# Test Grading Form

## Instructions/Orientation to the Grading Process

The purpose of this activity is to grade the quality of a test using 13 dimensions. Each dimension is given a score ranging from 1 to 7, for a range of 13 to 91 total points. The test can then be graded based on its total score. For example, a test with 82 out of 91 points received 90% of the total possible points and would be given an "A" grade. Because the psychometric aspects of a test are crucial, specifically reliability and validity, it is suggested that if these two dimensions do not receive a rating of at least a 4, then the evaluation of the test be terminated and it receive no additional consideration.

It may be that some of the dimensions do not apply for a given test evaluation and therefore could be removed from the process with adjustments made to the scoring and grading process. For example, it may be that administration time or cost are not important and could be removed from the evaluation. Or if a test measures a single dimension (e.g., depression), then a review of the factor structure could be removed. (Although in such a case it might be interesting to perform a factor analysis to explore for sub-factors, but that is beyond the scope of this process.)

The following is provided to assist in your completion of this activity:

- Operational definitions for the 13 dimensions (Erbes et. al., 2004) (pages 2-3)
- Anchor ratings for scoring each dimension (pages 4-6)
- Rating sheet for entering your scores with notes (page 7)
- Grading table for the total score (page 8)
- Probability table for evaluating the statistics reported for the psychological assessment (page 9)

If possible, it would be more accurate to include at least two or even more psychologists to grade the test, this would allow for some measure of interrater reliability.

Regarding sources of information, besides the test manual and supplemental materials provided by the test publisher, the Mental Measurements Yearbook (MMY) and peer-reviewed journal articles along with a review of the test materials themselves should be utilized.

Ultimately every psychologist needs to decide what grade they are willing to accept when selecting a test to use. Although an "A" test would be optimal, given the nature and purpose of the testing program and the lack of any alternatives, it may be acceptable to use a "B" or even "C" test. One benefit of using this grading process is that if a "B" or "C" test is used, the psychologist can address the test's short-comings and urge for the necessary caution in interpreting and applying the test data in the clinical setting.



Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.

## Criteria and Operational Definitions (based on Erbes, et al).

### Psychometric Strength of the Test:

Demonstrated psychometric strength is essential and is the foundational element to evaluating the quality of a psychological test.

1. *Reliability*. The test has demonstrated acceptable test-retest, alternate form, split-half, or Kuder-Richardson/coefficient alpha (internal consistency) reliability.
2. *Validity*. The test has demonstrated acceptable criterion-related (concurrent or predictive), internal consistency, or convergent or discriminant validity.
3. *Factor structure*. If the test has multiple factors, the independence of these factors has been adequately demonstrated.
4. *Responsiveness to treatment effects*. The test is responsive to treatment effects when they exist, to clarify the degree of change in an individual client's psychological functioning over time. This can be assessed by demonstration of statistically and/or clinically significant change and by published effect sizes in outcome studies using the test.

### Scope of the Test:

The test should be relevant to populations served by the institution's mental health providers and cover domains that are deemed to be critical to measure, such as client's symptoms, functioning, well-being, and/or outcome within that mental health setting.

5. *Efficient Assessment of Clinical Domains*. The test adequately and efficiently assesses the clinical domains that are critical to the setting and assesses a minimal number, if any, of other irrelevant domains.
6. *Clinical Setting*. The test has demonstrated usefulness in the clinical setting for which it is being considered for use (e.g., inpatient or outpatient).
7. *Similar Clients.* The test has demonstrated application with the types of clients for which it is being consider for use (e.g., diagnostic groups, demographics).

### Administration Issues:

The test should be easily administered, cost effective, and allow for use of computerized data entry and analysis.

8. *Ease of Administration*. The test is efficiently administered with minimal staff time and resources and/or it is in the form of self-report.
9. *Ease of Scoring and Interpretation*. The test is easy and straightforward to score, and the findings and results from the test are easily interpreted by psychologists (e.g., related to common diagnoses, symptoms, or syndromes) and easily understood by clients and nonprofessional audiences.
10. *Time Efficient.* The test is brief enough to be completed in less than 20 minutes.
11. *Low cost*. Cost per administration is low.



Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.

12. *Availability of Computer Support*. Data from the test can be directly entered, scanned, or keyed into databases or spreadsheets to allow for scoring the test and for the aggregation and analysis of data from the clinical population.
13. *Validity indicator*. The test has scales for determining validity of responses (e.g., response consistency, social desirability).

Reference:

Erbes, C., Polusny, M.A., Billig, J., Mylan, M., McGuire, K., Isenhart, C., & Olson, D. (2004). Developing and Applying a Systematic Process for Evaluation of Clinical Outcome Assessment Instruments. *Psychological Services, 1,* 31-39.

## Examples of Rating Anchors

### Psychometric Strength of the Test:

*Reliability (test-retest, alternate form, split-half, Kuder-Richardson/coefficient alpha [internal consistency]).*
1 _ No data on reliability are available or the reliability studies are of poor quality.
3 _ Reliability estimates on most scales or total scale consist of one or two types of reliability and/or the reliability coefficients are not statistically significant at the .01 level (see Table 1).
5 _ Reliability estimates on most scales or total scale consist of one or two types of reliability, the reliability coefficients are statistically significant at the .01 level (see Table 1) but are lower than .80, or the internal consistency measure is less than .80.
7 _ Reliability data are available from multiple, relevant (i.e. like the target population) samples and multiple types of reliability, reliability coefficients are statistically significant and are above .80 (see Table 1) or the internal consistency measure is greater than .80.

*Validity (criterion-related [concurrent and predictive], internal consistency, convergent and discriminant).*
1 _ No data on validity are available or the validity studies are of poor quality.
3 _ Validity estimates on most scales or total scale consist of one or two types of validity and/or the validity coefficients are not statistically significant at the .01 level (see Table 1).
5 _ Validity estimates on most scales or total scale consist of one or two types of validity, the validity coefficients are statistically significant at the .01 level (see Table 1) but are lower than .40.
7 _ Validity data are available from multiple, relevant (i.e. like the target population) samples and multiple types of validity, validity coefficients are statistically significant and are above .40 (see Table 1).

*Factor Structure*
1_There are no factor analytic studies.
3_There are few factor studies and/or the studies have questionable designs or techniques (e.g., inadequate sample for the number of items).
5_There are factor studies with appropriate statistical designs and techniques but from samples that are dissimilar to the client population for which the test is being considered.
7_Factor analytic studies are available that support the test's theoretical structure or that suggest an alternative empirically-based structure that are from multiple and relevant samples and use appropriate statistical designs and techniques.

*Responsiveness to treatment effects*
1_There are no studies where the test has been used to demonstrate treatment effectiveness.
3_There are case studies or quasi-experimental studies treatments where the test identifies clinically significant differences (e.g. effect sizes) in the predicted directions between control and treatment groups from multiple and relevant samples.
5_ There are well-controlled, RCT studies using evidence-based treatments where the test identifies clinically significant differences in the form of effect sizes less than .80 in the predicted direction between control and treatment groups.


Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.

7_There are well-controlled, RCT studies using evidence-based treatments where the test identifies clinically significant differences in the form of effect sizes equal to or larger than .80 in the predicted direction between control and treatment groups from multiple and relevant samples.

**Scope of the Test:**

*Efficient Assessment of Clinical Domains*
1 _ Test does not assess any of the clinical domains of interest.
3 _ Test assesses some of the clinical domains of interest, but it also assesses many domains that are of little or no interest (i.e., superfluous).
5 _ Test assesses many of the clinical domains of interest, and it assesses few domains that are of little or no interest.
7 _ Test explicitly assesses all clinical domains of interest, and it assesses few if any domains that are of little or no interest.

*Clinical Setting*
1_There are no data regarding the type of setting in which the test has been used and/or the test has not been used in a setting like that for which it is being considered.
3_The test has been used in "nonclinical" settings (e.g., educational or occupational) and/or the test has been used in a setting similar to that which it is being considered (e.g., the test is being considered for use in a partial hospitalization setting but has been used only in outpatient settings).
5_The test has been used in the setting for which it has been considered, but there are insufficient demographic and clinical (e.g., diagnostic) information to make a specific comparison.
7_ The test has been used in the setting for which it has been considered, and there are sufficient demographic and clinical (e.g., diagnostic) information about that clinical setting to make comparisons.

*Similar Clients*
1_The test has not been used on the client population for which it has been considered.
3_The test has been used on a client population like that for which it is being considered (e.g., the test is being considered for use with adolescents but has only been used with adults).
5_ The test has been used on the client population to that for which it has been considered, but there are insufficient demographic and clinical (e.g., diagnostic) information to make a specific comparison.
7_ The test has been used on the client population to that for which it has been considered, and there are sufficient demographic and clinical (e.g., diagnostic) information about that client population.

**Administration Issues:**

*Ease of Administration*
1 _ There is little or no information regarding the administration of the test.
3 _ The test has insufficient or difficult instructions/format that would require significant staff explanation for administration.
5 _ The administration and scoring instructions are somewhat confusing or incomplete and would require staff effort and training to fully understand.
7 _ Test administration is nearly self-explanatory and easily understood, the administration is simple and straightforward, and there is minimal clinician effort required.


Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.

*Ease of Scoring and Interpretation*
1_The test relies on jargon and abstract concepts that may be difficult for clinicians to interpret and for clients to understand and/or the administration and/or scoring instructions for the test are complex, confusing, and/or involve several steps (e.g., transferring scores to multiple sheets) that makes the process onerous. Or the reading level is above that which would be appropriate for the population for which it is being considered.
3_With effort many clinicians would be able to understand the test's administration process and be able to interpret the results in a way that would be understood by many clients.
5_The test materials provide information and guidance to help the clinician understand the test's administration process and be able to interpret the results in such a way that the results would be understood by many clients.
7_The administration and interpretation information is straightforward and easy to understand by clinicians who can then easily interpret the results to clients and nonprofessionals.

*Time Efficient*
1_The test takes over an hour to complete.
3_ The test takes 40 to 60 minutes to complete.
5_ The test takes 20 to 40 minutes to complete.
7_ The test takes less than 20 minutes to complete.

*Low Cost*
1_The cost of the core test items (e.g., test kit and manual), test items that need to be replenished (e.g., answer and profile/interpretation sheets), and/or computer or scoring services are potentially prohibitive.
3_One of the three costs described above is low while the other two are potentially prohibitive.
5_Two of the three costs described above is low while one is potentially prohibitive.
7_All costs described above are low and would not restrict the test's use.

*Availability of Computer Support*
1_ There is no process by which test data can be directly entered (via scanning or keying in) into databases or spreadsheets or by which test scores can be calculated.
3_Test results need to be transferred from answer sheets into a database and for test scores to be calculated, but there are no provisions for additional statistics.
5_ Test data can be directly entered (via scanning or keying in) into databases or spreadsheets and test scores calculated, but there are no provisions for additional statistics.
7_Test data can be directly entered (via scanning or keying in) into databases or spreadsheets and test scores calculated and other statistics (e.g., percentiles or standard scores) can be generated. Or the process of administering and scoring the test is so simple that computerization is not needed.

*Validity Indicator*
1_There are no validity indicators
3_There are purported validity indicators, but they are theoretical in nature and have not been subjected to empirical analysis.
5_ There are validity indicators, and they have been subjected to empirical analysis but in a clinical setting and/or with a clinical population different from that which the test is being considered.
7_ There are validity indicators, and they have been subjected to empirical analysis in clinical settings and/or with clinical populations like that which the test is being considered.


Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.

## Ratings Sheet

| Dimension | Score | Notes |
|---|---|---|
| **Psychometrics** | | |
| Reliability (If Reliability rating <4, consider not using the test). | | |
| Validity (If Validity rating <4, consider not using the test). | | |
| Factor Structure | | |
| Responsiveness to treatment effects | | |
| Sub-score; if score less than 70% of the total possible points for this section, consider not using this test. | | |
| **Scope of the Test** | | |
| Efficient Assessment of Clinical Domains | | |
| Clinical Setting | | |
| Similar Clients | | |
| **Administration** | | |
| Ease of Administration | | |
| Ease of Scoring and Interpretation | | |
| Time Efficient | | |
| Low Cost | | |
| Availability of Computer Support | | |
| Validity Indicator | | |
| Total Score: | | |
| Average Score (based on number of dimensions used) | | |



Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.

Guide to “Grading” the test:
If all 13 dimensions are used, then the total score could range between 13 and 91. Therefore, using the percent of points received from the ratings, the test could be given a grade:

| % of 91 | Score | Grade |
|---|---|---|
| 90 | 82 | A |
| 80 | 73 | B |
| 70 | 64 | C |
| 60 | 55 | D |
| 50 | 46 | F |


Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.

Table 1.

**INSTRUCTOR NOTE:** Because even a small correlation coefficient could be statistically significant but essentially clinically meaningless, we need to be cautious about the test's psychometric features, given other measures of course, and not overstate our confidence in the test's reliability and validity. For example, looking at the table, a test's correlation coefficient of .19 with a measure of concurrent validity would be statistically significant at the .05 level with an N of 80. However, that is only about 3.6% of shared variance or variance in common between the test and the criterion. The use of this table to evaluate the statistics reported for the psychological assessment shows the importance of being aware of any study's sample size and the impact it has on the interpretation of the statistic.

One-Tailed Probabilities[a]

| N | .05 | .025 | .01 | .005 | .0005 |
|---|---|---|---|---|---|
| 5 | .80 | .88 | .93 | .96 | .99 |
| 6 | .73 | .81 | .88 | .92 | .97 |
| 7 | .67 | .75 | .83 | .87 | .95 |
| 8 | .62 | .71 | .79 | .83 | .93 |
| 9 | .58 | .67 | .75 | .80 | .90 |
| 10 | .55 | .63 | .71 | .77 | .87 |
| 11 | .52 | .60 | .69 | .73 | .85 |
| 12 | .50 | .58 | .66 | .71 | .82 |
| 13 | .48 | .55 | .63 | .68 | .80 |
| 14 | .46 | .53 | .61 | .66 | .78 |
| 15 | .44 | .51 | .59 | .64 | .76 |
| 16 | .43 | .50 | .57 | .62 | .74 |
| 17 | .41 | .48 | .56 | .61 | .73 |
| 18 | .40 | .47 | .54 | .59 | .71 |
| 19 | .39 | .46 | .53 | .57 | .69 |
| 20 | .38 | .44 | .52 | .56 | .68 |
| 22 | .36 | .42 | .49 | .54 | .65 |
| 24 | .34 | .40 | .47 | .51 | .63 |
| 26 | .33 | .39 | .45 | .50 | .61 |
| 28 | .32 | .37 | .44 | .48 | .59 |
| 30 | .31 | .36 | .42 | .46 | .57 |
| 40 | .26 | .31 | .37 | .40 | .50 |
| 50 | .23 | .28 | .33 | .36 | .45 |
| 60 | .21 | .25 | .30 | .33 | .41 |
| 80 | .19 | .22 | .26 | .29 | .36 |
| 100 | .17 | .20 | .23 | .26 | .32 |
| 250 | .10 | .12 | .15 | .16 | .21 |
| 500 | .07 | .09 | .10 | .11 | .15 |
| 1000 | .05 | .06 | .07 | .08 | .10 |

[a]One-tailed means the probability of a specific plus or minus correlation or greater. For the probability of an absolute correlation or greater, double the one-tailed probability.

N = number of cases. For partial correlation holding k variables constant, use N = N* - k, where N* is the number of cases for partial correlations.



Isenhart, C. (2020). *Test grading form.* [Assignment Description]. Phoenix, AZ. University of Arizona Phoenix School of Medicine.