Chapter 3: Translating HR Effects into Utility Metrics – An Issue of Communication

The major goal of this chapter is to explain how HR professionals can both estimate and communicate the effect of HR policies/practices on important business metrics to key decision makers. The target audience of these communications might include other HR professionals, but is most often line managers and executives whose primary responsibilities are to attend to these business metric. For example, I know I have not communicated well when a line manager asks "you clearly seem to think a correlation of .35 between the new selection system and subsequent job tenure is a good thing . . . how big does a correlation have to be before it is 'good'?"

Chapter 3 Goal: To learn how to communicate the expected value of HR policies and procedures in terms line managers can understand and related to important business decisions.

Line managers do not have to ask this question if they have also been told the expected effect of using a new selection system which exhibits criterion validity of r_{xy} = .35 is $3000 in increased profit per quarter for each individual hired. Alternatively, depending on the context, other operational metrics may be even more meaningful. For example, telling the line manager "with the old selection system we expected 60 out of every 100 people hired to perform the job adequately, while with the new selection system we expect 75 out of every 100 to reach adequate performance levels" translates a criterion validity of r_{xy} = .35 into an operationally meaningful scale or metric. In a tight labor market line managers intimately understand the implications of trying to meet deadlines and production goals with only 60% of new hires versus 75% performing adequately. Specifically, one would expect to have to hire 125 people under the old selection system to obtain the same 75 adequately performing newcomers that could have been found by hiring 100 people using the new selection system. The new selection system would

save the equivalent of 25% of the compensation budget for new employees (at least for as long as it would have taken to identify which 50 of the 125 did not perform adequately and terminate their employment). If three candidates are considered for each individual hired, the new selection system immediately impacts the line manager's life by reducing the number of job interviews a manager has to sit through from 375 to 300 (a time savings that becomes very real to line managers with packed calendars)!

Why is decision context important and how can it vary? These two examples show one fundamental difference - the latter decision situation involves a choice between competing HR systems (i.e., the new and old selection systems), while the former involves a choice between an HR system as some other capital investment. "Percent of every 100 hired expected to perform adequately" is a metric that can be used to compare competing HR selection systems, though it is not helpful in deciding whether to spend scarce budget dollars on a new selection system versus new tooling for the shop floor (or any other capital investment). The former example communicates expected effect of HR practices on an absolute scale (dollars) that permits direct "apples to apples" comparisons of expected payoffs of monies spent on HR systems to payoffs expected from alternative non-HR uses. While direct "apples to apples" comparisons on a common dollar metric are clearly more helpful in making a wider range of decisions, comparisons using operational measures like "percent expected to perform adequately" can be very useful in making more tactical choices among competing uses of HR budgets.

This chapter becomes even more relevant when business metric used in Chapter 2 are not directly meaningful in the context of the decision being made. Recall in Chapter 2's realtor example that expected realtor <u>total</u> sales volume might be more relevant to a relatively young firm whose critical short term goal is to obtain some minimal market share needed to survive.

Expected average sale price might be more strategically relevant to more mature firms which have turned their focus from survival to profitability. However, all too often HR professionals are faced with situations where the "best" metric available is not clearly related to critical business metrics. For example, what if our only measure of how well the new and old realtor training programs worked in Chapter 2 consisted of subsequent supervisory performance appraisal ratings received at the end of their first year of employment? If the realty firm is faced with spending $250k on a training program versus purchasing a new office information system, knowing the training system is expected to yield higher supervisor performance ratings doesn't help it decide whether to spend the money on training. I may be able to say that new realtors who receive training are expected to be rated 2 points higher by their supervisors, but what does that mean economically – measured in dollars – or operationally – measured in terms of percent expected to perform "adequately" - to the realty firm? If I buy a certificate of deposit at a local bank instead of spending it on the training system, a year later I will have $250k plus interest. What will I have if I spend it on a training system?

Again, one goal of this chapter is to emphasize how important, relevant business metrics are to HR policies and practices as well as describe ways of obtaining them. As noted in Chapter 2, discovery and use of relevant business metrics is highly dependent on creative "outside the typical business metric box" efforts of the HR professionals involved. Consider a vertically integrated retail and manufacturing organization examining HR systems used to select its retail store managers. Store profitability, revenue, and size information would probably be very useful in constructing strategically relevant business metrics. If only district (i.e., aggregate store data within geographic district) profit and revenue data is kept centrally, the diligent HR professional may have to spend part of 6-9 months traveling among geographically disperse district offices to

obtain store-level sales and profitability information needed to evaluate how a new store manager selection system impacts profit, sales revenue, or profit and sales revenue adjusted for store size. Having done this on a number of occasions myself, I can tell you that it provides ample motivation to become best friends with the firm's IT personnel. Usually a small modification of accounting or operational information record keeping and retrieval will eliminate (or minimize) the need for future labor intensive data acquisition efforts. The key is simultaneously identifying the array of business metrics that might be of interest and figuring out who the individuals are in accounting or MIS who can make change happen in the firm's data gathering and reporting systems so measures of those business metrics can be easily and accurately obtained.

Utility

A literature has evolved over the last 60 years that broadly examines how HR policies, practices, and interventions (indeed, how all organizational interventions) have operational and/or economic impact on organizations. Two different approaches from the "utility" literature will be discussed in here. The Taylor-Russell model was first presented in 1938 and is useful for comparing HR alternatives in terms of percent of individuals expected to perform adequately. The second, known as the Brogden-Cronbach-Gleser (BCG) model, was first presented by Brogden in the late 1940's (Brogden, 1947, 1948) and substantially extended in a textbook authored by Cronbach and Gleser in 1965. The BCG model estimates the dollar impact of HR alternatives. In fact, the Taylor-Russell model is a special case of the BCG model when the dollar business

What is adequate? Use of the Taylor-Russell model begs the question of what constitutes adequate performance. It is usually easy to identify what constitutes clearly adequate or inadequate performance. The nature of some tasks lend themselves to discrete adequate/inadequate cut offs – an apartment complex owner/manager either does or does not obtain enough monthly revenue to cover the mortgage and costs of operation from efforts to achieve an optimal combination of occupancy and rent level. At worst, one must rely on judgments of subject matter experts (SMEs) about what constitutes "adequate" performance.

metric (Y) has been recorded as either adequate or inadequate (e.g., when sales personnel have either meet or not meet a dollar sales goal). I will start with the simpler Taylor-Russell approach before building to the more complex BCG model. After getting the reader comfortable with both approaches, I will briefly discuss the distinction between tactical and strategic utility. I will end with a discussion of what hurdle prevented the BCG model from being used from 1949 until the early 1980's, and what was done to overcome that hurdle.

Taylor-Russell Model

Drs. Taylor and Russell (no relation) developed one of the first ways of measuring personnel selection utility (Taylor & Russell, 1939).[1] As before, I will first describe a real HR application of the Taylor-Russell model before describing its assumptions and limitation.

Imagine you are employed by the Federal Aviation Administration and are responsible for selecting individuals to replace air traffic controllers fired by President Regan in August, 1981. Because of a labor contract agreed to with the Professional Air Traffic Controller's Organization (PATCO), the only measure of air traffic controllers' job performance is an immediate supervisor's assessment of whether or not s/he is performing as a "fully functional" air traffic controller. Anyone not considered to be performing as a fully functional air traffic controller is immediately removed for his/her position for remedial training or termination. Further, virtually no air traffic controllers with more than five years of job tenure are ever identified as not performing at the fully functional level. Hence, while air traffic controller job performance may in fact vary along multiple dimensions represented by many distinct performance levels, for administrative purposes the union contract recognizes only two performance levels – fully versus not fully functional air traffic controller. Virtually all air

[1] The Taylor-Russell model can be used to assess the utility of any method used to select any resource – personnel or otherwise - used in production of a firm's goods or services.

traffic controllers identified as performing at non-fully functional levels do so in their first five years on the job.

As part of your responsibilities, a number of "experimental" personnel selection tests were administered to newly hired air traffic controllers over the last 5-7 years. "Experimental" means you gave the tests to all applicants at the same time any personnel selection tests in actual use were administered, i.e., before job offers were made or accepted. Applicants were not told these additional tests would not be used to make employment decisions, so applicant motivation was the same for the "real" and the "experimental selection tests. Two final assumptions for purposes of this semi-fictional example are 1) job offers were made "top-down," i.e., those with the highest scores on the personnel selection test in actual use were made job offers and 2) all job offers were accepted (i.e., not one turned down a job offer, causing the FAA to then make offer the same job to the applicant scoring next lower on the list). As a result of trial testing effort conducted between 1971 and 1976, in September, 1982, of our fictional example (one month after President Regan fired over 20,000 air traffic controllers for going on an illegal strike) you have seven columns of information in an Excel spreadsheet on 500 air traffic controllers hired between '71 & '76. The first column contains their scores on the personnel selection test actually used to select air traffic controllers from 1971 to 1976. Columns 2-4 columns contain test scores from three experimental personnel selection tests administered but not scored or used for those applicants. The fifth column indicates whether each air traffic controller was ever deemed to be performing at a non-fully functional level during her/his first five years of job tenure. The sixth column contains their dates of initial employment, while the seventh column contains the date they were rated as not performing at the fully functional level. Only 75 (15%)

had been rated as performing as non-fully functioning air traffic controllers (the seventh column was blank for the remaining 525).

The first thing we want to know is whether any of the test scores (experimental or not) predict whether the air traffic controller is rated "non-fully functioning" at some later point on the job, or what industrial psychologists call criterion validity. So, we open the spread sheet in Excel and initially create a "dummy" criterion variable Y in an eight column. All 75 rated as non-fully functional are assigned a "0," while everyone else is assigned a "1." We then choose four blank cells and use the COR function in Excel to calculate the Pearson product moment correlation coefficient, or "simple correlation r_{xy}" between the four personnel selection test scores (i.e., the actual test used and the three experimental tests) and Y. Let's say the following criterion validity correlations are derived:

- Current Selection Test: $r_{xy} = .15$
- Experimental Test 1: $r_{xy} = .20$
- Experimental Test 2: $r_{xy} = .25$
- Experimental Test 3: $r_{xy} = .30$

Point Biserial Correlation. Before computers, statistics had to be computed by hand using formula. The formula for the Pearson product moment correlation, or "simple" correlation r_{xy}, becomes arithmetically much simpler when either X or Y is "truly" dichotomous. This "simpler" simple correlation was called a point biserial correlation (who knows why). Since computers don't care whether the formula is arithmetically simpler or not, software is generally written only to estimate Pearson product moment correlations, even though authors will often refer to this output as point biserial when X or Y are dichotomous. I suspect it will take the death of my generation who were forced to learn statistics using hand computation before the "point biserial" label falls from use. Note, Y in this example would not be "truly" dichotomous if supervisors rated air traffic controllers' performance on a 1-9 point scale, then labeled all those with ratings of 3 or less as "non-fully functional" and all those with ratings of 4 or more as "fully functional." In this case the fully functional/non-fully functional distinction would not be "Ture" because it was artificially created, i.e., it could have been created by labeling those with ratings of 2 or less as non-fully functional and those with ratings of 3 or more as fully functional. Point biserial correlations behave in known ways when X or Y is truly dichotomous. They behave in unknown ways when X or Y are artificially dichotomized (tetrachoric correlations, something yet again different from r_{pb} behave better in this instance). As we will see in the next section, when X and Y are continuous, it would be simpler to use a different model (i.e., the Brogden-Cronbach-Gleser model) instead of the Taylor-Russell model, avoiding complexities caused by having to use tetrachoric

Our first reaction might be "wow, ET3 is correlated 2 times better with fully/non-fully functional status when compared to our existing test!" That sounds good, but most line

managers (and especially government politicians!) will have no idea what that means. So what does it mean?

