APPENDIX G:

TEST REVIEW FORM

CONTENTS

1 INTRODUCTION

Assessment Standards of South Africa (ASSA) have adopted the broad EFPA review framework with some adaptations for ecological validity for the South African context. Note specifically that the original EFPA numbering is included in square brackets in the various sections for reference. Most of the adaptations were done to provide specific guidelines in order to remove subjectivity in the rating of assessments. Certain rating elements to do with the peripheral aspects of the test (e.g. price, content of reports, and aesthetics of test material) were excluded from this review process.

It is difficult to set clear criteria for rating the technical qualities of an instrument. The notes provide some guidance on the values to be associated with inadequate, adequate, good and excellent ratings. However, the nature of the instrument, its area of application, the quality of the data on which the technical properties are based, and the types of decisions that it will be used for could all affect the way in which ratings are awarded. Under some conditions a reliability of 0.70 is fine; under others it would be inadequate. These guidelines are intended to set a minimum standard using an internationally accepted set of recommendations that will help guide test developers in what is expected when compiling evidence for their assessments.

For more information about ASSA and the test review process, please consult the Test Review Information Pack.

The key principles guiding the development of the review criteria are:

1. The review criteria should be clear, transparent and openly available
2. The review criteria should be objective and easy to use for ratings
3. The review criteria should specify the minimum standard for certification
4. The review criteria should provide clear guidelines to encourage developers to strive for excellence
5. The review criteria should allow for contextualization to be done relevant to the application of the test (i.e., not only insist on a South African population norm as a standard).
6. Only the categories deemed critical to determine the psychometric quality of the instrument are rated in this review process.

For the ASSA Review Form a **5-point rating scale** will be used as indicated in Table 1.

Table 1: TEST REVIEW RATING SCALE

ASSA Number rating	**ASSA Description**
N/A**	Not applicable**
0	No information provided
1	Inadequate
2	Adequate
3	Good
4	Excellent

**Note where sub-sections are not applicable these will be dropped from overall scoring so that the total score will be calculated using only the applicable completed sections.

PLEASE NOTE: The scoring system will be finalised after the criteria have been agreed upon and piloted with existing assessments to test the weighting and correctness of the scoring procedure.

A total score can be calculated based on relevant and required completed sections. The total score can be considered as a broad indication of **overall quality**. For those sections where a rating below the desired level has been given, further feedback should be provided by the reviewer to guide the responsible test developer on how to improve that aspect. Note that the number of points allocated to a sub-section will not necessarily reflect its importance in the overall evaluation process. The overall points per subsection can be used in future phases as cut scores for minimum standards of compliance with best practice.

2 RATING OF THE MEASURE

2.1 RELIABILITY INFORMATION [EFPA 10]

Test developers need to indicate the reliability approach followed (e.g., internal consistency, test-retest reliability, inter-rater reliability, equivalence or IRT-based) and the rationale in line with the intended purpose for which the test was developed.

Reliability refers to the degree to which scores are free from measurement error variance (i.e. a range of expected measurement error). For reliability, the guidelines are based on the need to have a small Standard Error for estimates of reliability. Guideline criteria for reliability are given in relation to two distinct contexts: the use of instruments to make decisions about groups of people (e.g. organisational diagnosis) and their use for making individual assessments. Reliability requirements are higher for the latter than the former. Other factors can also affect reliability requirements, such as the kind of decisions made and whether scales are interpreted on their own or aggregated with other scales into a composite scale. In the latter case the reliability of the composite should be the focus for rating not the reliabilities of the components.

When an instrument has been translated and/or adapted from a non-local context, one could apply reliability evidence of the original version to support the quality of the translated/adapted version. In this case, evidence of equivalence of the measure in a new language to the original should be provided. Without this it is not possible to generalise findings in one country/language version to another. For internal consistency reliability, evidence based on local groups is preferable as this evidence is more accurate and usually easy to get. For some guidelines with respect to establishing equivalence, see the introductory section on Validity. A guide of critical points for comment when an instrument has been translated and/or adapted from a non-local context is included in the EFPA Guidelines as the Appendix.

