PSYTECH
INTERNATIONAL

Norms for Psychometric Assessment

Understanding, choosing, and constructing the right norms

TABLE OF CONTENTS

Introduction 3
Background Information 3
Mean average 3
Standard deviation 3
The normal distribution 4
Moving from Raw to Normed Data 4
Choosing and Using Norms 6
The Right Norm for the Job 6
Generic, Specific, and Customised Norms 6
Broad, generic or general norm groups 6
Narrow or specific norm groups 6
Customised norm groups 6
Considerations for Specific and Custom Norms 7
Demographics 7
Context of testing 7
Industry-Specific Norms 7
Developing and Building Norms 7
Why and How is a Norm's Spread Important? 7
Clustered but varied in a meaningful way 8
Balanced 8
Basic Norming Challenges 10
Unclean data 10
Data duplication 10
Age of data 10
Restriction of Range – a Serious Norming Challenge 11
Why is Restriction of Range a Problem? 11
Biased (incorrect) mean average scores 11
Underestimation of score variation 11
Excessive skew 11
Restriction of Range and Test Validation 13
Norm Groups vs Validation Studies 13
Industry-Based Norms are a Blessing And A Curse 14
Restriction of Range and Industry or Role-Specific Norms 14
Avoid Using Industry-Based Norms for Selection 14
Ideal profiles as an aid to interpretation 14
When Customised or Industry-Based Norms can be Used 15
Research Projects 15
Comparison between groups 15
A Quick Word on using Norms across Countries or Cultures 16
Summary 17
Bibliography and Further Reading 17
Appendix I - Checklist for Building Custom Norms 18
Appendix II – VMI Norm Summary 19
Appendix III – OIP+ Norm Summary 47

INTRODUCTION

As people, we tend to define others in terms of their observable characteristics and behaviours. We make sense of this information by 'measuring people up' against one another, considering how people (including ourselves) differ or are similar to one another. Psychometric assessment is a formal version of this evaluative process, and involves taking a scientific approach to quantifying the abilities, traits, and characteristics of human beings. Many psychometric assessments make comparisons between people, and those that do so meaningfully rely on normative samples.

A normative sample (or 'norm') is a group of scores from other people who've completed an assessment, against which a respondent's score(s) on that particular assessment are considered. This norm provides a context or formal frame of reference which allows us to interpret the results in a meaningful way. For example, a score of 18/24 on a measure of social confidence might imply that a candidate is exceptionally socially outgoing. However, comparing this score to a group of other respondents' scores might reveal that our candidate is in fact slightly less socially confident than others. Having this sort of external reference offers a number of benefits:

- The norm provides a '**yardstick**' as to the average score of some particular group of people. (E.g., perhaps the person is less socially outgoing than the average *salesperson*)
- The norm indicates the **variability** of these scores around the average – do people all tend to score very similar to one another, or are their scores widely spread?
- Norms that are based upon **real-life** data (e.g., job applicants) rather than laboratory samples allow the norms to account for the effects of test-taker anxiety, time demands, etc.

BACKGROUND INFORMATION

Before talking about norms in any depth, it's important to clarify some basic statistical concepts such as *averages*, *standard deviations*, and the *normal distribution.*

Mean average

The mean average is an indication of the mid-point of your data-set (that is, the number around which other numbers are clustered). It is simply the *sum of all the scores* in which you're interested, *divided by the number of scores*. Note that there are actually three types of average (mean, median, mode), but for our discussion of norms we're only interested in the *mean* average.

Standard deviation

The standard deviation (SD) serves as a measure of how spread out are the data within the data set in which you're interested. The SD can be thought of as the *average distance of scores from their average*.

Example 1 - Mean, SD, and Normal Distribution

> After a recent physics assignment, the scores for a group of 10 students are as follows:
>
> *93, 78, 72, 75, 70, 61, 62, 73, 82, 86,*
>
> The *average* score for this group of students is 75.2 – as the total of the scores is 752 and 752/10 = 75.2. The *standard deviation* (SD) for this group of students is 10.03 – that is, on average, scores are approximately 10.03 points above or below the average of 75.02.

The normal distribution

The terms 'normative sample' and 'norm' refer to a *normal distribution*. A normal distribution occurs when scores or measures are clustered around a single central average score (known as being 'unimodal') in a particular way, which means that the further from the average a score might be the less-common that score. If graphed as a histogram, with raw scores on the horizontal (x-axis) and the frequency of these scores on the vertical (y-axis), this creates a 'bell-shaped curve', as demonstrated in Figure 1 - An example of a normal distribution Figure 1.

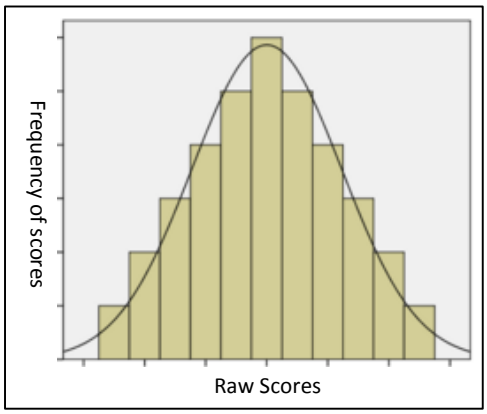


Figure 1 - An example of a normal distribution

In a normal distribution, approximately 68% (2/3) of scores are within one SD from the mean, 28% are between one and two SD of the mean (either above or below), and only 4% are two or more SD above or below the mean. Figure 2 demonstrates this but it is an important fact because, provided the norm is a valid representation of the population in question, the further a score is from the mean, the less likely a respondent is to obtain it.

