



Test Norms: All Are NOT Created Equal

HOGAN
ASSESSMENT SYSTEMS

T H E S C I E N C E O F P E R S O N A L I T Y

A Review of Best – and Worst – Practices

He got a 23!!! Can you believe it? Believe what? Individual test scores are meaningless without norms. Norms provide a frame of reference for interpreting test scores (Nunnally, 1967). As such, norms are vital for providing context for interpreting raw test scores. However, it is the *quality* of those norms that is of particular importance. By using accurate norms, one person can be compared to a suitable group of others and, relative to others, conclusions can be drawn about that person’s predicted future behavior.

This year, Hogan Assessment Systems (Hogan) completed a renorming of the Hogan Personality Inventory (HPI). The publication of the *Hogan Personality Inventory Manual (3rd Edition)* documents changes to the assessment, including the updated norms. Prior to 2007, norms were published in 1995. We observed some score shifts over the last 12 years that encouraged the renorming.

It is common for test publishers to spend resources on norming during initial assessment development. However, we suspect that it is less common for publishers to devote similar levels of resources to maintain current norms after they are initially established. If true, this is unfortunate, because test publishers have a professional responsibility to develop and maintain relevant norms.

To provide norms, test publishers collect data from “suitable and representative” people for whom the test is intended. Specifically, four standards address the norming sample that should: (a) consist of individuals for whom the test was intended and with whom examinees will be compared; (b) represent the referent population; (c) include a sufficient number of cases; and (d) be appropriately subdivided (Cronbach, 1984). Practical and professional considerations encourage publishers to establish and maintain norms. However, even these standards leave substantial room for interpretation. When it comes to norms, what sample size is “sufficient?” How can the sample be “appropriately subdivided?” What kinds of people are “suitable and representative?”

To answer these questions, the research team at Hogan conducted a review of several available assessments. This review captured information on dates norms were calculated and updated, sample sizes and sampling procedures, and metrics in which norms are presented. Table 1 contains this information. As seen, test publishers seem to disagree about the interpretation of professional standards. Specifically, disagreement appears in four major areas: (a) level of granularity; (b) sampling procedures and sample size; (c) data presentation; and (d) renorming efforts. We discuss these four areas next.

Level of Granularity. As seen in Table 1, some publishers/authors prefer a broad approach, using “global” or “general” norms. Although this approach is convenient during *initial* norm development, it is ineffective for general use. First, test scores are meaningless when compared against such broad norms. That is, it is unlikely that an applicant for a domestic Customer Service Representative job will be compared to a truly global applicant pool. Rather, it is likely that other applicants would be English-speaking Americans from the Customer Service industry. As such, more specific norms are more appropriate.

However, others take this advice to extremes, developing overly specific norms for any number of subgroups (e.g., male high school students nominated as National Science Fair delegates, female writers of children’s books). Although “drilling down” past global or general norms is appropriate for

some applications, there comes a point when the utility of these highly-specific norms disappears. In fact, as these extreme (but *real*) examples show, the reference group becomes so small that score interpretation is lost.

As illustrated, the most appropriate level of granularity likely falls between global and micro-level norms. Of course, the test's intended use should speak to the appropriate level of granularity for its norms. In most cases, this will fall somewhere between the population of the planet and male high school students nominated as National Science Fair delegates.

In our assessments, Hogan takes great care to develop appropriate norms. Hogan develops "general norms" as an initial frame of reference during the assessment's initial release. However, once sufficient data are collected, appropriate norms are calculated that are representative of the population to which individual test scores are compared. Table 1 presents this information for the HPI.

Sampling Procedures and Sample Size. The second issue facing test publishers/authors concerns sampling procedures and sample sizes used to develop norms. Simply stated, this issue comes down to, "How do we get a 'suitable and representative' norm group?" and "How many people do we need?" As seen in Table 1, test publishers offer different answers to these questions.

Some use "convenience samples" of whatever data are available to compute norms. Commonly used convenience samples include seminar participants, clients, trainees, students, and research participants. Although these data are plentiful and easily obtained, they are inappropriate for score interpretation. Specifically, it is improper to compare the test scores from job applicants to those of professionals from seminars or research participants.

A similar problem involves offering free testing – especially online – to as many individuals as possible to generate a large enough sample to calculate norms. This practice has become known as "Google-norming." As one might expect, this practice falls prey to the same pitfalls as gathering convenience samples. Specifically, it is unlikely that the population used to "Google-norm" a test will be "suitable or representative" of the individuals taking an assessment in an applied context.

