

Options for Adaptivity in Computer-Assisted Language Learning and Assessment

Robert Mislevy
University of Maryland

Carol A. Chapelle
Yoo-Ree Chung
Jing Xu
Iowa State University

Levy et al. (2006) framed the issue of adaptivity by describing the many ways that adaptivity can be constructed in assessments. They proposed a taxonomy which categorizes assessments by three dimensions with potential adaptive power—claim status, observations status, and the controlling parties of claims and observations. This paper interprets these dimensions in the taxonomy in the domain of language assessment by providing language tests in the current market as specific examples. By introducing a richer concept of adaptivity, it sheds light on the development of a new generation of language assessments with the help of technology.

Most language teachers and researchers have an idea of what adaptivity means because they are acquainted with a computer-adaptive language test. What we will call a traditional computer-adaptive test determines examinees' level of ability in reading comprehension, listening comprehension or general language proficiency, for example. In such a test, examinees are presented a sufficient number of items, one at a time, for the test to make a reliable estimate of their ability with respect to the construct that the test is intended to measure. After the examinee responds to the first item on the test, the program selects subsequent items based on their responses. In general, when examinees respond correctly, the test gives them a more difficult item. When they respond incorrectly, the program selects an easier item for the next one. The traditional computer-adaptive language test is efficient at arriving at an ability estimate for examinees on a construct because each examinee spends time responding to only those items that are of an appropriate level of difficulty. Such tests are particularly welcome as placement tests when a single score is needed quickly for placement into a course and as part of a proficiency battery in which limited time is available for obtaining a score for each part of the test.

The traditional computer-adaptive language test has served well over the past decades, but as Levy, Behrens and Mislevy (2006) point out, computer technology provides a wide

range of opportunities for educators wishing to develop instruction and assessment that can help learners in a variety of ways. New options, such as adaptivity, cannot be fully understood or explored using the concepts and language of past testing practices. It is limiting to equate a powerful concept such as adaptivity with the traditional computer adaptive test, which assesses a single ability through computer selection of test items. Levy et al. explain that in order to begin to understand the new options presented by computer technology, new language and concepts are needed for conceptualizing the process of assessment. Rather than terms such as “construct,” “item,” and “score,” for example, a richer vocabulary is needed to allow test developers to design assessments that take advantage of the capabilities of technology. Levy et al. have introduced such terms, which we use in this paper to demonstrate a range of options in computer-assisted language assessment.

REFRAMING COMPUTER-ADAPTIVE TESTING

Informally, the key questions that bear on adaptivity in any assessment are “What claims are to be made about students’ knowledge or skills?” “What is the evidence that will be gathered to support these claims?” “Do either the targeted claims or the kinds of evidence change over the course of the assessment?” “If they do, who gets to decide how they change?” Levy et al. formalize these notions in terms of a space for a complex set of assessment options by introducing three characteristics of testing that can vary in ways that create different types of adaptivity and therefore are suited to different uses. First, they use the expression “observation status,” the selection and presentation of items, which can be either fixed or adaptive. The common understanding of “an adaptive test” vs. “a linear test” reflects two different options for the observation status. Secondly, they point out that it is not only the items (or observations) that can be presented adaptively; the construct that the test measures can also adapt according to examinees’ performance. In Levy et al.’s terms, what the test measures (the construct or multiple constructs), is referred to in terms of a claim that is to be made about the examinee. They indicate that “claim status” can be fixed as it is in the traditional computer adaptive test (CAT) but it can also be adaptive. In a measurement model, the variables for observations and the variables for students that ground claims are called the frame of discernment (Shafer, 1976), and in any kind of adaptive test the frame of discernment evolves in response to students’ performance. The third dimension of adaptivity is the “locus of control,” which refers to who makes the decisions about how this happens, with regard to both observations and claims. The locus of control can be with the examiner as it is in the traditional CAT in which the adaptive routine for selection of the sequence of the items as well as the claim to be made about the examinee is controlled by the examiner.