This is where the Taylor-Russell model can help. While we will see later that the correlation has a lot going for it as a way of quickly communicating how well a test predicts job performance, most people, including most HR professionals and line managers, don't have a great intuitive understanding of what r_{xy} = .30 means. What Drs. Taylor and Russell did was develop a way to translate information contained in r_{xy} into something most folks could understand. What might that be? The answer is "likelihood that those selected using the test will perform the job adequately," or in the air traffic controller example, likelihood that the air traffic controller will perform as a fully functioning air traffic controller over her/his first 5 years on the job. Clearly, other things being equal (e.g., cost of the test), we would want to use the test that maximizes this likelihood. Again, other things being equal, simple examination of the criterion validity correlations reported above will always answer this question – the test with the highest criterion validity will always be expected to select a high proportion of air traffic controllers who perform at fully functional levels through their first five years of employment. Well, that lets us rank personnel selection tests, but does not tell us how much better one is than another. Imagine Experimental Tests 1, 2, and 3 (ET1, ET2, & ET3) cost $15, $25, and $50 per applicant, respectively. We get an increase of .05 in r_{xy} in going from ET1 to ET2 at the cost of an additional $10 per applicant. We get yet another increase of .05 in r_{xy} going from ET2 to ET3 for $25 more per applicant. Is it worth the extra $10 per applicant to use ET2 over ET1 and is it worth the extra $25 to use ET3 over ET2 (or extra $35 to use ET3 over ET1)?

Taylor and Russell helped us answer this question by showing us how to estimate the exact likelihood that applicants chosen with a test would perform adequately on the job (hold on

for a moment . . . I am getting to the explanation of how that happens, be patient). If a test with r_{xy} = .25 is expected to result in 67% of those selected performing adequately and selecting applicants at random is expected to result in only 50% performing adequately, then we know use of the test is likely to get us 17 more adequately performing employees out of every 100 hired when compared to random selection. While one would hope most organizations do not make job offers at random, percent of those randomly selected performing adequately does provide a nice common base line or point of reference.

Fortunately, Taylor and Russell's method also permits direct comparison between tests. If ET2 with r_{xy} = .25 resulted in 67% expected to perform adequately, while ET3 with r_{xy} = .30 resulted in 71% expected to perform adequately, we would be able to say that the extra $25 cost per applicant for ET3 is expected to result in an extra 4 fully functioning air traffic controllers out of every 100 hired. Approximately 200,000 individuals applied for the openings created by President Regan, of whom 20,000 needed to be hired. So, 200,000 x $25 = $5,000,000 more would be needed to use ET3 over ET2, resulting in 4% x 20,000 = 800 more fully functional air traffic controllers expected to be selected compared to use of ET2. Another way of saying it is that use of ET3 is expected to generate 800 more fully functioning air traffic controllers than ET2 at the cost of an extra $6250 each ($5,000,000 ÷ 800 = $6250). Now we are using language organizational decision makers can understand! However, the Federal Aviation Administration is a federal agency, so the HR decision maker's constituency might include both line managers and politicians. What "language" or business metric would be most meaningful to political decision makers?

Let's assume assessments of whether air traffic controllers are performing at fully functional levels are made without error – no fully functioning air traffic controllers are mistakenly rated as non-fully functioning, and vice versa. Further assume FAA investigators could identify correctly every time an air traffic controller made a mistake that made her/him "non-fully functional" and any loss of life caused by that mistake. Use of ET2 would be expected to result in 800 more air traffic controllers selected who would make a mistake revealing them as non-fully functional in their first five years on the job. If the average loss of life per air traffic controller mistake was k = 1.2 (a number I pulled out of thin air just for this example), then mistakes made by these 800 air traffic controllers in revealing themselves as non-fully functional would be expected to result in 800 x1.2 = 960 lives lost. We have now translated estimates of r_{xy} = .30 and r_{xy} = .25 for ET3 and ET2 (respectively) into an estimate of 960 lives not lost when ET3 is used for air traffic controller selection, a non-financial metric that even politicians should be able to understand!

Merging Languages. Of course, one could also estimate the cost incurred in using ET3 per expected life saved, effectively combining metrics of the manager ($) and politician (lives), i.e., $3759.40 ÷ 1.2 = $3132.83 per life saved. As goulish as this might sound, this type of computation is made in many other contexts. For example, engineers routinely estimate the expected accident and death rate per mile driven for alternate highway designs. In comparing two alternate designs for a long left hand turn one could easily estimate both the expected cost and expected loss of life per mile driven for each design. Taking a ratio of expected increase in cost divided by expected decrease in number of lives lost would yield expected increase in construction cost per traffic death avoided. Then comes the interesting part . . . when does it become too expensive to choose the better design? At $1,000 per life saved, $10,000, $100,000? How much money is reasonable to ask tax payers to fork over to save lives through more expensive highway construction or better air traffic controller selection? This is why this is a political decision!

Criterion validity, selection ratio, and base rate

Criterion Validity. So, now that I have yet again teased you with how an HR professional could use the Taylor-Russell model to evaluate the impact of competing selections systems, let's see exactly how this is done. Drs. Taylor and Russell showed how three pieces of information describing a personnel selection situation can be used to estimate the likelihood of adequate

performance. Not surprisingly, the selection test's criterion validity, or the r_{xy} correlation between applicants' test scores (X) and whether they subsequently perform adequately (Y), is one of these three pieces of information. The correlation r_{xy} directly reflects how well an applicant's test score (X) predicts whether s/he subsequently performs adequately on the job(Y). As noted above, "other things being equal," higher criterion validities result in higher likelihoods of adequate performance among those selected, though you can't know exactly how high the likelihood is of adequate performance without considering those other things. The "other things" that also impact likelihood of adequate performance when a selection test is used are the selection ratio and base rate.

Selection Ratio. The selection ratio is equal to the number of positions open divided by the number of applicants.[2] Number of positions open reflects the number of job offers the employer needs to make and have accepted to meet customer demand. Unfortunately, it has not always been clear when a person becomes a job applicant. In March, 2004, the Office of Federal Contract Compliance Programs and Equal Employment Opportunity Commission jointly issued guidelines regarding what constitutes an "internet applicant."[3] Most controversy surrounds how minimum qualifications are used and whether individuals not meeting minimum qualifications should be considered applicants. For the moment, let us assume this is a non-issue – we will revisit it in more detail in the chapter devoted to recruiting – and that a simple count of number of applications received will work (electronic or otherwise). Hence, if 100 individuals apply for 10 open positions, 10 ÷ 100 = .10 or a 10% selection ratio. Likelihood of performing adequately

[2] Note, we will assume all open positions can be filled. Situations where this is not true generally constitute extreme situations in where questions of utility are irrelevant. For example, when the number of applicants is less than the number of open positions, a selection test is not needed – job offers should be made to all applicants who are not clearly incapable of performing the job (e.g., applicants who state on their application that they would be unable to work the required schedule).

[3] Questions like "should people who simply point their browsers at an on-line application without filling it out be considered an applicant?" are answered in this new guideline.

goes up, other things being equal, as the selection ratio gets smaller. What this means in plain language is that the more applicants you have, the more likely it is that a personnel selection test will be able to find ones who can perform the job adequately.

Base Rate. Finally, base rate is typically defined as the percentage of applicants in the applicant population who can do the job adequately, or the percentage expected to perform adequately when job offers are randomly made to applicants. At one extreme, a base rate of 0 means no applicants are capable of performing the job adequately. At the other extreme, a base rate of 1.0 means every applicant is capable of performing adequately. Hopefully it should be clear that no selection test can improve on these base rates. If no applicants are capable of performing the job, the selection test is not going to find one who is. Similarly, if all applicants are able to perform the job adequately, the selection test is not tell you who can't do the job because they all can.

What is less obvious is that as the base rate approaches .50 (again, holding selection ratio and criterion validity constant), the opportunity for the test to affect the likelihood of selected applicants performing adequately goes up. This is because a selection test can contribute most in situations that are most uncertain – the firm is most unsure about who to select when there is a 50/50 chance an applicant will perform adequately (it will be wrong 50% of the time if it guesses either way). If the base rate is 60%, the firm can be sure of only being wrong on average 40% of the time if it guesses all applicants will perform adequately and hires all it needs randomly from the applicant pool. A selection test has the opportunity of minimizing selection mistakes the most in circumstances where the most mistakes are likely to be made, i.e., the closer the base rate is to 50%.[4]

[4] Statistically, the variance of a dichotomous variable (e.g., Y = adequate/inadequate performance ratings) is a function of p times 1 - p, where p and 1 - p are the proportion of adequate and inadequate performers. The

Examples. There is a formula originally published by Taylor and Russell (1939) that shows how criterion validity, selection ratio, and base rate are combined to tell us how likely it is that applicants selected with a test under those conditions will perform adequately on the job. The good news is that I will not reprint the formula here because Drs. Taylor and Russell were kind enough to publish tables containing the results one would obtain from this formula for many common combinations of criterion validity, selection ratio, and base rate. I will first use these tables (reprinted at the end of this chapter for your convenience) to show readers where the information used in the air traffic controllers example came from. I will then walk through some other common selection scenarios in which the Taylor-Russell tables can be useful.