Alternatively, other tests reviewed in Table 1 reflect relevant samples in their norms. Specifically, these norms are calculated so that the make-up of the workforce is properly represented in norms. That is, if workforce occupations are 20% management, 20% finance, 20% sciences, 20% healthcare, and 20% sales, these percentages will be as closely reflected in the normative sample as possible. Other criteria used to stratify samples include gender, ethnicity, and education level. Reviewing these procedures against the benchmark of "suitable and representative," one can see that representative and stratified norms are more appropriate frames of reference than convenience samples or "Google-norms."

Sample sizes represent another point of disagreement among test publishers listed in Table 1. Although Cronbach (1984) requires that "a sufficient number of cases" be included, no more concrete guidance is given on how many cases are "sufficient." In our review, sample sizes ranged from less than 200 to well over 100,000. Our own research indicates that, in general, scale statistics begin to stabilize after 400 cases or more are collected. However, fluctuations due to extreme responses, job categories, or other variables may still remain. For example, the data provided by ten extreme respondents has a much larger impact on a dataset of 200 than on a dataset of 100,000. As such, it is advisable that publishers construct norms using a large sample so that the effects of extreme responses are minimized.

For the 2007 renorming of the HPI, our research team calculated norms representative of the labor force and stratified by job category, gender, and race/ethnicity. The final sample size of our norming dataset included 156,641 cases, the largest norming dataset observed in Table 1. However, our view is that the chief feature of the

norming process is not the sample size but the representativeness with which the samples were chosen for aggregation.

Data Presentation. The third point of disparity noted in Table 1 is the variety of formats used to present norms. Not all publishers/authors even provide their norms in technical documents, citing concerns about “proprietary information.” Although we understand this argument, we find that the majority of publishers with psychometrically sound instruments have no reasonable explanation for not sharing these data.

In our review, we found three formats used to present data, with norms typically given for two of the three formats. The first is simply the presentation of raw scale scores. Although these scores are tied directly to the assessment, they are difficult to interpret because different assessments and scales have different total scores. For this reason, norms are presented in one of two other formats more easily understood.

As one alternative, some tests include norms as standardized scores. Standardized scores are expressed using a mean and a standard deviation, although these vary depending on the method of standardized used. For example, z-scores use a mean of 0 and standard deviation of 1. Alternatively, T-scores use a mean of 50 and standard deviation of 10. Sten scores use a mean of 5.5 and standard deviation of 2. Although standardized scores place assessment scores on a ranking metric, the problem with these formats is that the score ranges vary and, like raw scores, they are not commonly understood.

Percentile scores represent an alternative to standardized scores. Percentiles, like standardized scores, place an individual’s scores on a ranking metric where they are interpreted easily against others’ scores. Unlike z-scores (range -3 to +3), T-scores (range 20 to 80), and even sten scores (range 1 to 10), percentile scores use a 0 – 100% range, the most commonly understood and easily interpreted score format. For example, a raw scale score may correspond to a z-score of 1.1. However, it is difficult to interpret this standardized score. That same scale score may correspond to a percentile score of 85%, facilitating the easy interpretation that this person scores above 85% of others on that scale.

For our assessments, Hogan presents norm tables in technical manuals. These norms are presented in both raw score and percentile formats. With our assessments, we are transparent, disclosing normative data in a way that is easily understood by our partners, clients, test-takers, and the public.

Renorming Efforts. The most diverse area of information noted in our review involves norm maintenance and renorming. Existing norms may become obsolete for many reasons, but five factors can influence existing norms and signal a need for their evaluation, revision, and/or replacement:

1. Samples used to calculate existing norms may become outdated
2. New test-takers may be more familiar with the assessment type than previous test-takers
3. Individuals and groups asked to contribute to norming samples may change
4. The purpose and application of the assessment may change
5. The representation of the norming samples may change with demographic/occupational shifts

Due to these and other changes, assessment norms should be monitored, maintained, and revised, when necessary. No universal guidance is available on the frequency for updating norms because different assessments require renorming more or less often. However, if norms are to serve their intended purpose of providing an accurate context for score interpretation, even the frames of reference should be recalibrated occasionally. In fact, the 1999 *Standards for Educational and Psychological Testing* state in Standard 4.18 that “it is the publisher’s responsibility to assure that the test is renormed with sufficient frequency to permit continued accurate and appropriate score interpretation.”

As seen in Table 1, some test publishers take care to update their norms and describe these efforts in their technical documents. Others do not describe norm updates, with many of these being the same

that do not provide norms in technical documents. A third group states explicitly that they do not update their norms.

