g., graphology) is more prevalent in practice than we would hope (König, Klehe, Berchtold, & Kleinmann, 2010).

Thus, we conducted a very limited and exploratory simulation to test how our results might change under a less optimistic (but perhaps also possible, although hopefully not common) selection situation. Namely, instead of cognitive ability, conscientiousness, and biodata, we used social media (SM)-based assessment of job suitability, graphology, and vocational interest as predictors, along with structured interview. The predictors in the exploratory simulation have substantially lower criterion-related validity coefficients compared to the predictors that were used in the main

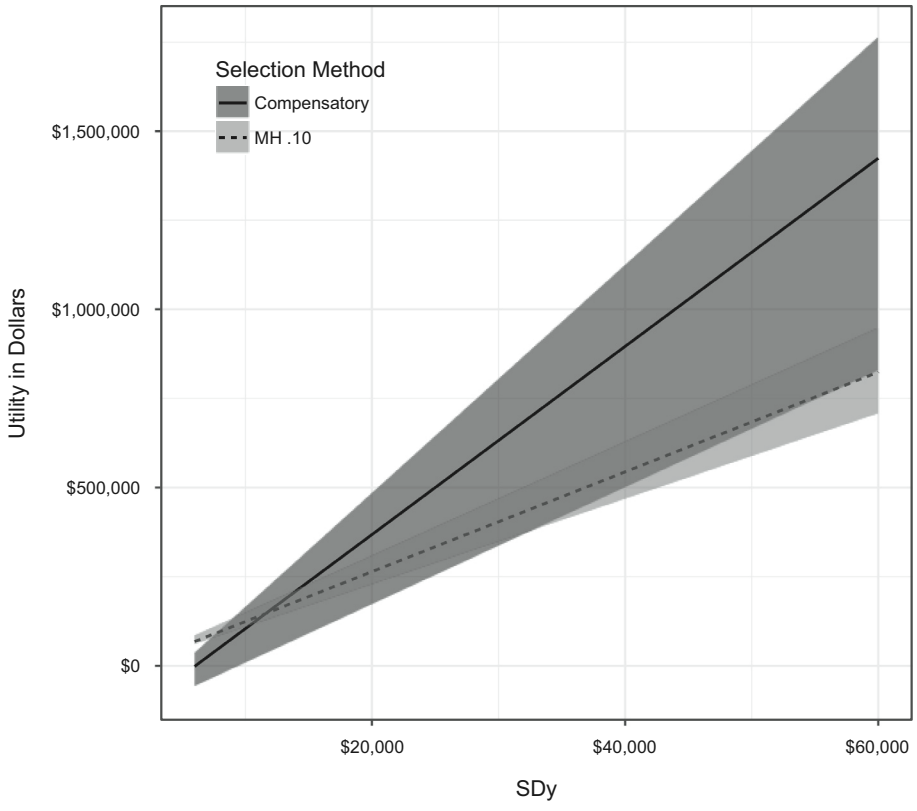

Figure 2. Range of overall selection utility values for compensatory model selection and multiple-hurdle model selection (at 10% of the cost of compensatory model selection) based on interquartile range of criterion performance score for applicants selected in SR = .40, *N* = 100, and cost per applicant = $1,600 condition.

study (r_{xy} = −.07 for SM assessment; Van Iddekinge, Lanivich, Roth, & Junco, 2016, .02 for graphology; Schmidt & Hunter, 1998, and .10 for vocational interest, as measured by the Strong Interest Inventory (Strong, 1965); Hunter & Hunter, 1984). We set the predictor intercorrelations to be small (r = .10) because there was little reason to believe that scores on these measures would be meaningfully correlated. We estimated the population correlation matrix by correcting the observed correlation matrix for measurement error: .93 for SM assessment (Van Iddekinge et al., 2016), .45 for graphology (Rafaeli & Klimoski, 1983), and .80 for vocational interest as measured by the Strong Interest Inventory (Donnay, Morris, Schaubhut, & Thompson, 2004). Then, we ran the simulation following the procedure in the main study.

As expected, with low validity predictors, the relative advantage of compensatory selection over multiple-hurdle selection in terms of mean criterion performance in the selected subgroup was substantially reduced ($\bar{z}_y$ = .99, .77, and .48 for compensatory model selection versus $\bar{z}_y$ = .50, .42, and .26 for multiple-hurdle model selection when SR = .10, .20, and .40, respectively). Consequently, multiple-hurdle selection was more likely to result in higher selection utility. For example, whereas multiple-hurdle selection at 10% of the cost of compensatory selection resulted in higher utility when SD_y was between $6,000 and $13,000 across selection ratios and sample sizes in the main study, with low validity predictors, multiple-hurdle selection at 10% of the cost of compensatory selection resulted in higher utility even when SD_y was as high as $29,000. Even when the cost per applicant was lower, multiple-hurdle selection resulted in higher utility to the extent that SD_y was low. Nevertheless, compensatory selection still produced higher overall utility than multiple-hurdle selection in most conditions (supplementary analysis results are available online at https://osf.io/9cp4w/), indicating that even when low-validity measures are used to make selection decisions, the gain in validity (and ensuing gain in utility) from conducting compensatory selection may outweigh the cost savings from conducting multiple-hurdle selection, especially to the extent that cost per applicant is lower and SD_y is higher.

Discussion

There are fundamental differences in how selection is conducted in compensatory selection and multiple-hurdle selection. Thus, the general choice of the selection model has important effect on which applicants are selected, how selection varies from sample to sample, and what the forecasted level of criterion scores would be for any

given construct – all of which affect selection utility. Namely, despite the psychometric advantages associated with compensatory model selection, practitioners may lean toward multiple-hurdle model selection because administering an entire predictor battery to every applicant is expensive and time-consuming. Any differences in the simulation results thus provide a basis for comparison between a more reliable versus a more cost-efficient selection method.

The simulation results showed that when organizations are selective, a multiple-hurdle selection approach imposes significant effects that cost organizations in terms of criterion performance in the selected subgroup, even where we simulated only two stages, and selection in the first stage involved a composite of three test scores. We suspect these effects would be even greater with more hurdles in the model, and with single tests at each hurdle. But even in our scenario for multiple hurdles, the mean criterion performance scores for applicants were consistently lower than those for applicants selected through compensatory selection.

Similarly, compensatory model selection generally resulted in higher level of overall selection utility, but the advantage was greater as SD_y increased. This means that compensatory selection should be especially advantageous in selections for high complexity jobs. Namely, there is greater variance in productivity and dollar value associated with SD_y in jobs that are higher in complexity (Hunter, Schmidt, & Judiesch, 1990), meaning that selection of best performers will yield greater utility in high complexity jobs than in low complexity jobs. Then, there should be greater upside in utility relative to the increase in costs of the selection process.

However, when the SD_y was relatively low, multiple-hurdle model selection resulted in higher utility as the cost of selection procedure increased. This was true even in the lowest selection ratio simulation condition (SR = .10), where we found a large effect difference in mean criterion performance between compensatory selection and multiple-hurdle selection. Thus, in selection for jobs that has a lower level of expected monetary return associated with criterion performance, multiple-hurdle model selection may be more advantageous than compensatory model selection in terms of overall selection utility, even though multiple-hurdle model selection is expected to result in lower level of mean criterion performance.

Results also showed that in order for multiple-hurdle model selection to produce higher utility than compensatory model selection, multiple-hurdle model selection has to cost substantially less than compensatory model selection. In our study, the utility for multiple-hurdle model selection exceeded that for compensatory model selection only when the cost of multiple-hurdle model selection was set at 30% of the cost of compensatory model selection or lower. It may not be too difficult for organizations to develop a multiple-hurdle model selection process that is considerably less costly than compensatory model selection, especially in any large-scale selection situations that are common in large enterprises. However, it should be noted that cost-saving strategies that involve undermining the psychometric properties of the predictor measurement scores (e.g., use of less reliable, less valid measures) should result in corresponding decline in the selection utility.

Limitations and Future Research

Several limitations of the current study should be noted as indicators of future research on the effect of selection methods on the effectiveness of predictor test scores for hiring the best applicants and ultimately, utility.

First, although there is an increasing need for HR practitioners to demonstrate the value of HR interventions to organizational performance (Cascio & Boudreau, 2011), research suggests that practitioners remain ambivalent about the usefulness of utility analysis (König, Bösch, Reshef, & Winkler, 2013). Generally, HR practitioners rightfully perceive utility analysis equations as relying on unrealistic assumptions (e.g., using random selection as the baseline model) and/or fail to take into account important variables that should be factored into the estimates.

We created simulation conditions that were sensitive and applicable to practice, where we chose a wide range of realistic parameter values as suggested in previous research. Nevertheless, a wider range of parameters could be incorporated. For example, there are costs associated with administering a selection procedure that we did not consider in our utility calculation (e.g., fixed costs, such as salary; variable costs, such as commissions going to high-performing employees; Cascio & Boudreau, 2011). In addition, for jobs where the point of dichotomization between successful and unsuccessful performance is a critical one (e.g., a nuclear power plant operator performing at a level of minimally acceptable competence), the base rate among job applicants is an important determinant of utility that informs the choice between compensatory and multiple-hurdle selection methods (i.e., higher composite score reliability should lead to higher utility when the base rate is low; cost savings in the selection procedure should lead to higher utility when the base rate is high).

All of this is to say that a number of other potential selection scenarios (such as other combinations of selection models and different predictors/criterion) are justifiable for conceptual, practical, and legal reasons. Thus, the specific simulation parameters and the resulting selection utility estimates in the current study should not be viewed as a definitive summary; rather, they should be viewed as an illustrative example of the effect that selection method

could have on selection utility under a relatively specific and limited set of selection situations. Although our results may not generalize to other scenarios, we at least provide the tools that can (hopefully) inspire others to explore other selection and utility scenarios further, with more detail and usefulness than is found in traditional utility tables.

Second, we considered utility in terms of monetary return associated with the selection decisions. Ultimately, organizations use employment tests to make more productive personnel selection decisions, but the effectiveness of measures within a predictor battery is also informed by other factors, such as fairness across subgroups, diversity, and legal exposure. This issue is fundamental and informative in the arenas of science, practice, and litigation, and there are often tradeoffs that can be quantitatively modeled (if not optimized) in large-sample selection systems (De Corte, Lievens, & Sackett, 2007). Future simulation work could consider examining the practical implications of selection methods in a broader framework that incorporates attention toward multiple outcomes of selection, which may also increase the acceptability of the utility analysis results (Macan, Lemming, & Foster, 2012).

Finally, our utility estimates were based on true overall job performance scores, which importantly provide useful estimates of the actual value of a selection system, yet organizational decisions about HR practices and individual employees are inevitably made on the basis of observed performance scores – and multiple dimensions of performance and utility as well. Even though we estimated predictor correlations with "true" performance ratings that were corrected for measurement error variance alone, in the real world, subjective performance measures can also be contaminated by various systematic effects that affect performance ratings, such as rater leniency or other rater idiosyncrasies, especially in the absence of rater training, that will affect the accuracy of utility estimates. Thus, although the current simulation is informative, all of these issues contribute to the real-world difficulties that organizations face when attempting to assess with accuracy the utility associated with any HR intervention.

Conclusion

It should be clear at this point that generally speaking, substantive, practical, and psychometric considerations are all important when deciding among different decision rules for personnel selection. The current simulation results provide an important illustration of the cost-reliability tradeoff between compensatory and multiple-hurdle selection. Specifically, our results suggest that compensatory selection may be especially beneficial for jobs where SD_y is high, and the cost of administrating the test battery to all the applicants is manageable; conversely, multiple-hurdle selection may be beneficial in selection situations where organizations do not need to be very selective, and the cost of test administration is relatively high. We found this pattern of results even when the criterion-related validities of predictor variables were low (and therefore, the advantage of compensatory selection was compromised). By providing these detailed simulation study results, one gains more practical insight on how personnel selection practices can affect the outcomes of selection. This is an important extension of understanding selection practices, beyond managers, researchers, and practitioners, who are typically relying on subjective judgment when assessing and combining ingredients of reliability and validity coefficients, sample sizes, and results from traditional utility tables to reach conclusions about the effectiveness of a selection system. As a profession, we can and should continue to do better than that.

References

Allen, N. L., & Dunbar, S. B. (1990). Standard errors of correlations adjusted for incidental selection. *Applied Psychological Measurement, 14*, 83–94. https://doi.org/10.1177/014662169001400109

Arthur, W. Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435–442. https://doi.org/10.1037/0021-9010.93.2.435

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37*, 64–76. https://doi.org/10.1037/h0061548

Cascio, W. F., & Boudreau, J. W. (2011). *Investing in people: Financial impact of human resource initiatives* (2nd ed.). Upper Saddle River, NJ: Pearson.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380–1393. https://doi.org/10.1037/0021-9010.92.5.1380

Donnay, D. A. C., Morris, M. L., Schaubhut, N. A., & Thompson, R. C. (2004). *Strong Interest Inventory manual: Research, development, and strategies for interpretation*. Mountain View, CA: CPP.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics, 35*, 1–27. https://doi.org/10.3102/1076998609346969

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions* (2nd ed.). New York, NY: Routledge.

Hambleton, R., & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*, 19–38. https://doi.org/10.1207/s15324818ame1001_2

Howardson, G. N., & Behrend, T. S. (2014). Using the Internet to recruit employees: Comparing the effects of usability expectations and objective technological characteristics on Internet recruitment outcomes. *Computers in Human Behavior, 31*, 334–342. https://doi.org/10.1016/j.chb.2013.10.057

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21*, 264–276. https://doi.org/10.1111/ijsa.12036

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*, 28–42. https://doi.org/10.1037/0021-9010.75.1.28

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27*, 179–182. https://doi.org/10.1007/BF02289635

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221–240. https://doi.org/10.1207/s15324818ame1703_1

König, C. J., Bösch, F., Reshef, A., & Winkler, S. (2013). Human resource managers' attitudes toward utility analysis. *Journal of Personnel Psychology, 12*, 152–156. https://doi.org/10.1027/1866-5888/a000090

König, C. J., Klehe, U. -C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment, 18*, 17–27. https://doi.org/10.1111/j.1468-2389.2010.00485.x

Lievens, F., & Harris, M. M. (2003). Research on Internet recruiting and testing: Current status and future directions. In C. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 18, pp. 131–165). Chichester, UK: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Mosley.

Macan, T., Lemming, M. R., & Foster, J. L. (2012). Utility analysis: Do estimates and format matter? *Personnel Review, 42*, 105–126. https://doi.org/10.1108/00483481311285255

Mendoza, J. L., Bard, D. E., Mumford, M. D., & Ang, S. C. (2004). Criterion-related validity in multiple-hurdle designs: Estimation and bias. *Organizational Research Methods, 7*, 418–441. https://doi.org/10.1177/1094428104268752

Mendoza, J. L., & Mumford, M. D. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational and Behavioral Statistics, 12*, 292–293. https://doi.org/10.3102/10769986012003282

Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology, 3*, 33–42.

Rafaeli, A., & Klimoski, R. J. (1983). Predicting sales success through handwriting analysis: An evaluation for the effects of training and handwriting sample content. *Journal of Applied Psychology, 68*, 212–217. https://doi.org/10.1037/0021-9010.68.2.212

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27*, 152–171. https://doi.org/10.1177/0146621602239476

R Core Team. (2017). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Roomsburg, J. D. (1989). *Utility as a function of selection ratio and base rate: An empirical investigation of military aviation selection.* Unpublished doctoral dissertation, University of Texas, Austin, TX, USA.

Roth, P. L., Switzer, F. S., Van Iddekinge, C. H., & Oh, I. S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology, 64*, 899–935. https://doi.org/10.1111/j.1744-6570.2011.01231.x

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology, 55*, 807–825. https://doi.org/10.1111/j.1744-6570.2002.tb00130.x

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*, 538–544. https://doi.org/10.1037/0021-9010.92.2.538

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49*, 549–572. https://doi.org/10.1111/j.1744-6570.1996.tb01584.x

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*, 573–605. https://doi.org/10.1111/j.1744-6570.2003.tb00751.x

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Shaffer, G. S., Saunders, V., & Owens, W. A. (1986). Additional evidence for the accuracy of biographical data: Long-term retest and observer ratings. *Personnel Psychology, 39*, 791–809. https://doi.org/10.1111/j.1744-6570.1986.tb00595.x

Sturman, M. C. (2000). Implications of utility analysis adjustments for estimates of human resource intervention value. *Journal of Management, 26*, 281–299. https://doi.org/10.1177/014920630002600206

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23*, 565–578.

Van Iddekinge, C. H., Lanivich, S. E., Roth, P. L., & Junco, E. (2016). Social media for selection? Validity and adverse impact potential of a Facebook-based assessment. *Journal of Management, 42*, 1811–1835. https://doi.org/10.1177/0149206313515524

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*, 224–235. https://doi.org/10.1177/00131640021970475

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574. https://doi.org/10.1037/0021-9010.81.5.557

Received May 23, 2017
Revision received January 15, 2018
Accepted January 16, 2018
Published online November 2, 2018

Jisoo Ock
14 Nanyang Drive
School of Social Sciences
Division of Psychology
#04-41, Nanyang Technological University
Singapore 637332
jisoo.ock@gmail.com

