11/3/2019



Selection Validation Report

Prepared by Jack Wychor

Table of Contents

[Executive Summary 2](#_Toc23518887)

[Analysis 3](#_Toc23518888)

[Predictor Analysis 3](#_Toc23518889)

[Cognitive Ability 3](#_Toc23518890)

[Integrity 3](#_Toc23518891)

[Boldness 3](#_Toc23518892)

[Leadership Skill 3](#_Toc23518893)

[Interview Scores 4](#_Toc23518894)

[Overall 4](#_Toc23518895)

[Criterion Analysis 5](#_Toc23518896)

[Overall 5](#_Toc23518897)

[Regression Validation 5](#_Toc23518898)

[Model 1 6](#_Toc23518899)

[Generalizability 7](#_Toc23518900)

[Adverse Impact 7](#_Toc23518901)

[Banding 7](#_Toc23518902)

[Model 2 7](#_Toc23518903)

[Generalizability 7](#_Toc23518904)

[Adverse Impact 7](#_Toc23518905)

[Performance 7](#_Toc23518906)

[Project Reflection 9](#_Toc23518907)

[Reliability 9](#_Toc23518908)

[Data Complications 10](#_Toc23518909)

[Adverse Impact 10](#_Toc23518910)

[Data Issues 10](#_Toc23518911)

[Hypotheses 10](#_Toc23518912)

# Executive Summary

In October of 2019, Wychorp received employee information from Globex. Information was collected and analyzed pertaining to 5 predictors (cognitive ability, integrity, boldness, leadership skill, and interview evaluation), 5 criteria (termination, supervisor ratings, performance index, discipline, and accidents), demographic information (sex), and position in the company (division and level) for each employee. Data was collected from 506 employees, 480 of whom had complete data in every field. Predictive selection procedures were constructed based on relationships between predictors and criteria. The key results were as follows:

* A model was constructed that showed an acceptable adverse impact ratio (0.85) and produced a strong validity coefficient (R2 = 0.35) utilizing predictor data in a backward regression model on a unit weighted criteria composite score of all criteria.
* Scores on the Wonderlic, integrity scale, boldness, and interview scores were the only significant predictors in the regression models.
* Because boldness is a measured using a Hogan assessment, it was weighted lower (50%) to prevent any bias it may have had given that Hogan assessments measure fit instead of performance. Boldness was also found to cause adverse impact when not weighted.
* Integrity was treated as a scale of the all integrity questions except question 3 because question 3 lowered internal consistency estimates when included.
* Interview scores tended to be negatively related to criteria composite scores in prediction models.
* Models created using this data tended to show strong cross validated generalizability.
* Reliability estimates for criteria were limited and more data may be needed in the future to compute more accurate validation estimates considering range restriction and/or low predictor or criterion reliability.
* Reliability estimates for predictors were, like the criteria, unobtainable with the exception of the integrity scale given a lack of data about predictor scores.
* Additional data relating to protected classes of employees may need to be collected to test for effects of the prediction model on other protected classes.
* Additional data relating to the importance of individual criteria would increase the confidence of validity estimates for prediction models.

# Analysis

## Predictor Analysis

Prior to being selected for hire, job candidates completed three self-report questionnaires with the purpose of measuring cognitive ability, integrity, and boldness. Additionally, hiring managers gave ratings of their interview with each candidate as well as rated candidates on their effectiveness in three areas related to leadership skill (task management, people management, and strategic thinking). There were five total possible criteria on which all five predictors were compared. The quality of each predictor was analyzed separately and recorded below.

### Cognitive Ability

Cognitive ability was measured using the 50-item Wonderlic test. It produced the strongest relationship with both supervisor ratings (*r* = 0.49) and performance index scores (*r* = 0.37), and was moderately related to discipline (*r* = 0.14) and accidents (*r* = -0.13). Given the nature of cognitive ability tests and the potential to cause adverse impact, great care must be taken before using one as a selection device. A one-way between subject’s ANOVA was conducted to determine whether Wonderlic scores differed between the two levels of gender (male and Female). There was not a significant difference at the *p* < 0.05 level between both conditions of gender *F*(1,504) < 0.01, *p* = 0.94. With this in mind, it is highly recommended that Globex obtain demographic data pertaining to the ethnicity or race applicants to thoroughly record any potential differences regarding protected classes and this measure.

### Integrity

A 7-item scale was used to assess integrity. Internal consistency estimates of the scale were recorded using Cronbach’s Alpha (α = 0.85). Estimated reliability increased when item 3 was dropped (α = 0.89). Therefore, item 3 was excluded from the analysis and scale scores for each applicant were obtained using the other 6 items. The integrity scale scores strongly predicted discipline (*r* = -0.32) and supervisor ratings (*r* = 0.28).