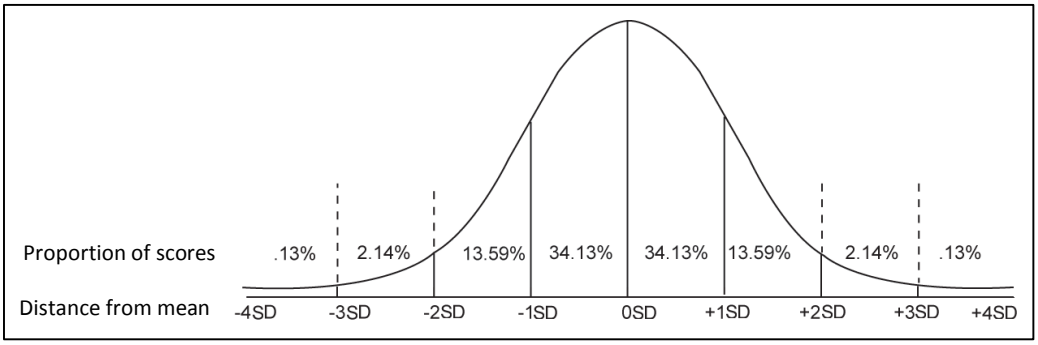


Figure 2 - Proportion of scores within particular sections of the normal distribution.

MOVING FROM RAW TO NORMED DATA

Having discussed the statistical foundation upon which the norming of data is performed, it is also important to understand the relationship between raw data and norms. Raw data, norms, and normed data are three different, albeit related things (illustrated in Example 2, below):

- **Raw datum/data** - the outcome or numerical result of some measurement – sometimes referred to as a *raw score*. In Example 2 (below) the first patient's height measurement is a raw datum.
- **Norm -** a collection of raw scores which are processed and grouped together to serve as a comparison group. In Example 2, the latter 200 patients' height measurements are assembled into a norm.
- **Normed data** - the result of mathematically comparing raw data against a norm. Although Example 2's approach to 'norming' patient 1's data isn't particularly formal, the determination that patient 1 is *taller than the average, but not exceptionally so* could be considered a very rough piece of normed data.

Example 2 –Patient Height

Imagine that a doctor needs to determine whether a male patient is tall, short, or of average height. She measures the man's height and finds it to be 184cm. This measurement represents a single, raw data-point. It is 'raw' because no transformations have been performed, nor has there been any formal attempt to interpret the score – so far, it's just a measurement. The doctor decides to use her other patients as a comparison, and measures the height of the next 200 males she treats (she is a very busy doctor!)

The doctor looks at the results of her 200 measurements, and finds that while the heights vary somewhat, they tend to be centred or clustered around a height of 177cm – this is the average (mean) height of the males in her comparison group. Since the original patient's height of 184cm is larger than this, he is considered *taller than the average*. The doctor also notes that two thirds of the measured male patients have heights which are within 6cm of 177cm, and with a few exceptions (less than 4%) the remaining third of patients are within 12cm of 177cm. So, while the original patient is taller than the average for this group, he is *not exceptionally taller*, since 184cm is within 12cm of 177cm (i.e., he is between 1 and 2 SD above the mean).

While this example deals with height, the same principle and process also applies to psychometric test results, such as those from a test of abstract reasoning (AR) ability, as demonstrated in Figure 3:

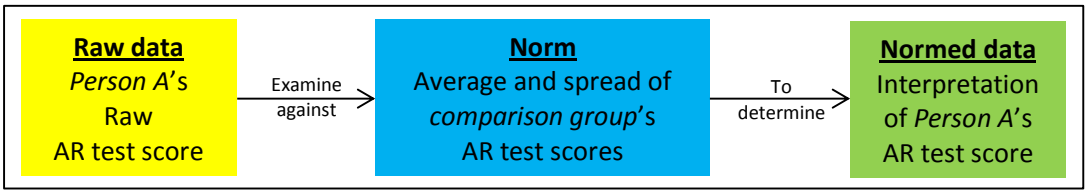


Figure 3 - Interpreting a candidate's raw test score by comparing it to a norm.

All modern assessment platforms will automatically perform the 'norming' of raw scores when scoring assessments. However the basic process is very straightforward and involves subtracting the mean from the respondent's score, dividing the result by the SD to give a z-score, and converting the z-score to stanine, sten, percentile, etc., to make the normed result easier to interpret, as demonstrated in Figure 4. Please note that the process for norming scores on 'adaptive' tests (e.g., Adapt-g) is conceptually similar but technically different.

$$\text{Normed score (Stanine)} = 5 + 2\left(\frac{raw\ score - norm\ mean}{norm\ SD}\right)$$

Figure 4 – Formula for converting a raw score to a normed stanine (standard nines) score.

