Using Norms in the Interpretation of Test Results


Introduction

As scientists in the field of psychology we seek to understand, predict and control human behaviour. In clinical, industrial and educational settings psychologists make attributions as to the cause, or likelihood, of behaviour based on evidence gathered from observational, interview, and assessment data.

Research into the process of making attributions and building theories or schema of people’s behaviours shows the process is fraught with difficulties, even with professionals involved (Meehl, 1954; Sawyer, 1966). Attributions are seldom made on the full range of available data as Heider (1958) or Kelly (1967) would have liked to believe. Difficulties include:

- characterising information on the basis of preexisting theories (Nisbett and Ross, 1980).
- people being unaware of the effects of small (Nisbett & Ross, 1980) or unrepresentative samples (Hamill, Wilson, & Nisbett, 1980).
- underutilising base rate information (Hamill et al., 1980).

Of all the problems of building our judgements about people the last point is arguably the easiest to control. Through using objective assessment techniques (after evaluation of the psychometric qualities of the tools) we can make judgements confident that the results are going to give a measure of the consistency of an individual’s behaviour (based on the reliability of the tool and validity of its predictions) and its uniqueness (based on the comparison of the result to appropriate groups).

This does not deny the usefulness of less structured information, rather it suggests that as psychologists we too can fall into the trap of weighting anecdotal or colourful information over more valid base rate data (Ginosar & Trope, 1980; Hamil et al., 1980; Taylor & Thompson, 1982). The normal distribution provides us with the tool to compare a standardised sample of behaviour with a large sample of that behaviour to give us a measure of relative propensity. The validity coefficient, as an index of the relationship between our sample of behaviour and the behaviour we are trying to predict, gives us the justification for this comparison.

As we compare a person’s aptitudes and attributes in relation to other people, scores on tests and questionnaires also need to be compared to relevant comparison groups. We do this through seeing how a person’s score sits in relation to other’s scores on a normal distribution, or bell curve. Norms are sets of data derived from groups of individuals, who have already completed a test or questionnaire. These norm groups enable us to establish where an individual’s score lies on a standard scale, by comparing that score with that of other people.

Within the field of occupational testing, a wide variety of individuals are assessed for a broad range of different jobs. Clearly, people vary markedly in their abilities and qualities, and therefore the norm group against which an individual is compared is of crucial importance. It is very likely that the conclusions reached will vary considerably when an individual is compared against two different groups; for instance, school leavers and managers in industry. For this reason, it is important to ensure that the norm groups used are relevant to the given group or situation that the data are being used for.

Norming Systems

There are a number of different norming systems available for use, which have strengths and weaknesses in different situations. These can be grouped into two main categories; rank order and ordinal.

Rank Order Systems

When a group of people are given a test or questionnaire we expect to observe a range of different scores as people differ in their abilities or personal qualities. This spread of results allows us to arrange people in a rank order scale according to their performance. When an individual’s score is subsequently compared with this scale, we can give a
percentile score which represents the percentage of the comparison group that the individual has scored above. For instance, a score corresponding to the 75th percentile means that an individual score or response is greater in magnitude than 75% of the norm group in question. Someone who has scored at the 30th percentile has performed or responded in a way that is higher than 30% of the norm group. The 50th percentile is equivalent to the average of the scale.

Percentiles have the disadvantage that they are not equal units of measurement. For instance, a difference of 5 percentile points between two individual’s scores will have a different meaning depending on its position on the percentile scale, as the scale tends to exaggerate differences near the mean and collapse differences at the extremes. Accordingly percentiles must not be averaged nor treated in any other fashion mathematically.

However, they have the advantage that they are easily understood and can be very useful when giving feedback of results to candidates or discussing results with managers.

**Ordinal Systems**

To overcome the problems of interpretation implicit with rank order systems, various types of standard scores have been developed.

The basis of standard scores is the Z-score which is based on the mean and standard deviation. It indicates how many standard deviations above or below the mean a score is. A Z-score is merely a raw score which has been changed to standard deviation units.

The Z-score is calculated by the formula:

\[
Z = \frac{X - \overline{X}}{SD}
\]

where:
- \(Z\) = standard score
- \(X\) = individual raw score
- \(\overline{X}\) = mean score
- \(SD\) = standard deviation

Usually when standard scores are used they are interpreted in relation to the normal distribution curve. It can be seen from Figure 1 that Z-scores in standard deviation units are marked out on either side of the mean. Those above the mean are positive, and those below the mean negative in sign. By the calculation of the Z-score it can be seen where the individual’s score lies in relation to the rest of the distribution.

