EFPA REVIEW MODEL FOR THE DESCRIPTION AND EVALUATION OF PSYCHOLOGICAL TESTS[1]
TEST REVIEW FORM AND NOTES FOR REVIEWERS[2]

It is IMPORTANT that reviewers read all notes carefully whilst carrying out the review.

Original version compiled and edited by Dave Bartram
Updated and revised by Patricia Lindley, Dave Bartram and Natalie Kennedy, April 2004, May 2005[3]
Current version 3.42: September 2008

For details of the Review Procedure and any local modifications to the review content and criteria, consult you local Psychological Association. The present document provides for procedures that employ two reviewers for each test review, with a third person to oversee the review (the 'Consulting Editor') and a Senior Editor who is responsible for ensuring uniformity of application of the criteria across reviews. Local arrangements may result in some of these functions being combined.

EFPA recommends that the evaluations in these reviews are directed towards qualified practising test users, though they should also be of interest to academics and specialists in psychometrics and psychological testing.

Users of this document and its contents are required by EFPA to acknowledge this source with the following text:

> *"The EFPA Test Review Criteria were largely modelled on the form and content of the British Psychological Society's (BPS) test review criteria and criteria developed by the Committee of Test Affairs (COTAN) of the Dutch Association of Psychologists (NIP). Dave Bartram and Patricia Lindley originally developed the BPS criteria and review procedures for the UK Employment Service and later expanded them for use by the BPS. Arne Evers edited the Dutch rating system for test quality. EFPA is grateful to the BPS and the NIP for permission to build on their criteria in developing the European model. EFPA is also grateful to Dave Bartram, Arne Evers, and Patricia Lindley for their contributions to the development of this model. All intellectual property rights in the original BPS and NIP criteria are acknowledged and remain with those bodies."*

---

[1] The EFPA Standing Committee on Tests and Testing has endorsed these notes for reviewers and the related review format. Member Psychological Associations may use them as the basis for their own instrument review procedures. The intention of making this widely available is to encourage the harmonisation of review procedures and criteria across Europe. Comments on these documents are welcomed in the hope that the experiences of instrument reviewers will be instrumental in improving and clarifying the processes.

[2] This document has been produced from a number of sources, including the BPS Test Review Evaluation Form (NPAL, and BPS Steering Committee on Test Standards); the Spanish Questionnaire for the Evaluation of Psychometric Tests (Spanish Psychological Association) and the Rating System for Test Quality (Committee on Testing of the Dutch Association of Psychologists). Some of the content has been adapted with permission from: BPS Books Reviews of Level B Assessment Instruments for use in Occupational Assessment, Notes for Reviewers: Version 3.1. December 1998: Copyright © NPAL, 1989, 1993, 1998.

[3] The present version is an integration of two separate documents (the Review Form and the Notes for Reviewers). In addition the contents have been edited and amended in the light of use by the BPS for their online test reviews. Version 3.4 was accepted as the replacement for all previous versions by the General Assembly of EFPA in July 2005.

Section 1:
Description of the Instrument: General Information & Classification

The first section of the form should provide the basic information needed to identify the instrument and where to obtain it. It should give the title of the instrument, the publisher and/or distributor, the author(s), the date of original publication and the date of the version that is being reviewed.

Sections 1.1 through 1.9 should be straightforward. They are factual information, although some judgement will be needed to complete information regarding content domains.

EFPA 3.2 reference		
	Reviewer 1:	
	Reviewer 2:	
	Consulting Editor:	
	Senior Editor:	
	Senior Update Editor: **(only required for updates)**	
	Update Editor: **(only required for updates)**	
	Date of Current Review:	
1.1	**Instrument Name (local version):**	
	Short Version Test Name:	
1.2	**Original Test Name (if the local version is an adaptation:**	
1.4	**Authors of the Original Test:**	
1.3	**Authors of the Local Adaptation:**	
1.7	**Local Test Distributor/Publisher:**	
1.8	**Publisher of the original version of the Test (if different to current Distributor/Publisher:**	
1.9.1	**Date of publication of current revision/edition:**	
1.9.2	**Date of publication of adaptation for local use:**	
1.9.3	**Date of publication of original Test:**	

General Description of the Instrument Short stand-alone non-evaluative description (200-600 words)

This section should contain a concise non-evaluative description of the instrument. The description should provide the reader with a clear idea of what the instrument claims to be - what it contains, the scales it purports to measure etc. It should be as neutral as possible in tone. It should describe what the instrument is, the scales it measures, general points of interest or unusual features and any relevant historical background. This section may be quite short (c 200-300 words). However, for some of the more complex multi-scale instruments, it will need to be longer (c 300-600 words). It should be written so that it can stand alone as a description of the instrument. As a consequence it may repeat some of the more specific information provided in response to sections 1.1 - 1.29.

This item should be answered from information provided by the publisher and checked for accuracy.

Section 2: Classification

1.10.1	**Contents Domain** *(please tick those that apply)* You should identify the content domains specified by the publisher. Where these are not clear, this should be indicated and you should judge from the information provided in the manual (standardisation samples, applications, validation etc) what the most appropriate answers are for 1.10.1.	☐ Scholastic attainment ☐ General ability ☐ Verbal ability ☐ Numerical ability ☐ Spatial ability ☐ Non-verbal ability ☐ Perceptual speed ☐ Memory ☐ Manual skills/dexterity ☐ Personality – Trait ☐ Personality – Type ☐ Personality – State ☐ Cognitive Styles ☐ Motivation ☐ Values ☐ Interests ☐ Beliefs ☐ Disorder and pathology ☐ Group function ☐ Family function ☐ Organisational function, aggregated measures, climate etc ☐ School or educational function ☐ Other: (describe below)
1.10.2	**Intended or Main Area (s) of Use.** *(please tick those that apply)* You should identify the intended areas of uses specified by the publisher. Where these are not clear, this should be indicated and you should judge from the information provided in the manual (standardisation samples, applications, validation etc) what the most appropriate answers are for 1.10.2.	☐ Psycho-clinical ☐ Psycho-neurological ☐ Forensic ☐ Educational ☐ Work and Occupational ☐ Counselling, advice, guidance and career choice ☐ General health, life and well-being ☐ Sports and Leisure ☐ Other (describe below)
1.10.3	**Intended Mode of Use (conditions under which the instrument was standardised and validated)** (t*ick one box*) This section is important as it identifies whether the instrument has been designed with the intention of it being used in unsupervised or uncontrolled administration conditions. This item should be answered from information provided by the publisher and checked for accuracy.	☐ Unsupervised administration without control over the identity of the test taker and without full control over the conditions of administration (e.g. open access Internet delivered test; test available for purchase from bookstores) ☐ Controlled by unsupervised administration. Control over conditions (timing etc) and some control of identify of the test taker (e.g. tests administered over the Internet but only to known individuals - password restricted access)