The hope for computer-assisted assessment is that the power of the computer might be applied to the need for an expanded set of test uses, but for this ideal to become reality, at the conceptual level, test developers need to be able to see ways of moving beyond the traditional CAT that is useful for placement and proficiency testing. This paper aims to work toward this goal by illustrating some of the options for adaptivity of existing

language tests thereby expanding the potential for designing future language tests to suit their specific purposes. The actual and hypothetical tests that we discuss are displayed in Table 1, which shows the position of each in a three dimensional space delineated by observation status across the top, claim status along the vertical, and locus of control on the third dimension. The third dimension is shown in this two-dimensional figure by dividing each of the positions of the vertical and horizontal. We will discuss each of these tests in turn, describing the ways that their observation, and claim status each make them either fixed or adaptive, and discussing the differences in examiner vs. examinee control for each.

CLAIM STATUS

The claim for a test refers to the statement that is to be made about the examinee on the basis of observations of his or her performance on the test. A claim might be that an examinee has strong reading comprehension ability or that the examinee is able to write an effective essay using the conventions of standard written English. In each of these.

Table 1. Examples of language assessments with a variety of types of adaptivity

Observation Status		Fixed		Adaptive	
		Examiner Controlled	Examinee Controlled	Examiner Controlled	Examinee Controlled
Fixed	Examiner Controlled	(1) CET-4 WT; RCAA (Jamieson et al., this volume)	(2) Transparent Language Test	(3) ACT EPT	(4) Hypothetical grammar test with item-level feedback
	Examinee Controlled	(5) ←	(6) nonsensical	(7) →	(8) →
Adaptive	Examiner Controlled	(9) SOPI	(10) nonsensical	(11) OPI	(12) Hypothetical CGT based on CCT (Zhang, this volume)
	Examinee Controlled	(13) FETN; S-TOPIK	(14) nonsensical	(15) DIALANG	(16) Hypothetical online language test site

cases, the claim refers to one ability or multiple abilities of the examinee, but it would not be difficult to imagine a test in which different claims about language ability are made about examinees depending on their performance. For example, what if an examinee performs poorly on the part of the test requiring recognition of correct grammatical forms and is therefore not required to spend time completing the essay. In this case, the claim for some examinees would be limited to grammar, whereas for others, the claim would be about writing ability including a claim about grammar knowledge. In other words, adaptive claims are possible depending on the examinee's performance. The tests in Cells 1-4 in Table 1 contrast with the rest of the examples in that the former are intended for making a fixed claim or claims about the examinees, like the traditional CAT does, whereas the latter may result in adaptive claims.

Tests with Fixed Claims

Table 1 shows examples of tests with fixed claims in Cells 1 through 4. In Cell 1, the College English Test Band 4 Written Test (CET-4 WT)ⁱ is a high-stakes test intended for certifying college students' general English proficiency in China. The test makes a single claim that is fixed based on the examiner's choice of intended test inferences. The test consists of six sections—Writing, Skimming and Scanning, Listening, Cloze, Reading in depth, and Translation—and they are intended for assessing four language aspects selected by the examiner: writing, reading (skimming and scanning), listening, and integrated language ability which is comprised of intensive reading, Chinese-English translation, and vocabulary and structure. Based on an examinee's performance in these four language aspects, a total scaled score and a percentile rank—as compared to other examinees—are reported as the indicator of his/her general English proficiency. The observables of the CET-4 WT are also fixed and controlled by the examiner. As a paper-and-pencil test, the CET-4 WT presents the same test items in a predetermined order to all examinees. The task types used in the test include essay writing, true and false, cloze, multiple-choice, and fill-in-the-blank. This test contains no adaptive elements so each examinee is given the same amount of time to complete each section.

