Reading passages supports the content validity of the test.

Construct validity is defended in two ways. First, the manual makes the argument that IRT analyses indicated that "none of the listening items in Form A and three of the listening items in Form B were misfitting. In the reading section, five of the Form A items and one of the Form B items were misfitting.... This high degree of fit provides evidence for the construct validity of both Form A and Form B of the APT" (p. 13). Second, the manual discusses the correlations between Listening and Reading for Form A (.72) and Form B (.53) (but oddly, it does not discuss the correlation between forms for each skill).

COMMENTARY. We noted above a number of problems with the reporting of statistic analyses. In addition, the taped material was unclear for some of the items because of static, especially for items taken from the radio. Sometimes in dialog items, the voices of the speakers are almost identical, making it hard to answer questions about what Speaker A or Speaker B meant (e.g., Items 10 and 12 on Form A). Moreover, a number of inference items appear to test knowledge of math rather than knowledge of Arabic. Some items could lead to situations in which a student does understand the Arabic but does not have time to do the necessary calculations (e.g., Item 28 on Form A).

The APT also has a number of positive features in the statistical explanations: Seven helpful examples are provided to illustrate the statistical concepts; the descriptive statistics are clearly explained along with how to use them for sound decision making; the norm group sample is adequately described; reliability is clearly explained and includes standard errors of measurement and estimate; and adequate content validity arguments are provided.

In addition, both the Listening and Reading are based on authentic and interesting materials. The Listening items (a) include a diversity of accents in Standard Arabic pronunciation throughout the Gulf region, the Levant, Egypt, and North Africa; (b) reflect the full range of situations in the taped voices so that listeners better understand the intended meanings; and (c) tap into a good mixture of cultural expressions and proverbs to test knowledge of cultural appropriateness.

SUMMARY. All in all, the APT is a challenging test of overall Arabic language proficiency. Although it has some problems in terms of the explanations of the statistical analyses and a few problems of item quality, on balance, it provides a practical external objective assessment tool that can usefully be used to evaluate Arabic language students and Arabic programs. In addition, the APT can provide a good index for students and teachers of Arabic to gauge where they are relative to the ACTFL Arabic Proficiency Guidelines. As such, it can provide curricular direction, especially for newly founded Arabic programs, and will allow for comparisons among Arabic programs nationwide.

REVIEWERS' REFERENCES


Review of the Arabic Proficiency Test by
ANTONY JOHN KUNNAN, Professor, TESOL Program, Charter College of Education, California State University—Los Angeles, Los Angeles, CA:

DESCRIPTION. The Arabic Proficiency Test (APT) consists of a total of 100 four-option multiple-choice items in two sections, Listening Comprehension and Reading Comprehension, each with 50 items.

The Listening Comprehension section is in three parts, to be completed in 50 minutes: Part one consists of 15 short, spoken Arabic utterances; each utterance is followed by a spoken question or statement about it in English. Part two consists of 20 items based on a variety of short spoken passages in Arabic including radio and TV news items, commercials, simple announcements, brief conversations, and short descriptions and narratives. Part three consists of 15 items based on a variety of passages including radio and TV commentaries, parts of lectures, poetry readings, jokes, proverbs, and interviews. Test takers have a few seconds after listening to the recorded items to choose the correct response from the options printed in English. The Arabic voices on the audiotape are native speakers of Arabic speaking Modern Standard Arabic in a natural pacing and intonation. All the listening materials are authentic although some of the passages are recorded for clarity and others are taken directly from radio broadcasts.
The Reading Comprehension section is in three parts to be completed in 60 minutes: Part one consists of 15 short common printed texts including street signs, messages, invitations, and advertisements. Part two consists of 20 passages of varying length featuring news items, announcements, biography, correspondence, narration, and description. Each statement or passage in these two sections is followed by a statement or question in English along with four choices in English. Part three consists of 15 challenging passages featuring discussion, argumentation, and supported opinion. Each passage is followed by questions and four choices in Arabic.

DEVELOPMENT. The Arabic Proficiency Test (APT) was developed in 1992 by the University of Michigan and the Center for Applied Linguistics (CAL) and is part of the CAL’s Arabic language testing program that includes the Arabic Speaking Test and the Arabic Writing Test. The APT is designed to measure the ability to understand spoken and written standard Arabic by American and other English-speaking learners of Arabic. The major focus of the test is on assessing the test takers’ ability to function in typical real-life language-use situations. The test is intended for test takers at proficiency levels from Novice High to Superior according to the ACTFL Arabic Proficiency Guidelines developed by a team of professors of Arabic in 1989. Test developers classified each item in the Listening Comprehension and Reading Comprehension sections as to the type of comprehension task required to answer the item correctly. Three broad areas were used: understanding the main idea, understanding a fact or detail, or understanding an inference.