CHOOSING AND USING NORMS

THE RIGHT NORM FOR THE JOB

For our interpretation of a respondent's test scores to be genuinely meaningful, we need to choose the correct group against which to interpret these results. When choosing a norm, it's important to think about what information you wish to obtain from the respondent's test results. If other people have also completed the same test for the same reason, then we would also compare them against the same norm. Example 3 demonstrates a situation where different norms could produce different interpretations:

Example 3 - The Typing Test

> Image that you were trying to interpret the results of someone who's just completed a computerized typing test. Trying to interpret their test scores against a norm group of data-entry professionals is likely to yield a very different interpretation than if we'd compared them against a norm group of primary school children. If the respondent had applied for a position which required them to quickly and accurately enter information using a keyboard then we would want to use the *data-entry professional* norm group. If they were a student in a computing class and we wanted to judge their performance against other school-age computing students, then the *school children* norm would be a better choice.

With regards to assessments completed for selection and employee development, the norm should be chosen based upon the requirements of the position applied for, the posting sought after, etc. This allows for easy and reliable comparison across candidates and consistent interpretation of results across selection cycles. In this context, norm groups should not be selected based upon demographic characteristics of the respondent (e.g., ethnicity, age, gender), nor should multiple norms be used to interpret a single respondent's results on a particular test.

GENERIC, SPECIFIC, AND CUSTOMISED NORMS

Since a norm is only as informative as the test results it contains, it is important to think of norms in terms of whether they are generic or specific, and if specific, whether they are customised to a particular purpose:

Broad, generic or general norm groups

These (often large) norms typically contain the results of all respondents who've completed the assessment. Norm membership may be restricted in terms of nationality, recency of results, etc. but the norm is effectively a broad, non-specific group of respondents. Examples could include *International Respondents* or *New Zealand Applicants*.

Narrow or specific norm groups

These contain the results of a particular group of respondents, who were chosen for inclusion into the norm based upon one or more important characteristics. For occupational tests some common characteristics include:

- The **sector** in which the respondents work (e.g., *public service*)
- The **industry** in which the respondents work (e.g., *hospitality industry workers*)
- The **level of job** (i.e., seniority) of respondents (e.g., *managerial professional*)
- The **type of job** held by respondents (e.g., *process worker*)

Specific norms sometimes also serve as a 'cross-section' of these characteristics, e.g., *public service managers*.

Customised norm groups

These are norm groups that are customized to a particular project or user, and are often very specific. A customised norm group might include only respondents who've applied for some particular job (e.g., *Case-Managers*) or applicants for positions at a particular level within a given organization (e.g., *Company X's Managerial Staff*). Their specificity allows customized norms to provide a narrower and arguably more relevant interpretation of results than an equivalently sized 'generic' norm. However, this specificity also inhibits the sample size, limits the contexts under which the norm can be used, and can result issues of *restriction of range* (discussed in a later section).

CONSIDERATIONS FOR SPECIFIC AND CUSTOM NORMS

When choosing a specific or customized norm it is crucial to consider the how well the norm group reflects the wider population of people it's intended to represent. Two key considerations are:

1) Respondent and norm-group demographic details (qualities *and* quantities)
2) The context in which testing is being completed

Demographics

The demographic makeup of the comparison group should approximate, both quantitatively and qualitatively, the demographics of the wider population in which you're interested. This means that the comparison-group should be a representative sample of your staff, country's population, etc. both in terms demographic qualities and the proportion of the population possessing these qualities. For example, if 30% of your organisation's employees are aged between 25 and 30, then approximately 30% of your comparison group should also be aged between 25 and 30.

Context of testing

If tests are being completed for employee selection then these results should be interpreted against a group of respondents who'd also completed tests as part of a selection process. Likewise, if an ability/aptitude test is completed under supervised and time-bound conditions then the norm group should be comprised of people who also completed the testing under supervised, time-bound conditions.

In addition, the comparison group's scores should equate with what you'd expect of this population. For instance, a comparison group of experienced lawyers should score reasonably well on a test of legal knowledge. If not, it may be that the sample of respondents was biased in some way – such as if only law students were to have completed the assessments. If the sample was not biased, the particular assessment in question may be a poor choice for that specific testing context – for instance, senior lawyers may score poorly on a legal knowledge test (intended for recent graduates) which places high importance on the speed of responding.

INDUSTRY-SPECIFIC NORMS

Industry-specific norms are arguably the most common form of specific norms and can prove particularly helpful in regards to interpreting results on 'general-purpose' assessments (e.g., GRT1, 15FQ+) within specific contexts. For instance, if a candidate were to complete a reasoning ability assessment as part of their application for a junior accountant role in a large accountancy firm, their results could be evaluated against a norm group comprised of other candidate accountants. This would provide clear insight as to how our candidate has performed relative to applicants for similar roles in similar organisations. Contrast this with a generic/general norm, which would only tell us how the candidate has performed with regards to anyone who'd completed the test – leaving the user to make their own judgment as to how the candidate performed relative to other applicants in their industry, which may very well turn out to be incorrect.

If the accountancy firm from the previous example were assessing high volumes of candidates for their junior accountancy roles they might also consider constructing their own customised norms. This would allow the firm to evaluate new candidates against the pool of previous candidates. Being able to quickly know that a particular candidate has scored above the average for accountant candidates at that particular accountancy firm affords much simpler and reliable interpretation and decision-making than merely knowing that the candidate has scored above the average respondent.