For example, one publisher included in the current review states in their technical documents that “no attempt has been made to ‘update’ certain of the previously established norms, such as the college norms, to accommodate minor fluctuations that may have occurred. Rather, the earlier normative samples have been maintained as fixed reference groups to provide for a continuity of score interpretation.” Although some advocate this view, ours is that norms cannot be considered “fixed” reference groups when the test populations those norms reflect are continuously changing.

Hogan monitored HPI score shifts since previous norms were published in 1995. Changes were sufficient in number and magnitude to renorm the assessment this year. This effort included a representative sample of working adults chosen proportionately to match U.S. Department of Labor occupational categories and stratified by ethnicity and gender. We take a similar approach for all our assessments, monitoring, maintaining, and updating norms to ensure the continued applicability of our tests to a changing population.

Conclusions

Norms provide an important frame of reference for score interpretation. Our review reveals substantial variation on how to construct, present, monitor, and update norms. For example, although professional standards require maintenance of norms, publishers take different approaches to do so. Norms may be too broad or specific, sampling procedures and sample sizes may be insufficient, norms may not be available publicly, and publishers may fail to renorm their assessments. Not all assessments are created equal. However, even among sound assessments, not all *norms* are created equally well. Simply having norms is not sufficient; it is the quality of those norms that differentiates accurate interpretation.

Table 1. Assessment Review

Organization	Assessment	Manual Date	Norms Computed	Norms Updated	Approximate Sample Size	Sampling Procedure	Norms in Document	Raw Scores	Standardized Scores	Percentile Ranks
Hogan Assessment Systems	HPI	2007	1995	2005	156,641	Representative & stratified	YES	YES	NO	YES
American Management Association	DISC	1999	Not Listed	Not Listed	2,150	Convenience sample	NO	YES	NO	YES
Hogrefe	FOCUS	2004	2002 - 2004	Not Listed	4,000 - 40,000	Convenience sample	YES	YES	YES	NO
Test Agency, Ltd.	MDQ	2004	Not Listed	Not Listed	10,000	International sample of managers & professionals	YES	YES	YES	NO
Hogrefe	MPQ	1996	Not Listed	Not Listed	Not Listed	Convenience sample	YES	YES	YES	NO
SHL	OPQ	2006	1998 - 1999	2005	270 - 17,000	Stratified for English and translated subgroups	YES	YES	YES	NO
Sigma Assessment Systems	PRF	2001	Not Listed	Not Listed	3,000	Stratified sample of U.S. & Canadian college students	YES	YES	NO	YES
Sigma Assessment Systems	SFPQ	2000	Not Listed	Not Listed	1,000	Random sample of U.S. & Canadian census respondents	YES	YES	NO	YES
PFS	TDI	2007	2003 - 2004	Not Listed	1,000	Employed professionals & college students	YES	YES	NO	YES
Psychometrics Canada, Ltd.	WPI	2001	Not Listed	Not Listed	6,000	Matched sample of males & females	YES	YES	YES	NO
The Psychological Corporation	GPP-1	1993	1991	Not Updated	500 - 5,000	Representative & stratified	YES	YES	NO	YES
Watson-Glaser	Watson Critical Thinking Appraisal	1980	Not Listed	Not Listed	175 - 1,800	Student and employee samples	YES	YES	NO	YES
Consulting Psychologists Press	CPI 260	2005	1996 - 2002	2002 - 2004	1,300 - 6,000	National samples from U.S., U.K., France & Italy	YES	YES	YES	NO
Consulting Psychologists Press	CPI	2002	2002	1996	6,000	Matched sample of males & females & various subgroups	YES	YES	NO	NO
Wonderlic	PCI	2002	Not Listed	Not Listed	4,500	National sample with occupational & educational subgroups	YES	YES	NO	YES
Psychological Assessment Resources	SDS	1987	1970 - 1978	1985	2,500	Convenience sample	YES	YES	NO	YES
Wonderlic	ERI	1993	Not Listed	Not Listed	Not Listed	Not Listed	NO	N/A	N/A	N/A
Minnesota Report: Personnel Selection System	MMPI - 2	1989	Not Listed	1989	Not Listed	National sample with various subgroups	NO	N/A	N/A	N/A
Consulting Psychologists Press	MBTI	2003	1998	1998	3,000	Representative sample of U.S. population	NO	N/A	N/A	N/A

Table 1. Assessment Review (Continued)

