Validity

Whereas reliability is concerned with how accurate and precise a questionnaire is, validity looks at how relevant it is. Does the questionnaire actually measure what it claims to measure? Studies relating to the validity of personality questionnaires provide information about the degree to which the questionnaire measures what it was designed to measure.

Types of validity

There are four main types of validity, each looking at a different aspect of how useful and relevant the questionnaire will be in practice:

- Face validity
- Content validity
- Construct validity
- Criterion validity

Face validity

Face validity is the extent to which the questionnaire appears or looks as if it measures what it claims to measure. For example, a test that is a general test of numerical ability would appear to have high face validity if the content of the test is a series of numerical questions. A crude measure of face validity would be the extent to which somebody could open a questionnaire booklet or have a look at a series of items and guess what the questionnaire is trying to measure. If they could guess accurately, then the questionnaire would be considered to have high face validity.

When we talk about the FIRO questionnaire, we are looking at three areas of interpersonal need. You can clearly see from the questions that these areas are related. The questions are about how you interact with people and how they interact with you, therefore the FIRO questionnaire can be said to have high face validity.

This area of validity is the least important for the FIRO instrument – the more important issue is whether the person completing the questionnaire feels comfortable, and this can be achieved with thorough administration.

Content validity

Content validity is the extent to which the content of the questionnaire is suitable for measuring what it claims to measure. For example, a test that claims to be a measure of general numerical ability would have a low content validity if all the questions in the test were about long multiplication. However, this content would be more appropriate if the test claimed to be a more specific measure of long multiplication. Asking an expert who is familiar with the subject area of the questionnaire whether the items are relevant and appropriate most easily assesses content validity. For example, a mechanical aptitude test could be developed and the content could be checked by somebody who is familiar with the area of mechanics being measured – to verify that the relevant aspects of the skills are being assessed.

In relation to the FIRO questionnaire, content validity looks at whether the set of items that make up each scale adequately cover the area being considered. Schutz originally researched the content validity of the FIRO questionnaire during the development and construction of the questionnaire. He believed that if the theory behind the questionnaire construction was valid (Guttman scaling), the content validity could be assumed given that the scales were achieving the set standards of reproducibility.

Construct validity

This looks at whether the questionnaire is appropriate for measuring a particular psychological construct. Such constructs are typically defined by psychological theory. If it can be demonstrated that a test or questionnaire successfully measures such an underlying theoretical construct, the questionnaire can be said to have construct validity.

As the FIRO instrument is based on Schutz's theory of interpersonal needs, construct validity is particularly important in establishing its credibility. Construct validity is typically measured in one of two ways:

1. Comparisons with other personality questionnaires

2. Criterion-related validity

1. Comparisons with other personality questionnaires

When using comparisons with other personality questionnaires to assess construct validity, it is important to remember that the two questionnaires have not been designed to measure the same construct, therefore the relationship between them will not be very strong. This is borne out in the correlations that we would expect. For this reason we look for correlations that are significant, that is, unlikely to have occurred by chance.

Table 5.3 (overleaf) shows the kind of relationships that would be predicted on the basis of the FIRO and MBTI® theories underlying the respective measures. The MBTI instrument is based on Jung's model of psychological Type, which looks at how you are energised, the information you prefer to gather, your decision-making style and your approach to the external world. As expected, Extraversion was related to higher scores on Expressed and Wanted Inclusion and Expressed and Wanted Affection in the FIRO-B instrument. Thinking was significantly related to higher scores on Expressed Control, while Feeling correlated significantly with Expressed Affection. Please see the EDS for further data relating to the correlations between the MBTI and the FIRO-B instruments.

Table 5.3 Correlations between FIRO-B scale scores and MBTI Step I continuous scores

UK general population sample (n=1,512)

Myers-Briggs Type Indicator®	eI	wI	eC	wC	eA	wA
Extraversion–Introversion	-0.41**	-0.38**	-0.13**	0.07**	-0.36**	-0.27**
Sensing–iNtuition	0.12**	0.19**	0.18**	0.02	0.10**	0.03
Thinking–Feeling	0.10**	0.10**	-0.24**	0.18**	0.25**	0.23**
Judging–Perceiving	0.02	0.07**	0.00	-0.02	0.00	0.00

Significant at: *p<0.05, **p<0.01

Table 5.4 Correlations between FIRO Business scale scores and MBTI Form M Continuous Scores

Myers-Briggs Type Indicator®	eInv	wInv	eInf	wInf	eCon	wCon	Total Inv	Total Inf	Total Con	Total e	Total w	Overall
Extraversion–Introversion	-.43**	-.30**	-.24**	.06	-.33**	-.29**	-.35**	-.46**	-.25**	-.42**	-.16**	-.42**
Sensing–iNtuition	-0.2	-.02	.02	-.03	-.04	.01	-.02	-.02	-.02	-.02	-.01	-.02
Thinking–Feeling	0.00	.02	-.20**	.20**	.10*	.16**	.15**	-.05	.17**	.01	-.03	.06
Judging–Perceiving	.03	.05	.02	.02	.09*	.06	.09	.07	.06	.05	.04	.08

Significant at: *p<0.05, **p<0.01 Negative correlations are associated with E, S, T and J; positive correlations are associated with I, N, F and P.

Reference: *FIRO Business: Technical Guide.* Nicole A Herk, Richard C. Thompson, Michael L. Morris, Nancy A Schaubhut (2009, CPP, Inc.).

A further study conducted by Schnell et al[3] supports these conclusions, and found that Extraversion was related to higher scores on all dimensions of the FIRO-B questionnaire except Wanted Control. They further looked into the correlations between overall need scores and a preference for Extraversion or Introversion. More recent research data is published in Table 2.6 of the EDS. This reinforces the fact that Extraversion–Introversion is a broad concept; in their article, Schnell & Hammer state that "because of the FIRO-B's focus on interpersonal needs, lower overall results of the FIRO-B can be expected with Introverts".

Tables 5.5 and 5.6 show the psychological Types with the highest and lowest mean scores for each FIRO-B cell. Within each cell in Table 5.5, Types are listed in descending order, with the Type with the highest need score listed first. Within each cell in Table 5.6, Types are listed in ascending order, with the Type with the lowest need score listed first. It can be seen that MBTI Types that combine Extraversion and Feeling tend to exhibit consistently high Inclusion and Affection needs, but not necessarily particularly high Control needs. Conversely, Types that combine Introversion and Thinking tend to exhibit low needs in all the categories except Expressed Control.

Table 5.5: Ranking of psychological Type with the highest mean scores within FIRO-B dimensions

	Inclusion	**Control**	**Affection**
Expressed	ESFP ENFP	INTJ ENTJ	ESFJ ENFP ENTP
Wanted	ENFP ESFJ INFJ ENTJ	INFJ ISFJ ENFJ	ESFJ ENFP INFP ENTP ESFP

[3] Schnell, Hammer, Fitzgerald, Fleenor & Van Velsor (1994), reported in Schnell & Hammer's article 'Integrating the FIRO-B with the MBTI' (1997) – see reference list for details.

Table 5.6: Ranking of psychological Type with the lowest mean scores within FIRO-B dimensions

	Inclusion	Control	Affection
Expressed	ISTP ISTJ	INFP ISFJ ESFP	INTP ISTP
Wanted	ISTJ	INTP	ISTP INTP

Table 5.7 looks at the correlations between FIRO-B scores and the 16pf® factor scores. This table reports the four highest significant correlations for each FIRO-B scale; the full data table can be found in Table 2.10 of the EDS.

Many of the strongest correlations with the ***16pf instrument*** are with FIRO-B Expressed behaviours, particularly Expressed Inclusion and Expressed Affection. The table on the next page shows that there are similar patterns in the results between these two scales, and also between Wanted Inclusion and Wanted Affection. These data suggest that these FIRO scales show clear links with personality traits that influence the way in which individuals relate to others.

The Primary Factor found to correlate most highly with Expressed Control is Dominance, which is to be expected given the behaviours associated with that scale. This suggests that those who score highly on Expressed Control are likely to be those who will want to express their opinions and influence others towards their own way of thinking and doing things. For further interpretation and links between the FIRO-B and 16pf instruments, see Chapter 2 of the EDS.

® 16pf and the 16pf logo are trade marks and registered trade marks of the Institute for Personality and Ability Testing, Inc. (IPAT). IPAT is a subsidiary of Performance Assessment Network, Inc. (PAN).

Table 5.7: Correlations between FIRO-B and 16pf factor scores

Inclusion		Control		Affection	
Expressed	Wanted	Expressed	Wanted	Expressed	Wanted
Self-Reliance *(-0.47)*	Liveliness *(0.40)*	Dominance *(0.40)*	Apprehension *(0.18)*	Privateness *(-0.42)*	Warmth *(0.30)*
Liveliness *(0.41)*	Self-Reliance *(-0.35)*	Social Boldness *(0.24)*	Dominance *(-0.17)*	Warmth *(0.37)*	Privateness *(-0.31)*
Social Boldness *(0.39)*	Privateness *(-0.27)*	Openness to Change *(0.15)*	Social Boldness *(-0.09)*	Liveliness *(0.33)*	Self-Reliance *(-0.25)*
Warmth *(0.36)*	Warmth *(0.25)*	Tension *(0.13)*	Perfectionism *(0.09)*	Social Boldness *(0.33)*	Liveliness *(0.24)*

Correlating FIRO-B scores and the ***Adjective Checklist (ACL)*** choices produces examples of correlations that further support the construct validity of the FIRO-B questionnaire.

Table 2.11 in the EDS reports the output of a study comparing the FIRO-B scales with ACL data. The results from this study show correlations between the FIRO-B scales and ACL data that supports the constructs of the individual scales. The data shows correlations between the Inclusion scales and items such as:

- Outgoing
- Quiet (-)
- Sociable
- Talkative

The Control scales correlate with items such as:

- Aggressive
- Assertive
- Opinionated
- Outspoken

The Affection scales correlate with items such as:

- Talkative
- Sociable
- Enthusiastic
- Cold (-)

Table 5.8 shows correlations between FIRO Business and Big Five scales based on the ACL.

FIRO Business Scale	Extraversion	Agreeableness	Conscientiousness	Openness	Neuroticism
Expressed Involvement	.40**	.27**	.10*	.13**	.15**
Wanted Involvement	.20**	.16**	.03	.07	.02
Expressed Influence	.38**	-.04	.08*	.25**	-.04
Wanted Influence	-.25**	-.01	-.25**	-.21**	-.23**
Expressed Connection	.30**	.31**	.05	.11**	.09*
Wanted Connection	.16**	.21**	-.04	.01	-.09*

Note: N = 586; *p<.05, **p<.01.

Reference: FIRO Business: Technical Guide. Herk, N., Thompson, R., Morris, M., Schaubhut, N. (2009, CPP, Inc.).

For this study, the Big Five are measured by the NEO Personality Inventory. There are correlations to the following:

- Expressed Involvement to Extraversion and Agreeableness
- Wanted Involvement to Extraversion
- Expressed Influence to Extraversion and Openness
- Wanted Influence is negatively correlated to Extraversion, Conscientiousness, Openness and Neuroticism
- Expressed Connection is correlated to Extraversion and Agreeableness
- Wanted Connection is correlated to Agreeableness.

2. Criterion-related validity

The second method of establishing construct validity is criterion-related validity. Criterion-related validity assesses the extent to which questionnaire scores are correlated with criterion measures, which are typically behavioural measures. Criterion-related validity does not depend on the validity of another instrument; instead it looks directly at behaviour. The aim is to discover which questionnaire scores can be used to predict future performance. In practice, criterion-related validity can be determined in one of two ways: concurrent validity and predictive validity.

Concurrent validity

Concurrent validity is the simultaneous measuring of the questionnaire scores and the criterion measures – for example, when questionnaire scores are correlated with current performance on the job. The advantage of this method is that it is quick and relatively easy to carry out. However, because this method involves testing people currently in the job, there will be many other potential influences on their performance, such as their experience since joining the organisation. This means that concurrent validity studies are a less accurate way of researching the link between questionnaire scores and job performance when compared to predictive methods.

Predictive validity

When questionnaires are being used for selection and recruitment, the ultimate aim of using the questionnaire is to make advance predictions about future performance. The only way to test this is to carry out what is called a predictive validation study. This begins by testing a group of job candidates on the questionnaire(s) being investigated; the questionnaire scores are then put to one side without being used in the selection decision. After a period of time, once these people have been in the job long enough for their performance to be legitimately measured, the criterion scores for those same individuals should be collected. The questionnaire scores and criterion scores can then be correlated to see to what extent the questionnaire scores were predictive of future performance. Such a study is time consuming, and may be expensive and difficult to carry out. Often a concurrent validation study will be carried out as a quicker, but less accurate, indication of criterion-related validity.

Criterion-related research studies

There is a wealth of information published about the application of the FIRO-B instrument in a variety of contexts. For the purposes of this section, a selection of these validity studies has been outlined below to illustrate examples of criterion-related validity.

Leadership

O'Brien and Kabanoff (1981) conducted a study looking at group leader effectiveness. They used the Inclusion and Control scales to allocate people to groups and found that the groups that were matched as to their interpersonal needs structures were more productive based on ratings of group performance.

Leadership and influence

Gluck (1983) conducted a study examining groups of law students and the degree to which they wanted others to have control over them. They found that law students differed from the general population in how much they were willing for others to influence them. This study found that the lowest scores in the sample were for Expressed and Wanted Control. The results of this study were discussed in relation to the role of a lawyer and the requirement to work independently, and often to take on collaborative roles with clients.

Team compatibility

Fischer et al (1995) conducted a study looking at team compatibility using both the 16pf and FIRO-B instruments. They explored the concept of 'group warmth' as a derivative of the FIRO-B Inclusion and Affection scales. Their study provided evidence of increased team compatibility and linked their concept of 'group warmth' to the commercial effectiveness of teams.

Emotional climate

Vraa (1974) assigned graduate students to groups based on their Wanted Inclusion scores, and then measured the 'emotional climate' as the groups progressed. Regardless of the composition of the group, warmth increased with time. Hostility started at a high level in the high Wanted Inclusion group and decreased over time, while the opposite happened in the low Wanted Inclusion group. Vraa also found that the tendency to leave the group occurred less often in the high Wanted Inclusion group.

A note about statistical significance

When analysing results from validity studies, we need ways of checking how clearly the prediction or hypothesis was supported. Sometimes, what looks like a promising trend could really be a chance outcome. It would be wonderful if validity studies came out with neat, obvious results, but this is not usually the case. Even where there is a clear prediction, we never expect things to work perfectly.

Because human behaviour is complex, we need to do more than just look at the information we obtain in our study – we need to use statistical analysis to identify patterns. While you do not need to know how to do such analyses (unless you decide to do research yourself), it is important to be able to interpret analyses that have been done for you.

Tests of statistical significance are ways of analysing data, all of which seek to help us decide whether the patterns appearing in our data are likely to be ‘real’, or whether they may have arisen by chance. A ‘real’ pattern would be one that is not just true for the particular individuals in our sample, but which would also hold good if we sampled much more widely.

It is possible to calculate the probability that a result could have been obtained by chance alone. You will not have to do any calculations, but you need to understand what the stars mean. Here it is in shorthand:

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

Translating the top line into English, its full meaning is: a single star by a result shows that the probability (p) that the result occurred by chance alone is less than (<) five in 100 (0.05) (which is the same as one in 20).

Two stars show that the probability that the result occurred by chance is less than one in 100, while for three stars this is less than one in 1,000.

Another way to say this is that the result is “significant at the 0.05 (or 0.01 or 0.001) level”. Usually, we are happier the more significant the results are (the more stars there are), because it means we have some interesting patterns to look at. The choice of 0.05 as the basic level of significance is arbitrary but fairly standard – we do not usually consider results worth paying attention to if the probability that they occurred by chance is greater than one in 20.

Note, however, that the absence of a significant result does not mean there definitely is no effect or pattern – only that we cannot confidently conclude that there is one from the study in question. Also, the degree of statistical significance does not tell you how big the effect is, or whether it is an effect of great practical significance. With very large samples, we can often detect very subtle patterns and be very confident that we would find that subtle pattern again if we sampled more widely. However, the actual effect may be very small. The relationship between smoking in pregnancy and birth weight of a child would be a good example. There is a correlation, as the more cigarettes smoked during pregnancy the lower the birth weight tends to be, but the correlation is very small, as smoking is far from the only determining factor. Nevertheless, with large samples the effect is statistically significant, and in this case would probably be considered practically important as well.

A final word about validity

Validity studies are being conducted all the time and we have only been able to touch upon some of the more important areas in this chapter. The FIRO-B and FIRO Business questionnaires are well researched and validated instruments and have been researched across the world in a wide variety of contexts. With their basic validity well established, there is always scope for additional investigation and research. In particular, research looking at different frames of reference with increased sample sizes and comparisons across different cultural groups is something researchers in this field are keen to investigate.