Also in Cell 1, the Readiness Check and Achievement Assessment (RCAA) developed by Jamieson et al. (this volume) are low-stakes tests which have fixed claims as well as fixed sets of observations controlled by the examiner. The RCAA tests are intended to help teachers and learners at university-level intensive English programs to understand areas of language knowledge that learners need to work on. The Readiness Check test and the Achievement test. The two tests are used for different pedagogical purposes. The first is used to check students' readiness for in-class instruction and activities while the second is to assess students' learning outcomes after a class. The constructs of these two tests are fixed and are determined by the examiners based on their analysis of what students have studied, what they are going to study in their language classes, and the difficult language aspects that previous students reported. The observations of the tests are also fixed and selected by the examiners as students enrolled in the language learning program are always presented the same test items. However, based on the test results, students are provided with individualized remedial materials. Although RCAA and CET4-WT fall in

the same cell, they are different in terms of the number of claims the test can make. In CET-4 WT, only a single claim is made about examinees' general language proficiency. By contrast, RCAA makes multiple claims in both vocabulary and grammar knowledge (Jamieson et al., this volume). From this example, we can see tests with fixed claims can have more than one claim and that tests in a single cell can be for high stakes or low stakes.

Cell 2 contains a low-stakes counterpart to the CET-4. Delivered on the Web, the Transparent Language English Proficiency Test (TLEPT)ⁱⁱ for native Spanish speakers allows examinees to choose the area of their general English proficiency to be tested. In this sense the observations are examinee-controlled. Once that initial choice is made, both the claims and observations are fixed and examiner controlled. The test consisting of four sections is intended to make claims about three language aspects: grammar knowledge, vocabulary knowledge, and reading ability. Two grammar sections of the test assess an examinee's ability to manipulate sentence elements (verbs, adjectives, prepositions, conventions, modifiers, and function words) as well as to recognize erroneous sentence elements. A vocabulary section, on the other hand, assesses an examinee's ability to select and use appropriate words in a given context. Finally, the reading section assesses an examinee's referring, inferring, and summarizing ability in reading. The score of each section is reported individually in terms of the percentage of items answered correctly. Though the test items of the four sections are predetermined by the examiner, the examinee has the freedom to choose the way to proceed through the test and to select the order in which parts of the test and individual multiple choice items are completed. The examinee is also free to spend as much or as little time on the test as he or she wishes. Even though this test is not adaptive in terms of claims or observations, it provides for some elements of examinee choice in how the test is completed.

As noted in the introduction, the most familiar adaptive tests lie in Cell 3. The kind of adaptivity that characterizes this cell is evident in the ACT ESL Placement Test (ACT-EPT).ⁱⁱⁱ The ACT-EPT is a medium-stakes test intended for placing postsecondary students into appropriate ESL courses in the United States. It has fixed, examiner controlled claims and adaptive, examiner controlled observations. The test consists of three modules and each is intended to make a claim about one aspect of language ability selected by the examiner: grammar/usage, reading, and listening. The grammar/usage module assesses an examinee's ability to recognize and manipulate sentence elements (verbs, subjects and objects, modifiers, function words, conventions, and word formation), and sentence structure and syntax. The reading module assesses an examinee's referring and reasoning ability in reading and the listening module an examinee's ability to understand explicitly and implicitly stated information in speech. The score on each module is reported in terms of five levels ranging from near-beginner to near-native speaker. Detailed proficiency descriptors are also provided to define the things that a typical student at each proficiency level can do. The observations on the ACT-EPT are adaptive based on each learner's language ability and are controlled by the adaptive procedures in the test which were designed by the examiner. Like the traditional CAT,

the ACT-EPT presents multiple-choice test items to an examinee based on his or her previous responses and thus routes the examinee to the appropriate levels of test items until a sufficient number of items has been given at the appropriate level.