		☐ Supervised and controlled administration. Test administration under the control of a qualified administrator or proctor ☐ Managed administration. Test administration only provided through specified testing centres (e.g. licensing and certification assessment programmes)
1.10.4	**Description of the populations for which the test is intended:** This item should be answered from information provided by the publisher. For some tests this may be very general (e.g. adults), for others it may be more specific (e.g. manual workers, or boys aged 10 to 14). This section should contain the stated populations. Where these may seem inappropriate, this should be commented on in the Evaluation section of the review.	
1.10.5	**Number of scales and brief description of the variable or variables measured by the instrument** This item should be answered from information provided by the publisher. Please indicate the number of scales (if more than one) and provide a brief description of each scale if its meaning is not clear from its name.. Reviews of the instrument should include discussion of other derived scores where these are commonly used with the instrument and are described in the standard documentation - e.g. 16PF criterion scores and Adjustment Specification Equation scores - but not scales which are `add-ons' - e.g. the use of 16PF or OPQ scores to generate Belbin team-type measures.	
1.11	**Items format** *(select one)* This item should be answered from information provided by the publisher. Note that it is important not to confuse multiple choice same scale with multiple choice different scale item formats. The latter, ipsative, formats require test takers to make choices between sets of two or more items drawn from *different* scales. For 'Forced-choice mixed-scale alternatives,' the test taker has to select which of two scales are more like them or with which of two statements they most agree. For multiple choice version, there may be three or more statements drawn from an equivalent number of different scales. Typically these statements may have to be ranked or the most- and least-like-me options selected.	☐ Open ☐ Multiple choice, same scale alternatives ☐ Bipolar adjectives ☐ Likert ratings ☐ Forced choice, mixed scale alternatives (ipsative) – see Notes for explanation. ☐ Multiple choice, mixed scale alternatives (ipsative) – see Notes for explanation. ☐ Adjective pair of sets, mixed scales (ipsative) ☐ Other (describe below)
1.12	**No of Test Items:** This item should be answered from information provided by the publisher. If the instrument has several scales, make clear whether you are indicating the total number of items or the number of items for each scale. Where items load on more than one scale, this should be documented.	

1.13	**Administration Mode(s):** This item should be answered from information provided by the publisher. If any special pieces of equipment (other than those indicated in the list of options, e.g. VCR, tape recorder) are required, they should be described here. In addition, any special testing conditions should be described. `Standard testing conditions' are assumed to be available for proctored/ supervised assessment. These would include a quiet, well-lit and well-ventilated room with adequate desk-space and seating for the necessary administrator(s) and candidate(s).	☐ Interactive Individual Administration ☐ Supervised Group Administration ☐ Computerised locally-installed application – supervised/proctored ☐ Computerised web-based application – supervised/proctored ☐ Computerised locally-installed application – unsupervised/self-assessment ☐ Computerised web-based application – unsupervised/self-assessment ☐ Other (indicate)
1.14	**Response Mode:** This item should be answered from information provided by the publisher. If any special pieces of equipment (other than those indicated in the list of options, e.g. VCR, tape recorder) are required, they should be described here. In addition, any special testing conditions should be described. `Standard testing conditions' are assumed to be available for proctored/ supervised assessment. These would include a quiet, well-lit and well-ventilated room with adequate desk-space and seating for the necessary administrator(s) and candidate(s).	☐ Oral Interview ☐ Paper & Pencil ☐ Manual operations ☐ Computerised ☐ Other (indicate)
1.15	**Time required for administering the instrument:** This item should be answered from information provided by the publisher. The response to this item can be broken down into a number of components. In most cases, it will only be possible to provide general estimates of these rather than precise figures. The aim is to give the potential user a good idea of the time investment associated with using this instrument. Do NOT include the time needed to become familiar with the instrument itself. Assume the user is experienced and qualified. Preparation time (the time it takes the administrator to prepare and set out the materials for an assessment session). • Administration time per session: this includes the time taken to complete all the items and an estimate of the time required to give instructions, work through example items and deal with any debriefing comments at the end of the session. • Scoring: the time taken to obtain the raw-scores. • Analysis: the time taken to carry out further work on the raw scores to derive other measures and to produce a reasonably comprehensive interpretation (assuming you are familiar with the instrument). • Feedback: the time required to prepare and provide feedback to a candidate. It is recognised that time for the last two components could vary enormously - depending on the context in which the instrument is being used. However, some indication or comments will be helpful.	Preparation: ☐ Administration: ☐ Scoring: ☐ Analysis: ☐ Feedback: ☐

1.16	**Indicate whether different forms of the instrument are available** (genuine or pseudo-parallel forms, short versions, computerised versions, etc). If computerised versions do exist, describe briefly the software and hardware requirements: For Section 1.16, report whether or not there are alternative versions (forms) of the instrument available and describe the applicability of each form for different groups of people. In some cases, different forms of an instrument are meant to be equivalent to each other - i.e. alternate forms. In other cases, various forms may exist for quite different groups (e.g. a children's form and an adult's form). Where more than one form exists, indicate whether these are equivalent/alternate forms, or whether they are designed to serve different functions - e.g. short and long version; ipsative and normative version. Some, instruments may be partly or fully computerised or available in computerised versions. For each of the four `stages' of the assessment process, indicate the options available from the supplier. Note that CBTI packages, if available, should be indicated.	