2.1.1 Data provided about reliability [EFPA 10.1]

0 - No information given
1 - Only one overall reliability coefficient given
2 - At least one type of reliability coefficient given (for each scale and/or subscale)
3 - Reliability coefficients for a number of different groups (for each scale or subscale)
4 - Standard error of measurement given for a number of different groups (for each scale or subscale)

2.1.2 Interpretation of reliability evidence

0 - No interpretation done
1 - Mentions reliability evidence only
2 - Presents evidence of reliability as appropriate for the measure
3 - Discusses evidence of reliability in relation to the context
4 - Presents critical exposition of all aspects of reliability as applicable to the measure

2.1.3 Reliability coefficients are reported with samples that [EFPA 10.2.4]

N/A - Not applicable
1 - Do not match intended test takers
2 - Match the intended test takers

2.1.4 Internal consistency [EFPA 10.2]

Internal consistency reliability is the degree to which the items in a scale appear to measure a similar construct. The use of internal consistency coefficients is not sensible for assessing the reliability of speed tests, heterogeneous scales (also mentioned empirical or criterion-keyed scales; Cronbach, 1970), effect indicators (Nunnally & Bernstein, 1994) and emergent traits (Schneider & Hough, 1995). In these cases, all items concerning internal consistency should be marked ‘not applicable’. It is also biased as a method for estimating reliability of ipsative scales. Alternate form or retest measures are more appropriate for these scale types.

2.1.4.1 Sample size [EFPA 10.2.1]

N/A – Not applicable
0 – No information provided
1 – Single sample; inadequate sample size (n < 100)
2 – Single sample; fair sample size (100 ≤ n < 300)
3 – Single or multiple samples; at least one n > 500
4 – Multiple samples; n > 500 for each sample

2.1.4.2 Kind of coefficients reported [EFPA 10.2.2]

- ○ Not applicable
- ○ Coefficient alpha or KR-20
- ○ Lambda-2
- ○ Greatest Lower bound
- ○ Omega (Factor analysis)
- ○ Theta (Factor analysis)
- ○ Lambda (Factor analysis)
- ○ Other, describe

2.1.4.3 Size of coefficients [EFPA 10.2.3]

N/A – Not applicable
0 – No information provided
1 – Majority of scales $r < 0.70$
2 – Majority of scales $0.70 \leq r < 0.80$
3 – Majority of scales $0.80 \leq r < 0.90$
4 – Majority of scales $r \geq 0.90$

2.1.5 Test-retest reliability – Temporal stability [EFPA 10.3]

Test-retest refers to relatively short time intervals (e.g. < 3 months), whereas temporal stability refers to longer intervals in which more change is acceptable (e.g. > 6 months). Both aspects are relevant, particularly for tests to be used for predictions over longer periods. To assess the temporal stability, more than one retest may be required. The use of a test-retest design is not sensible for assessing the reliability of state measures (a high test-retest coefficient would invalidate the state character of a test). In this case all items concerning test-retest reliability should be marked 'not applicable'.

2.1.5.1 Sample size [EFPA 10.3.1]

N/A – Not applicable
0 – No information provided
1 – Single sample; inadequate sample size (n < 100)
2 – Single sample; fair sample size (100 ≤ n < 300)
3 – Single or multiple samples; at least one n > 500
4 – Multiple samples; n > 500 for each sample

2.1.5.2 Size of coefficients [EFPA 10.3.2]

N/A – Not applicable
0 – No information provided
1 – Majority of scales $r < 0.60$
2 – Majority of scales $0.60 \leq r < 0.70$
3 – Majority of scales $0.70 \leq r < 0.80$
4 – Majority of scales $r \geq 0.80$

2.1.5.3 Data provided about test-retest interval [EFPA 10.3.3]

N/A – Not applicable
0 – No information provided
1 - Brief mention is made of test-retest interval

2 - Some discussion or justification for test-retest interval
3 - Detailed discussion of test-retest interval within context
4 - Critical consideration of test-retest interval in relation to the test, its assumptions and theory within the context of use

2.1.6 Equivalence reliability (Parallel or Alternative forms) [EFPA 10.4]

Equivalence reliability refers to evidence that the constructs measured in parallel or alternative forms are similar, and that the forms can essentially be used interchangeably. Usually the same sample will complete both versions of the test and the results are compared. This is not applicable for tests that do not have different forms.

2.1.6.1 Sample size [EFPA 10.4.1]

N/A – Not applicable
0 – No information provided
1 – Single sample; inadequate sample size (n < 100)
2 – Single sample; fair sample size (100 ≤ n < 300)
3 – Single or multiple samples; at least one n > 500
4 – Multiple samples; n > 500 for each sample

2.1.6.2 Are assumptions for parallelism met? [EFPA 10.4.2]

Note that a test can be considered to be a parallel test if in the same group, mean scores and variances are similar for both forms of the test.