Note that nine tables were published by Drs. Taylor and Russell in 1939, one corresponding for base rates = .10, .20, .30, .40, .50, .60, .70, .80, & .90 (labeled "Percent of Employees Considered Satisfactory"). Each table column corresponds to selection ratios of .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, & .95. Finally, each row corresponds with 21 different criterion validities (r_{xy}) ranging from .05, .10, . . . to .90 & .95. So, in the air traffic controller example above, I said the FAA knew in advance that randomly selecting air traffic controllers from the applicant pool would yield about 50% who would end up performing at a fully functional level, so base rate = .50 and we find the table titled "Proportion of Applicants Considered Satisfactory = .50." Recall there were 200,000 applications for 20,000 openings, so the selection ratio was 20,000/200,000 = .10, so we go to the second column in the main body of the table that has the designated selection ratio ".10" above it. Finally, recall ET2 and ET3 were criterion validity correlations of r_{xy} = .25 and .30, respectively. So, we run our thumb down the .10 column, stopping at the rows labeled .25 and .30 on the far left side. For ET2, the Selection

---

highest product of p(1 – p) occurs when p = .50. The greater the variance in Y, the more help X can be in predicting Y. This becomes obvious in the next section of this chapter when we consider selection situations where performance is measured on a continuous scale, e.g., sales volume generated by retail sales clerks.

Ratio = .10, and r_{xy} = .25 location in the Base Rate = .50 table brings our finger to the .67 value in the body of the table. Hence, this is how I was able to say above that out of every 100 air traffic controllers hired, ET2 was expected to select 67 who performed at a fully functional level (though in introducing this example, I failed to mention there are 10 applicants for every opening - SR = .10 - and 50% of the applicants are likely to be able to perform the job). For ET3, the Selection Ratio = .10 and r_{xy} = .30 location in the Base Rate = .50 table brings our finger to the .71 value in the body of the table. This why I again said in my introduction to the example that, out of every 100 air traffic controllers hired, ET3 is expected to select 71 who perform at a fully functional level (when there are 10 applicants for every opening (SR = .10) and 50% of the applicants are likely to be able to perform the job).

Finally, recall I earlier said the difference between these two expected success rates is .71 - .67 = .04 and 20,000 new air traffic controllers were being hired, ET2 was expected to select .67 x 20,000 = 13,400 new air traffic controllers who will perform at the fully functional level their first five years on the job. ET3 was expected to select .71 x 20,000 = 14,200 new air traffic controllers who will perform at the fully functional level their first five years on the job, or 800 more than ET2. See, it wasn't that hard. No equations . . . not even a Greek letter!

Of course, someone still has to make a final judgment call as to whether the higher cost of ET 3 ($25 per test for a total of an extra $5,000,000 needed in the testing budget) is justified. Are the additional 800 fully functioning air traffic controllers worth the extra $6250 each once costs? Are the extra 960 passengers expected not to die because of air traffic controller mistakes worth the $5,000,000 extra needed to use ET3 instead of ET2? Hopefully you will all agree that these questions are much more palatable to line managers, executives, and CEOs than "is the extra r_{xy} = .05 gained by using ET3 worth $5,000,000 more than the cost of ET2?" However, ask

me whether saving an expected 960 passenger lives is worth an extra $5,000,000 – making judgment calls like this is why line managers, executives, and CEOs make the big bucks! In my experience, HR is lucky to be asked to the table when these kinds of decisions are made, as they usually involve trade-offs with non-HR expenditures. For example, if the $5,000,000 needed for ET3 could alternatively be used for a new piece of equipment, I can guarantee HR will not be at the table if all it can say is "$5,000,000 spent on a new air traffic controller selection test – ET3 – will get us a criterion validity correlation that is r_{xy} = .30, which is .05 higher than the next best test." I have literally heard line managers say "boy, you HR people sure talk funny" in response to statements like this. The good news is that HR might be invited into the conversation if it can say "$5,000,000 extra for the best air traffic controller selection test will get us 800 more fully functional air traffic controllers vs. the next best selection test. They would replace 800 non-fully functional air traffic controllers that the next best test would have selected, saving on average 960 passenger lives in their first 5 years on the job."

Unfortunately, while a lot better than spouting "for $5,000,000, criterion validity r_{xy} will go up .05," it still leaves line managers with an apples to oranges kind of comparison. Comparing the dollar figure for the net present value of an equipment capital investment to 800 more fully functional air traffic controllers (or 960 few passenger deaths) does portray each option in a way line managers can understand. It does not, however, use the same underlying metric for comparison. The next section introduces a method for estimating how much value a new selection system is expected to add in dollar terms.

Brogden-Cronbach-Gleser

The goal of this section is to start with concepts and arithmetic that students are comfortable with from high school and, hopefully, overcome intimidation and confusion caused by complex terms and Greek symbols typically used in traditional statistical treatments of this topic. If HR professionals are to

credibly communicate the expected dollar value added of an HR policy or procedure, they must overcome their aversion to statistical jargon and communicate what they find to line managers, executives, and CEOs in a way that does not confuse and intimidate them!

Step 1: Remember High School?

Recall the formula for a straight line that most of you were first exposed to in high school geometry classes, i.e.:

$$y = a + bx$$

Equation 1

where,

- y = the dependent variable, or typically the criterion performance measure of interest
- x = the independent variable, or typically the thing we hope predicts our performance measure
- a = the "y intercept," of where the line crosses the y axis of a graph where x = 0
- b = the slope, or "rise over run," of the line, i.e., the ratio of vertical line length divided by horizontal line length when a right triangle is created on any portion of the line.

First, let's change this equation slightly. Let's substitute the symbols b_0 for a, b_1 for b, and $y_\$$ for y. In this way we go from Equation 1 above (and reprinted below) to Equation 2.

$$y = a + bx$$

$$\hat{y}_\$ = b_0 + b_1 x$$

Equation 2

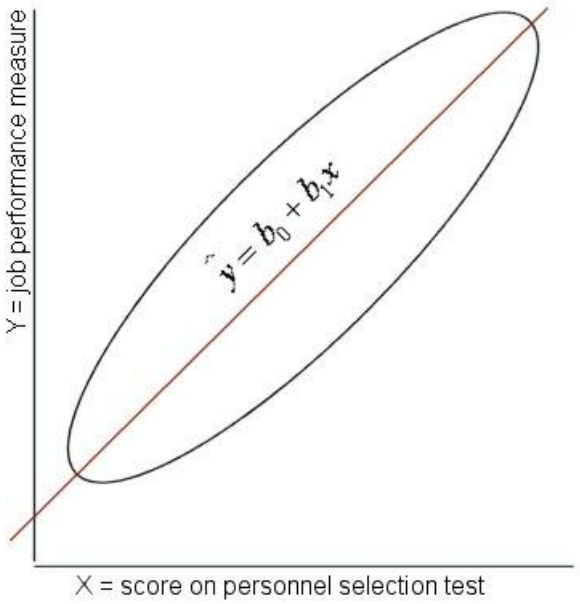
Then, let's add an e after the x, or . . .

$$y_\$ = b_0 + b_1 x + e$$

Equation 3

Here, "e" stands for random error. While our original formulae (Equations 1 & 2) described the points that fall on a straight line, Equation 3 describes points that fall around a straight line. For example, Figure 1 below shows a straight line passing through a cloud of points in the shape of an ellipse, or as my daughter used to say, a hot dog on a stick!

Figure 1



Equations 1 & 2 are formulae describing the values of x and y that fall on the stick. The stick crosses the y axis at b_0. Equation 3 is a formula that describes points that make up the content of the ellipse or hot dog around the stick. Each point in the cloud represents an individual applicant's score X on some personnel selection test and her/his subsequent job performance measure Y, while the cloud of all points can represent some sample or population. The Pearson product moment correlation (r_{xy}) between X and Y directly reflects how "fat" the ellipse-shaped cloud of points is around the line, or how much prediction error e_i we can expect between $\hat{y}_i$ and y_i.[5] If we don't know yet how someone is going to perform on the

[5] In fact, $r_{xy} = \sqrt{1 - \left(\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-1} \right)}$, or the square root of one minus the sum of squared errors of prediction. If we had not taken the square root we would have had $r_{xy}^2 = 1 - \left(\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-1} \right)$. r_{xy}^2 is often called the "coefficient of determination" and can be interpreted as the percentage of variance in Y explained by X.

job (which we can't know before s/he is hired), then one guess or estimate of how s/he might perform on the job would be the $\hat{y}_{\$}$ value obtained from plugging the applicant's x score into Equation 2. The letter "e" in Equation 3 is called "error," because while our estimate $\hat{y}_{\$}$ from Equation 2 might be a good guess, it is not likely to be exactly the level of job performance obtained by that applicant later on the job. Note, as $y_{\$}$ is the actual performance attained by that applicant, $\hat{y}_{\$} - y_{\$} = e$, or the "error" by which our original predicted job performance $\hat{y}_{\$}$ was off from the applicants actual job performance $y_{\$}$. Ordinary least squares regression analyses give us the formula for the "best" fitting straight line (i.e., Equation 2), where "best" means the formula for the straight line $\hat{y}_{\$} = b_0 + b_1 x$ (Equation 2) that minimizes how big the sum of all squared errors (e^2) will be across people in the sample. This is where the "least squares" portion of the "ordinary least squares" label comes from!

Step 2: Standardize x

Recall our ultimate goal is to know what the financial impact in dollar terms is when a firm uses a personnel selection system in hiring employees. To do this, let's assume our measure of job performance is already in dollar terms (a big assumption, but bear with me). Examples of this might include sum of all sales made during a week/month/quarter by selected sales personnel, profit obtained from retail operations managed by selected store managers, or sum of all outstanding debts paid during a week/month/quarter for each call center collection agent selected. Starting with Equation 2 reprinted below . . .

$$\hat{y}_{\$} = b_0 + b_1 x$$

Equation 2

Let's first standardize all the applicants' test scores, i.e., take each applicant's original score on the test, subtract the mean, and divide by the standard deviation, or . . .

$$z_i = \frac{x_i - \bar{x}}{SD_x}$$

Equation 4

where:

- x_i = the test score earned by applicant i.
- z_i = the "standard" or z score for applicant i.
- $\bar{x}$ = the sample average or mean, typically of all applicants obtained in some sample.
- SD_x = the standard deviation of x_i around $\bar{x}$, or $SD_x = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{N-1}}$

Equation 2 modified to reflect the fact that x is now standardized becomes . . .

$$\hat{y}_\$ = b_0 + b_1 z$$

Equation 5

Step 3: Some final substitutions

Finally, let's switch gears a little, modifying Equation 5 to show the financial impact expected from use of the test in selecting a group of applicants. Specifically, we want to know the expected value of $\hat{y}_\$$ for all applicants hired. Recall the words "expected value" in statistics typically refers to the arithmetic average, or mean. So, modifying Equation 5 again we have . . .

$$E(\hat{y}_\$) = E(b_0) + E(b_1)E(z_s)$$

Equation 6

Where E($\hat{y}_\$$) means the "expected value of performance y in dollar terms." Also note that the letter "s" is now subscripted to the letter z. This occurs because the only value obtained by the firm comes from applicants who are actually selected, hence, the subscript "s" stands for "selected." As "expected value" typically means "average," so $E(y_\$) = \bar{y}_\$$ and $E(z_s) = \bar{z}_s$. So, substituting these values into Equation 6 yields . . .