The benefits and problems associated with industry-specific norms are discussed further in a subsequent section.

DEVELOPING AND BUILDING NORMS

WHY AND HOW IS A NORM'S SPREAD IMPORTANT?

Not all norms are created equal – sample size is important however a large norm isn't necessarily a high-quality norm. In fact in many situations a sample size larger than 200 is likely to generate diminishing returns for the test user. What is more important is that the scores are distributed (spread out) in a meaningful way - that is, the variability of scores can be useful in interpreting test results. Most normed assessments are intended for scores which are normally distributed, which means that the scores are:

Clustered but varied in a meaningful way

This entails two separate ideas:

- The majority of respondents should have scored at or near a single point (the average), with results being progressively less common the further they are from the average. Contrast this with scores that are evenly distributed across the possible range (a 'platykurtic' distribution), which would lead to differences in scores being understated.
- However, if the data is insufficiently varied due to the majority of your respondents having scored exceptionally similarly to one another (a 'leptokurtic' distribution) then score differences will be greatly exaggerated when the norm is considered.

Balanced

The number of scores above and below the average needs to be approximately equivalent. If there are extreme scores which 'skew' the mean average upwards or downward, then differences in scores on one side of the average will be exaggerated, and differences in scores on the other side of the average will be understated.

Table 1 provides examples of a normal distribution, along with examples of some non-normal distributions. If the data within the comparison group fails to meet these requirements then test results will be misinterpreted, both in regards to the comparison group, and other test respondents, making comparison between applicants unreliable.

Please refer to Appendix I for a checklist of key requirements when assembling custom norms.

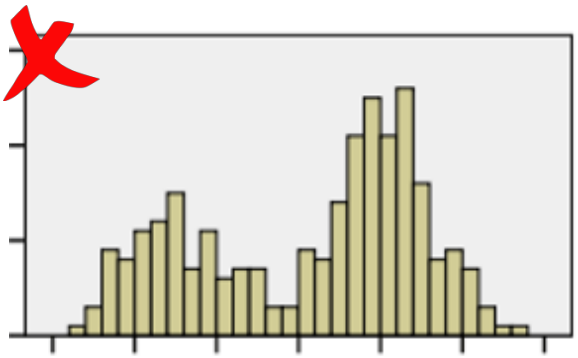
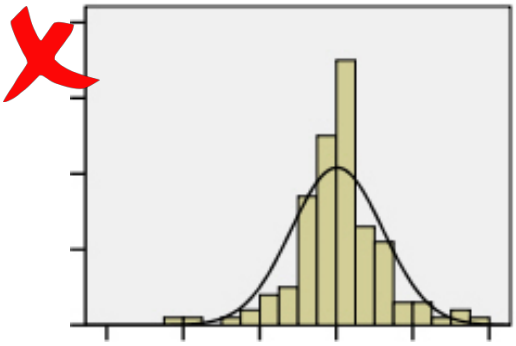
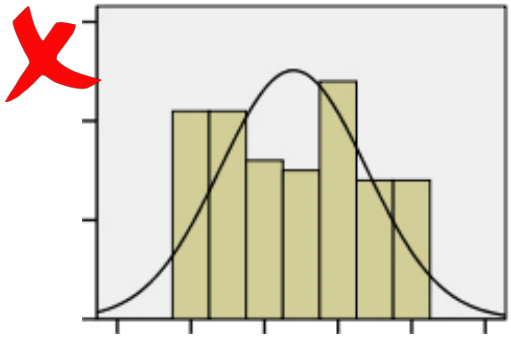
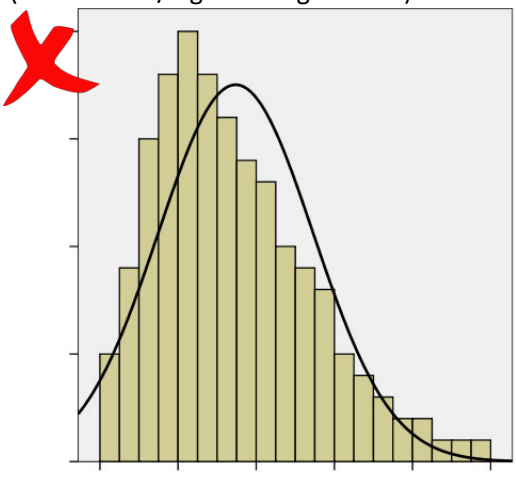
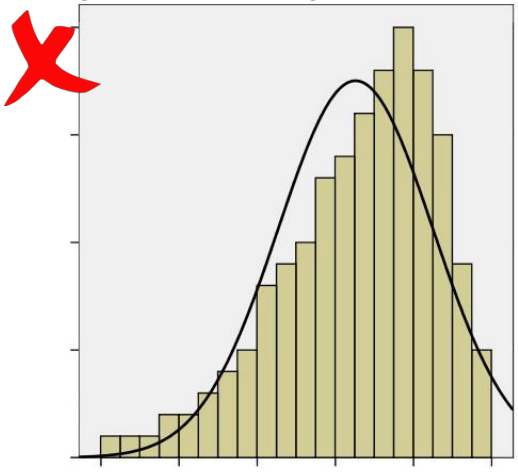
A normal distribution (clustered and balanced about a single average) 	**Two 'average points'** (bimodal) 
Not enough variation/spread (Leptokurtic/peaked) 	**Not enough clustering/too evenly spread** (Platykurtic/flat) 
Positive Skew (floor effects/high-scoring outliers) 	**Negative Skew** (ceiling effects/low-scoring outliers) 

Table 1 – Examples of normal and non-normal distributions. Note that the horizontal (x) axis represents test-scores, and the vertical (y) axis represents the frequency or quantity of respondents attaining this score. A normal distribution has been overlain as a black line for all distributions, except the bimodal example.