An examinee-controlled counterpart to the traditional CAT is adaptive with examinee controlled observations. In other words, it is the examinee who selects which items to complete on the test in order to obtain a score. Although we could not find an existing language test as an example for Cell 4, we can imagine a hypothetical Cell 4 grammar test, which might be useful for instruction and learning. Such a grammar test would provide test takers with feedback on their performance on each item as illustrated by the program developed by Choo and Kim (this volume). In addition to providing feedback for test takers, let us say this grammar test allows test takers to use the feedback they receive on each item to help them in selecting the difficulty level of the following item. When the test is finished, a total test score is computed on the basis of the difficulty level of selected test items that were answered correctly. The claim of such a grammar test is fixed by the examiner because it is intended to make a single claim about examinees' grammar ability, producing a single test score. At the same time, such a test inference is chosen by the examiner while the observations are adaptively determined by the examinee. Accordingly, although the test item pool is pre-determined by the examiner, individual examinees may encounter different items during test taking depending on their own observations and judgments. In this scenario, the hypothetical grammar test is a low-stakes assessment, which may be useful for instruction.

All of these tests with fixed claims have claims that are defined by the examiner, who developed the test to measure a construct or constructs such as reading comprehension. But as we saw, a single claim about reading comprehension or vocabulary knowledge, for example, can be arrived at through a fixed set of observations that is invariant across examinees regardless of their performance (Cells 1 and 2), or it can be made on the basis of observations selected adaptively (Cells 3 and 4). Even when claims and observations are fixed, some element of learner control can come into play in settings where examinees have choices about what to be tested on and the order they wish to complete the items (Cell 2). Adaptivity can be controlled by a set algorithm designed by the examiner to select items on the basis of prior performance by examinees (Cell 3), or it can be controlled by the examinees themselves (Cell 4). In the examples from Cells 1 through 4, we saw the most traditional linear (Cell 1) and adaptive (Cell 3) tests, but we also saw tests that expand the test use into instruction and learning by providing choice and immediate feedback about performance to the learner.

Tests with Adaptive Claims

In many tests, examinees' performance is interpreted to make claims about different constructs at different levels. For example, beginners may demonstrate performance that provides a basis for claims about vocabulary and pronunciation alone, whereas at an advanced level, performance would allow for claims about these aspects of language in addition to rhetorical knowledge. Tests yielding claims about more than a single aspect of

language have the potential for adjusting the constructs tested during the testing process. In a language test, the student model variables concern the aspects of language ability to be assessed while the observable variables concern learner behaviors which provide evidence for levels of proficiency in these language aspects. In the example given above, the frame of discernment reflects choices of the test constructs among vocabulary knowledge, pronunciation, and rhetorical knowledge and such choices are made based on learner performance (observations). Now both the student model variables and the observable variables can play a role in adjusting claims and yielding interpretations of learner performance in an intertwined manner as an adaptive-claim language assessment proceeds. At this point, it may be appropriate to draw clear relationships between the number of claims and the adaptivity of claim status before moving on to discuss examples of multiple-claim assessments.

It is witnessed earlier that the relationship between the dimensionality (or multiplicity) of claims and the adaptivity of claim status is not of a one-to-one relationship. Fixed claims may comprise one or more claims, but this is not true for adaptive claims. In this regard, Levy et al. (2006) point out three general properties of assessment as follows:

- (a) Adaptivity of claims in assessments entails multiple claims;
- (b) Univariate-claim (i.e., single-claim) assessments are inevitably fixed; and
- (c) Multiple-claim assessments can be either fixed or adaptive (p. 6).

These relationships between multiplicity and adaptivity of claims in assessments are shown in Figure 1. Note that single-claim assessments are always fixed, but not vice versa. As a single-claim language assessment focuses on only one aspect of language ability, adaptivity in claim status, which is by and large constructed through multiple observations, does not play a role in such univariate assessments. On the other hand, multiple-claim assessments are either fixed or adaptive. As seen below, a fixed multiple-claim assessment can involve either fixed or adaptive observations but always addresses the same set of claims, adaptive multiple-claim assessment may involve multiple observations as ‘the hypotheses of interest that are investigated may change as new information (from observation) is brought to bear’ (Levy et al., 2006, p. 5).