$$\bar{y}_{\$} = E(b_0) + E(b_1)\bar{z}_s$$

Equation 7

Of course, we can't stop here! We have to substitute something in place of $E(b_0)$ and $E(b_1)$. Why? Because ultimately we want to be able to actually put numbers into the right side of Equation 7 to derive an estimate of the financial impact, i.e., dollar utility, of using the personnel selection system. We can calculate $\bar{z}_s$ simply by standardizing the selection test scores of all applicants, then averaging just those scores of the individuals who were actually selected (hence the subscript s). When no selection system is used (i.e., if applicants had been chosen at random), $\bar{z}_s$ is expected to be the same as the average of z scores for all applicants. Since the average of all z scores is always 0 (prove this for yourself using Equation 5 when the individual's score x_i happens to be equal to the average), when $\bar{z}_s = 0$ then $E(b_1)\bar{z}_s = 0$ too, and what is left - $E(b_0)$ - will be the average dollar performance of individuals selected at random from the applicant pool. The symbol for expected or average dollar performance for everyone in the applicant pool is $\mu_{\$}$, so we can substitute $\mu_{\$}$ for $E(b_0)$ in Equation 7.

$$\bar{y}_{\$} = \mu_{\$} + E(b_1)\bar{z}_s$$

Equation 8

Finally, the value of $E(b_1)$ can be obtained from use of multiple regression software (e.g., the regression function in Excel). This is the regression coefficient or beta weight associated with x (as opposed to the "constant," which is the estimate of $E(b_0)$). For reasons that will become clear, it is often useful to substitute for $E(b_1)$ also. Specifically, the sample regression coefficient or slope estimate is also defined as follows . . .

$$b_1 = r_{xy}\left(\frac{SD_y}{SD_x}\right)$$

Equation 9

where:

- r_{xy} = the simple Pearson product moment correlation between test scores on the personnel selection test x and the measure of job performance y.
- SD_y = the standard deviation of job performance measured in dollars
- SD_x = the standard deviation of all applicant's test score performance

However, recall that we standardized applicant test scores in Equation 4 to create the z variable used in Equation 5. So, instead of $b_1 = r_{xy}\left(\frac{SD_y}{SD_x}\right)$, $b_1 = r_{xy}\left(\frac{SD_y}{SD_z}\right)$. The standard deviation of z scores is $SD_z = 1.0$ (again, prove it to yourself by setting x_i equal to a value that is one standard deviation higher than the mean, i.e., making $x_i = \bar{x} + SD_x$). Substituting 1 for SD_x, Equation 8 becomes $b_1 = r_{xy}SD_y$.

So, substituting $\mu_\$$ for $E(b_0)$ and $r_{xy}SD_y$ for b_1 in Equation 8 we get . . .

$$\bar{y}_\$ = \mu_\$ + r_{xy}SD_y\bar{z}_s$$

Equation 10

. . . where $\bar{y}_\$$ is the average dollar value of the work accomplished by those selected. Subtracting out the cost of testing (C) an applicant we get an even better estimate of total dollar value added per applicant selected of . . .

$$\bar{y}_\$ = \mu_\$ + r_{xy}SD_y\bar{z}_s - C$$

Equation 11

Making a final change to reflect the number of applicants selected (N_s) and tested (N_a) we get the total dollar value added from N_s newcomers selected from N_a applicants:

$$N_s\bar{y}_\$ = N_s(\mu_\$ + r_{xy}SD_y\bar{z}_s) - N_aC$$

or

$$U_{total} = N_s(\mu_\$ + r_{xy}SD_y\bar{z}_s) - N_aC$$

Equation 12

Note, Equations 10 to12 focus on the <u>total</u> dollar value added from work performance of those selected using some personnel selection system. They do not tell us how much of that performance was

due to use of the personnel selection system. The portion of the total dollar value added by those selected that is due to the personnel selection system is usually called the utility of that selection system. The utility or dollar value added to the firm due to use of the personnel selection system by the N_s individuals selected can be estimated by subtracting $\mu_\$$ from both sides of Equations 11 & 12. Recall $\mu_\$$ is the dollar value of work performance the firm expects to get when it chooses applicants at random (i.e., what it would have received without use of the selection test), hence, $\bar{y}_\$ - \mu_\$$ is equal to the dollar performance gain resulting from use of the selection procedure, or . . .

$$N_s(\bar{y}_\$ - \mu_\$) = N_s r_{xy} SD_y \bar{z}_s - N_a C$$

Equation 13

Equation 13 is often written as . . .

$$\Delta U = N_s r_{xy} SD_y \bar{z}_s - N_a C$$

Equation 14

. . . where ΔU is the expected change in utility in dollar terms expected due to use of the personnel selection system to select N_s new hires from N_a applicants.

In sum, Equation 14 tells us the net dollar impact a selection system has, while Equation 12 equals the gross or total expected dollar impact of selecting N_s new hires from N_a applicants. Cronbach and Glaser (1965) extended Brogden's (1949) model to two-stage and multi-stage selection, fixed treatment selection, placement, and classification decision situations (as one might imagine, the formulae get more complicated). Fixed

Taylor-Russell Revisited. As I noted at the beginning of the chapter, the Taylor-Russell model is really just a special case of the BCG model. Taylor-Russell is simpler since Y is measured as "adequate" or "inadequate" performance. The portion of the BCG that captures dollar value added is $r_{xy} SD_y \bar{z}_s$ has three parts that correspond directly to the criterion validity, base rate, and selection ratio components of the Taylor-Russell model. Specifically, criterion validity r_{xy} is identical for both approaches. SD_y is the BCG model's way of capturing base rate - recall our base rate discussion described how it reflected the degree of uncertainty or variability in Y, and base rate = .50 had the highest variability. SD_y in the BCG model is a direct measure of variability in Y when Y is a continuous (as opposed to dichotomous) variable. Finally, $\bar{z}_s$ is the BCG parallel of the selection ratio. Specifically, as selection ratio goes down (i.e., the smaller the proportion of applicants selected), $\bar{z}_s$ goes up. Prove this to yourself by drawing a bell shaped curve on a piece of paper with 3-4 vertical lines through it representing "cut" scores (applicants above the cut score are hired, those below are not). The selection ratio will go down as cut scores move from left to right. Since z scores get larger moving left to right, the average z score of those falling to the right of the cut score (i.e., the average of those selected, $\bar{z}_s$) goes up as the cut score moves from right to left. In fact, $\bar{z}_s$ can be arithmetically derived from the selection ratio and the height of the normal curve at the cut score if test scores are normally distributed.

treatment selection, described in more detail below, involves use of this model in situations where the HR manager assigns people to "treatments." Personnel selection situations are typically referred to as random effects treatments, as applicants assign themselves scores based on their answers to test questions (i.e., actual scores observed in a group of applicants are the result of random sampling from the population of all individuals who might have applied). In the realtor training example from Chapter 2, the HR director decided who would receive new versus old training and, hence, assigned (or fixed) employees in one training assignment versus another (where X = 0 for old realtor training and X = 1 for new realtor training). Importantly, this means any discrete action taken by the firm whether it involves HR policies and practices or not with the purpose of influencing $y_\$$ can be assessed in terms of its expected dollar utility.

Discrete changes such as introduction of a new training program require some relatively simple changes in Equations 12 and 14. Evaluating the impact of a new training program on $y_\$$ (where $y_\$ = y_{\$_{post}} - y_{\$_{pre}}$) would typically involve some pilot project in which the new training was given to personnel in randomly selected unit, followed by a comparison of average performance gain in this group to average performance gain achieved by some other group which received the old training program. One could apply the BCG model by simply correlating x (where x = 0 for those assigned to a control group and x = 1 for those assigned to the new training procedure) and $y_\$$ ($r_{xy_\$}$) and using it in Equations 12 and 14 above. Alternately, $r_{xy_\$}$ is the only part of Equations 12 & 14 affected by whether "x" is 0 = control group, 1 = training group or "x" is the score on some selection test ranging from 0-100. If we dropped r_{xy} out of Equations 12 &14, we would need to replace it with something that reflected how much more $\overline{y}_\$$ performance was obtained from the group receiving the new training program. Of course, the average performance difference $\overline{y}_{\$_{trained}} - \overline{y}_{\$_{control}}$ would capture this, but it couldn't be that simple! We have to "standardize" the $\overline{y}_{\$_{trained}} - \overline{y}_{\$_{control}}$ difference score by dividing it by its standard deviation, basically the

Different arithmetic, same "d." In my never ending attempt to make things simple, let me show yet another, perhaps easier way to get "d." Recall from the end of Chapter 2's realtor training example that we ultimately calculate a "z test statistic" to determine whether we can/cannot reject our initial conservative position that the new training program didn't work better. Below are two equations showing how to get the "d" statistic from the "z" statistic:

$$d = \frac{2z}{\sqrt{n_1 + n_2 - 2}}$$

$$d = \frac{z(n_1 + n_2)}{\sqrt{n_1 + n_2 - 2}\sqrt{n_1 n_2}}$$

where n_1 = sample size receiving new training program and n_2 = sample size of control group.

The first one is appropriate when $n_1 = n_2$, while the second one should be used when $n_1 \neq n_2$. Regardless of how "d" is arrived at arithmetically, all of these formulas result in the same "d" statistic. The "d" statistic is often referred to as a "standardized difference score" (Rosnow & Rosenthal, 1991; Rosenthal & Rosnow, 1996). The standardized difference score is often used to quickly describe a personnel selection test's "adverse impact potential" when computed as follows:

$$d = \frac{\bar{x}_{white} - \bar{x}_{black}}{\sqrt{\frac{s^2_{white} + s^2_{black}}{2}}}$$

Typical d scores for general cognitive ability tests range hover between .75 and 1.00. Actual adverse impact will occur due to a test's d score, the shape of x_{black} and x_{white} score distributions, and the selection ratio. However, "other things being equal," the bigger the standardized difference in black and white test scores, the more likely it is that actual use of the test in making personnel selection decisions will lead to adverse impact (i.e., violation of the 4/5^{ths} rule).

same thing we did with standardized z scores in Chapter 2, only this time we use a "pooled" our estimate of standard deviation by averaging estimate of $s^2_{trained}$ and $s^2_{control}$, or $d = \frac{\bar{y}_{\$_{trained}} - \bar{y}_{\$_{control}}}{\sqrt{\frac{s^2_{trained} + s^2_{control}}{2}}}$, (Rosnow & Rosenthal, 1991; Rosenthal & Rosnow, 1996).