BASIC NORMING CHALLENGES

While the concept of norms seems simple enough, constructing or even choosing good norms can be challenging. As discussed earlier, a norm should have at least 200 candidates, scores must be clustered around a single central/average score, and the spread of results will need to be meaningful. Some common issues which need to be addressed when constructing norms are:

Unclean data

Not all test data is relevant or even valid. For instance, psychometric tests might be completed for demonstration or trial purposes, and fatigued or frustrated respondents may randomly respond to test items without having read the questions. This 'unclean' data needs to be excluded from your norm as it isn't meaningful.

Data duplication

If respondent data is duplicated within a particular dataset this reduces the variability (SD) of the dataset and can bias the average score. Consequently, it is important that each respondent's data only appear once within a given dataset. The exception to this is when respondents have genuinely completed the assessment on two or more occasions, and practice effects are unlikely to have influenced their scores. For example, when considering results of assessments in which items are randomly drawn from an item pool (e.g., CRTBi, GRTi).

Age of data

Data can be obsolete if it is old enough to no longer reflect the population it is intended to represent. Two examples of where this may have occurred is where a large-scale social event has occurred (e.g., war or extreme natural disasters) which may affect people's responses to preference measures, or the well-known *Flynn effect* in which IQ scores appear to be progressively trending upwards over time.

RESTRICTION OF RANGE – A SERIOUS NORMING CHALLENGE

Another concern is that of restriction of the range of results. Restriction of range is said to have occurred when membership in the norm group is *biased* in some way that was unintentional or not part of the criteria for inclusion into the norm. This is particularly relevant to assessments used in employee selection, as in many selection processes candidates need to pass screening criteria (e.g., curriculum vitae, application form, initial interview) before they are invited to complete psychometric assessments. Consequently the pool of applicants completing the assessments have been 'filtered' and are no longer a representative sample of the wider applicant population.

Higher scores tend to be more desirable than lower scores for many assessments which are used for selection and developmental purposes – although this is not always the case. As a consequence, restriction of range tends to be most easily conceptualised as the exclusion of data from respondents who've scored towards the lower-end of a distribution. However, it is important to realize that the restriction (i.e., exclusion of data) *can occur at either or both end(s) of a distribution*.

WHY IS RESTRICTION OF RANGE A PROBLEM?

Restriction of range has a number of substantial implications which limit or even negate the usefulness and validity of a norm. In short, examining range-restricted data only gives us 'part of the picture', which leads to incorrect interpretation of test results and poor selection decisions. Some specific detrimental outcomes of a restriction in range (i.e., the incorrect exclusion of data) are:

Biased (incorrect) mean average scores

The exclusion of data from the sample results in the mean average being calculated as higher (or lower) than it actually is. Consequently, candidates' scores may be perceived as being below (or above) the mean, when the opposite is in fact true.

Underestimation of score variation

The culling of data from one or both ends of the distribution leads to the spread of results appearing much narrower than it actually is, leading to a smaller apparent standard deviation. Consequently candidate's scores may appear more extreme (i.e., further from the mean) than they actually are.

Excessive skew

The exclusion of (typically) low scores serves to 'unbalance' the distribution, and introduces false 'floor-effects' (i.e., a synthetic minimum score which lots of respondents appear to have achieved). The resulting positive skew:

- Exaggerates the scores of those who've scored below the true average
- Understates the scores of those scoring above the true average
- Scores harshly those who'd scored in the average range (i.e., within 1 SD of the mean)

Example 4 provides a detailed example of how restriction of range might arise, and its impact on the interpretation of results.

Example 4 - Restriction of Range

A large organization wishes to build a norm to help them select recent university graduates for a fast-track development programme, by using data from real candidates. Once the norm is available, the organization will determine whether to interview candidates based on their (1) grade-point average (GPA); and (2) score on a test of general reasoning ability (once the new norm is ready). When developing the norm the organization could test all candidates, or it could choose to only assess those candidates whose GPAs exceed a particular minimum. Figure 5 below demonstrates both options, with the red line denoting a hypothetical minimum GPA score the organization might implement, and the blue line denoting a hypothetical minimum score on the reasoning ability test:

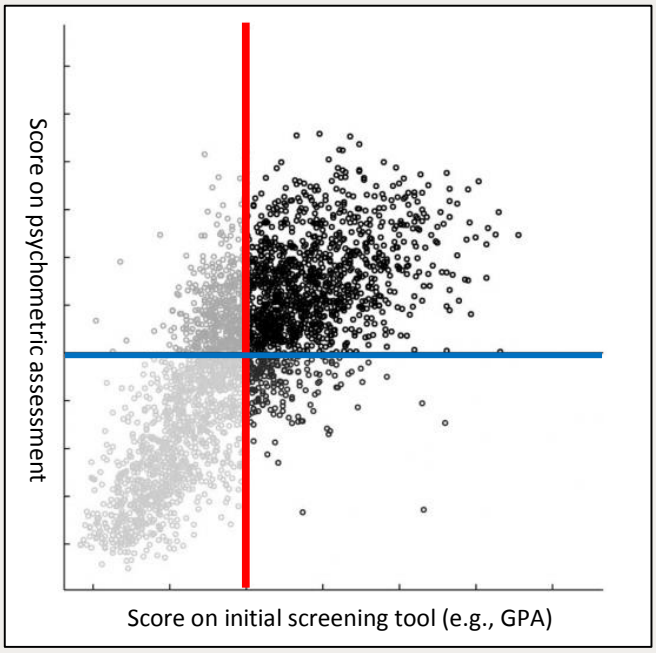


Figure 5 – Decreased test score variability as the result of screening candidates prior to test administration. Note that the red line indicates the minimum criteria on the screening tool. The blue line denotes a hypothetical minimum acceptable score on the subsequent psychometric test.

This figure suggests that if all candidates were to complete the reasoning ability test, scores would be observed to range broadly. However, if only candidates with GPAs above the minimum score were invited to test, the 'greyed' data-points to the left of the thick-red line would not have been gathered. The result of this would be as follows:

(1) The mean **average appear to be much higher**, making the interpretation of the test 'harsher' on candidates – a candidate would need to score much more highly to be in the 'average range'.
(2) The observed **range of scores would be much smaller**, meaning the SD would be under-estimated, and a given candidates' distance from the mean (relative to the norm group) would be exaggerated.
(3) The company's chosen minimum score would **seem much easier to achieve** than it actually is. In this example, with restricted range the minimum score would appear to be well below the 'biased' mean average score

RESTRICTION OF RANGE AND TEST VALIDATION

One additional consideration is the impact that restriction of range can have on validating the assessment (i.e., determining the relationship between test results and job performance). Since selection decisions tend to be based at least partly on assessment results, the organization will only have performance data for people who've achieved the minimum or desired score on the test. As a result, the true relationship between test scores and job performance can be attenuated, or completely misrepresented since there are no performance data for people who've scored below the minimum or desired score.

Figure 6 below provides an example of this situation, in which the green line denotes the strong positive relationship between test scores and job performance that would be observed if all candidates who'd completed tests were hired. If a minimum score were introduced (as denoted by the vertical red line) the observed relationship between test scores and job performance would appear much weaker, and is denoted by the orange line.

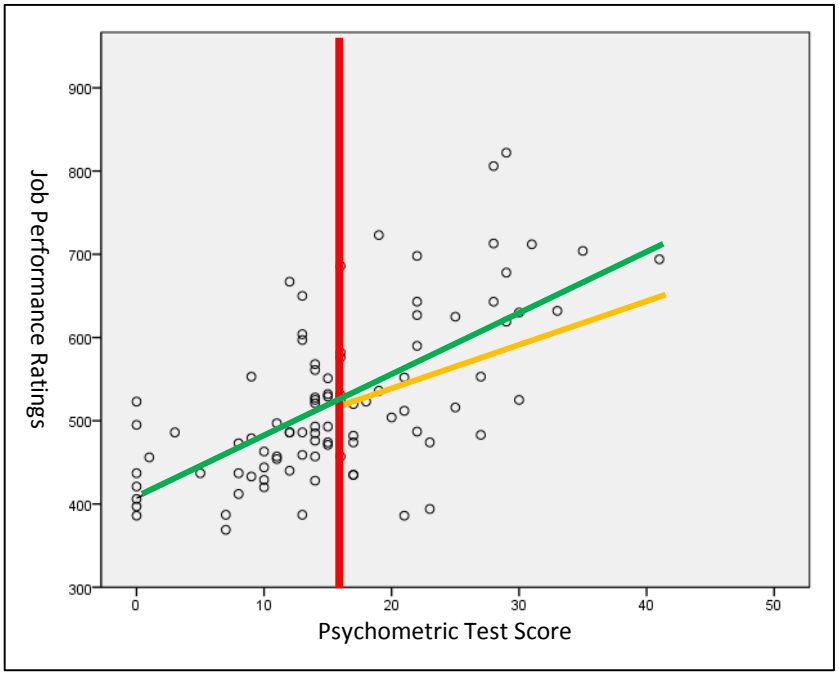


Figure 6 - Hypothetical relationship between psychometric test scores and job performance ratings. Note that green line denotes actual relationship, and the orange line denotes percieved relationship if range is restricted.