N/A – Not applicable
0 – No information provided
1 – Many large effect sizes and significant differences in mean scores and variance
2 – Mostly small effect sizes with significant differences in mean scores and variance
3 – Mostly negligible effect sizes with significant differences in mean scores and variance
4 – Mostly no significant differences in mean scores and variance

2.1.6.3 Size of coefficients [EFPA 10.4.3]

N/A – Not applicable
0 – No information provided
1 – Majority of scales $r < 0.70$
2 – Majority of scales $0.70 \leq r < 0.80$
3 – Majority of scales $0.80 \leq r <- 0.90$
4 – Majority of scales $r \geq 0.90$

2.1.7 IRT based method [EFPA 10.5]

Different IRT methods have different indicators of reliability. This type of reliability is not required, unless the method of test development and measurement is purely based on the IRT method. The guidelines provided here are very general, and in some cases, the reviewer may have to apply judgement in the rating of the coefficients.

2.1.7.1 Sample size [EFPA 10.5.1]

Depending on the item response model – minimum values for "adequate" are suggested to be:
200 for 1-parameter studies
400 for 2-parameter studies
700 for 3-parameter studies
These are based on Parshall, Davey, Spray, and Kalohn (2001). These values apply to dichotomous models, but can be of some guidance for the reviewer when polytomous models are used for which the sample sizes may be smaller.

N/A – Not applicable
0 – No information provided
1 – Single sample; inadequate sample size (n < 200/400/700)
2 – Single sample; fair sample size (n > 200/400/700)
3 – Single or multiple samples; at least one n > 200/400/700
4 – Multiple samples; n > 200/400/700 for each sample

2.1.7.2 Kind of coefficients reported [EFPA 10.5.2]

- ○ N/A – Not Applicable
- ○ Reliability of the estimated latent trait
- ○ Rho
- ○ Information function
- ○ Person separation index
- ○ Others, describe:

2.1.7.3 Size of coefficients [EFPA 10.5.3]

Both guidelines for reliability coefficients (including rho) as for the information function are given. The guidelines for the information function are based on those for reliability coefficients since Information = $1/SE^2$, and given some often made assumptions, $r = 1 - SE^2$. Note that SE and information values are dependent on the value of the latent trait and that each test has a range within which the information value is optimal. The rating should not a priori be based on this optimal value, but on the information value of the score or range of scores that are of specific importance (e.g., critical scores). For these scores the information value may be optimal, but not necessarily so. If there are no such scores, the rating should be based on the mean information value (see also Reise & Havilund, 2005). Because there is not much experience with these rules-of-thumb, we advise raters to use these rules with care.

N/A – Not applicable
0 – No information provided
1 – Majority of scales r < 0.70; information < 3.33
2 – Majority of scales *r* < 0.70 ≤ 0.80; ≤ information < 5.00
3 – Majority of scales *r* < 0.80 ≤ 0.90; < information < 10.00
4 – Majority of scales *r* ≥ 0.90; information ≥ 10.00

2.1.8 Inter-rater reliability [EFPA 10.6]

Inter-rater reliability refers to the consistency with which different raters score the same test. If the scoring of a test involves no judgmental processes (e.g. simply summing the scores of multiple-choice items), this type of reliability is not required and all items concerning inter-rater reliability should be marked 'not applicable'. Note that although inter-rater reliability may not apply to the whole test, it may apply to one or more subtests (e.g. some subtests of an intelligence test).

2.1.8.1 Sample size [EFPA 10.6.1]

N/A – Not applicable
0 – No information provided
1 – single sample; inadequate sample size (n < 100)
2 – single sample; fair sample size (100 ≤ n ≤ 200)
3 – single or multiple samples; at least one n > 200
4 – multiple samples; n > 200 for each sample

2.1.8.2 Kind of coefficients reported [EFPA 10.6.2]

- ○ N/A – Not Applicable
- ○ Percentage Agree
- ○ Coefficient Kappa
- ○ Intra Class Correlation
- ○ Coefficient Iota
- ○ Other (please describe)

2.1.8.3 Size of coefficients [EFPA 10.6.3]

N/A – Not applicable
0 – No information provided
1 – Majority of scales $r < 0.60$
2 – Majority of scales $0.60 \leq r < 0.70$
3 – Majority of scales $0.70 \leq r < 0.80$
4 – Majority of scales $r \geq 0.80$

2.1.9 Reviewers comments, evaluation & recommendation on reliability [EFPA 10.8]

Consolidation of all comments that the reviewer has made throughout this section on reliability. Add overall impressions here with recommendations.

