国際言語テスト学会　実施ガイドライン
日本言語テスト学会　訳
(パート１のみ、JLTA 執行部翻訳　2018 年 5 月１４日意訳版)

A. 質の保証されたテストを実施するための基本事項

1. テスト全体、および大問・小問が測定の対象としている言語能力ーすなわち構成概念―を明示すること。
2. テストの結果から目的や用途に応じて、妥当な能力推定をするために必要となる情報を提供すること。
    - 妥当性とは、テスト得点から対象能力を推定する際の正確さを指す。例えば、ビジネス・コミュニケーションの英語の運用能力を測定することを目的としたテストの場合、ビジネス・コミュニケーションにおいて英語を使う能力が構成概念である。
    - 対象能力―構成概念―が正しく測定できる程度により、その推定は妥当であるということができる。妥当性を検証するためには、テスト開発者は、構成概念を定義し、構成概念の要素を明示しなければならない。
    - つまりテスト得点の推定および解釈が妥当であるためには、テストが測定対象としている構成概念をできる限り正確に記載する必要がある。
3. テストの目的、用途に応じた信頼性を保証すること。
    - 信頼性とは、テスト結果の一貫性を指し、テストが実施された特定の時間、環境以外の言語使用場面にテスト結果の一般化が可能かを示す。

B. テスト設計者（test designer）とテスト作成者（test writer）の責任

1. テストの設計においては、テストの目的を決定し明示すること。
2. テスト設計者は、測定対象とする構成概念を定義し、テストとしてどう具体化するかを明示すること。
3. テストおよびテスト・タスクを細目（specifications）として詳細に記述すること。
4. テスト・タスクおよび各テスト項目は、事前テストを実施する前に編集すること。事前テスト（pre-test）を実施できない場合は、テスト実施後結果を報告する前に分析し、機能しない（malfunctioning）または不適合の（misfitting）タスクや項目は、受験者に報告するにあたって得点から除外すること。
5. 手作業で採点する必要のあるテストにおいては、採点手順のガイドラインおよび採点基準を準備すること。受験者のパフォーマンスについて信頼性の確保された採点を行うために、使用前には十分に検証をすること。
6. 採点者は採点に当たって訓練を受けた者であること。採点者間信頼性（inter-rater reliability）と採点者内信頼性（intra-rater reliability）を公表すること。
7. テスト資材（test materials）は、安全な場所に保管したうえで、すべての受験者に対して公平性が保たれ、特定の受験者が不利益を被ることがないようにすること。
8. テスト実施においては、受験者全員が公平に受験できるよう、十分に注意を払うこと。
9. 採点手順を遵守し、誤りが全くないよう一定の手順をとって得点を処理すること。

10. テスト結果の報告においては、受験者およびテスト得点を利用する担当者が容易に理解できるように記載すること。

C. 利害関係の大きい（high-stakes）テストを開発し実施する機関の義務

- 入学試験、資格認定試験、その他いわゆる利害関係の大きいテストを開発、実施する大学、学校、資格認定団体等の機関は、最新の言語テスティングの理論および実践に精通しているテスト設計者と項目作成者を活用すること。対象言語の非母語話者が作成したテスト項目については、当該言語に高度な能力をもつ第三者が検証を行うこと。
- **テスト受験者と関連する利害関係の保持者への責任**
- **テスト実施前**
  - あらゆる受験者を想定して、テストの目的、構成概念、妥当性、信頼性、採点の手順、採点基準、報告の方法等に関する情報を提供すること。
- **テスト実施時**
  - テスト実施にあたっては、どの受験者も不利にならないように十分配慮すること。
  - 受験者が皆同じ受験上の注意を受け、受験時間および補助資料等において公平性が保たれるようにすること。
  - テスト実施の資材（materials）を管理し、テスト監督者は実施手順について十分理解し、テスト運営者によってモニターされていること。
  - テスト実施の際に何らかの問題が起こった場合、対応策と合わせて速やかに公表すること。
  - スピーキングテストの実施にあたっては、受験者、対話者（interlocutor）、採点者が安定した環境で受験、実施、採点ができるよう十分な環境を整えること。
- **採点時**
  - テスト機関は、すべての受験者の答案および成果を正確に採点し、データベースに正確に入力すること。採点のプロセスを定期的にモニターして、計画した通り処理すること。
- **その他の留意事項**
  - 全受験者が皆、完全に同じテストやテストの版を受けるとは限らないこともある。その際には、実施したテストや版が比較可能であることを保証すること。
  - 複数の版を実施する場合には各版の信頼性推定値を公表すること。