Using "d" instead of "r_{xy}," Equations 12 & 14 become, respectively:

$$U_{total} = N_s \mu_{\$} + dN_s SD_y \bar{z}_s - N_a C$$

Equation 15

and;

$$\Delta U = dN_s SD_y \bar{z}_s - N_a C$$

Equation 16

Other discrete organizational changes such as new business strategies could also be assessed this way. For example, former CEO Jack Welsh is widely known to have imposed a strategic goal of being ranked number 1 or 2 in each business arena General Electric participated in. What if one wanted to assess the impact of a different strategic initiative, say, focusing on innovation and change? Perhaps a group of managers of GE business units might argue that their units would perform better if they looked more like 3M, a company devoted to innovation and creating new products – historically one of 3M's strategic goals is to have at least 40% of its revenue come from products developed and introduced to the market in the last 5 years. Our fictional GE business unit heads might argue such a strategy

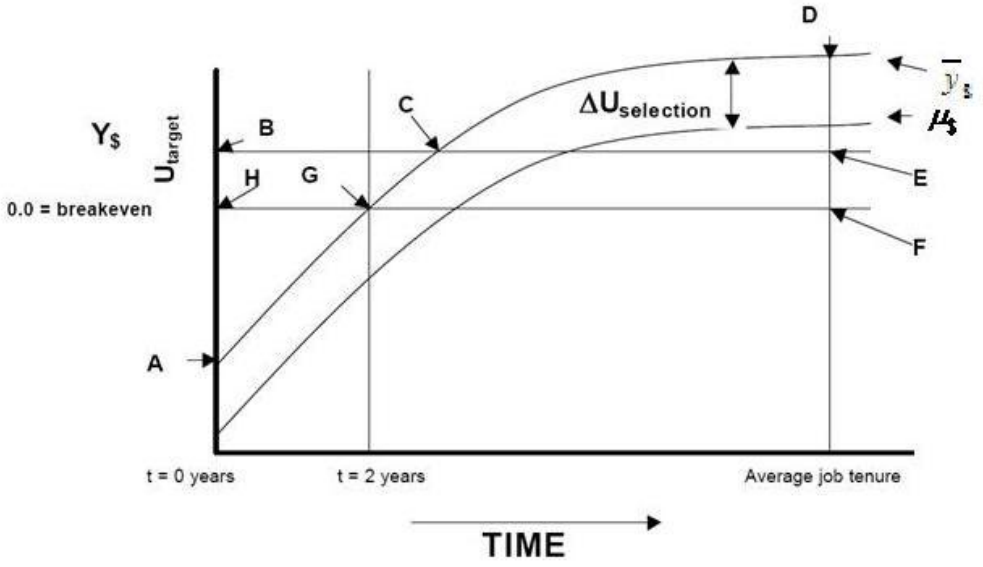
would be more appropriate for their units, yielding higher profits (though not sales) than a strategy focused on ranking #1 or #2 in sales.

After computing the z statistic to test whether the two strategies yield equal profit and/or sales ($H_o : \Delta\bar{y}_{\$_{new\ strategy}} - \Delta\bar{y}_{\$_{control}} = 0$) and converting it to d, one would then calculate SD_y from entire sample of GE division profit/sales figures. Average z score of those hired ($\bar{z}_s$) $H_o : \Delta\bar{y}_{\$_{new\ strategy}} - \Delta\bar{y}_{\$_{control}} = 0.$

Strategic Utility

Figure 2 below graphically portrays both *total* utility (U_{total}) and *incremental* utility ($\Delta U_{selection}$) over time for two groups of individuals recruited to purchase a new fast food franchise when $y_\$$ is measured as franchise profit. Group 1 are individual franchisees selected using some pre-existing franchisee selection system. Group 2 are individual franchisees selected using some new and improved franchisee selection system. Both groups average around 6 years as owners before exiting the business. Group 1's projected profit trajectory over its ownership tenure is captured by the line marked $\mu_\$$, while Group 2's projected profit trajectory is captured by the line marked $y_\$$ (line AGCD, the curved line immediately above $\mu_\$$). $\Delta U_{selection}$ will always be positive when the new franchise selection system has a higher criterion validity (r_{xy}) compared to the existing selection process and equal to the difference between the two curved lines ($\bar{y}_\$$ and $\mu_\$$) at any point in time.

Figure 2: Strategic Utility


D
$\bar{y}_s$
$\Delta U_{selection}$
C
Y_S
U_{target}
B
μ_s
H
G
E
0.0 = breakeven
F
A
t = 0 years
t = 2 years
Average job tenure
TIME

If an HR executive at the food franchisor's home office just looked at $\Delta U_{selection}$, s/he would always decide in favor of the new selection system. I call this a traditional "tactical" application of the BCG model. However, notice a couple of other characteristics of Figure 2, most notably, the financial "break even" point and the strategic profitability goal U_{target}. Finally, let's assume that the average franchisee has working capital to stay in operation 2 years before s/he must make a profit or exit the business. Under the existing selection system, $\mu_{\$}$ does not reach the breakeven point at the 2 year mark. This suggests that only some small portion (< 50%) of franchisees selected under the existing system survive past year 2. Under the new and improved selection system, the average franchisee is expected to just reach the breakeven point at about the time his/her working capital runs out (i.e., point G at year 2 on line $\bar{y}_{\$}$). Further, franchisees selected using existing methods are expected to reach strategic profitability goals somewhere in their 4^{th} or 5^{th} years of operation, while franchisees selected using the new and improved system are expected to reach profitability goals sometime shortly before the end of year 3. Importantly, a "strategic" view of utility suggests the franchiser should not consider selling franchises unless U_{total} is expected to reach some breakeven and/or target level within some specified time period. Note that multiple circumstances besides the selection system might impact U_{total}. For example, targeted recruiting efforts that increase applicant quality should increase U_y and U_{total}. Alternatively, decreasing franchise cost or fees would permit franchisees to retain a larger pool of initial working capital, hence extending the amount of time they can survive between start-up and the breakeven or target profit levels.

A Fly in the Ointment: Standard Deviation of Performance in Dollar Terms

So, why didn't the BGC model take the HR world by storm? Why is this probably the first time any HR professionals reading this have ever heard of it? Well, the answer lies in two parts. The first part is SD_y, or the standard deviation of performance measured in dollars. Very simply, no one could come up with a measure of SD_y that anyone had great confidence in until the late 1970's and early 1980's (Cronbach & Gleser, 1965, moaned about it on p. 121 of their

text). That explains the first 25-35 years of BCGs anonymity. Once some reasonable means of estimating of SD_y came on line, the BCG model faced an uphill battle against typical reactions people have to anything that looks like statistics. Tell the truth - before your introduction to utility analysis in this chapter, would you have found the equation $\Delta U = N_s r_{xy} SD_y \bar{z}_s - N_a C$ warm, friendly, attractive, or appealing? Would it have inspired you to put yourself on the line with a hard dollar estimate of an HR system's payoff to your General Manager? I thought not. If we both are equally honest, you probably don't find it warm, friendly, attractive, or appealing after reading the material in this chapter, though you are hopefully less intimidated by it. In all likelihood and in spite of the examples I provided above, until you actually run through some sample applications relevant to actual problems you face justifying HR expenditures, you still won't embrace it as a powerful HR tool. Well, I have the entire rest of this book to change your mind! Until then, and for the remainder of this chapter, I will discuss how the field overcame the SD_y speed bump and some of its interesting implications.

"Please don't ever let anyone see this." This is what the V.P. of HR at a major Fortune 50 firm told me when he read my report in 1986. I had just starting helping him develop and implement a general manager selection system (described in Russell, 2001). I thought he might appreciate a white paper describing the expected dollar impact of the system we were designing to select internal candidates for the lowest level executive position with P & L responsibility. I carefully documented all my assumptions, why they were conservative, and proudly estimated the firm should expect an addition $5M in annual profit for each position filled using the new system. The V.P. read the report carefully, asking me a handful of clarification questions that showed he fully understood the report content. Then he said "Craig, I have never seen anything like this. We should always do estimates like this before launching any and all HR programs. However, my colleagues in line management, in the controller's office, and especially the CEO will never believe anyone could ever estimate the dollar value contributed by one of our 'people' systems. If they see this report, the HR unit and I will lose all credibility – credibility that has been hard earned over a long period of time. Please don't ever let anyone see this." I honored, though never forgot, his request. It was one of the many experiences I have had in over 30 years that motivated the writing of this book.

What is SD_y? Let us revisit exactly what SD_y is again before describing ways in which in it can be measured. Let's start with an equation, specifically,

$$SD_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

Equation 17

Where y_i is a measure of the dollar value of performance contributed to the firm by employee "i" and "n" is the number of employees affected by the HR policy or practice. Labor economics says the demand for labor doesn't exist without customer demand for the product or service being produced. Unfortunately, asking customers how much of the price they pay for a new car was due to the value of the design engineer's contribution to operation of the thrust bearing in the automatic transmissions is likely to get you a blank look! In all but the simplest production processes consumers are not good judges of value contributed by each factor of production, including labor.

A slightly conservatively low estimate of y_i might be the employee's salary.[6] Marginal revenue product theory in labor economics indicates employees' salaries will always be equal to or somewhat lower than the actual value placed by the firm on the employees' work contributions (Mahoney, 1979). Average salary for 100 people doing the same job might be a good conservative estimate of the dollar value added by the average employee in that job ($\bar{y}$). Unfortunately, as anyone knows who has actually held a job, individual salaries paid to those 100 employees don't necessarily accurately reflect each one of those individual's contributions to the firm. Believe it or not, there are people who are over and under paid relative to the "true" value of their contributions to the product or service being produced! This becomes less of a problem if we assume the amount by which they are over or under paid relative to the value of their contributions is randomly distributed across employees (it almost has to be, since the firm cannot afford to pay more than average $\bar{y}$ across all 100 employees – over paid individuals have to be

[6] At the same time, it is a high estimate of the dollar value placed by the employee on the leisure s/he would otherwise engage in if s/he had not decided to work instead. Marginal revenue product theory in labor economics indicates employees' salaries will always be equal to or somewhat lower than the actual value placed by the firm on the employees' work contributions (Mahoney, 1979).

balanced out by underpaid individuals, and as long as both are randomly distributed over the entire range of salaries, the overall shape or dispersion of the salary distribution should not be influenced). In that case, we could assume each employee's salary y_i might be a reasonable index of her/his dollar value to the firm. One estimate of SD_y might be obtained by simply computing the average salary ($\bar{y}$) then plugging in each employee's salary y_i to Equation 16 above.

If the job was a key or benchmark job (i.e., it had an external labor market from which new employees were hired), a coarser estimate of SD_y could be obtained from wage and salary survey information. In this case, y_i would be wage reported by each participating firm and $\bar{y}$ would be the average of those y_i.