Section 3: Measurement & Scoring

1.17	**Scoring procedure for the test:** This item should be completed by reference to the publisher's information and the manuals and documentation. Bureau services are services provided by the supplier - or some agent of the supplier - for scoring and interpretation. In general these are optional services. If scoring and/or interpretation can be carried out ONLY through a bureau service, then this should be stated in the review - and the costs included in the recurrent costs section.	☐ Computer scoring with direct entry of responses by test taker ☐ Computer scoring with manual entry of responses from the paper response form ☐ Computer scoring by Optical Mark Reader entry of responses from the paper response form ☐ Simple manual scoring key – clerical skills only required ☐ Complex manual scoring – requiring training in the scoring of the instrument ☐ Bureau-service – eg. Scoring by the company selling the instrument ☐ Other (describe below)
1.18	**Scores:** This item should be completed by reference to the publisher's information and the manuals and documentation. Brief description of the scoring system to obtain global and partial scores, correction for guessing, qualitative interpretation aids, etc).	
1.19	**Score Transformation for standard scores:**	☐ Normalised – scores obtained by use of normalisation look-up table ☐ Not-normalised – scores obtained by linear transformation

1.20	**Scales used** (*tick all that apply)* This item should be completed by reference to the publisher's information and the manuals and documentation.	*Percentile Based Scores* ☐ Centiles ☐ 5-grade classification: 10:20:40:20:10 centile splits ☐ Deciles, and other equi-percentile classifications *Standard Scores* ☐ Z-scores ☐ IQ deviation quotients etc (e.g. mean 100, SD=15 for Weschler or 16 for Stanford-Binet) ☐ College Entrance Examination Board (e.g. SAT mean=500, SD=100) ☐ Stens, ☐ Stanines, C Scores ☐ T-Scores ☐ Other (please describe)

Section 4:
Computer Generated Reports

Note that this is purely *descriptive*. Evaluations of the reports would be part of the Evaluation section of the review

For instances where there are multiple generated reports available please complete items 1.21 – 1.22 for each report (please copy pages if necessary). This classification system could be used to describe two reports provided by a system, for example, Report 1 may be intended for the test taker or other un-trained users, and Report 2 for a trained user who is competent in the use of the instrument and understands how to interpret it.

1.21	**Are Computer Generated Reports available with the instrument?** If the answer to 1.21 is 'YES' then the following classification should be used to classify the types of reports available. For many instruments, there will be a range of reports available. Please complete a separate form for each report	☐ Yes (complete sections below) ☐ No (move to section 1.23)
1.21.0	**Name or description of Report:** *(See introduction to this section)*	
1.21.1	**Media:***(select one)* Reports may consist wholly of text or contain text together with graphical or tabular representations of scores (e.g. sten profiles). Where both text and data are presented, these may simply be presented in parallel or may be linked, so that the relationship between text statements and scores is made explicit.	☐ Text only ☐ Unrelated text and graphics ☐ Integrated text and graphics ☐ Graphics only

1.21.2	**Complexity:** *(select one)* Some reports are very simple, for example just substituting a text unit for a sten score in a scale-by-scale description. Others are more complex, involving text units which relate to patterns or configurations of scale scores and which consider scale interaction effects.	☐ Simple (For example, a list of paragraphs giving scale descriptions) ☐ Medium (A mixture of simple descriptions and some configural descriptions) ☐ Complex (Contains descriptions of patterns and configurations of scale scores, and scale interactions)
1.21.3	**Report Structure**:*(select one)* Structure is related to complexity.	☐ Scale based (where the report is built around the individual scales) ☐ Factor based (where the report is constructed around higher order factors - such as the 'Big Five' for personality measures. ☐ Construct based - where the report is built around one or more sets of constructs (e.g. in a work setting these could be such as team types, leadership styles, tolerance to stress etc) which are linked to the original scale scores. ☐ Criterion based where the reports focuses on links with empirical outcomes (e.g. training potential, job performance, absenteeism etc). ☐ Other
1.21.4	**Sensitivity to Context** *(select one)* When people write reports they tailor the language, form and content of the report to the person who will be reading it and take account of the purpose of the assessment and context in which it takes place. A report produced for selection purposes will be different from one written for guidance or development; a report for a middle-aged manager will differ from that written for a young person starting out on a training scheme and so on.	☐ One version for all contexts ☐ Pre-defined context-related versions ☐ User definable contexts and editable reports
1.21.5	**Clinical-actuarial** *(select one)* Most reports systems are based on clinical judgment. That is, one or more people who are `expert-users' of the instrument in question will have written the text units. The reports will, therefore, embody their particular interpretations of the scales. Some systems include actuarial reports where the statements are based on empirical validation studies linking scale scores to, for example, job performance measures.	☐ Based on clinical judgement of one expert ☐ Based on empirical/actuarial relationships ☐ Based on clinical judgement of group of experts

1.21.6	**Modifiability** (*select one)* The report output is often fixed. However, some systems will produce output in the form of a file that can be processed by the user.	☐ Not modifiable (fixed print-only output) ☐ Limited modification (limited to certain areas e.g. biodata fields) ☐ Unlimited modification (e.g. through access to Word Processor document file)
1.21.7	**Degree of Finish** *(select one)* A related issue is the extent to which the system is designed to generate integrated text - in the form of a ready-to-use report - or a set of `notes', comments, hypotheses etc. The latter is clearly of far more use when the text is available to the user in modifiable form and can form the basis for the user's own report. In many cases, reports are designed to a very high standard of presentation, having a `published' appearance and quality.	☐ Publication quality ☐ Draft Quality
1.21.8	**Transparency** *(select one)* Systems differ in their openness or transparency to the user. An open system is one where the link between a scale score and the text is clear and unambiguous. Such openness is only possible if both text and scores are presented and the links between them made explicit. Other systems operate as 'black boxes', making it difficult for the user to relate scales scores to text.	☐ Clear linkage between constructs scores and text ☐ Concealed link between constructs, scores and text ☐ Mixture of clear/concealed linkage between constructs, scores and text
1.21.9	**Style and Tone** *(select one)* Systems also differ in the extent to which they offer the report reader guidance or direction. Some are stipulative: 'Mr X is very shy and will not make a good salesman...'. Others are designed to suggest hypotheses or raise questions: 'From his scores on scale Y, Mr X appears to be very shy. If this is the case, he could find it difficult working in a sales environment. This needs to be explored further with him.'	☐ Directive ☐ Guidance ☐ Other
1.21.10	**Intended Recipients** (*select all that apply)* Reports are generally designed to address the needs of one or more categories of users. Users can be divided into four main groups: *a)* *Qualified users.* These are people who are sufficiently knowledgeable and skilled to be able to produce their own reports based on scale scores. They should be able to make use of reports that use technical psychometric terminology and make explicit linkages between scales and descriptions. They should also be able to customize and modify reports. *b)* *Qualified system users.* While not competent to generate their own reports from a set of scale scores, people in this group are competent to use the outputs generated by the system. The level of training required to attain this competence will vary considerably, depending on the nature of the computer reports (e.g. trait-based versus competency-based, simple or complex) and the uses to which its reports are to be put	☐ Qualified test users ☐ Qualified system users ☐ Test takers ☐ Third Parties