The APT may be used for the following purposes: admission, placement, or exemption to/from an Arabic study program; application for a scholarship or an appointment; competency testing upon exit from an Arabic program; certification of Arabic language proficiency for occupational purposes; and evaluation of an Arabic instructional program. Two forms, Forms A and B, are currently available.

TECHNICAL. According to the APT test interpretation manual (herein referred to as the manual) developed by CAL, test takers receive scale scores for each section of the test in addition to raw scores based on the number of correct answers. The APT scale score more accurately reflects a test taker’s proficiency than the number of correct answers for three reasons: first, scale scores can be compared across the two forms of the test; second, scaled scores are based on the proficiency shown by the test taker in terms of the difficulty of the items the test taker answers correctly; and finally, the scale is interpreted in a normative manner based on a norming sample of 206 North American test takers studying Arabic at the college level.

In terms of the validity of APT score interpretations, the manual reports on evidence from content, concurrent, and construct validity studies. In terms of content validity, the content and functional domains can be grouped into three categories based on the ACTFL Guidelines: Intermediate Level—Survival includes simple daily tasks such as shopping for food, clothes, and incidental necessities. The Advanced Level—Tourist Resident in Arab society includes understanding social conversation on everyday topics and reading short, descriptive narrations. The highest level, the Superior Level—Foreign Professional in Arab Society, includes oral and written language on matters that are of general interest to educated native speakers in authentic situations. However, although the authenticity of the content and function domains in the APT is commendable, the complete dependence on the multiple-choice response format diminishes the overall authenticity of the APT.

In terms of concurrent validity, the manual reports a study (N = 429) in which the reported level of Arabic study was used as an external criterion. Average and median APT scores increase as amount of study increases. This provides evidence for the concurrent validity of the APT as a measure of proficiency in Modern Standard Arabic.

In terms of evidence from construct validity, the manual reports the use of the Item Response Theory (IRT) model to develop and score the APT. The fit statistics provided by the IRT analysis offer evidence of the extent to which unidimensionality exists. The infit mean square is one of the most useful statistics in a test like the APT that has a wide range of proficiency. This statistic has an expected value of 1.00 and items with +1.20 and above and .80 are considered misfitting the model. Using this criterion, in the Listening section, none of the items in Form A and three in
Form B were misfitting; in the Reading section, five items in Form A and one in Form B were misfitting. This shows that a high number of items fit the IRT model providing evidence for the construct validity of both Form A and Form B of the APT.

In terms of reliability, the IRT equivalent of the KR-20 reliability is provided for each section of each form of the APT. The reliability estimates are: Form A Listening section .77; Reading section .81; and Form B Listening section .62 and Reading section .83. These estimates suggest that the APT has high internal consistency reliability. In terms of the precision of measurement of the APT, this is estimated by the standard error of measurement (SEM). The SEM estimates (rounded to the nearest whole number) are: Form A Listening section 10, Reading section 7; and Form B Listening section 10, Reading section 8. These estimates are used in constructing confidence intervals such as that the “true score” of a test taker lies, with a 67% certainty within +1 or -1 standard error of the given score. So, if a given score is 95 on Form A of the APT reading, the true score could lie between 88 and 102 (95 minus 7 and 95 plus 7).

COMMENTARY. The value of the APT as an instrument to measure the general proficiency in listening and reading attained by English-speaking learners of Arabic is well established through authentic content and function domain tasks and reports of studies that provide evidence for the validity and reliability of the APT test score interpretation. However, a few additional measures would have strengthened the case for the APT. These are: (a) incorporating alternative response formats such as true-false, matching, and gap-filling so that task authenticity is enhanced and any bias due to the multiple-choice format is minimized; (b) conducting a content review of test items for test bias and test performance analysis for differential item/test functioning (DIF/DTF) for salient test-taking groups such as gender and academic major (for subject matter) as suggested in the fairness chapter in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999); and (c) reporting on studies with test accommodations such as extended time for test takers with disabilities as required of testing agencies by the Americans with Disabilities Act of 1990.

SUMMARY. The APT is a well-designed test for its purpose and the reported validity and reliability studies should give test users confidence regarding the usefulness of this instrument. The APT test interpretation manual is an additional tool that test users can consult prior to purchasing and administering the test. However, there are two important concerns to which the test developers of the APT should pay attention: the complete dependence on the multiple-choice response format both in terms of lack of authenticity as well as any bias created by it and the lack of studies on test fairness specifically in terms of test bias and DIF/DTF.

REVIEWER’S REFERENCE