Wage and Salary Surveys. Firms typically report average and median salaries of current employees within key jobs and the high, low, and mid-points of the job's salary range. Any of these might be used as y_i in deriving SD_y, though each SD_y estimate would be best used for different utility estimates. For example, a state economic agency might want to estimate the dollar impact of a state-wide training initiative targeting entry level production jobs for an industry it hopes to attract in greater numbers to the state. The best y_i used to estimate SD_y in this instance would be the low end of salary ranges reported in a W&S survey. Alternatively, a local community college attempting to estimate utility of some custom course offered for incumbents in these jobs would get a more appropriate estimate of SD_y using salary mid-points as y_i. The bottom line is that y_i used to compute SD_y will depend on which population of current or future employees the HR policy or practice is targeting.

A reliable estimate of SD_y might be obtained this way for jobs with reasonable numbers of incumbents or firms contributing to a wage and salary survey (e.g., n > 30). Unfortunately, at best 15-20% of the jobs within a firm have external labor markets needed for wage and salary surveys. Further, most jobs don't have 30 or more incumbents in individual firms. How do you estimate SD_y when one or two people hold the job in question? In other words, how do you estimate SD_y when there are only 1 or 2 y_i (or none)?[7] This is where application of the BCG model stalled for lack of good SD_y estimates between the late 1940's and late 1970's. Some

[7] One might ask why anyone would care what SD_y was for jobs with only 1-2 incumbents. Firms still need HR systems for those jobs (e.g., recruiting, selection, etc.). They often purchase pre-made HR tools from consulting firms for these jobs. So, consulting firms are interested in estimating SD_y for those jobs. Further, higher level jobs tend to have fewer incumbents even though they can have much greater dollar impact on the firm (e.g., SD_y for CEOs is going to be much higher than SD_y for janitorial positions). Hence, even though there may only be a handful

efforts were mounted using human resource accounting methods in the mid-1960's, but were abandoned due to a host of problems that could not be overcome (Roche, 1965).

40 Percent Rule. Schmidt and Hunter (1983) recommended using 40% of average salary as an SD_y estimate. Using this rule, they noted that about 57% of the cost of all goods and services produced in the U.S. is due to the cost of labor, then SD_y expressed as a percentage would be .40 x 57% ≈ 20%. They argued that use of standard deviation in performance as a percentage of costs, or SD_p = 20%, could be substituted for SD_y in the BCG model, yielding a utility estimate expressed in terms of the expected percentage performance increase. Subsequent review of 85 studies examining work output and work samples by Hunter, Schmidt, and Judiesch (1990) found SD_p averaged 15% for low complexity jobs, 25% for medium complexity jobs, 39% for sales jobs (97% for life insurance sales), and 45% for high complexity jobs. Hence, the "40% rule" is generally thought to be a conservative estimate of SD_y. I would recommend its use for coarse utility estimates at best – I would not use it if I needed precise estimates of value added by and HR system for comparison with net present value estimates of alternative capital investments (Boudreau, 1991).

Occupation vs. Job SD_y. Any given firm might gain much greater value from a job than other firms. For example, two truck driver jobs may otherwise be identical in terms of equipment, route, delivery schedule, etc. However, one truck driver may haul computer components between manufacturers while the other delivers cow manure to fertilizer processing plants. Since customer demand for computers tends to be higher than consumer demand for fertilizer, value and revenue generated from delivery of computer components will be higher than value and revenue generated from delivery of manure. Hence, estimation of SD_y for an occupation like budget analyst by Schmidt et al. (1979) is not useful for any individual firm's utility estimate. Schmidt et al.'s SD_y estimate might be useful for a Budget Analyst Professional Association in estimating the dollar impact of its continuing educational courses.

Direct Estimation. Schmidt, Hunter, McKensie, and Muldrow (1979) developed a method using subjective, or "rational" estimates of SD_y obtained from subject matter experts. So, who is a "subject matter expert" on the value of labor contributions? It should be someone

---

of incumbents in a job, high versus low performance by those incumbents could cause huge swings in dollar value to the firm.

intimately familiar with all resources contributing to the production process as well as how they might substitute for or complement one another. Schmidt et al. (1979) used two sets of immediate supervisors to estimate SD_y for the occupations of Budget Analyst and Computer Programmer. The rationale was pretty straight forward. It all hinged on the assumption that dollar value of performance for these two occupations is normally distributed.[8] Values one standard deviation above the average occur close to the 85th percentile, while values one standard deviation below average occur close to the 15th percentile. Schmidt et al. asked supervisors to estimate how much dollar value individuals performing at the 15th, 50th, and 85th percentiles would bring to the firm. In this way, two separate estimates of SD_y could be obtained as follows:

- 1st SD_y estimate: 85th – 50th percentile values
- 2nd SD_y estimate: 50th – 15th percentile values

Results would at least partially support our assumption that performance was normally distributed if the two SD_y estimates are not significantly different. If both SD_y estimates are reasonably close, averaging the two will yield a better SD_y estimate than either original value. Of course, there are all sorts of ways to enhance the quality of these estimates. For example, Bobko, Karren, and Parkington (1983) found use of a narrative definition of an "average" worker's dollar value added to the firm placed next to where supervisors were asked to make their estimates of the 50th percentile value resulted in less variability in those estimates. Bobko et al. (1983) found direct estimates of SD_y using 50th percentile anchors were very close to actual sales dollar SD_y for sales positions. Other evaluation techniques like use of consensus discussion by supervisory raters have also been found to improve agreement among ratings.

[8] This is clearly not correct in most jobs, as there is some minimum level of performance below which firms will not continue to employ an individual. Consequently any true performance distribution will be "left truncated" in the sense that you never see values below some minimum floor level.

Does the direct estimation method solve the SD_y problem holding back wide spread use of the BCG model? Well, kind of. Subsequent studies have shown mixed results it how well its SD_y estimates converge with independent SD_y estimates generated other ways. Further, Bobko, Shetzer, and Russell (1991) found that type of subject matter expert and context within which percentile estimates were made influenced SD_y estimates. Bobko et al. (1991) asked students and heads of faculty recruiting committees to estimate the dollar value of 85th and 50th percentile performing business school faculty members. Half of the subject matter experts were asked for percentile estimates of the value for a faculty member who had decided to quit and take a position on some other faculty. The other half were asked for percentile estimates of the value of a new faculty member who had just accepted a position on their faculty.[9] Curiously, SD_y estimates were highest for students estimating the value of faculty members who are voluntarily leaving, while heads of faculty recruiting committees gave higher SD_y estimates when rating newly hired faculty members. Both students and other faculty members (who are usually heads of faculty recruiting committees) can be thought of as "consumers" of services provided by faculty members. The point readers should take away here is that value is in the eye of the beholder. The fact that different constituencies or "customers" of a job might have different views of the job value should not be too surprising (e.g., students' views of faculty values are typically limited to teaching dimensions, while faculty views generally extend to research and service). I am not aware of any direct estimation studies done in which traditional retail customers were asked for estimates, though it might be possible for jobs in which >90% of the tasks involve customer service that the customer actually sees performed.

[9] 15th percentile estimates were not requested because the authors felt it subject matter experts would find it nonsensical to estimate the dollar value of a newly hired 15th percentile performing faculty member, as this person is not likely to be hired to begin with.

In sum, the direct estimation procedure at times yields SD_y estimates that converge with independent SD_y estimates, though it is not clear when it will happen. This gives me confidence that, at best, direct estimates of SD_y used in the BCG model can be expected to yield coarse approximations of utility. For example, if I had 20 subject matter experts providing 15th, 50th, and 85th percentile dollar value estimates, I would be reasonably confident that true SD_y estimate will be somewhere between the 4th lowest and 17th highest individual's average SD_y values (having done this many times, I have seen too many goofy low and high SD_y ratings, so loping off the low and high estimates should yield a reasonable range). I could use the 4th and 17th highest SD_y estimates in the BCG model to compute two separate utility estimates for a selection system with some known criterion validity r_{xy} and $\bar{z}_s$. Clearly, I would feel better if the two utility estimates (ΔU in Equation 14 above) are very close together. Unfortunately, most of the time they are very far apart. Hence, my choice of the word "coarse" in describing direct SD_y estimates – I would not count on ΔU in increased revenue from a selection system for to qualify for a business loan!

CREPID. No, this is not an STD. CREPID stands for "Cascio-Ramos Estimate of Performance in Dollars" and is yet another method of estimating SD_y. Yes, there are actually two guys named Cascio and Ramos responsible for this label (see Cascio & Ramos, 1986). For those few of you versed in the Factor Comparison method of job evaluation, you will see parallels in CREPID. CREPID involves a minimum of eight steps to estimate each incumbent's y_i, $\bar{y}$, and SD_y:

1. Determine the principle tasks and activities (PTs) performed in the job. One rule of thumb recommended by Cascio and Ramos (1986) is that this includes all PTs that make up at least 10% of total job time. Following this rule of thumb,

CREPID could not be used when 2 tasks make up 10% and 16 tasks make up 5% of the job.

2. Acquire ratings of how often each PT is performed and its relative importance. Ratings can be made on any scale (e.g., 1-5 point scales) as long as the same scale is used for both frequency and importance.[10] Use of different scales would cause whichever one had the large scale to have more influence on each incumbent's y_i estimate.
3. Multiply importance rating by frequency rating.
4. Take average pay for target job and allocate dollars across PTs in proportion to the product of importance and frequency ratings from Step 3. So, if a job had 3 PTs and the importance X frequency rating = 10 for each PT, 33% of the average salary would be allocated to each PT. If importance X frequency ratings were 10, 20, and 20, respectively, 20% of the average salary would be allocated to PT1 while 40% would be allocated to PT2 and PT3.
5. Rate each incumbent's current performance on each PT using a 0 – 200 point scale, then divide by 100 to put it on a 0 – 2.0 scale.
6. For each incumbent, multiply each 0 – 2.0 point PT performance rating by the dollar value assigned to that PT in Step 4 above.
7. Add up all the products of all PT performance ratings and PT dollar values you just got in Step 6 for each incumbent – this will be a "best" estimate of the dollar value of each incumbent's current job performance (y_i).

[10] Cascio and Ramos (1986) used a 0-7 point scale that, in my opinion, could cause problems. Specifically, since importance and frequency ratings are multiplied, a critically important task at a nuclear energy plant that is performed very infrequently could result in 0 x 7 = 0 in Step 3, and $0 would be allocated to that PT in Step 4. Since by definition each PT is a "principle" task, it should not be possible by definition for PTs to be rated "0" on importance or frequency. Hence, I have used 1-7 point scales in CREPID applications I have been involved in.