	(low stakes or high stakes). c) *Test Takers.* The person who takes the instrument will generally have no prior knowledge of either the instrument or the type of report produced by the system. Reports for them will need to be in language that makes no assumptions about psychometric or instrument knowledge. d) *Third parties.* These include people - other than the candidate - who will be privy to the information presented in the report or who may receive a copy of the report. They may include potential employers, a person's manager or supervisor or the parent of a young person receiving careers advice. The level of language required for people in this category would be similar to that required for reports intended for Test Takers.	
1.22	**Do Distributors offer a service to correct and/or develop computerised reports?** *(select one)*	☐ Yes ☐ No

Section 5:
Supply, Condition and Costs

This defines what the publisher will provide, to whom, under what conditions and at what costs. It defines the conditions imposed by the supplier on who may or may not obtain the instrument materials. If one of the options does not fit the supply conditions, provide a description of the relevant conditions

1.23	**Documentation provided by the Distributor as part of the test package** *(select all that apply)*	☐ User Manual ☐ Technical (psychometric) manual ☐ Supplementary technical information and updates (eg. local norms, local validation studies etc) ☐ Books and articles of related interest ☐ Combinations of the above (specify)
1.24	**Methods of Publication** *(select all that apply)* For example, technical manuals may be kept up-to-date and available for downloading from the Internet, while user manuals are provided in paper form or on a CD.	☐ Paper ☐ PC - Diskettes ☐ PC – CD/ROM ☐ Internet download ☐ Live internet (instrument runs in a web browser) ☐ Other (specify)
Sections 1.25- 1.27 cover costs. This is likely to be the section that is most quickly out of date. It is recommended that the supplier or publisher is contacted as near the time of publication of the review as possible, to provide current information for this section.		
1.25.1	**Start-Up Costs.** Price of a complete set of materials (all manuals and other material sufficient for at least one sample administration). Specify how many candidates could be assessed with the materials obtained for start-up costs, where these costs include materials for recurrent assessment This item should try to identify the 'set-up' cost. That is the costs involved in obtaining a full reference set of materials, scoring keys and so on. It only includes training costs if the instrument is a 'closed' one - where there will be an <u>unavoidable</u> specific training cost, regardless of the prior	

	qualification level of the user. In such cases, the training element in the cost should be made explicit. The initial costs do NOT include costs of general-purpose equipment (such as computers, cassette tape recorders and so on). However, the need for these should be mentioned. In general, define: any special training costs; costs of administrator's manual; technical manual(s); specimen or reference set of materials; initial software costs etc.	
1.25.2	**Recurrent Costs:** Specify, where appropriate, recurrent costs of administration and scoring separately from costs of interpretation. (see 1.26 – 1.27) This item is concerned with the ongoing cost of using the instrument. It should give the cost of the instrument materials (answer sheets, non-reusable or reusable question booklets, profile sheets, computer usage release codes or 'dongle' units etc.) per person per administration. Note that in most cases, for paper-based administration such materials are not available singly but tend to be supplied in packs of 10, 25 or 50. Itemise any annual or per capita licence fees (including software release codes where relevant), costs of purchases or leasing re-usable materials, and per candidate costs of non-reusable materials.	
1.26.1	**Prices for Reports generated by user installed software:**	
1.26.2	**Prices for Reports generated by postal/fax bureau service:**	
1.26.3	**Prices for Reports by Internet Service:**	
1.27	**Prices for other bureau services: correcting or developing automatic reports:**	
1.28	**Test-related qualifications required by the supplier of the test** *(select all that apply)* 1.28 concerns the user qualifications required by the supplier. For this section, where the publisher has provided user qualification information, this should be noted against the categories given. Where the qualification requirements are not clear this should be stated under 'Other' *not* under 'None'. 'None' means that there is an explicit statement regarding the lack of need for qualification.	☐ None ☐ Test specific accreditation ☐ Accreditation in general achievement testing: measures of maximum performance in attainment ☐ Accreditation in general ability and aptitude testing: measures of maximum performance in relation to potential for attainment ☐ Accreditation in general personality and assessment: measures of typical behaviour, attitudes and preferences ☐ Other (specify)

1.29	**Professional qualifications required for use of the instrument** *(select all that apply.* 1.29 concerns the user qualifications required by the supplier. For this section, where the publisher has provided user qualification information, this should be noted against the categories given. Where the qualification requirements are not clear this should be stated under 'Other' *not* under 'None'. 'None' means that there is an explicit statement regarding the lack of need for qualification.	☐ None ☐ Practitioner psychologist with qualification in the relevant area of application ☐ Practitioner psychologist ☐ Research psychologist ☐ Non-psychologist academic researcher ☐ Practitioner in relevant related professions (therapy, medicine, counselling, education, human resources etc) ☐ Holder of BPS Certificate of Competence in Occupational Testing Level A ☐ Holder of BPS Certificate of Competence in Educational Testing Level A ☐ Holder of BPS Certificate of Competence in Occupational Testing Level B ☐ Other (indicate)

Section 6:
Evaluation of Test Materials

Potentially there are four sources of information that might be consulted in carrying out this evaluation:

1. The manual and /or reports that are supplied by the publisher for the user:
   a) These are always supplied by the publisher /distributor before the instrument is accepted by the office and form the core materials for the review.

2. Open information that is available in the academic or other literature:
   a) This is generally sourced by the reviewer and the reviewer may make use of this information in the review and the instrument may be evaluated as having (or having not) made reference to the information in its manual.

3. Reports held by the publisher that are not formally published or distributed:
   a) The distributor/publisher may make these available at the outset or may send them when the review is sent back to the publisher to check for factual accuracy. The reviewer should make use of this information but note very clearly at the beginning of the comments on the technical information that "the starred rating in this review refers to materials held by the publisher/ distributor that is not [normally] supplied to test users". If these contain valuable information, the overall evaluation should recommend that the publisher should publish these reports and/or make them available to test purchasers

4. Reports that are commercial in confidence:
   a) In some instances, publishers may have technically important material that they are unwilling to make public for commercial reasons. In practice there is very little protection available for intellectual property to test developers (copyright law being about the only recourse). Such reports might cover the development of particular scoring algorithms, test or item generation procedures and report generation technology. Where the content of such reports might be important in making a judgement in a review, the BPS could offer to undertake to enter into a non-disclosure agreement with publisher. This agreement would be binding on the reviewers and editor. The reviewer could then evaluate the information and comment on the technical aspects and the overall evaluation to the effect that "the starred rating in this review refers to materials held by the publisher/ distributor that have been examined by the reviewers on a commercial in confidence basis. These are not supplied with the manual."