2.2 VALIDITY INFORMATION

Validity is the extent to which a test measures what it claims to measure. In the literature many types of validity are differentiated. For example, Foxcroft and Roodt (2019) mention three broad categories of validity evidence, covering 11 different types of validity coefficients. Borsboom, Mellenbergh, and Van Heerden (2004) state that a test is valid for measuring an attribute if variation in the attribute causally produces variation in the measured outcomes. Hence, they maintain that a differentiation between types a validity is not relevant.

It is not necessary for a test developer or publisher to present all of these debates or provide evidence for all validity types. Some aspects of validity may not be necessary for some types of tests. What is important is for there to be some evidence of validity checking. The methods used for this must be rigorous and well supported in the literature and the results should ultimately support the use of the test for the South African context. It may be that some types of validity are supported whilst others aren't. Presenting this evidence is good and demonstrates rigour. In sum it is necessary for the test developer/publisher to present the validity evidence arguing for the utility of the test on the evidence-based outcomes of the validity investigations.

When an instrument has been translated and/or adapted from a non-local context, evidence of equivalence of the measure in a new language to the original should be presented. Without this it is not possible to generalise findings in one country/language version to another. Some examples of equivalent evidence include:

- Invariance in construct structure – e.g. via factor structure or correlation with standard measures.
- Similar criterion related validity – e.g. similar profile of correlations of a multi-scale instrument with independent external criterion – such as ratings of job competencies.
- Items show similar patterns of scale loadings e.g. items correlate in the same pattern with other scales; strongest/weakest loading items are similar in original and new languages.
- Bilingual candidates have similar profiles in two languages (c.f. alternate form reliability).

Validity generalisation needs stronger evidence when translating tests across linguistic families (e.g. from a European to an African language). In such a situation equivalence is under greater threat because of the differences in language structure and cultural differences. However, validity generalisation might be inferred from evidence of validity invariance in previous translations when a test has been translated into multiple languages. For instance, if an English test has already been translated into Sesotho, isiZulu and Setswana and has been shown to have equivalence in these languages.

2.2.1 Content validity

Content validity refers to the degree to which a test adequately covers the behavioural domains or constructs of interest being measured (e.g. does a video clip in a Situational Judgement Test reflect the behaviours related to the competency being measured). This type of evidence is deemed to be less important in psychometric assessments, but is critical for the evaluation of assessment centre exercises and other skills-based assessments where the content of the scales must be judged to be reflective of the construct or skill to be measured.

2.2.1.1 Content validity processes followed

N/A – Not applicable
0 – No information provided
1 –Some description provided of process followed.
2 –Clear description provided of the process followed and construct domains measured consistent with the context.
3 – Clear description provided of the process followed using best practice guidelines consistent with the context.
4 – Clear description provided of the process followed using best practice guidelines and completion of a substantive study with clear links between the content domain and items consistent with the context.

2.2.2 Construct validity [EFPA 11.1]

The purpose of construct validation is to find an answer to the question whether the test actually measures the intended construct or, partly or mainly, something else. Common methods for the investigation of construct validity are exploratory or confirmatory factor analysis, item-test correlations, comparison of mean scores of groups for which score differences may be expected, testing for invariance of factor structure and item-bias (DIF) for different groups, correlations with other instruments which are intended to measure the same (convergent validity) or different constructs (discriminant validity), Multi-Trait-Multi-Method research (MTMM), IRT-methodology and (quasi-)experimental designs.

2.2.2.1 Information about construct validity presented

N/A – Not applicable
0 – No information provided
1 – Inadequate evidence provided for construct validity
2 – Adequate evidence provided for at least one aspect of construct validity within context
3 – Adequate evidence provided for at least two or more aspects of construct validity within context
4 – Good evidence provided for multiple aspects of construct validity within context

PLEASE NOTE: The following types of analysis are not all compulsory. If a factor analysis was done, use the guidelines to evaluate the quality, if Rasch analysis was done, use the guidelines to evaluate the quality. If the analysis technique was not used, mark as 'not applicable'.