8. Finally, calculate the average ($\bar{y}$) and standard deviation (SD_y) of y_i across all incumbents.

Basically, CREPID uses subject matter experts (e.g., some combination of seasoned incumbents and supervisors) to decompose total job value (y_i) into value obtained from the job's key or principle tasks, duties, and responsibilities. After rating an incumbent's performance on these key tasks, performance ratings are combined with key task dollar values to estimate the dollar value received by the firm from the incumbent's performance of that task. Adding dollar value received across all key tasks yields an estimate of the incumbent's total dollar value delivered through her/his task performance y_i. Doing this across all n incumbents in any particular job permits estimation of $SD_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$.

At least three things can derail CREPID estimates of SD_y: poor identification of PTs, poor ratings of PT's relative importance/frequency, and poor ratings of incumbent performance.[11] A number of interesting assumptions and characteristics behind CREPID are also not immediately obvious. First, notice an incumbent receiving performance ratings of 100 on all PTs will end up with y_i equal to the average salary for the position. An incumbent receiving the highest performance rating (200) on each PT will end up with $y_i = 2(\bar{y})$ and an incumbent receiving the lowest performance rating (0) will end up with $y_i = 0$. For most jobs, 0 and $2(\bar{y})$ will fall well outside the job's salary range. For example, if the salary range is ± 15% around $\bar{y}$ (or $\bar{y} \pm .15\bar{y}$), an incumbent with performance ratings of 200 on all PTs would be thought to be adding value equal to $2\bar{y}$, even though at best s/he is earning a salary that was 85%

[11] Traditional problems with performance appraisal rating errors tend not to be as severe using CREPID, since they are not feed back to incumbents and not used to actually influence incumbent employment conditions (e.g., via merit pay increases, promotions, etc.).

lower (i.e., the top of the salary range, or $1.15\bar{y}$). So, use of a 0 – 200 point performance rating scale may yield an inaccurately wide range of y_i, and inaccurately high CREPID $SD_{y.}$

Second, as with any data-based method, CREPID cannot not yield accurate SD_y estimates if enough data is not present. Some minimum number of incumbents must exist in order to get a reasonable estimate of SD_y – I would not be confident with $n < 20$. Finally, CREPID estimates SD_y among current employees in the job. What if all current employees in a retail sales job are highly motivated, seasoned veterans who all perform at the highest levels? CREPID would estimate each incumbent's y_i to be very high and almost equal to one another, causing SD_y to be very small. Recall the purpose of the BCG model is to estimate expected dollar value added from use of a personnel selection test to decide which applicants to hire. If we used a selection test to replace these seasoned workers once they retired, the best estimate of SD_y would be one derived from CREPID applied to the new workers' performance on each PT. If the current distribution of current incumbent performance is not highly similar to the future performance distribution of newly hired incumbents who had been screened by the selection test, CREPID SD_y will be inaccurate.

Summary

We covered a lot of ground in this chapter. Subsequent chapters will build and extend examples of how the Taylor-Russell and BCG procedures can be used. Hopefully the reader now knows that it is not impossible or even real difficult to estimate how HR interventions might affect important business metrics. Clearly precision commonly associated with cash management accounting procedures has not been attained: while I would be real confident two bank tellers' cash drawers had different quantities of various denominations in them using traditional cash management accounting techniques, I would not be equally confident the dollar

value added to the bank by those two cashiers was exactly equal to the difference in their y_i estimates derived from a CREPID SD_y estimate. Other things being equal, precision in measurement is usually better. Fortunately, existing precision available from the Taylor-Russell approach, BCG model, and various methods of estimating SD_y is often enough for managers' decision making needs.

References

Bobko, P., Karren, R, & Kerker, S.P. (1987). Systematic research needs for understanding supervisory-based estimates of SD_y in utility analysis. Organizational Behavior and Human Decision Processes, 40, 69-95.

Bobko, P., Shetzer, L., & Russell, C.J. (1991). On the robustness of judgments of SDy in utility analysis: Effects of frame and presentation order. Journal of Occupational Psychology, 64, 179-188.

Boudreau, J.W. (1991). Utility analyses for decisions in human resource management. In M.D. Dunnette and L.M Hough (eds.), Handbook of Industrial and Organizational Psychology (Vol.2, 2nd ed., pp. 621-745). Palo Alto, CA: Consulting Psychologists Press.

Brogden, H.E. (1949). When testing pays off. Personnel Psychology, 2, 171-185.

Cascio, W.F. & Ramos, R.A. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. Journal of Applied Psychology, 71, 20-28.

Cronbach, L.J. & Gleser, G.C. (1965). Psychological Tests and Personnel Decisions (2nd Ed.). Urbana: University of Illinois Press.

Mahoney, T.A. (1979). Compensation and reward perspectives. Homewood, IL: Irwin.

Roche, W.J. Jr. (1965). A dollar criterion in fixed-treatment employee selection. In L.J. Cronbach and G.C. Gleser (eds). Psychological Tests and Personnel Decisions (2nd Ed., pp. 254-267). Urbana: University of Illinois Press.

Rosenthal, R. & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis (2nd ed.). New York: McGraw Hill.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. Psychological Methods, 1, 331-340.

Russell, C.J. (2001). A longitudinal study of top-level executive performance. Journal of Applied Psychology, 6, 510-517.

Russell, C.J., Colella, A., & Bobko, P. (1993). Expanding the context of utility: The strategic impact of personnel selection. Personnel Psychology, 46, 781-801.

Schmidt, F.L. & Hoffman, B. (1973). An empirical comparison of three methods of assessing the utility of a selection device. Journal of Industrial and Organizational Psychology, 1, 1-11.

Sturman, M.C. (2003). Utility analysis: A tool for quantifying the value of hospitality human resource interventions; utility analysis can be used to assess the effectiveness of human resource interventions (among other uses). Cornell Hotel & Restaurant Administration Quarterly, April.

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.

Table 1: Taylor-Russell Tables

Proportion of Employees Considered Satisfactory = .20
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20
.05	.23	.23	.22	.22	.21	.21	.21	.21	.20	.20	.20
.10	.26	.25	.24	.23	.23	.22	.22	.21	.21	.21	.20
.15	.30	.28	.26	.25	.24	.23	.23	.22	.21	.21	.20
.20	.33	.31	.28	.27	.26	.25	.24	.23	.22	.21	.21
.25	.37	.34	.31	.29	.27	.26	.24	.23	.22	.21	.21
.30	.41	.37	.33	.30	.28	.27	.25	.24	.23	.21	.21
.35	.45	.41	.36	.32	.30	.28	.26	.24	.23	.22	.21
.40	.49	.44	.38	.34	.31	.29	.27	.25	.23	.22	.21
.45	.54	.48	.41	.36	.33	.30	.28	.26	.24	.22	.21
.50	.59	.52	.44	.38	.35	.31	.29	.26	.24	.22	.21
.55	.63	.56	.47	.41	.36	.32	.29	.27	.24	.22	.21
.60	.68	.60	.50	.43	.38	.34	.30	.27	.24	.22	.21
.65	.73	.64	.53	.45	.39	.35	.31	.27	.25	.22	.21
.70	.79	.69	.56	.48	.41	.36	.31	.28	.25	.22	.21
.75	.84	.74	.60	.50	.43	.37	.32	.28	.25	.22	.21
.80	.89	.79	.64	.53	.45	.38	.33	.28	.25	.22	.21
.85	.94	.85	.69	.56	.47	.39	.33	.28	.25	.22	.21
.90	.98	.91	.75	.60	.48	.40	.33	.29	.25	.22	.21
.95	1.00	.97	.82	.64	.50	.40	.33	.29	.25	.22	.21
1.00	1.00	1.00	1.00	.67	.50	.40	.33	.29	.25	.22	.21

Proportion of Employees Considered Satisfactory = .30
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30	.30
.05	.34	.33	.33	.32	.32	.31	.31	.31	.31	.30	.30
.10	.38	.36	.35	.34	.33	.33	.32	.32	.31	.31	.30
.15	.42	.40	.38	.36	.35	.34	.33	.33	.32	.31	.31
.20	.46	.43	.40	.38	.37	.36	.34	.33	.32	.31	.31
.25	.50	.47	.43	.41	.39	.37	.36	.34	.33	.32	.31
.30	.54	.50	.46	.43	.40	.38	.37	.35	.33	.32	.31
.35	.58	.54	.49	.45	.42	.40	.38	.36	.34	.32	.31
.40	.63	.58	.51	.47	.44	.41	.39	.37	.34	.32	.31
.45	.67	.61	.55	.50	.46	.43	.40	.37	.35	.32	.31
.50	.72	.65	.58	.52	.48	.44	.41	.38	.35	.33	.31
.55	.76	.69	.61	.55	.50	.46	.42	.39	.36	.33	.31
.60	.81	.74	.64	.58	.52	.47	.43	.40	.36	.33	.31
.65	.85	.78	.68	.60	.54	.49	.44	.40	.37	.33	.32
.70	.89	.82	.72	.63	.57	.51	.46	.41	.37	.33	.32
.75	.93	.86	.76	.67	.59	.52	.47	.42	.37	.33	.32
.80	.96	.90	.80	.70	.62	.54	.48	.42	.37	.33	.32
.85	.99	.94	.85	.74	.65	.56	.49	.43	.37	.33	.32
.90	1.00	.98	.90	.79	.68	.58	.49	.43	.37	.33	.32
.95	1.00	1.00	.96	.85	.72	.60	.50	.43	.37	.33	.32
1.00	1.00	1.00	1.00	1.00	.75	.60	.50	.43	.38	.33	.32

Proportion of Employees Considered Satisfactory = .40
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.40	.40	.40	.40	.40	.40	.40	.40	.40	.40	.40
.05	.44	.43	.43	.42	.42	.42	.41	.41	.41	.40	.40
.10	.48	.47	.46	.45	.44	.43	.42	.42	.41	.41	.40
.15	.52	.50	.48	.47	.46	.45	.44	.43	.42	.41	.41
.20	.57	.54	.51	.49	.48	.46	.45	.44	.43	.41	.41
.25	.61	.58	.54	.51	.49	.48	.46	.45	.43	.42	.41
.30	.65	.61	.57	.54	.51	.49	.47	.46	.44	.42	.41
.35	.69	.65	.60	.56	.53	.51	.49	.47	.45	.42	.41
.40	.73	.69	.63	.59	.56	.53	.50	.48	.45	.43	.41
.45	.77	.72	.66	.61	.58	.54	.51	.49	.46	.43	.42
.50	.81	.76	.69	.64	.60	.56	.53	.49	.46	.43	.42
.55	.85	.79	.72	.67	.62	.58	.54	.50	.47	.44	.42
.60	.89	.83	.75	.69	.64	.60	.55	.51	.48	.44	.42
.65	.92	.87	.79	.72	.67	.62	.57	.52	.48	.44	.42
.70	.95	.90	.82	.76	.69	.64	.58	.53	.49	.44	.42
.75	.97	.93	.86	.79	.72	.66	.60	.54	.49	.44	.42
.80	.99	.96	.89	.82	.75	.68	.61	.55	.49	.44	.42
.85	1.00	.98	.93	.86	.79	.71	.63	.56	.50	.44	.42
.90	1.00	1.00	.97	.91	.82	.74	.65	.57	.50	.44	.42
.95	1.00	1.00	.99	.96	.87	.77	.66	.57	.50	.44	.42
1.00	1.00	1.00	1.00	1.00	1.00	.80	.67	.57	.50	.44	.42