### Boldness

Boldness was measured using standardized T-scores of responses to Hogan Development Survey. Relationships with supervisor ratings (*r* = 0.25), performance index (*r* = 0.19), and discipline (*r* = 0.19) were generally moderate. Like cognitive ability, levels of boldness (similar to the construct extraversion) are likely to differ among gender and/or other protected classes. To test this, a one-way ANOVA was conducted to determine whether scores on boldness differed between the two levels of gender (male and female). There was a significant difference at the *p* < 0.05 level between the two conditions of gender *F*(1,504) = 55.49, *p* < 0.01. While a significant difference among levels of a protected class on a predictor measure do not guarantee adverse impact, it is nonetheless very likely. This may be in part due to the nature of Hogan assessments: results are not based on the highest score representing the “best” score. Instead, higher scores on a Hogan assessment usually represent positive outcomes in some ways and negative in others. The implications of this result are discussed further in the Regression Validation section.

### 

### Leadership Skill

Leadership skill was assessed as a composite of applicants’ ability to participate in a leaderless group discussion on three constructs – Managing tasks, managing people, and thinking strategically – and rated on a 5-point Likert scale. Inter-item correlations between all three items were very limited (all |*r*|< 0.09) and show that each construct should be considered individually. Correlations between managing tasks and criterion measures were moderate at best (all *r* < 0.17) suggesting that it should be safe to ignore. Managing people was strongly related to at least one measure (supervisor ratings; *r* = 0.26) and strategic thinking was strongly related to performance index (*r* = 0.30) and supervisor ratings (*r* = 0.22).

### Interview Scores

Interview scores were obtained through hiring manager ratings of applicant interviews rated on a 10-point Likert scale. Due to the lack of information regarding the number of managers who rated the applicants and which manager rated which applicant, reliability estimates were not able to be captured. Although interviews were structured, it is unknown whether manager ratings were based on a behaviorally anchored rating scale or other standardized measure. Finally, interview score correlations with criterion measures were limited (all *r* < 0.20). For these reasons, it is suggested that full information regarding scoring personnel be provided and utilized in the future for concerns relating to inter-rater reliability.

### Overall

Pearson Product-Moment Correlations were used to assess the relational strength between each predictor and each criterion. Figure 1 demonstrates the strength of each predictor-criterion relationship. Higher numbers and darker colors represent stronger relationships.

*Figure 1.* *Correlation Matrix of Predictor and Criterion Measures.*

A screenshot of a cell phone

Description automatically generated

## Criterion Analysis

### Overall

All five criterion measures were combined using the unit weighting technique. Termination was dichotomized to indicate only whether a person was fired with 1 = they are still employed and 0 = they were fired. Accidents, discipline, and supervisor ratings were scaled into z-scores while supervisor ratings were already z-scores and scaling a terminated as a dichotomized variable is illogical. Because accidents were not directly measuring manager performance but instead performance of their individual contributors, it received a weighting penalty of 0.75. It is also important to note that 21 applicants did not receive supervisor ratings. For these individuals, supervisor rating scores were imputed as 0 or the mean after being scaled. Removing these cases from the analysis was considered, but each of these individuals had been terminated. This would effectively remove over 25% of terminated cases, severely affecting termation’s quality as a criterion.

## Regression Validation

Backward stepwise regression was used to measure valid relationships between predictor measures and the combined criterion score described in Criterion Analysis. Predictor measures include scores on the Wonderlic, boldness, managing people, strategic thinking, interview scores, and the integrity scale. Given the large number of predictors and lack of theoretical evidence, interactions were excluded from this analysis. Data was and split into a training section with 75% of the data (n = 380) and a holdout set of 25% of the data (n = 126) and partitioned randomly (e.g. there was no systematic method for choosing which group an applicant was put into).

## Model 1

Results of the first regression model are summarized in Tables 1, 2, and 3.

Table 1

*Backward Regression of Composite Criterion Scores on Predictor Measures Step 1*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Predictor** | **Beta-Weight** | **B-Weight** | **Std. Error** | **t value** | **p** |
| (Intercept) | -0.29 | -1.49 | 0.11 | -2.69 | 0.01\* |
| Wonderlic | 0.19 | 0.03 | 0.02 | 8.79 | < 0.01\* |
| Boldness | 0.25 | 0.02 | 0.03 | 9.19 | < 0.01\* |
| Managing People | -0.01 | -0.05 | 0.02 | -0.66 | 0.51 |
| Strategic Thinking | -0.01 | -0.07 | 0.02 | -0.59 | 0.56 |
| Integrity Scale | 0.13 | 0.12 | 0.03 | 4.01 | < 0.01\* |
| Interview | -0.08 | -0.05 | 0.02 | -3.63 | < 0.01\* |

*Note.* \* *p* < 0.05. R2 = 0.33, *F*(6,373) = 32.23, *p* < 0.01.

Table 2

*Backward Regression of Composite Criterion Scores on Predictor Measures Step 2*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Predictor** | **Beta-Weight** | **B-Weight** | **Std. Error** | **t value** | **p** |
| (Constant) | -0.26 | -1.70 | 0.1 | -2.58 | 0.01 |
| Wonderlic | 0.18 | 0.02 | 0.02 | 10.46 | < 0.01\* |
| Boldness | 0.25 | 0.02 | 0.03 | 9.22 | < 0.01\* |
| Integrity Scale | 0.12 | 0.10 | 0.03 | 3.96 | < 0.01\* |
| Interview | -0.08 | -0.05 | 0.02 | -3.81 | < 0.01\* |

*Note.* \* *p* < 0.05. R2 = 0.34, *F*(4,375) = 48.37, *p* < 0.01.

Table 3

*Selection Ratios of Males and Females for Model 1*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sex** | **Applied** | **Selected** | **Ratio** | **Impact Ratio** |
| Male | 246 | 84 | 0.34 | 0.47 |
| Female | 260 | 42 | 0.16 |  |

### Generalizability

The predictors for Model 1 include scores on the Wonderlic, boldness, integrity scale, and interviews using 75% of applicants (n = 380). Fitting this model to the holdout sample revealed that the estimated shrinkage is negligible (ΔR2 < 0.01). Therefore, this model should have no issues generalizing.

### Adverse Impact

Although the original model produced a robust R-squared value (R2 = 0.39), using this technique with the top-down approach would create disparate impact on female applicants. Table 3 shows that the impact ratio produced by model 1 (0.47) falls far below the acceptable standard (0.80).

### Banding

To correct for potential selection biases scores may have had, sliding and fixed bands in conjunction with Model 1 were tested. Because reliability estimates were only obtainable for internal consistency of the integrity subscale (α = 0.89), 0.89 was used as a general reliability estimate for determining band size. Ideally, reliability estimates for the Wonderlic, boldness, integrity scale, and interview scores would be collected to create a more representative composite reliability estimate. Both sliding and fixed bands helped to correct for potential bias Model 1 may have for certain protected classes but failed to meet the minimum legal standard (0.80).

## Model 2

Given the significant difference between males and females on the boldness scale, a new model was tested scaling boldness as a predictor by 0.50 given the nature of Hogan assessments discussed previously. The results are summarized in Tables 5, 6, and 7.

### Generalizability

Cross validation procedures were identical to those used for Model 1 and found shrinkage to be small (ΔR2 = 0.03).

### Adverse Impact

Tables 5 and 6 demonstrate that giving boldness a lower weight creates a model that meets legal standards regarding hiring ratios when using a top-down procedure (0.85) and that boldness is the sole predictor causing adverse impact.

### Performance

Table 7 shows that using Model 2 produces only a small decrease in validity estimates (ΔR2 = 0.04) when compared to Model 1 and is the recommended model to use for selecting applicants.

See Table 5 for details on Model 2’s regression equation. Additionally, Figure 2 and Table 8 provide expectancy information on 75th percentile, 50th percentile, and using no selection tool. Most importantly, Model 2 produced robust true positive and true negative selection rates (38%; 78%) at the 75th percentile, displaying its effectiveness as a selection tool when compared to the base rates (25%; 75%).

Table 5

*Regression Results of Predictors from Model 1 on Composite Criteria Scores Weighing Boldness at 0.50*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Predictor** | **Beta-Weight** | **B-Weight** | **Std. Error** | **t value** | **p** |
| (Intercept) | -0.25 | -0.66 | 0.09 | -2.66 | 0.01\* |
| Wonderlic | 0.19 | 0.02 | 0.02 | 11.02 | < 0.01\* |
| Interview | -0.07 | -0.04 | 0.02 | -3.19 | < 0.01\* |
| Integrity Scale | 0.12 | 0.11 | 0.03 | 4.17 | < 0.01\* |
| Boldness (\* 0.50) | 0.45 | 0.45 | 0.05 | 8.57 | < 0.01\* |

*Note.* \* *p* < 0.05. R2 = 0.35, *F*(4,375) = 51.10, *p* < 0.01.

Table 6

*Selection Ratios for Males and Females for Model 2*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sex** | **Applied** | **Selected** | **Ratio** | **Impact Ratio** |
| Male | 260 | 60 | 0.23 | 0.85 |
| Female | 246 | 66 | 0.27 |  |

Table 7

*Model 1 and Model 2 Comparisons*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **R2** | **Shrinkage Estimate** | | **Residual Df** | **RSS** | **DF** | **Sum of Squares** |
| 1 | 0.39 | < 0.01 | 375 | | 38.08 |  |  |
| 2 (Boldness Weighted) | 0.35 | 0.03 | 375 | | 39.50 | 0 | -4.38 |

*Note.* F and *p* values are not applicable given that both models have identical degrees of freedom.

*Figure 2. Expectancy Scatter gram of Model 2.*

A screenshot of a cell phone

Description automatically generated

*Note.* Each dot represents an applicant. Larger dots represent two applicants with identical scores. Green dots represent the people selected if 25% of applicants were taken based on predicted scores. Blue dots and green dots represent the people selected if 50% of applicants were selected. Red dots represent all people below 50% on the cutoff score. Numbers between dotted lines represent the number of dots within their rectangular section. TP = True Positive Rate. TN = True Negative Rate. SR = Selection Ratio.

Table 8

*Expectancy Table for Model 2*

|  |  |  |
| --- | --- | --- |
|  | **Actual Score** |  |
| **Predicted Score** | Top 25% | Bottom 75% |
| Top 25% | 50 | 81 |
| Bottom 75% | 82 | 292 |

# Limitations

### Reliability

First, reliability estimates were not able to be calculated for criteria because information about interrater reliability, whether there were alternate forms, temporal stability estimates, and internal consistency were not provided/calculable for any predictor except for the integrity scale. Leadership skill was measured using three items that had very poor internal consistency and thus each of the three predictors needed to be considered separately. If they were all measuring the same thing, however, we would expect that they should produce a high internal consistency, so the leadership skill is highly suspect as a construct. The lack of reliability estimates also made creating band sizes very difficult given that the equation requires a reliability estimate.

Criterion reliability were unobtainable as well. Ideally, any type of reliability estimate for any of the criteria would help in the case that a reliability correction to validity is needed.

### Data Complications

Next, there are 21 NAs in the supervisor rating column which would be fine to omit except that all of them were terminated which would reduce the usability of being terminated as part of the criteria composite.

Additionally, termination really wants to be a trichotomous variable and could find very complex results along the lines of employees who left on voluntarily scored too high on predictor measures while employees who left involuntarily scored too little and employees who were not terminated were in the middle. A linear regression does not allow for trichotomous variables though so any effects of being too high (e.g. voluntary) and two low (e.g. involuntary) on predictor scores are undetectable. In addition, being dummy coded as 1 or 0 means that scaling the dichotomous terminated or not variable is nonsensical leading to the awkward issue of having 4 scaled predictors and one that is not. Theoretically this effect should be negligible, however, even if it is confusing.

Finally, the standardized beta-weight for interview on the criteria composite was negative and significant. This means that people scored well on interviews tended to perform poorly. It was still used as a predictor in the model because it was significant data, but this is a very messy issue for the company. Training on using standardized scales could help this in the future.

### Adverse Impact

The biggest concern regarding adverse is the effect of Wonderlic on race/ethnicity. Although we were able to show there is evidence that boldness is a BFOQ based on the large amount of unique variance it can explain, the Wonderlic is highly likely to discriminate based on race/ethnicity but we have no data to show that it does/does not create adverse impact and because it is one of the strongest predictors, it would be unwise to throw it out. On top of this, attempts to use fixed and sliding bands did not reduce adverse impact ratios enough.

### Data Issues

Something in this dataset created a scenario where the average R2 of a multiple regression model using a random sample of 25% of the data was *larger* than the average R2 of a multiple regression model using 75% of the data across 1000 random 75%/25% splits of the data without replacement. This meant that shrinkage estimates were in some cases negative and that it was hard to interpret.

### Hypotheses

There were no criteria that were stated as being more important than others. This means that unit weighting was use out of necessity to combine criteria. Ideally, there would be some indication of which criteria are more important for which job so that they could be weighted accordingly.

This leads into the second point that there are different levels and divisions of the job and we don’t know which levels or divisions the company is trying to hire for. It may be the case that the selection tool works better for certain jobs than others. In this scenario, we would want to create a selection method for each job individually. Given the sample sizes, the top level only has 96 employees. While this should be enough to create a regression model, there is not enough data to have both a training and holdout sample. This means that even if a separate regression model was created for each level of the employees, generalizability of some models would not be able to be estimated and would be considered questionable at best.

**Other**

Due to space constraints, I did not want to give each criteria its own individual analysis and instead lumped them all into one paragraph. One issue that could also have popped up is that although measure of employee behavior (e.g. discipline and accidents) corrected for team size by becoming ratios, we don’t know if this corrected enough. It could be the case that teams of 200 still have higher average *ratios* of accidents than teams of 20, which contaminates the measure, but because we don’t know their team size we will never know. On top of this, it is not clear if discipline is a good or bad thing. For example, if a manager had a low ratio, it could indicate that they have a good team (a good outcome) or that they were unsuccessful in detecting that people on their team were taking unethical actions (a bad outcome). Therefore, it was up in the air as to whether it should be used as a criterion or not.