Norm Groups vs Validation Studies

Norm groups and groups of candidates or scores used to validate a test (or even the norm itself) are two different, independent concepts. Scores from the **norm group** are used to *interpret* scores by acting as a comparison. Whereas scores from the **validation study** are used to verify that the assessment is correctly measuring the ability, trait, etc. For occupational tests, this typically involves comparing test scores (normed or raw) against some measure of job or training performance to determine the '*criterion validity*' (either *predictive* or *concurrent* depending on whether test scores were from applicants or current incumbents). When developing a test or norm it is also common to evaluate the '*content validity*' of a test by comparing respondent's scores from a validation study against their scores on another assessment which measures either a similar construct ('*convergent validity*') or an unrelated construct ('*divergent validity*').

INDUSTRY-BASED NORMS ARE A BLESSING AND A CURSE

RESTRICTION OF RANGE AND INDUSTRY OR ROLE-SPECIFIC NORMS

Due to their narrow and intentionally biased nature, industry and role-specific norms are particularly susceptible to issues arising from restriction of range. The reason for this is that industry and role-specific norms are typically based upon real-world data obtained from respondents who were successful in their application for a particular role. Given that the organisation's decision to hire was likely based upon the very assessment results which the respondent is contributing to the norm group, this biases the dataset substantially. Even when non-successful candidate's data is included in a norm, these respondents have usually been filtered (i.e., invited to complete assessments) based on some sort of screening criteria, which also biases the sample substantially.

AVOID USING INDUSTRY-BASED NORMS FOR SELECTION

It is strongly recommended that specific (e.g., industry or role-based norms) not be used to interpret assessments used for selection decision-making purposes. This is due to the very high likelihood and substantial negative impact of restriction in range, along with the challenges inherent in developing high quality specific or custom norms. While it might be useful to know how a particular candidate measures up against others in your industry, the bias inherent in a highly-specific comparison group will likely lead to incorrect and indefensible decisions. Consequently, it is best to base selection decisions on broad, general norms.

There are a variety of real and serious risks associated with using highly-specific norms for selection decision making. The two most common are incorrectly rejecting a candidate – whereby the candidate is unfairly excluded and the organization misses out on a quality employee; or incorrectly hiring a candidate who fails to perform. More serious examples might include a forestry worker who *just* meets the minimum requirements in terms of safety-attitudes due to having his results interpreted against a poor-quality or restricted norm, but in reality is very relaxed about safety. Another example may be a volunteer fire-fighter whose verbal comprehension is greatly exaggerated by an inappropriate norm, and in reality struggles to understand written instructions and makes a serious and fatal mistake when responding to a stressful, time-pressured situation.

Ideal profiles as an aid to interpretation

One way to provide some of the benefits of a specific, industry or role-based norm, without risking bias or error due to high restriction of range is to compare normed results against an 'ideal profile'. This entails determining what the test scores from an 'ideal' role candidate would look like, and then using this ideal profile as a second, highly-specific yardstick against which to examine the candidate's profile. The 'fit' of the candidate's profile against that of the ideal can then inform decision-making, and can highlight potential strength/development areas which might warrant further attention during the selection or on-boarding processes.

The GeneSys assessment platform features tools and reports to help you build and apply ideal profiles. However it is important that the ideal be well-informed, and based upon sound data and theory, rather than intuition. In addition it is not enough to only examine high performers' test scores – you must also consider the test scores of poor or mediocre performers. This helps to avoid a restriction in range (based on performance scores) and allows you to be more certain of which traits and abilities are important for identifying high-potential candidates. This is demonstrated in

Example 5 *Below(The creation of ideal profiles is beyond the scope of this paper, but for more information please contact your nearest Psytech distributor http://www.psytech.com/contact/.

Example 5 - Ideal-Profiles Based on What Seperates High and Low Performers

The organization discussed in Example 4 decides that they want to develop an ideal profile to help with selecting candidates for their fast-track graduate programme. A recruiter examines the profiles of applicants who went on to perform very highly in the programme and finds that the applicants tended to profile as being *socially confident*, *highly-assertive*, *data-driven* and *detail-oriented;* and all had above-average *verbal reasoning* and *abstract reasoning* scores. The recruiter concludes that the ideal profile should include all six characteristics.

However, one of her colleagues suggests that some of these tendencies might have arisen due to chance or been unintentionally 'selected for' by an earlier step in the selection process. So the recruiters examine this 'high performer' profile against the profiles of applicants who struggled in the programme. They observe that these candidates tended to profile as being less *socially-confident* and *detail-oriented* than their high-performing colleagues, and have lower *abstract reasoning* scores overall. Since these three characteristics seem to separate the high-performers from average or poor performers, the recruiter decides that the ideal candidate will need to be *socially-confident*, *detail-oriented*, and have an above average *abstract reasoning* score.

WHEN CUSTOMISED OR INDUSTRY-BASED NORMS CAN BE USED

Despite the limitations and risks associated with specific (industry, role-based, etc.) and custom norms there are a number of contexts in which they are immensely useful. These include:

Research Projects

Specific norms, such as those focusing on members of a particular industry, are appropriate to use when testing has been performed as part of a research project. Unlike testing performed for selection purposes, test results in research projects typically aren't used to make high-stakes decisions that can impact greatly on respondents, the organization, or the wider-public. In addition, the restriction in range can be controlled for statistically during the analysis of research data. This is demonstrated in Example 6:

Example 6 - Using a Police-specific norm in a research context

A researcher wishes to examine whether a tendency to empathise with others influences Police detectives' satisfaction with their work. The researcher has a group of detectives complete a personality test and interprets their results against a norm group comprised entirely of Police officers. He recognizes that this is a biased sample, and accounts for this when statistically analysing the relationship between personality scores and job satisfaction ratings.