Explanation of Star Ratings

All sections are scored using the following rating system where indicated by: [rating]. Detailed descriptions giving anchor-points for each rating are provided.

Where a [ ] rating is provided on an attribute that is regarded as critical to the safe use of an instrument, the review will recommend that the instrument should not be used, except in exceptional circumstances by highly skilled experts or in research.

The instrument review needs to indicate which, given the nature of the instrument and its intended use, are the critical technical qualities. It is suggested that the convention to adopt is that ratings of these critical qualities are then shown in bold print.

In the following sections, overall ratings of the adequacy of information relating to validity, reliability and norms are shown, by default, in bold.

Any instrument with one or more [ ] or [*] ratings regarding attributes that are regarded as critical to the safe use of that instrument, shall not be deemed to have met the minimum standards.

Entry on rating form	*EFPA standard rating*	*UK Review representation*	*Explanation*
[n/a]	[n/a]	[n/a]	This attribute is not applicable to this instrument
0	[-]	[None]	Not possible to rate as no, or insufficient information provided
1	[-1]	[*]	Inadequate
		[**]	NOT NOW USED
3	[0]	[***]	Adequate or Reasonable
4	[1]	[****]	Good
5	[2]	[*****]	Excellent
		[N.r.i.o.r] * (for updates only)	Item was not rated in original review

In this section a number of ratings need to be given to various aspects or attributes of the documentation supplied with the instrument (or package). The term `documentation' is taken to cover all those materials supplied or readily available to the qualified user: e.g. the administrator's manual; technical handbooks; booklets of norms; manual supplements; updates from publishers/suppliers and so on.

Suppliers are asked to provide a complete set of such materials for each Reviewer. If you think there is something which users are supplied with which is not contained in the information sent to you for review, please contact you Consulting Editor.

Items to be rated n/a or 0 to 5 (half ratings are acceptable) **Rating**

Quality of the explanation of the rationale, the presentation and the quality of information provided: (This overall rating is obtained by using judgment based on the ratings given for items 2.1 – 2.8)		
2.1	**Overall rating of the Quality of the explanation of the rationale:** (This overall rating is obtained by using judgment based on the ratings given for items 2.1.1 – 2.1.5)	
2.1.1	i) Theoretical foundations of the constructs:	
2.1.2	ii) Test development procedure:	
2.1.3	iii) Thoroughness of the item analyses and item analysis model:	
2.1.4	iv) Explanation of content validity:	
2.1.5	v) Summary of relevant research:	

2.2	**Adequacy of documentation available to the user (user and technical manuals, norm supplements etc):** (This rating is obtained by using judgment based on the ratings given for items 2.2.1 – 2.2.6) ***For Section 2.2, the following 'benchmarks are provided for an 'excellent' (*****) rating.*** The focus here is on the quality of coverage provided in the documentation accessible to qualified users. Note that section 2.2. is about the comprehensiveness and clarity of the documentation available to the user (user and technical manuals, norm supplements etc.) in terms of its coverage and explanation. In terms of the quality of the instrument as evidenced by the documentation, areas in this section are elaborated on under: 2.1, 2.3, 2.9, 2.10 and 2.11.	
2.2.1	Rationale: [see rating 2.1] Well-argued and clearly presented description of what is designed to measure and why it was constructed as it was.	
2.2.2	Development: Full details of item sources, piloting, item analyses, comparison studies and changes made during development trials.	
2.2.3	Standardisation: Clear and detailed information provided about sizes and sources of standardisation sample and standardisation procedure.	
2.2.4	Norms: Clear and detailed information provided about sizes and sources of norms groups, conditions of assessment etc.	
2.2.5	Reliability: Good explanation of reliability and a comprehensive range of internal consistency and retest measures provided with explanations of their relevance, and the generalisability of the assessment instrument	
2.2.6	Validity: Good explanation of validity with a wide range of studies clearly and fairly described.	
2.3	**Quality of the Procedural instructions provided for the user:** (This overall rating is obtained by using judgment based on the ratings given for items 2.3.1 – 2.3.7)	
2.3.1	For test administration: Clear and detailed explanations and step-by-step procedural guides provided, with good detailed advice on dealing with candidates' questions and problem situations.	
2.3.2	For test scoring, norming etc: Clear and detailed information provided, with checks described to deal with possible errors in scoring	
2.3.3	For interpretation and reporting: Detailed advice on interpreting different scores, understanding normative measures and dealing with relationships between different scales, with plenty of illustrative examples and case studies	
2.3.4	For providing feedback and debriefing test takers and others: Detailed advice on how to present feedback to candidates	
2.3.5	For providing good practice issues on fairness and bias: *Detailed information reported about sex and ethnic bias studies, with relevant warnings about use and generalisation of validities*	
2.3.6	Restrictions on use: Clear descriptions of who should and who should not be assessed, with well-explained justifications for restrictions (e.g. types of disability, literacy levels required etc)	
2.3.7	References and supporting materials: Detailed references to the relevant supporting academic literature and cross-references to other related assessment instrument materials.	
Quality of the materials: (This overall rating is obtained by using judgment based on the ratings given for items 2.4 – 2.8)		
2.4	General quality of test materials (test booklets, answer sheets, test objects, software, etc):	
2.5	Test quality of the local adaptation (if the test has been translated and adapted into the local language):	
2.6	Ease with which the test taker can understand the task:	
2.7	Ease with which responses or answers can be made by the test taker:	
2.8	Quality of the items:	

Reviewers' comments on the documentation: (comment on rationale, design, test development and acceptability)

Section 7:
Evaluation of Norms, Reliability and Validity

General guidance on assigning ratings for these sections

It is almost impossible to set clear criteria for rating the technical qualities of an instrument. Under some conditions a reliability of 0.70 is fine; under others it would be inadequate. A criterion-related validity of 0.20 can have considerable utility in some situations, while one of 0.40 might be of little value in others. For these reasons, summary ratings should be based on your judgement and expertise as a reviewer and not simply derived by averaging sets of ratings.

These notes provide some guidance on the sorts of values to associate with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which reliability and validity estimates are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded.