The standard score is very important when comparing scores from different scales within the questionnaire. Before these scores can be properly compared they must be converted to a common scale such as a standard score scale. These can then be used to express an individual’s score on different scales in terms of norms.

One important advantage in using the normal distribution as a basis for norms is that the standard deviation has a precise relationship with the area under the curve. One standard deviation above and below the mean includes approximately 68% of the sample. From Figure 1 it can be seen that Z-scores can be rather difficult and cumbersome to handle because most of them are decimals and half of them can be expected to be negative.

To remedy these drawbacks various transformed standard score systems have been derived. These simply entail multiplying the obtained Z-score by a new standard deviation and adding it to a new mean. Both of these steps are devised to eradicate decimals and negative numbers.

**T Scores (Transformed Scores)**

The T-score is a linear transformation of the Z-score, based on a mean of 50 and standard deviation of 10. T-scores have the advantage over Z-scores that they do not contain decimal points nor positive and negative signs. For this reason they are used more frequently than Z-scores as a norm system, particularly for aptitude tests. A T-score can be calculated from a Z-score using the formula:

\[
T = (Z \times 10) + 50
\]
**Stens** (Standard Tens)

The Sten (standard ten) is a standard score system commonly used with personality questionnaires. It is based on a transformation from the Z-score and has a mean of 5.5 and a standard deviation of 2. Sten scores can be calculated from Z-scores using the formula:

\[
\text{Sten} = (Z \times 2) + 5.5.
\]

As the name suggests, stens divide the score scale into ten units. Each unit has a band width of half a standard deviation except the highest unit (Sten 10) which extends from 2 standard deviations above the mean, and the lowest unit (Sten 1) which extends from 2 standard deviations below the mean.

Stens have the advantage that they are based on the principles of standard scores and that they encourage us to think in terms of bands of scores, rather than absolute points. With stens these bands are sufficiently narrow not to mark significant differences between people, or for one person over different personality scales, while at the same time guiding the user not to over-interpret small differences between scores.

The relationship between Stens, T-Scores and Percentiles is shown in the chart of the normal distribution curve (Figure 1).

**Figure 1 - Normal Distribution Curve**

---

**Choosing Norm Groups**

The importance of New Zealand based norms lies in two places; the face validity and construct validity. When providing feedback to a respondent, giving the information in the context of a group they feel they should be
compared with is very important for the person’s acceptance of the results. The results will be less meaningful when compared to an inappropriate group. Results from our research to date suggests some very real differences between New Zealand, Australia, and the United Kingdom in the areas of personality and aptitudes.

Norm group size should be determined by the standard deviation of the sample and the reliability of the instrument. A rough rule of thumb is that for instruments which meet the gold standard of reliability (.75), norm groups should be made up of 100 people or more and should be directly relevant to the purpose of the test.

As discussed previously, the purpose of norms is to provide baserate information about the likelihood of a skill or behaviour being displayed by the individual. In order for baserate information to be meaningful the comparison group needs to be as similar to the individual tested and the circumstances in which they were measured. For example, comparing the score of a degree manager with 16 year old school leavers is unlikely to give useful base rate information as to the ability to solve problems.

In-house norms provide the most directly relevant base rate information as the recipient of the information will be very familiar with the prevalence of the behaviour being measured. The disadvantage of norms produced in-house is that the comparability with the whole population is lost. For example, are the people in your own institution more or less disabled than the whole population, are your applicants better or worse than those of other companies.

The table below outlines the tests currently commercially available, which have New Zealand norms.

<table>
<thead>
<tr>
<th>NZCER</th>
<th>Publicly Available NZ Norms (sample size in brackets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACER</td>
<td>PL/PQ: Form 5 (460), Form 6 (412) Students&lt;br&gt;BL/BQ: Form 7 (424), Teachers College (184), University (1,083)&lt;br&gt;B90: Tertiary Psychology and Education Students (876) J2 to Form 2 (400)&lt;br&gt;Burt Word Reading Test: J2 to Form 2 (1,000)&lt;br&gt;PAT Mathematics: Standard 2 to Form 4 (1,000)&lt;br&gt;PAT Reading: 8 to 14 year old students (8,016)&lt;br&gt;PAT Study Skills: Primary and Intermediate Schools (21,478)&lt;br&gt;PRETOS: Standard 2 to Form 2 Students (4,920)&lt;br&gt;SHEIK: Secondary School Students (665)&lt;br&gt;SPELL Write: 4.5, and 6 Form Students (500)&lt;br&gt;Standard Progressive Matrices: Primary and Intermediate Students (1,250)&lt;br&gt;Standard 2 to Form 5 Students (3,174)</td>
</tr>
</tbody>
</table>