2.2.2.1.1 Factor analysis [EFPA 11.1.2]

N/A – Not applicable
0 – No information provided
1 - Structure not supported
2- Can demonstrate that structure reflects intended measurement model, using best practice guidelines within the context.
3 - Can demonstrate that structure reflects intended measurement model, using more than one type of factor analysis, according to best practice guidelines within the context.
4 - Replicable factor structures across multiple samples, using best practice guidelines within the context

2.2.2.1.2 Mean score differences for relevant groups [EFPA 11.1.5]

For example, pupils in grade 8 are expected to score higher than pupils in grade 6 on a test for numerical proficiency; children with the diagnosis ADHD should score higher on a test for hyperactivity than children not diagnosed with ADHD; salespersons should score higher on a test for commercial knowledge than the average working population. Even though the results are in the expected direction, this kind of research usually is inconclusive with respect to the construct validity of the test. However, the value of this kind of research is that when the expected differences are not shown, this would raise strong doubts about the construct validity of the test.

N/A – Not applicable
0 – No information provided
1 – Group comparisons are not relevant to the construct/score differences are not as expected
2 – Group comparisons are relevant to the construct and score differences are as expected
3 – Score differences are as expected and effect sizes are reported

4 – Score differences are as expected, effect sizes are reported, and implications for interpretation contextualised

2.2.2.1.3 Correlations with similar constructs (convergent validity). [EFPA 11.1.6]

N/A – Not applicable
0 – No information provided
1 – Use of a test that is not established/no clear comparison on target construct
2 – Correlation between two target tests > |.3|
3 – Correlation between two target tests > |.4|
4 – Correlation between two target tests > |.5|

2.2.2.1.4 Correlations with different constructs (discriminant validity) [EFPA 11.1.7]

N/A – Not applicable
0 – No information provided
1 – Correlation between two target tests > |.3|
2 – Correlation between two target tests > |.2|
3 – Correlation between two target tests > |.1|
4 – No correlation between two target tests

2.2.2.1.5 Rasch analysis

Rating scale items with fit values above 1.40 (underfit) or below 0.60 (overfit) should be excluded from the assessment (Wright, Linacre, Gustafson, & Martin-Löf, 1994). However, overfit is less of a threat to the scale validity than underfit, so a rating of "adequate" only includes underfitting values. Wright et al. (1994) also suggested different ranges for different types of scale, with the most strict being 0.8-1.2 for high-stakes MCQ and the least strict for clinical observation (0.5-1.7). The most recent literature suggests that items with good fit will generally have scores ranging between .70 and 1.35 (Linacre, 2015).

N/A – Not applicable
0 – No information provided
1 – Item fit statistics (INFIT MNSQ) > 1.40, negative point-biserial correlation
2 – Item fit statistics (INFIT MNSQ) < 1.40
3 – Item fit statistics 0.60 < (INFIT MNSQ) < 1.40
4 – Item fit statistics 0.70 < (INFIT MNSQ) < 1.35

2.2.2.2 Adequate sample sizes [EFPA 11.1.10]

N/A – Not applicable
0 – No information provided
1 – single sample; inadequate sample size (n < 100)
2 – single sample; fair sample size (100 ≤ n < 300)
3 – single or multiple samples; at least one n > 500

4 – multiple samples; n > 500 for each sample

2.2.2.3 How old are the studies? [EFPA 11.1.12]

It is difficult to formulate a general rule for taking the age of the research into account. For tests that intend to predict behaviour in rapidly changing environments, 15-year-old research may be almost useless, whereas for other tests 20-year-old (or even older) research may still be relevant.

Number of years: ____

2.2.3 Criterion validity [EFPA 11.2]

Criterion-related evidence of validity (concurrent and predictive validity) refers to studies where real-world criterion measures (i.e. not other instrument scores) have been correlated with scales. Predictive studies generally refer to situations where assessment was carried out at a 'qualitatively' different point in time to the criterion measurement - e.g. for a work-related selection measure intended to predict job success, the instrument would have been carried out at the time of selection - rather than just being a matter of how long the time interval was between instrument and criterion measurement. Studies can also be 'post-dictive', for example, where scores on a potential selection test are correlated with job incumbents' earlier line manager ratings of performance. Basically, evidence of criterion validity is required for all kinds of tests. However, when it is explicitly stated in the manual that test use does not serve prediction purposes (such as educational tests that measure progress), criterion validity can be considered 'not applicable'.