In order to provide some idea of the range and distribution of values associated with the various scales that make up an instrument, enter the *number of scales* in each section. For example, if an instrument being used for group-level decisions had 15 scales of which 5 had retest reliabilities lower than 0.6, 6 between 0.6 and 0.70 and the other 4 in the 0.70 to 0.80 range, this would be entered as:

Stability:

Median stability:

[ ]	No information given.
[5]	Inadequate (e.g. r < 0.6).
[6]	Adequate (e.g. 0.6 < r < 0.70).
[4]	Good (e.g. 0.70 < r < 0.80).
[ ]	Excellent (e.g. r > 0.80)

It is realised that it may be impossible to calculate actual median figures in many cases. What is required is your best estimate, given the information provided in the documentation. There is space to add comments. You can note here any concerns you have about the accuracy of your estimates. For example, in some cases, a very high level of internal consistency might be commented on as indicating a `bloated specific'.

For each of the possible ratings example values are given *for guidance only* - especially the distinctions between `Adequate', `Good' and `Excellent'. It is recognised that what is a `good' value for one instrument may be unrealistic for another. (For example, other things being equal we would expect a higher internal consistency for long scales than for short ones).

Where NORMS are concerned, the guidelines for sample sizes need to take into account the type of norms being used. If they claim to be representative general population norms, then the sample size will need to be quite large even to be `adequate'. If they are occupation-specific norm groups, smaller sample sizes may be `adequate'.

Careful consideration needs to be given to the suitability of international (same language) norms. Where these have been carefully established from samples drawn from a group of countries, they should be rated on the same basis as nationally based (single language) norm group.

For most purposes, samples of less than 150 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=150 is 0.082 of the SD - or just better than one T-score point.

For VALIDITY, guidelines on sample sizes are based on power analysis of the sample sizes needed to find moderate sized validities if they exist. Concurrent and predictive validity refer 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.

Construct validity includes correlations of scales from similar instruments. The guidelines on construct validity coefficients need to be interpreted flexibly. Where two very similar instruments have been correlated (with data obtained concurrently) we would expect to find correlations of 0.60 or more for `adequate'. Where the instruments are

less similar, or administration sessions are separated by some time interval, lower values may be adequate. When evaluating construct validity care should be taken in interpreting very high correlations. Where correlations are above 0.90, the likelihood is that the scales in question are measuring exactly the same thing. This is not a problem if the scales in question represent a new scale and an established marker. It would be a problem though, if the scale in question were meant to be adding useful variance to what other scale already measure.

When judging overall validity, it is important to bear in mind the importance placed on construct validity as the best indicator of whether a test measures what it claims to measure. In some cases, the main evidence of this could be in the form of criterion-related studies. Such a test might have an 'adequate' or better rating for criterion-related validity and a less than adequate one for construct validity. In general, if the evidence of criterion-related validity or the evidence for construct validity is at least adequate, then, by implication, the overall rating must also be at least adequate. It should not be regarded as an average or as the lowest common denominator.

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. classification of people into categories) 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 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 the focus for rating not the reliabilities of the components.

Items to be rated n/a or 0 to 5 (half ratings are acceptable)	**Rating**
Evaluation of Technical Information - Overall Adequacy: (This overall rating is obtained by using judgment based on the ratings given for items 2.9 – 2.11) *It is best to complete this rating after you have completed sections 2.9-2.11* This section is concerned with the nature and quality of the technical information that is presented in the available documentation. It is not concerned with how clearly or how well the information is presented. *Please make sure that you base your ratings on the information readily obtainable by the general user of the assessment instrument rather than an academic or specialist*. Where other information is available about an instrument's qualities (e.g. American manuals not published in this country, articles in the psychological literature) this should be referred to in the *EVALUATION* section of the review.	

Norms or reference group information

2.9	**Overall adequacy:** (This overall rating is obtained by using judgment based on the ratings given for items 2.9.1 – 2.9.5 **do not simply average numbers to obtain an overall rating.)** Ratings can be defined, by entering number of scales that meet the following criteria, and then judging the rating from the mode of the distribution. Where an instrument is designed for use with out recourse to norms or reference groups, the 'not applicable' category should be used rather than 'no information given'	[ ]
2.9.1	Appropriateness for local use, whether local or international norms: [n/a] Not applicable [ ] No information given. [*] Not locally relevant (e.g. inappropriate foreign samples). [***] Local general population sample or non-local norms that could be used with caution. [****] Local country samples or relevant international samples with good relevance for intended application. [*****] Local country samples or relevant international samples drawn from well-defined samples from the relevant application domain.	[ ]
2.9.2	Appropriateness for intended applications: [n/a] Not applicable [ ] No information given. [*] Norm or norms not adequate for intended applications. [***] Adequate general population norms and/or range of norm tables. [****] Good range of norm tables. [*****] Excellent range of sample relevant, age-related and sex-related norms with information about other differences within groups (e.g. ethnic group mix).	[ ]
2.9.3	Sample sizes: [n/a] Not applicable [ ] No information given. [*] Inadequate samples (e.g. less than 150). [***] Adequate samples (e.g. 150-300). [****] Large samples (e.g. 300-1000). [*****] Very large samples (e.g. 1000+).	[ ]
2.9.4	Procedures used in sample selection: (select one) - No information is supplied - Representative of population [summarise criteria below] - Incidental - Random	
2.9.5	Quality of information provided about minority/protected group differences, effects of age, gender etc: [n/a] Not applicable [ ] No information given. [*] Inadequate information. [***] Adequate general information, with minimal analysis. [****] Good descriptions and analyses of groups and differences [*****] Excellent range of analyses and discussion of relevant issues relating to use and interpretation.	[ ]

2.9.6 **Reviewers' comments on the norms:** Brief report about the norms and their history, including information on provisions made by the publisher/author for updating norms on a regular basis