| SHL New Zealand | Advanced Managerial Tests Managers and Professionals (414 - 484)<br>Applied Technology Series Apprentices and Graduate Engineers (117)<br>Critical Reasoning Test Battery Administrators and technicians (130 - 523)<br>Customer Contact Series Customer service and sales staff (129)<br>General Occupational Interest Inventory Blue collar workers off work for longer than 1 year (122)<br>Inventory Technology Test Series Information Technology programmers and analysts (100)<br>Inventory of Management Competencies Managers and professionals (135)<br>Management and Graduate Item Bank Graduates and managers (918 - 1,526)<br>Occupational Personality Questionnaires Graduates, Banking Managers, Managers (3,950)<br>• Concept Model 4.2 / 5.2 Customer service and sales staff (121)<br>• Customer Contact Styles 5.2 Administrative and clerical staff (256)<br>• Factor Model 4.2 Human resource professionals (162)<br>• Images Manual and technical staff (273)<br>• Work Styles Administrative and Clerical Staff (136 - 1,130)<br>Personnel Test Battery Apprentices and manual workers (175 - 344)<br>Technical Test Battery Production line workers (108 - 2,658)<br>Work Skills Series |
The Need for Responsible Instrument Use

Whilst it is widely accepted that well-constructed psychometric assessments provide objective information about a candidate and have been shown in general to lead to better and fairer employment decisions, there can be performance differences between ethnic or gender groups. This is especially common where socio-economic conditions impact on the educational opportunities available to particular groups, or where a candidate is not a native speaker of the language in which the assessment tool is presented.

A problem arises when a significant difference is found between the average performance of different ethnic groups or men and women on the assessment. In the absence of validation evidence there is likely to be a presumption that the group with the lower average performance was being indirectly discriminated against. That is, if an unjustifiable entry standard is set and demanded of all applicants, the lower scoring group would find it harder to comply with the requirement and, hence, would be indirectly discriminated against.

Positive validation evidence of an assessment instrument generally justifies the use of the instrument and rules out the possibility of unfair discrimination. By showing that those who perform poorly on the assessment also perform poorly on the job, a positive validation result confirms that rejecting low scoring candidates is reasonable. The greater the degree of disparate impact resulting from the use of a psychometric instrument, the higher the validity should be to justify its use.

There remains the possibility that overall validity is masking cases where an instrument has poorer or no predictive validity for some groups or that group differences in assessment scores are not reflected in job performance. Extensive research into these issues in the United States, covering many types of tests and a wide range of occupational fields, has indicated that such scenarios are extremely rare, if they exist at all, when best practice has been followed (Hunter & Hunter, 1985; Hartigan & Widgor, 1989). There is a lack of published studies in this area for other countries and more work still needs to be undertaken.

Some experts argue that where ethnic or gender group differences on assessment scores exceed group differences in job performance, separate norm tables for each group should be used for evaluating scores (Uniform Guidelines, 1985). Use of separate norms in these circumstances has not been tested in New Zealand courts or tribunals, but it certainly would not be justifiable in any other circumstances. In all cases the availability of direct or relevant instrument validation data means that discriminatory practices can be avoided.

It should be remembered that group differences relate to average performance. Even where there are substantial group differences there will be members of the lower scoring group who have better results than many people from the higher scoring group and vice versa. Furthermore, job success does not generally depend on a single ability or preferred style and assessment tools do not have perfect predictive power. Therefore, on occasion, those with poorer results on an assessment will do better in a job than an assessment result may suggest. For this reason it is preferable to interpret assessment results together with other available information.

In that the aforementioned patterns exist, it is particularly important that appropriate guidelines are followed to avoid improper use of psychometric tools. Considerations of fairness are important in themselves, particularly when the legal implications under the Human Rights Act 1993 for engaging in discriminatory practices in the selection and promotion of employees are taken into account.

Conclusions

We generally understand people’s behaviour in relation to theories we build about the world. These normative theories encompass the concept of relativity, but even as professionals we underutilise important information in forming these attributions. Gathering and using base-rate information is the easiest and arguable the most effective way to correct for these errors of judgement. Gathering objective information and interpreting it by comparing them with a relevant population provides us with important base-rate information, if the objective information has a proven relationship with what we are trying to measure (validity). Norms need to be relevant to what we want to use the results for, for example work related samples in work contexts and educational samples in educational contexts. New Zealand norms are an important aspect in the effective interpretation and understanding of test scores.

Separate norms can be used for different gender or ethnic groups if it is proven that differential validity exists, that is the relationship between test scores and behaviour is different for the groups concerned.
References