N/A – Not applicable
0 – no information provided
1 – Inadequate evidence provided for criterion validity
2 – Adequate evidence provided for at least one aspect of criterion validity within context
3 – Adequate evidence provided for at least two or more aspects of criterion validity within context
4 – Good evidence provided for multiple aspects of criterion validity within context

2.2.3.1 Description of the type of criterion study / information presented (concurrent / predictive) [EFPA 11.2.1]

○ Concurrent study
○ Predictive study
○ Both

2.2.3.2 Sample sizes [EFPA 11.2.2]

N/A – Not applicable
0 – No information provided
1 – single sample; inadequate sample size (n < 100)
2 – single sample; fair sample size (100 ≤ n < 300)
3 – single or multiple samples; at least one n > 500
4 – multiple samples; n > 500 for each sample

2.2.3.3 Quality of the criterion measure used [EFPA 11.2.3]

N/A – not applicable
0 – No information provided
1– Criterion is not clearly described
2 – Criterion is objective, has variance, relates to target domain
3 – Reliability of criterion is reported and is acceptable
4 – Criterion has good reliability evidence and representation of the criterion construct

2.2.3.4 Strength of the relation between test and criterion scores. [EFPA 11.2.4]

N/A – not applicable
0 – No information provided
1 – $r < 0.20$
2 – $0.20 \leq r < 0.35$
3 – $0.35 \leq r < 0.50$
4 – $r \geq 0.50$

2.2.3.5 How old are the criterion validity studies? [EFPA 11.2.5]

It is difficult to formulate a general rule for taking the age of the research into account. For tests that intend to predict behaviour in rapidly changing environments, 15-year-old research may be almost useless, whereas for other tests 20-year-old (or even older) research may still be relevant.

Number of years: _______

2.2.4 Reviewers comments, evaluation & recommendation on validity [EFPA 11.3]

2.3 Bias and equivalence information

According to the Society for Industrial and Organizational Psychology's (2018) guidelines, "bias refers to systematic error in a test score that differentially affects the performance of different groups of test takers" (p.23). They distinguish between measurement bias (which has to do with a systematic error in scores between groups) and predictive bias (which has to do with systematic error in the prediction-criterion relationship for different groups). While the authors caution against the overinterpretation of differences between groups, it remains good practice to investigate these differences and account for them if necessary.

In considering the whole issue of equivalence, it would be useful to follow Van de Vijver and Poortinga's (2005) classification:

- <u>Structural / functional equivalence</u>
  - There is evidence that the assessment measures the same psychological constructs across groups. This is generally demonstrated by showing that patterns of correlations between variables are the same across groups.
- <u>Measurement unit equivalence</u>
  - There is evidence that the measurement units are the same, but there are different origins across groups (i.e. individual differences found in group A can be compared with differences found in group B, but the absolute raw scores for A and B are not directly comparable without some form of re-scaling).
- <u>Scalar / Full score equivalence</u>
  - The same measurement unit and the same origin (i.e. raw scores have the same meanings and can be compared across groups).

2.3.1 Evidence of factor structure invariance across relevant groups [EFPA 11.1.4]

N/A – Not applicable
0 – No information provided
1 – No separate factor analysis performed for different groups
2 – Separate factor analysis for different groups, compared to target rotation (tucker values)
3 – Exploratory Structural Equation Modelling (ESEM) with structural equivalence
4 – ESEM full equivalence (configural, metric & scalar)

2.3.2 Investigation into differential item functioning for different sample groups

N/A – Not applicable
0 – No information provided

1 – Relevant groups not compared
2 – Relevant groups compared, using best practice principles, and DIF findings are explained in context
3 – Relevant groups compared, using best practice principles, minimal evidence for DIF explained with implications for interpretation of scores
4 – Relevant groups compared, using best practice principles, no evidence for DIF

2.3.3 Evidence of similarities of scores provided for different sample groups

N/A – Not applicable
0 – No information provided
1 – Relevant groups indicated but not compared
2 – Relevant groups compared and score differences are explained with implications for interpretation of scores
3 – Relevant groups compared, minimal score differences exist and are explained with implications for interpretation of scores
4 – Relevant groups compared, and no score differences reported

2.3.4 How old are the studies? [EFPA 11.2.5]

It is difficult to formulate a general rule for taking the age of the research into account. For tests that intend to predict behaviour in rapidly changing environments, 15-year-old research may be almost useless, whereas for other tests 20-year-old (or even older) research may still be relevant.

Number of years: ____

2.3.5 Reviewers comments, evaluation & recommendation on bias and equivalence

2.4 NORMS

To give meaning to a raw test score two ways of scaling or categorizing raw scores can be distinguished (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

First, a set of scaled scores or norms may be derived from the distribution of raw scores of a reference group. This is called norm-referenced interpretation.