Organization	Assessment	Manual Date	Norms Computed	Norms Updated	Approximate Sample Size	Sampling Procedure	Norms in Document	Raw Scores	Standardized Scores	Percentile Ranks
Center for Creative Leadership	CISS	1992	Not Listed	Not Listed	5,000	Convenience sample	YES	YES	YES	NO
Theodore Millon	MCMI-II	1987	Not Listed	Not Listed	1,300	Randomly selected clinical patients	YES	YES	NO	YES
Center for Creative Leadership	COS	1990	1988	Not Listed	2,800	Convenience sample	YES	NO	YES	NO
Center for Creative Leadership	CLI	1991	1988	1990	1,700 - 7,400	Convenience sample	YES	NO	YES	NO
S.F. Checkosky & Associates	Accurater	1991	Not Listed	Not Listed	6,000	Convenience sample	YES	YES	NO	YES
Profiles International	Profile XT	2006	2005	1992 - 1998	116,000	Convenience sample	NO	YES	YES	NO
Pearson Assessments	GZTS	1976	1950s - 1970s	Not Listed	15,000	100 specific subgroup norm datasets	YES	YES	YES	YES
Previsor	GPI	2000	Not Listed	Not Listed	300 - 460	Domestic & international convenience samples	NO	N/A	N/A	N/A
iPAT, Inc.	16PF	1994	2002	1988 - 1993	10,000	Representative & stratified	YES	YES	YES	YES
Ashton & Lee	HEXACO-PI	2004	Not Listed	Not Listed	2,415	Not Listed	NO	YES	NO	NO
Psychological Assessment Resources	NEO-PI-R	1992	1991	1989	1,000 (Form S)	Representative sampling of projections for 1995 U.S. census	YES	YES	NO	YES
PsyTech International	15FQ+	2002	1999 - 2000	Not Listed	1,200	Representative sampling	NO	YES	YES	NO
Psychological Services	EAS	2001	Not Listed	Not Listed	Not Listed	Representative of occupational & educational groups	NO	YES	NO	YES
PsyTech International	JTI	2005	Not Listed	Not Listed	Not Listed	Not Listed	NO	N/A	N/A	N/A
PsyTech International	OPP	2006	Not Listed	Not Listed	1,900	Male/Female samples and educational & occupational subgroups	NO	N/A	N/A	N/A
Harcourt Assessment	Raven's Progressive Matrices	1994	1992 - 1993	1979	600 - 1,400	U.S. & British general norms and educational & occupational subgroups	YES	YES	NO	YES
Goldberg et al.	IPIP	N/A	N/A	N/A	N/A	NO NORMS	N/A	N/A	N/A	N/A

References

- American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- American Management Association (1999). *AMA DISC survey*. Center for Applied Research, Inc.
- Bakker, S., & Macnab, D. (2001). *Work Personality Index (WPI) user's manual*. Edmonton, AB, Canada: Psychometrics Canada, Ltd.
- Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ technical manual (Version 2.0)*. New York: SHL Group.
- Borofsky, G. L. (1993). *User's manual for the Employee Reliability Inventory (ERI) screening system*. Libertyville, IL: Wonderlic Personnel Test, Inc.
- Briggs-Myers, I., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (2003). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator (3rd ed.)*. Palo Alto, CA: Consulting Psychologists Press.
- Cameron, A. (1996). *MPQ technical and user manual: Manchester Personality Questionnaire Range*. Oxford, UK: Hogrefe – The Test Agency, Ltd.
- Cameron, A. (2004). *FOCUS: Focus on Characteristics Underlying Style*. Oxford, UK: Hogrefe – The Test Agency, Ltd.
- Cameron, A. (2004). *MDQ user's manual: Management Development Questionnaire*. Oxford, UK: The Test Agency, Ltd.
- Campbell, D. (1990). *Campbell Organizational Survey (COS)*. Minneapolis, MN: National Computer Systems, Inc.
- Campbell, D. (1991). *Campbell Leadership Index (CLI)*. Minneapolis, MN: National Computer Systems, Inc.
- Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB-SCII Strong-Campbell Interest Inventory: Form T325 of the Strong Vocational Interest Blank (3rd ed.)*. Stanford, CA: Stanford University Press.
- Campbell, D. P., Hyne, S. A., & Nilsen, D. L. (1992). *Manual for the Campbell Interest and Skill Survey (CISS)*. Minneapolis, MN: National Computer Systems, Inc.
- Checkosky, S. F. (1991). *AccuRater: Office skills assessment battery*. Liverpool, NY: S. F. Checkosky & Associates, Inc.
- Childs, R., & McDonald, A. (2007). *Type Dynamics Indicator (TDI) technical summary*. Maidenhead, UK: Profiling for Success.
- Conn, S. R., & Rieke, M. L. (1994). *16PF technical manual: Fifth edition*. Champaign, IL: The Institute for Personality and Ability Testing, Inc.
- Costa, P. T., & McCrae, R. R. (1992). *NEO-PI-R professional manual: Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Lutz, FL: Psychological Assessment Resources, Inc.
- Cronbach, L. J. (1984). *Essentials of psychological testing (4th ed.)*. New York: Harper & Row.

- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.
- Gordon, L. V. (1993). *Gordon Personal Profile-Inventory (GPP-I) manual* (1993 Rev. ed.). San Antonio, TX: The Psychological Corporation; Harcourt Brace & Company.
- Gough, H. G., & Bradley, P. (2002). *California Psychological Inventory (CPI) manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
- Gough, H. G., & Bradley, P. (2005). *California Psychological Inventory-260 (CPI-260) manual*. Mountain View, CA: Consulting Psychologists Press.
- Guilford, J. S., Zimmerman, W. S., & Guilford, J. P. (1976). *The Guilford-Zimmerman Temperament Survey (GZTS) handbook: Twenty-five years of research and application*. San Diego, CA: Edits Publishers.
- Hammer, A. L. (Ed.) (1996). *MBTI applications: A decade of research on the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.
- Hathaway, S. R., McKinley, J. C., & Butcher, J. N. (1989). *Minnesota Multiphasic Personality Inventory – 2*. Minneapolis, MN: National Computer Systems, Inc.
- Hogan, R., & Hogan, J. (2007). *Hogan Personality Inventory manual* (3rd ed.). Tulsa, OK: Hogan Assessment Systems.
- Hogan, R., Hogan, J., & Warrenfeltz, R. (2007). *Hogan guide*. Tulsa, OK: Hogan Assessment Systems.
- Holland, J. L. (1972). *Professional manual for the Self-Directed Search (SDS): A guide to educational and vocational planning*. Palo Alto, CA: Consulting Psychologists Press.
- Holland, J. L. (1985). *The Self-Directed Search (SDS) professional manual* (1985 ed.). Odessa, FL: Psychological Assessment Resources.
- Holland, J. L. (1987). *1987 manual supplement for the Self-Directed Search*. Odessa, FL: Psychological Assessment Resources.
- International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (<http://ipip.ori.org/>). Internet Web Site.
- Jackson, D. N., Paunonen, S. V., & Tremblay, P. F. (2000). *Six Factor Personality Questionnaire (SFPQ) manual*. Port Huron, MI: Sigma Assessment Systems.
- Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO personality inventory: Two new facet scales and an observer report form. *Psychological Assessment, 18*, 182-191.
- Millon, T. (1987). *Manual for the Millon Clinical Multiaxial Inventory – II (MCMI – II)* (2nd ed.). Minneapolis, MN: National Computer Systems, Inc.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
- Profiles International, Inc. (2006). *Profile XT technical manual*. Waco, TX: Profiles International, Inc.

- PsyTech International. (2002). *Fifteen Factor Questionnaire (15FQ+) technical manual*. Bedfordshire, U.K.: Psychometrics, Ltd.
- PsyTech International. (2005). *Jung Type Indicator (JTI) technical manual*. Bedfordshire, U.K.: Psychometrics, Ltd.
- PsyTech International. (2006). *Occupational Personality Profile (OPPro) technical manual*. Bedfordshire, U.K.: Psychometrics, Ltd.
- Raven, J. (1994). *Occupational user's guide: Raven's Progressive Matrices & Mill Hill Vocabulary Scale*. United States: Harcourt Assessment, Inc.
- Ruch, W. W., Stang, S. W., McKillip, R. H., & Dye, D. A. (2001). *Employee Aptitude Survey (EAS) technical manual* (2nd ed., Version 2.2). Glendale, CA: Psychological Services, Inc.
- Russell, M., & Karol, D. (2002). *16PF administrator's manual: Fifth edition with updated norms*. Champaign, IL: The Institute for Personality and Ability Testing, Inc.
- Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, 53, 153-193.
- Sigma Assessment Systems (2001). *Personality Research Form (PRF)*. Port Huron, MI: Sigma Assessment Systems.
- Watson, G., & Glaser, E. M. (1980). *Watson-Glaser Critical Thinking Appraisal: Forms A and B manual*. San Antonio, TX: The Psychological Corporation; Harcourt Brace & Company.
- Wonderlic, Inc. (2002). *Personal Characteristics Inventory (PCI)*. Libertyville, IL: Wonderlic Personnel Test, Inc.