Validity

2.10	**Overall Adequacy:** (This overall rating is obtained by using judgment based on the ratings given for items 2.10.1 – 2.10.2.4. **Do not simply average numbers to obtain an overall rating. Usually this will be equal to either the Construct Validity or the Criterion-Related validity, whichever is the greater)**	
2.10.1	**Construct Validity - Overall Adequacy** (This overall rating is obtained by using judgment based on the ratings given for items 2.10.1.2 – 2.10.1.6. **Do not simply average numbers to obtain an overall rating.)**	[ ]
2.10.1.1	Designs used: *(tick as many as are applicable)* - No information is supplied - Correlations with other instruments and performance criteria - Intra-scale (item-rest correlations) - Differences between groups - Matrix Multitrait-Multmethod - Exploratory Factor Analysis - Confirmatory Factor Analysis - Experimental Designs - Other (indicate)	
2.10.1.2	Sample sizes: [ ] No information given. [*] One inadequate study (e.g. sample size less than 100). [***] One adequate study (e.g. sample size of 100-200). [****] More than one adequate or large sized study. [*****] Good range of adequate to large studies.	[ ]
2.10.1.3	Procedure of sample selection: *(select one)* - No information is supplied - Representative of population [summarise criteria below] - Incidental - Random	
2.10.1.4	Median and range of the correlations between the test and other similar tests: [ ] No information given. [*] Inadequate ($r < 0.55$). [***] Adequate ($0.55 < r < 0.65$). [****] Good ($0.65 < r < 0.75$). [*****] Excellent ($r > 0.75$)	[ ]
2.10.1.5	Quality of instruments as criteria or markers: [ ] No information given. [*] Inadequate information given. [***] Adequate quality [****] Good quality. [*****] Excellent quality with wide range of relevant markers for convergent and divergent validation.	[ ]
2.10.1.6	Differential Item Functioning (DIF) analyses: [N/A] Not applicable	[ ]
2.10.2	**Criterion–related Validity - Overall Adequacy** (This overall rating is obtained by using judgment based on the ratings given for items 2.11.1 – 2.10.4.2. **Do not simply average numbers to obtain an overall rating.)**	[ ]
2.10.2.1	Description of the criteria used and characteristics of the populations:.(*tick as many as are applicable)* - Concurrent - Predictive - Post-dictive	

2.10.2.2	Sample sizes: [ ] No information given. [*] One inadequate study (e.g. sample size less than 100). [***] One adequate study (e.g. sample size of 100-200). [****] One large or more than one adequate sized study. [*****] Good range of adequate to large studies.	[ ]
2.10.2.3	Procedure of Sample selection*: (select one)* - No information is supplied - Purposive or representative [summarise criteria below] - Incidental - Random	
2.10.2.4	Median and range of the correlations between the test and criteria: [ ] No information given. [*] Inadequate ($r < 0.2$). [***] Adequate ($0.2 < r < 0.35$). [****] Good ($0.35 < r < 0.50$). [*****] Excellent ($r > 0.50$)	[ ]
2.10.3 **Reviewers' comments on validity:**		

Reliability

2.11	**Overall Adequacy:** *(This overall rating is obtained by using judgment based on the ratings given for items 2.11.1 – 2.10.3.2.* ***Do not simply average numbers to obtain an overall rating.)*** For some instruments, internal consistency may be inappropriate (broad traits or scale aggregates), in which case place more emphasis on the retest data. In other cases (state measures), retest reliabilities would be misleading, so emphasis would be place on internal consistencies. In relation to reliability criteria, two main types of application are considered. Instrument that are designed for individual assessment require higher levels of reliability for practical effectiveness than those used to make decision on groups of people. In the suggested values given below, the first relates to instruments intended for making group decisions (e.g. selection sift tools) while the second set of values relates to those intended for scale-by-scale individual assessment.	[ ]
2.11.1.	Data provided about reliability: *(select one)* - Only one reliability coeffcient given - Only one estimate of standard error of measurement given - Reliability coefficients for a number of different groups - Standard error of measurement given for a number of different groups	
2.11.1	*Internal consistency:*	
2.11.1.1	Sample size: [ ] No information given. [*] One inadequate study (e.g. sample size less than 100). [***] One adequate study (e.g. sample size of 100-200). [****] One large or more than one adequate sized study. [*****] Good range of adequate to large studies.	[ ]
2.11.1.2	Median of coefficients: [ ] No information given. [*] Inadequate (e.g. r < 0.7) [***] Adequate (e.g. r =0.7 to 0.79) [****] Good (e.g. r = 0.8 to 0.89) [*****] Excellent (e.g. r>0.9)	[ ]
2.11.2	*Test retest stability:*	
2.11.2.1	Sample size: [ ] No information given. [*] One inadequate study (e.g. sample size less than 100). [***] One adequate study (e.g. sample size of 100-200). [****] One large or more than one adequate sized study. [*****] Good range of adequate to large studies.	[ ]
2.11.2.2	Median of coefficients: [ ] No information given. [*] Inadequate (e.g. r < 0.6 [***] Adequate (e.g. r = 0.6 to 0.69) [****] Good (e.g. r = 0.7 to 0.79) [*****] Excellent (e.g. r >0.8)	[ ]
2.11.3	*Equivalence reliability:*	
2.11.3.1	Sample size: [ ] No information given. [*] One inadequate study (e.g. sample size less than 100). [***] One adequate study (e.g. sample size of 100-200). [****] One large or more than one adequate sized study. [*****] Good range of adequate to large studies. [N/A] Not applicable	[ ]
2.11.3.2	Median of coefficients: [ ] No information given. [*] Inadequate (e.g. r < 0.6) [***] Adequate (e.g. r =0.6 to 0.69) [****] Good (e.g. r = 0.7 to 0.79) [*****] Excellent (e.g. r>0.8) [N/A] Not applicable	

Reviewers' comments on Reliability: - Comment on confidence intervals for reliability co-efficient - Provide Spearman-Brown equivalents

2.12 Section 8: Quality of Computer Generated Reports:

For each of the following attributes, some questions are stated that should help you make a judgment, and a definition of an 'excellent' [*****] rating is provided.