Comparison between groups

Highly specific norms can also be used to examine how one or more groups of respondents relate to each other, such as when an organization wishes to see how its employees 'measure up' against their wider industry. This information could then be used to inform learning and development programmes, refine selection processes, focus organizational development initiatives, etc., as in Example 7 below:

Example 7 - Using an industry-specific norm to inform an organisational development programme

A local web-design company is trying to distinguish itself from its competitors in terms of their designer's willingness to solicit and consider client's ideas and advice. They recognize that this can require a 'balancing act' between designers considering others' advice, all-the while still being able to assert their own point of view when needed. So that they can determine what sort of training and resources to provide their staff, they have them complete a personality assessment which examines these traits. To make the results more relevant, they then interpret the staff results against a 'Creative industries' norm.

A QUICK WORD ON USING NORMS ACROSS COUNTRIES OR CULTURES

There are situations and occasions which may require the use of an international or cross-cultural norm. For instance, an organization might receive employment applications from a number of different countries. Rather than using a separate norm for each country, the organization could use a broad, global norm to interpret candidate results so that they can compare all candidates against one another. This international norm should be constructed using as wide a sample (qualitatively speaking) as possible – and the proportion of people possessing these qualities needs to mirror that of the wider global population. The qualitative breadth of the norm group effectively 'averages out' the inter-cultural differences, allowing for consistent interpretation of results regardless of the location of the candidate's testing.

International norms can also be beneficial when working with tests that are only relevant to a particularly narrow subset of the population, such as those working in roles with very specific job requirements (e.g., air-traffic controllers). In these situations, were national-level norms to be used, there may not be enough respondents within each norm to guarantee that the norms meet the statistical and qualitative requirements discussed in earlier sections (e.g., statistical normality, a single 'average' point, etc.). Once again, the most important requirement for this international norm is that it is as qualitatively broad as possible.

SUMMARY

In the context of psychometric assessment, a 'norm' is a comparison group against which a respondent's score(s) on an assessment are interpreted. 'Raw' results are then 'normed' by mathematically evaluating them against the norm. Norms can be generic, specific to some particular role or context, or even customised to a particular project or organisation's needs. Norms used internationally or cross-culturally should be as broad as possible. The respondents included in the norm should be qualitatively similar to the wider population they represent (e.g., Australian Factory workers). The scores within the norm (i.e., the comparison data) need to be normally distributed and quantitatively similar to that of the group against which a respondent is to be considered.

In the context of psychometric testing for selection purposes, the respondent pool has typically been 'filtered' by the earlier ('screening') stages of the selection process. This means that the range of possible results in the norm is restricted, and the distribution of results misrepresented. This 'restriction of range' is particularly prevalent in industry and role-specific norms and can cause gross misinterpretation of results. As a consequence, highly specific norms should only be used for research and data-comparison purposes, and not for selection decision-making. Instead, selection decisions should be informed by scores normed against broad, generic norms. 'Ideal' profiles can also be used to facilitate more role-specific interpretations of scores without the risks inherent in using highly-specific norms.

Understanding and selecting norms is one of the most important steps in psychometric test administration. If done poorly it can lead to incorrect and indefensible decisions, a negative impact on organizational performance, and undesirable outcomes for test respondents. The GeneSys assessment platform includes tools to help you easily examine the statistical quality of the available norms, however the end user must also bear in mind the appropriateness and relevance of a particular norm to their testing situation. The remainder of this document is intended to help the end user, but for further information please visit the Psytech blog at http://blog.genesysonline.net/ or contact your nearest Psytech distributor at (http://www.psytech.com/contact/).

BIBLIOGRAPHY AND FURTHER READING

Escalona, S. K. (1982) Growing up with the threat of nuclear war: Some indirect effects on personality development. *American Journal of Orthopsychiatry, 52*(4), 600-607.

Flynn, J. R. (1987) Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171-191.

http://blog.genesysonline.net/custom-norms-getting-it-right/

APPENDIX I - CHECKLIST FOR BUILDING CUSTOM NORMS

☑	**General Requirements**
☐	The sample of sufficient **size** to guarantee an actual spread of results
☐	The sample contains only **genuine** assessment data (i.e., no data from trials or samples)
☐	Each respondent **appears only once** in the sample, unless they have multiple, genuine attempts which aren't influenced by practice effects
☐	The sample's range of demographic qualities mirrors that of the wider population (**qualitatively representative**)
☐	The proportion of people within the sample having those qualities mirrors that of the wider population (**quantitatively representative**)
☐	The sample's members sat the assessment(s) in same **context** as future respondents (Were tests timed and/or supervised?)
☐	There is no reason to believe the results of your sample will differ from the wider population

☑	**Technical Requirements**
☐	Data is centred around a single central average (**unimodal**)
☐	Data isn't clustered at the top or bottom ends of the distribution (**unskewed**)
☐	Scores become less common the further they are from the average (**normally distributed**)
☐	Wherever possible, **restriction of range** has been avoided by filtering respondents as little as possible