D. 公に入手可能なテストの開発、実施について

1. 当該テストが対象としている受験者について、詳細に解説すること。
2. 測定対象とした構成概念が一般人に理解できるように解説すること。
3. テスト受験者および使用者が、テストおよびその結果を適正に活用できるよう、妥当性、信頼性、バイアスの可能性についてわかりやすく解説すること。
4. 結果の報告に際しては、テスト使用者が結果を検討して、受験者の能力について正しく推測し、結果を適正に使うことができるように解説すること。

5. 当該テストについて、誤解を招くような誤った説明は行わないこと。
6. 受験者向けハンドブック等で以下の諸点を周知すること。
   6.1 テストおよび測定に関する概念を、わかりやすく解説すること。
   6.2 当該テストの目的に応じた信頼性および妥当性を、わかりやすい言葉を使って解説すること。
   6.3 採点の手順を説明すること。複数の異なる版のテストを実施した場合には、各版の難易度等が等しく、一貫性が保たれていることを、確認の方法と共に説明すること。
   6.4 テスト結果の適切な解釈の仕方を説明し、結果の精度について限界があればそれを合わせて解説すること。

E. テスト結果の利用者の責任

- テスト結果を参照し何らかの意思決定を行うために、以下の諸点を確認すること。

1. 信頼性と妥当性の確保されたテストの結果を使うこと。
2. 構成概念が意思決定の目的に合致していることを確認すること。
3. テスト結果は受験者個人の測定対象能力を完璧には測定しておらず、限界があることを認識すること。
4. 測定値の標準誤差（standard error of measurement）を考慮する。
5. 意思決定を公正に正確に行って記録に残し、それを説明できるようにしておくこと。

F. その他の特記事項

- 集団規準準拠テスト（Norm-referenced assessment）
  - テスト使用者が、規準となる集団が当該の受験者の能力を測るために適切かどうかを判断できるように、規準となる母集団の特徴を記載すること。
- 目標基準準拠テスト（Criterion-referenced assessment）
  - 基準の適切さについては測定対象領域の専門家に確認を求めること。
  - 目標基準準拠テストの信頼性と妥当性を決定するのに相関（correlation）は適切な方法ではない。適切な検証方法を使うこと。
- コンピュータ適応型テスト
  - 標本サイズ（標本の大きさ、sample size）は項目応答理論（Item Response Theory）における推定値の安定性を保証するのに十分であること。
  - コンピュータ適応型テストの原理、および紙版テストとの違い等について情報を提供し、受験者および利害関係の保持者（stakeholders）に資すること。

International Language Testing Association
Guidelines for Practice
(Part 1 only)

A. Basic Considerations for good testing practice in all situations

1. The test developer's understanding of just what the test, and each sub-part of it, is supposed to measure (its construct) must be clearly stated.
2. All tests, regardless of their purpose or use, must provide information which allows valid inferences to be made. Validity refers to the accuracy of the inferences and uses that are made on the basis of the test's scores. If, for example, the test purports to be measuring the ability to use English in business communication, the inferences based on the test score are valid to the degree that the test does in fact measure that ability. However, since the ability to use English in business communication is a construct. The test developer must spell out just what that construct is or what it consists of. The test score inference or interpretation can be valid only if the test construct offers as accurate as possible a picture of the skill or ability it is supposed to measure.
3. All tests, regardless of their purpose or use, must be reliable. Reliability refers to the consistency of the test results, to what extent they are generalizable and therefore comparable across time and across settings.

B. Responsibilities of test designers and test writers

1. Test design should include a determination and explicit statement of the test's intended purpose(s).
2. A test designer must decide on the construct to be measured and state explicitly how that construct is to be operationalized.
3. The specifications of the test and the test tasks should be spelled out in detail.
4. The work of the task and item writers needs to be edited before pretesting. If pretesting is not possible, the tasks and items should be analysed after the test has been administered but before the results are reported. Malfunctioning or misfitting tasks and items should not be included in the calculation of individual test takers' reported scores.
5. Information guides on scoring (also known as grading or marking schemes) must be prepared for test tasks requiring hand scoring. These guides must be tried out to demonstrate that they permit reliable evaluation of the test takers' performance.
6. Those doing the scoring should be trained for the task and both inter and intra-rater reliability should be calculated and published.

7. Test materials should be kept in a safe place and handled in such a way that no test taker is allowed to gain an unfair advantage over other test takers.
8. Care must be taken to ensure that all test takers are treated in the same way in the administration of the test.
9. Scoring procedures must be carefully followed and score processing routines checked to make certain that no mistakes have been made.
10. Reports of the test results should be presented in such a way that they can be easily understood by test takers and other stakeholders.

C. Obligations of institutions preparing or administering high stakes examinations

- Institutions (colleges, schools, certification bodies etc) developing and administering entrance, certification or other high stakes examinations must utilize test designers and item writers who are well versed in current language testing theory and practice. Items written by non-native speakers of the language being tested must be checked by someone with a high level of competence in the language.
- **Responsibilities to test takers and related stakeholders**
- **Before the test is administered**
  - The institution should provide all potential test takers with adequate information about the purposes of the test, the construct (or constructs) the test is attempting to measure and the extent to which that has been achieved. Information should also be provided as to how the scores/grades will be allocated and how the results will be reported.
- **At the time of administration**
  - The institution shall provide facilities for the administration of the test that do not disadvantage any test taker. Test administration materials should be carefully prepared and proctors trained and supervised so that each administration of the test can be uniform, ensuring that all test takers receive the same instructions, time to do the test, and access to any permitted aids. If something occurs that calls into question the uniformity of the administration of the test, the problem should be identified and any remedial action to be taken to offset the negative impact on the affected test takers should be promptly announced.
  - In the case of speaking tests, the facilities shall be capable of proper invigilation and oversight, providing a safe and secure environment in professional surroundings for raters/interlocutors and for test takers.
- **At the time of scoring**

- The institution shall take the steps necessary to see that each test taker's test paper is scored/graded accurately and the result correctly placed in the data-base used in the assessment. There should be ongoing quality control checks to ensure that the scoring process is working as intended.

- **Other considerations**
  - If a decision is to be made on candidates who did not all take the same test or the same form of a test, care must be taken to ensure that the different measures used are in fact comparable.
  - If more than one form of the test is used, inter-form reliability estimates should be published as soon as they are available.

D. Obligations of those preparing and administering publicly available tests

They should:

1. Make a clear statement as to what groups the test is appropriate for and for which groups it is not appropriate.
2. Make a clear statement of the construct the test is designed to measure in terms a layperson can understand.
3. Publish validity and reliability estimates and bias reports for the test along with sufficient explanation to allow potential test takers and test users to decide if the test is suitable in their situation.
4. Report the results in a form that will allow test users to draw the correct inferences from them.
5. Refrain from making any false or misleading claims about the test.
6. Publish a handbook for test takers which:
   - 6.1 Explains the relevant measurement concepts so that they can be understood by non-specialists.
   - 6.2 Reports evidence of the reliability and validity of the test for the purpose for which it was designed.
   - 6.3 Describes the scoring procedure and, if multiple forms exist, the steps taken to ensure consistency of results across forms.
   - 6.4 Explains the proper interpretation of test results and any limitation on their accuracy.

E. Responsibilities of users of test results

Persons who utilize test results for decision making must:

1. Use results from a test that is sufficiently reliable and valid to allow fair decisions to be made.
2. Make certain that the test construct is relevant to the decision to be made.

3. Clearly understand the limitations of the test results on which they will base their decision.
4. Take into consideration the standard error of measurement (SEM) of the device that provides the data for their decision.
5. Be prepared to explain and provide evidence of the fairness and accuracy of their decision making process.

F. Special considerations

In norm-referenced testing:

- The characteristics of the population on which the test was normed must be reported so that test users can determine if this group is appropriate as a standard to which their test takers can be compared.

In criterion-referenced testing:

- The appropriateness of the criterion must be confirmed by experts in the area being tested.
- Since correlation is not a suitable way of determining the reliability and validity of criterion referenced tests, methods appropriate for such test data must be used.

In computer adaptive testing:

- The sample sizes must be large enough to ensure the stability of the Item Response Theory (IRT) estimates.
- Test takers and other stakeholders must be informed of the rationale of computer adaptive testing and of the difference between paper and pencil tests and computer adaptive tests.

Note: See the following URL for the entire Guidelines.
http://www.iltaonline.com/page/ITLAGuidelinesforPra