Items to be rated 0 – 5 (half ratings are acceptable)		***Rating***
2.12	**Overall adequacy of computer generated reports:** (This overall rating is obtained by using judgment based on the ratings given for items 2.12.1 – 2.12.7. **Do not simply average numbers to obtain an overall rating.)**	[ ]
2.12.1	**Scope or coverage** Reports can be seen as varying in both their breadth and their specificity. Reports may also vary in the range of people for whom they are suitable. In some cases it may be that separate tailored reports are provided for different groups of recipients. • *Does the report cover the range of attributes measured by the instrument?* • *Does it do so at a level of specificity justifiable in terms of the level of detail obtainable from the instrument scores?* • *Can the 'granularity' of the report (i.e. the number of distinct score bands on a scale that are used to map onto different text units used in the report) be justified in terms of the scales measurement errors?* • *Is the report used with the same populations of people for who the instrument was designed? (e.g. Groups for whom the norm groups are relevant, or for whom there is relevant criterion data etc).* [*****] Excellent fit between the scope of the instrument and the scope of the report, with the level of specificity in the report being matched to the level of detail measured by the scales. Good use made of all the scores reported from the instrument.	[ ]
2.12.2	**Reliability**	

	• *How consistent are reports in their interpretation of similar sets of score data?* • *If report content is varied (e.g. by random selection from equivalent text units) is this done satisfactorily?* • *Is the interpretation of scores and differences between scores justifiable in terms of the scale measurement errors?* [*****] Excellent consistency in interpretation and appropriate warnings provided for statements, interpretation and recommendations regarding their underlying errors of measurement	[ ]
2.12.3	**Relevance or validity** The linkage between the instrument and the content of the report may be either explained within the report or be separately documented. Where reports are based on clinical judgement, the process by which the expert(s) produced the content and the rules relating scores to content should be documented. • *How strong is the relationship between the content of the report and the scores on the instrument? To what degree does the report go beyond or diverge from the information provided by the instrument scores?* • *Does the report content relate clearly to the characteristics measured by the instrument?* • *Does it provide reasonable inferences about criteria to which we might expect such characteristics to be related?* • *What empirical evidence provided to show that these relationships actually exist?* It is relevant to consider both the construct validity of a report (i.e. the extent to which it provides an interpretation that is in line with the definition of the underlying constructs) and criterion-validity (i.e. where statements are made that can be linked back to empirical data). [*****] Excellent relationship between the scales and the report content, with clear justifications provided.	[ ]
2.12.4	**Fairness, or freedom from systematic bias** • *Is the content of the report and the language used likely to create impressions of inappropriateness for certain groups?* • *Does the report make clear any areas of possible bias in the results of the instrument?* • *Are alternate language forms available? If so, have adequate steps been taken to ensure their equivalence?* [*****] Excellent, clear warnings and explanations of possible bias, available in all relevant user languages	[ ]
2.12.5	**Acceptability** This will depend a lot on the complexity of the language used in the report, the complexity of the constructs being described and the purpose for which it is intended. • *Is the form and content of the report likely to be acceptable to the intended recipients?* • *Is the report written in a language that is appropriate for the likely levels of numeracy and literacy of the intended reader?* [*****] Very high acceptability, well-designed and well-suited to the intended audience	[ ]
2.12.6	**Practicality** Practicality issues also affect acceptability. The main practical advantage of computer-generated reports is that they save time for the person who would otherwise have to produce the report. When that person is not the end-user, the practicality arguments may be harder to make. • *How much time does each report save the user?* • *How much time does each report take to read and use?* [*****] Excellent in terms of efficiency and value.	[ ]
2.12.7	**Length** This is an aspect of Practicality and should be reflected in the rating given for this. More specifically this provides an index of the ratio of quantity of output to input. The number of scales on which the report content is based are regarded as the input, and the number of report pages (excluding title pages, copyright notices etc) are regarded as the output. *To calculate this index***,** count up the number of scales, including derived scales and composite scales (e.g. for personality measures, higher order factor scales, scales for team types, leadership styles etc may be derived from the base scales). 1. Divide the total number of pages by the number of scales. 2. Multiple this ratio by 10 and round the result to the nearest integer. Generally values greater than 10 are likely to indicate reports that may be over long and over-interpreted. E.g.: Development Report - 8/7 x 10 = 11.42.	

2.12.8	**Reviewers' comments on Computer Generated Reports:** The evaluation can consider additional matters such as whether the reports takes account of checks on consistency of responding, response bias measures (e.g. measures of central tendency in ratings) and other indicators of the confidence with which the person's scores can be interpreted. Comments on the complexity of the algorithms can be included. For example, whether multiple scales are considered simultaneously, how scale profiles are dealt with etc. Such complexity should, of course, be supported in the manual by a clear rationale. Judging computer-based reports is made difficult by the fact that many suppliers will, understandably, wish to protect their intellectual property in the algorithms and scoring rules. In practice, sufficient information should be available for review purposes from the technical manual describing the development of the reporting process and its rationale, and through the running of a sample of test cases of score configurations. Ideally the documentation should also describe the procedures that were used to test the report generation for accuracy, consistency and relevance.

Section 9:
Final Evaluation:

3.4	**Evaluative Report of the Test:** This section should contain a concise, clearly argued judgement about the instrument/product. It should describe its pros and cons, and give some general recommendations about how and when it might be used - together with warnings (where necessary) about when it should not be used. The evaluation should cover topics such as the appropriateness of the instrument for various assessment functions or areas of application; any special training needs or special skills required; whether training requirements are set at the right level; ease of use; the quality and quantity of information provided by the supplier and whether there is important information which is not supplied to users. Include comments on any research that is known to be under way, and the supplier's plans for future developments and refinements etc.

	Conclusions:	
4.0	**Recommendations** (*select one)* The relevant recommendation, from the list given, should be indicated. Normally this will require some comment, justification or qualification. A short statement should be added relating to the situations and ways in which the instrument might be used, and warnings about possible areas of misuse. **All the characteristics listed below should have ratings of either [n/a, [2], [4], [5] if an instrument is to be 'recommended' for general use (box 5 or 6)**. [2.9] Norms and reference groups [2.10.1] Construct validity [2.10.2] Criterion–related validity [2.11] Reliability–overall [2.12] Computer generated reports If any of these ratings are [ ] or a [*] the instrument will normally be classified under Recommendation 1, 2, 3, or 4 or it will be classified under 'Other' with a suitable explanation given.	☐ 1 Research only tool. Not for use in practice. ☐ 2 Only suitable for use by an expert user under carefully controlled conditions or in very limited areas of application ☐ 3 Suitable for supervised use in the area(s) of application defined by the distributor by any user with general competence in test use and test administration ☐ 4 Requires further development. Only suitable for use in research ☐ 5 Suitable for use in the area(s) of application defined by the distributor, by test users who meet the distributor's specific qualifications requirements ☐ 6 Suitable for unsupervised self-assessment in the area(s) of application defined by the distributor ☐ 7 Other

5.	Notes References and Bibliography

You should check the standard sources for reviews for each instrument (e.g. `Buros' and Test Critiques). You should add details of any references cited in your Evaluation and list references to any other reviews that you know of. Where relevant, you can add a brief annotation (no more than 50 words or so) to each one concerning the conclusions drawn in that review. Indicate any other sources of information that might be helpful to the user.

See 1.10.5	Constructs Measured: