HOGAN INTERNATIONAL NORM DEVELOPMENT GUIDELINES
Introduction

Assume a person receives a raw score of 23 on a personality scale measuring Ambition. What does this score mean? Without a basis for comparison, the score means little. Norms provide a context for interpreting scores because they allow us to compare individual scores with those of a relevant group (Nunnally, 1967). For example, to understand an executive’s Ambition score, comparing it with the scores of a group of other executives would indicate how ambitious the person is relative to his or her peers. This group of scores is called a norm. However, the quality of the norms affects the score’s interpretation.

This paper outlines Hogan’s policies and procedures for creating international norms for translated assessments. We base these guidelines on twenty years of research, the experience of translating our tools into thirty languages, and a review of the literature regarding best practices.

Considerations

Despite the importance of norms, there is little consensus regarding how to create them. A number of articles describe how to create a single norm (e.g., Deshmukh, 2004; Henningham, J. P., 1996; Pélissolo & Lépine, 2000), but few provide guidance on choosing suitable normative samples or what issues to consider when creating a norm. The primary source on norm development is the Standards for Educational and Psychological Testing (Standards: American Educational Research Association, 1999), which suggests considering the composition and relevance of normative samples to determine their appropriateness. Both of these factors impact the accuracy and meaningfulness of score interpretations.

The adequate composition of a norm depends on the sample size and the degree to which it represents the comparison population or group. For example, if the norm is intended to represent salespeople in Mexico, the sample should contain data from multiple organizations and industries (e.g., retail, pharmaceutical, automotive) and from the many geographic regions of Mexico. Further, the total number of cases (individuals) in the norm must be large enough to represent the target population. Unfortunately, the literature provides very little guidance about the sample sizes needed for norm creation. The Standards only state that norm samples should be of “sufficient size” (p. 51). Research shows that stability and representativeness in normative samples can be achieved with one hundred cases (Tett et al., in press), but only for specific jobs in specific organizations. Norms representing multiple jobs across multiple organizations should require larger samples. Hogan generally requires at least 500 cases for any norm.
The alignment between the comparison group and the purpose(s) for which the scores are to be used determines a norm’s relevance. Imagine, for example, that a wireless phone provider in Paris, France, wants to screen applicants for a customer service job. Scores based on a global norm representing multiple languages and countries may not be relevant because only French-speaking people are likely to apply. Thus, a norm based on French-speaking, working adults is more relevant. This simple example indicates that test users should consider the relevance of a norm for its intended application.

Types of Norms

For each of our translations, we create norms that represent working populations, which we call local norms. Hogan publishes two types of local norms: itinerant and general. We create itinerant norms (Butcher & Garcia, 1978) from relatively small samples; they serve as precursors to general norms, which represent a cross-section of the relevant workforce. Hogan also publishes subgroup norms for specific purposes (e.g., managerial norms). Finally, some test publishers provide global norms, which combine data from multiple languages and cultures.

**Itinerant Norms.** When Hogan initially publishes a translated assessment for commercial use, we often lack sufficient data to calculate norms that represent the entire workforce. Consequently, we rely on data from convenience samples and developmental projects to create itinerant norms. We require at least 500 cases to calculate itinerant norms. In keeping with the Standards, we report all available demographic information, including age, gender, ethnicity, and job type. Although itinerant norms are not as robust as general norms, they are useful until we have enough data to calculate general norms.

**General Norms.** Hogan considers creating general norms when data are available from at least 2,000 cases. The larger dataset is essential in order to create norms that closely resemble the composition of a given country’s workforce. Typically, we rely upon labor statistics from the target country to create norms that match the distribution of jobs within the workforce as closely as possible. We may lack any data for certain industries despite their prevalence in the general workforce (e.g., agricultural workers). Therefore, our general norms do not necessarily represent all segments of the population but do represent all of those in which the assessment and norm are likely to be used.

In developing general norms, Hogan uses multiple stratification variables—characteristics that are used to organize population data. We create norms that match the target population on each stratification variable as closely as possible. Although stratification variables may vary from country to country, job categories are almost always the first level of classification. For example, if the Brazilian workforce contains 20% managers, we would create a normative sample that also contains 20% managers. Ethnicity commonly forms the second stratum. Some countries have less workforce diversity, fewer concerns over subgroup differences, or legislation with fewer requirements than those in the U.S. Civil Rights Act of 1964 (see Myors et al., 2008 for review). In such cases, stratification by
ethnicity may be unnecessary. Similarly, gender typically forms the final stratum, but stratifying normative data by gender may not be necessary for certain countries. Nevertheless, Hogan always stratifies data by gender, as the datasets tend to be close to population proportions.

**Subgroup Norms.** Although general norms represent the working population and are relevant for most applications, some users request norms that represent only subsets of the population. For example, a user may want to compare a manager’s score with other managers in Poland. In such cases, Hogan creates subgroup norms while paying attention to composition and relevance. To calculate norms for Polish managers, Hogan would require at least 500 cases representing multiple organizations and industries in Poland. Other examples include regional, industry, and client-specific norms. Sample composition requirements vary as a function of the intended reference group. When created appropriately, these norms can provide useful supplemental information, and are especially helpful in developmental contexts.

**Global Norms.** Some test publishers create global norms by combining data from multiple countries and languages into a single dataset. This appeals to multinational companies that are attracted by the simplicity of using one norm for all applications, but Hogan avoids creating multi-language norms because of inherent inaccuracies that make interpretations from this type of norm less meaningful. Score distributions from multiple, single-language datasets can differ from one another for many reasons, including sample, translation, and cultural differences (Meyer & Foster, in press). Although test developers can try to ensure that each sample – data from a single country or language – is the same size and composition, translation differences are inevitable, and observed cultural differences should not be eliminated.

**Implementing & Updating Norms**

One last consideration concerns how and when norms should be updated. Score distributions may change when we replace an itinerant norm with a general norm, or during regular norm maintenance. However, when updating norms, it is important to consider the potential effects of the change. For developmental applications, feedback may change when norms are updated. Such changes have even greater implications for personnel selection. Prior to a change, an applicant’s scores may fall short of the norm-based cutoff scores (minimums/maximums). However, with new norms, the same raw score may pass the cutoff scores. For example, assume that a raw score of 23 on Ambition is equal to a percentile score of 47% based upon an old norm; after a minor norm change, the same score is equal to a percentile score of 52%. If the cutoff score for a particular job is 50%, under the old norms a person with a raw score of 23 on Ambition would not be recommended for selection but would be recommended under the new norms. This example illustrates the importance of examining the impact
of norm changes on existing profiles and choosing an appropriate time to implement new norms so that active selection projects are not adversely affected.

Updating norms can be difficult, and some test publishers do not do it. However, the Standards state that

“If a publisher provides norms for use in test score interpretation, then so long as the test remains in print, it is the publisher’s responsibility to assure that the test is renormed with sufficient frequency to permit continued accurate and appropriate score interpretations.”

(p. 59)

Therefore, despite the difficulties, Hogan updates norms whenever sufficient data are available for any language or country. We update general norms when we have new data from previously under-represented sectors of the population and we work with users of each form to implement new norms appropriately.
REFERENCES