Proportion of Employees Considered Satisfactory = .50
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.05	.54	.54	.53	.52	.52	.52	.51	.51	.51	.50	.50
.10	.58	.57	.56	.55	.54	.53	.53	.52	.51	.51	.50
.15	.63	.61	.58	.57	.56	.55	.54	.53	.52	.51	.51
.20	.67	.64	.61	.59	.58	.56	.55	.54	.53	.52	.51
.25	.70	.67	.64	.62	.60	.58	.56	.55	.54	.52	.51
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52	.51
.35	.78	.74	.70	.66	.64	.61	.59	.57	.55	.53	.51
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53	.52
.45	.85	.81	.75	.71	.68	.65	.62	.59	.56	.53	.52
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54	.52
.55	.91	.87	.81	.76	.72	.69	.65	.61	.58	.54	.52
.60	.94	.90	.84	.79	.75	.70	.66	.62	.59	.54	.52
.65	.96	.92	.87	.82	.77	.73	.68	.64	.59	.55	.52
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55	.53
.75	.99	.97	.92	.87	.82	.77	.72	.66	.61	.55	.53
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55	.53
.85	1.00	.99	.97	.94	.88	.82	.76	.69	.62	.55	.53
.90	1.00	1.00	.99	.97	.92	.86	.78	.70	.62	.56	.53
.95	1.00	1.00	1.00	.99	.96	.90	.81	.71	.63	.56	.53
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56	.53

A second important concept to be included is the proportion of variable costs of performance denoted by *V*. Job performance, *y*, here is understood as the sales value of the product or service of an employee (Greer & Cascio, 1987). The contribution margin of job performance *y* is computed by *y* (1+*V*). Thus, the standard deviation of the contribution margin of job performance is calculated by $SD_y(1+V)$.

Incorporating, furthermore, taxes, denoted by TAX, we arrive at

$$\Delta U = N_S r_{xy} SD_y(1+V)\bar{z}_x(1-TAX) - C_f(1-TAX) - C_v N_a/SR(1-TAX). \tag{10}$$

We include discounting as a further concept. Since money can be invested to earn interest, financial analysis discounts future earning and costs to consider these potential investment returns. Given the interest rate *i* discounting ΔU over a period of *T* time units (mostly years) yields:

$$\begin{aligned}\Delta U = &\sum_{t=1}^{T} \frac{1}{(1+i)^t} N_a\, r_{xy}\, SD_y(1+V)\bar{z}_x(1-TAX) \\ &- \sum_{t=1}^{T} \frac{1}{(1+i)^{t-1}} C_f(1-TAX) \\ &- \sum_{t=1}^{T} \frac{1}{(1+i)^{t-1}} C_v\, N_a/SR(1-TAX)\end{aligned} \tag{11}$$

In the first line of (11) the returns from the personnel program are estimated. In the second line the variable costs, and in the third the fixed costs of the program are given.

The impact of personnel selection programs is usually of long duration. Since productivity is often variable during this time, Boudreau (1983b) suggested dividing the duration of the intervention program into a number of intervals. These are time periods in which productivity changes occur. Furthermore, the model parameters may have different values in each of the time periods. This is an essential feature in cases where, for example, a predictor has a variable predictive validity.

By dividing the duration into different periods, it is also possible to encompass the flow of employees. Assuming that N_{at} persons will be selected at the beginning of period *t* and N_{lt} employees selected by the program in a former period will leave the organization after time period *t*, the valid number of employees during this time period, N_t, amounts to

$$N_t = \sum_{j=1}^{t} (N_{aj} - N_{lj}).$$

When we include the above suggestions, the following utility model results:

$$\begin{aligned}\Delta U = &\sum_{t=1}^{T}\sum_{j=1}^{t}(N_{aj} - N_{lj}) r_{xpt}\, SD_{pt}(1 + V_t)\,\bar{z}_{xt}(1/(1+i_t)^t)(1 - TAX_t) \\ &- \sum_{t=1}^{T}(N_{at}/Q_t)C_{vt}(1/(1+i_t)^{t-1})(1 - TAX_t) \qquad (12) \\ &- \sum_{t=1}^{T}(C_{ft}(1/(1+i_t)^{t-1})(1 - TAX_t)\end{aligned}$$

Summarizing of all parameters of (12) yields:

ΔU	Utility of the personnel selection program in time period $1 \dots T$
t	Time period. This could be a duration of up to a complete year and indicates the particular t-th year after the commencement of the program.
T	Impact duration of the program.
N_t	The number of employees in the organization in time period t who have been selected through the program.
N_{at}	The number of employees selected in time period t.
N_{lt}	The number of employees who left the organization in time period t.
r_{xyt}	Validity of the selection instrument in time period t. This is the product-moment correlation of the predictor x and the sales value of the job performance y within the applicant's population.
$\bar{z}_t$	mean predictor score within the selected group in time period t
SR_t	Selection rate in time period t.
$\lambda(SR_t)$	Ordinate of the normal distribution of the selection rate in time period t.
SD_{yt}	Standard variation of the sales value of the job performance in the applicant population in time period t.
V_t	Proportion of the variable costs of the job performance for the organization in time period t.
i_t	Interest rate in time period t.
TAX_t	Tax rate in time period t.
C_{vt}	Variable costs of the personnel selection program per applicant in time period t.
C_{ft}	Fixed costs of the program including those incurred for the development, implementation and evaluation of the program in time period t.

The above model (12) refers to the utility of an introduced intervention compared to a situation without an intervention. Considering personnel selection we mostly have to compare different treatments, i.e. selection strategies because random selection occurs very seldom. In this case r_{xyt}, C_{ft} and C_{vt} refer to the difference of the validity, variable costs, fixed costs, resp. of a certain selection

program in relation to an alternative selection. Then ΔUillustrates the incremental value of a personnel selection program over an alternative program. This is the incremental utility produced when the utility of a personnel selection program is compared to an alternative program.

Model (12) still represents state of the art. However, some further complements have been proposed. Murphy (1986) derived formulas for correcting the average test score $\overline{z}_x$if selectees decline offers and lower scoring candidates must be accepted. Three cases representing realistic circumstances are considered: (a) offers are declined at random, (b) offers are declined by the highest scoring applicants, and (c) test scores are related to the probability of accepting an offer.

Tziner, Meir, Dahan & Birati (1994) take the inflation rate as a further parameter into account. This parameter behaves completely analogous to discounting. If the inflation rate is denoted by *f*, the discounting factor $1/(1+i)^t$ has to be multiplied by $1/(1+f)^t$. It is not necessary to include the inflation rate as a separate factor. We can combine both factors to an adjusted interest $i_a = i + f + if$, since $1/(1+i)^t \cdot 1/(1+f)^t = 1/(1+i+f+if)^t$. If the inflation rate is included we will speak of an adjusted interest rate i_a.

From http://www.mpr-online.de/issue4/art2/node2.html

The AssiStat - Statistical Formula Calculator

Reviewed in Personnel Psychology, Vol. 49, No. 2, Summer, 1996, pp. 534-536

http://www.assess.com/xcart/product.php?productid=216&cat=0&page=1

$d = M_1 - M_2 / \sigma_{pooled}$

$\sigma_{pooled} = \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]}$

In practice, the pooled standard deviation, σ_{pooled}, is commonly used (Rosnow and Rosenthal, 1996).

The pooled standard deviation is found as the root mean square of the two standard deviations (Cohen, 1988, p. 44). That is, the pooled standard deviation is the square root of the average of the squared standard deviations. When the two standard deviations are similar the root mean square will be not differ much from the simple average of the two variances.

$d = 2t / \sqrt{(df)}$

or

$d = t(n_1 + n_2) / [\sqrt{(df)}\sqrt{(n_1 n_2)}]$

d can also be computed from the value of the *t* test of the differences between the two groups (Rosenthal and Rosnow, 1991). . In the equation to the left "df" is the degrees of freedom for the *t* test. The "n's" are the number of cases for each group. The formula without the n's should be used when the n's are equal. The formula with separate n's should be used when the n's are not equal. $d = 2r / \sqrt{(1 - r^2)}$ *d* can be computed from r, the ES correlation.

The interpretation of Cohen's d

Cohen's Standard	Effect Size	Percentile Standing	Percent of Nonoverlap
	2.0	97.7	81.1%
	1.9	97.1	79.4%
	1.8	96.4	77.4%
	1.7	95.5	75.4%
	1.6	94.5	73.1%
	1.5	93.3	70.7%
	1.4	91.9	68.1%
	1.3	90	65.3%
	1.2	88	62.2%

Cohen (1988) hesitantly defined effect sizes as "small, *d* = .2," "medium, *d* = .5," and "large, *d* = .8", stating that "there is a certain risk in inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" (p. 25).

Effect sizes can also be thought of as the average percentile standing of the average treated (or experimental) participant relative to the average untreated (or control) participant. An ES of 0.0 indicates that the mean of the treated group is at the 50th percentile of the untreated group. An ES of 0.8 indicates that the mean of the treated group is at the 79th percentile of the untreated group. An effect size of 1.7 indicates that the mean of the treated group is at the 95.5 percentile of the untreated group.

Effect sizes can also be interpreted in terms of the percent of nonoverlap of the treated group's scores with those of the untreated group, see Cohen (1988, pp. 21-23) for descriptions of additional measures of nonoverlap.. An ES of 0.0 indicates that the distribution of scores for the treated group overlaps completely with the distribution of scores for the untreated group, there is 0% of nonoverlap. An ES of 0.8 indicates a nonoverlap of 47.4% in the two distributions. An ES of 1.7 indicates a nonoverlap of 75.4% in the two distributions.

	1.1	86	58.9%
	1.0	84	55.4%
	0.9	82	51.6%
LARGE	0.8	79	47.4%
	0.7	76	43.0%
	0.6	73	38.2%
MEDIUM	0.5	69	33.0%
	0.4	66	27.4%
	0.3	62	21.3%
SMALL	0.2	58	14.7%
	0.1	54	7.7%

	0.0	50	0%