Second, standards may be derived from a domain of skills or subject matter to be mastered (domain-referenced interpretation) or cut scores may be derived from the results of empirical validity research (criterion-referenced interpretation). With the latter two possibilities raw scores will be categorised in two (for example 'pass' of 'fail') or more different score ranges, e.g. to assign patients in different score ranges to different treatment programs, to assign pupils scoring below a critical score to remedial teaching, or to accept or reject applicants in personnel selection.

Only the selected section will apply for scores. Select the relevant type of norm and rate only the applicable section:

N/A
NORM-BASED INTERPRETATION
DOMAIN-REFERENCED INTERPRETATION
CRITERION-BASED INTERPRETATION

2.4.1 NORM-BASED INTERPRETATION

2.4.1.1 Norms appropriate for local use [EFPA 9.1.1]

N/A – Not applicable
0 – No information provided
1 – Samples used are not relevant (e.g., inappropriate foreign samples)
2 – Local or global samples used with some relevance to the application domain
3 – Local or global samples with good relevance for intended application
4 – Local or global samples drawn from well-defined populations from the relevant application domain

2.4.1.2 Norms appropriate for the intended applications. [EFPA 9.1.2]

N/A – Not applicable
0 – No information provided
1 – Norms are not appropriate for the intended application
2 – Norms provided are appropriate for the intended applications

3 – Range of norms provided that are appropriate for the intended applications
4 – Range of norms provided that are appropriate for the intended applications and are contextualised.

2.4.1.3 Sample size overall [EFPA 9.1.3]

For most purposes, samples of less than 200 test takers will be too small, as the resolution provided in the tails of the distribution will be very small. The SE_{mean} for a z-score with N = 200 is 0.071 of the SD - or just better than one T-score point. Although this degree of inaccuracy may have only minor consequences in the centre of the distribution the impact at the tails of the distribution can be quite big (and this may be the score ranges that are most relevant for decisions to be taken on basis of the test scores). If there are international norms then in general, because of their heterogeneity, these need to be larger than the typical requirements of local samples. Generally high-stakes use is where a non-trivial decision is based at least in part on the test score(s). As a guideline, 100 fewer subjects can be used if the assessment is going to be used in low stakes situations only.

N/A – Not applicable
0 – No information provided
1 – Less than 300
2 – 300 – 399
3 – 400 – 999
4 – Over 1000

2.4.1.4 Sample size (continuous/inferential norming) [EFPA 9.1.4]

Continuous norming procedures have become more and more popular. They are used particularly for tests that are intended for use in schools (e.g. group 1 to 8 in primary education) or for a specific age range (e.g. an intelligence test for 6 to 16-year-olds). Continuous norming is more efficient as fewer respondents are required to get the same amount of accuracy of the norms. Bechger, Hemker, and Maris (2009) have computed some values for the sizes of continuous norm groups that would give equal accuracy compared to classical norming. When eight sub-groups are used N = 70 (8x70) gives equal accuracy compared to Ns of 200 (8x200) with the classical approach; N = 100 (x8) compares to 300 (x8) and N = 150 (x8) to 400 (x8). In these cases, the accuracy on the basis of the continuous norming approach is even better in the middle groups, but somewhat worse in the outer groups. Apart from the greater efficiency, another advantage is that, based on the regression line, values for intermediate norm groups can be computed. However, the approach is based on rather strict statistical assumptions. The test author has to show that these assumptions have been met, or that deviations from these assumptions do not have serious consequences for the accuracy of the norms.

However, more recent recommendations by Kranszler and Floyd (2013) sets basic minimum sample sizes for these norm blocks (particularly in cognitive and neuropsychological applications) to an absolute minimum of 30 cases. However, Norfolk et al. (2015) recommend that this minimum standard be revised to 50 in light of the required accuracy of the norm, in line with Zhu and Chen’s (2011) findings.

N/A – Not applicable
0 – No information given
1 – Less than 50 (for at least 5 age groups)
2 – 50 - 69
3 – 70 - 99
4 – 100 or more

2.4.1.5 Procedures for sample selection [EFPA 9.1.5]

N/A – Not applicable
0 – No information provided
1 – Listed but no clear description of procedure used
2 – Some description of the procedure used
3 – Clear description of the procedure used
4– Procedure used is critically discussed in relation to context

2.4.1.6 Stratification/representativeness of the norm sample [EFPA 9.1.6]

N/A – Not applicable
0 – No information provided
1 – No stratification done when needed for intended application
2 – Stratification according to some relevant key variables for the intended application
3 – Stratification done according to relevant key variables for the intended application and contextualised
4 – Stratification done according to all relevant key variables for the intended application and contextualised, using random sampling strategies

2.4.1.7 How old are the norm studies [EFPA 9.1.8]?

N/A – Not applicable
0 – No information provided
1 – Norms 20 years or older
2 – Norms between 10 and 19 years old
3 – Norms between 5 and 9 years old
4 – Norms less than 5 years old

2.4.2 DOMAIN-REFERENCED INTERPRETATION

2.4.2.1 Expert judgement – judges appropriately selected and trained? [EFPA 9.2.1.1]

N/A – Not applicable

0 – No information provided
1 – Inadequate description of selection procedure
2 – Motivation provided for how judges were selected
3 – Judges selected against explicit criteria
4 – Judges selected against explicit criteria and represent relevant diverse views or groups where possible

2.4.2.2 Expert judgement – number of judges used adequate? [EFPA 9.2.1.2]

N/A – Not applicable
0 – No information provided
1 – Less than two judges
2 – Two judges
3 – Three judges
4 – Four or more judges

2.4.2.3 Expert judgement – critical score for size of inter-rater agreement coefficient [EFPA 9.2.1.5]

N/A – Not applicable
0 – No information provided
1 – *r* <.60
2 – .60 ≤ *r* < .70
3 –.70 ≤ *r* < .80
4 – *r* ≥ .80)

2.4.2.4 How old are the normative studies? [EFPA 9.2.1.6]

N/A – Not applicable
0 – No information provided
1 – Norms 20 years or older
2 – Norms between 10 and 19 years old
3 – Norms between 5 and 9 years old
4 – Norms less than 5 years old

2.4.3 CRITERION-REFERENCED INTERPRETATION

No explicit guidelines can be given as to which level of relationship is acceptable in setting the critical score, not only because what is considered ‘high’ or ‘low’ may differ for each criterion to be predicted, but also because prediction results will be influenced by other variables such as base rate or prevalence. Therefore, the reviewer must rely on his/her expertise for his/her judgement. Also, the composition of the sample used for this research (is it similar to the group for which the test is intended, more heterogeneous, or more homogeneous?) and the size of this group must be taken into account.

2.4.3.1 Rationale used in developing critical scores

N/A – Not applicable
0 – No information provided
1 – Listed but no clear description of rationale
2 – Rationale clearly described
3 – Rationale clearly described with appropriate evidence
4 – Rationale clearly described with appropriate evidence and contextualised

2.4.3.2 How old are the normative studies? [EFPA 9.2.2.2]

N/A – Not applicable
0 – No information provided
1 – Norms 20 years or older
2 – Norms between 10 and 19 years old
3 – Norms between 5 and 9 years old
4 – Norms less than 5 years old

2.4.4 Reviewers comments, evaluation & recommendation on norms [EFPA 10.8]

2.5 Reference list

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering* [On the use of continuous norming]. Arnhem, The Netherlands: Cito.

Borsboom, D., Mellenbergh, G., & Van Heerden, J. (2004). The concept of validity. *Psychological Review 111*(4), 1061-71. DOI: 10.1037/0033-295X.111.4.1061

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Foxcroft, C., & Roodt, G. (2019). *Introduction to psychological assessment in the South African context (5th ed.)*. Cape Town, South Africa: Oxford University Press.

Kranzler, J. H., & Floyd, R. G. (2013). *Assessing intelligence in children and adolescents: A practical guide.* New York, NY: Guilford Press.

Linacre, J. M. (2015). *Winsteps® Rasch measurement computer program User's Guide.* Beaverton, Oregon: Winsteps.com

Norfolk, P. A., Farmer, R. L., Floyd, R. G., Woods, I. L., Hawkins, H. K., & Irby, S. M. (2015). Norm block sample sizes: A review of 17 individually administered intelligence tests. *Journal of Psychoeducational Assessment, 33*, 544–554. doi:10.1177/0734282914562385

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Parshall, C. G., Spray, J. A., Davey, T., & Kalohn, J. (2001). *Practical Considerations in Computer-based Testing.* New York: Springer Verlag.

Reise, S. P., & Havilund, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Measurement, 84*, 228-238.

Schneider, R. J., & Hough, L. M. (1995). Personality and industrial/organizational psychology. In C. L. Cooper & I. T. Robertson (Eds.), International Review of Industrial and Organizational Psychology, 10, 75-129.

Society for Industrial and Organizational Psychology (SIOP). (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Bowling Green, OH: SIOP.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

Wright, B. D., Linacre, J. M., Gustafson, J.-E., Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370-371.

Zhu, J., & Chen, H.-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29*, 570-580.