Examiner-controlled adaptive claims

Cells 9 through 16 of Table 1 provide examples of tests with adaptive claims. From these examples, we see the choices of the test constructs can be either made by the examiner or the examinees. Three examples illustrate tests whose adaptive claims are controlled by the examiner. In Cell 9, the Simulated Oral Proficiency Interview (SOPI),^{iv} a speaking proficiency test that yields a single score which can be used for a variety of purposes, has adaptive, examiner-controlled claims and fixed, examiner-controlled observations. The SOPI consists of four parts: warm-up, level checks, probes, and wind-down. After responding to warm-up questions, an examinee’s proficiency level is evaluated via tasks designed for level-checking and observation. Trained raters make claims about the examinee’s oral proficiency by listening to his or her recorded responses to given

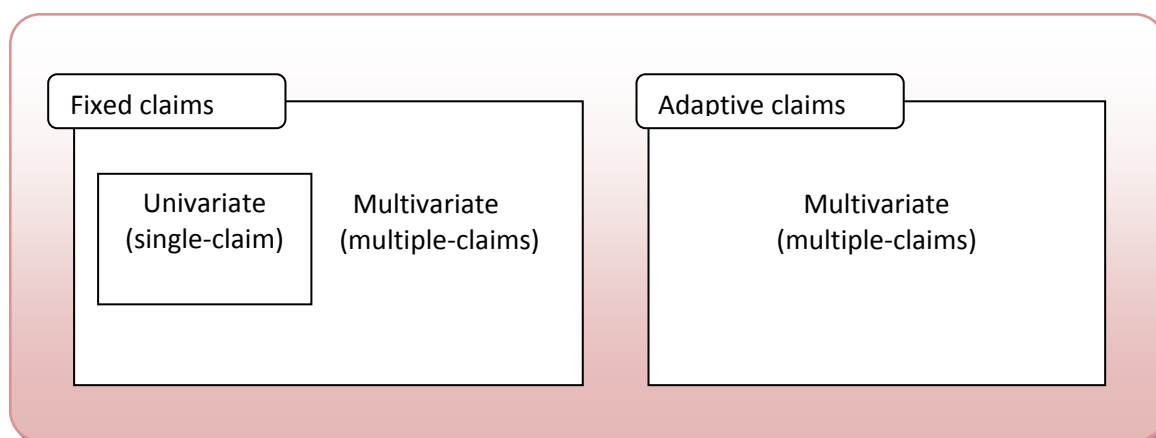


Figure 1. Relationships between the dimensionality of claims and the adaptivity of claim status

prompts in regard to different language functions described in the ACTFL Guidelines for speaking^v, which classifies proficiency levels into Novice, Intermediate, Advanced, and Superior. Scores for the SOPI range from Novice-mid to Superior and the claims associated with each of these levels differ. For example, the Novice level makes claims about examinees' vocabulary, oral fluency and the complexity of speech while the Superior level makes claims about the accuracy, pragmatic competence, and interactive strategies. Whereas the claims vary by level, the observables of the SOPI are fixed and controlled by the examiner. Developed as a semi-direct equivalent test to the face-to-face Oral Proficiency Interview (OPI), the SOPI delivers test items from a recorded tape and a test booklet. Since all test items are identical across the test takers, the SOPI can be administered to a group of examinees in a language lab setting simultaneously.

In Cell 11, the ACTFL Oral Proficiency Interview (OPI)^{vi} is a face-to-face or telephone-mediated oral test. Like the SOPI it has adaptive, examiner-controlled claims, but because it is given as a one-on-one interview, the observations can be chosen adaptively by the examiner, as well. Claims about an examinee's oral proficiency are made adaptively during the test administration by a trained interviewer, who also rates the examinee's performance with another trained rater. The interviewer starts with items targeting a certain proficiency level and adjusts the difficulty level of test items as the test goes on, by changing topics and language functions, on the basis of the interviewer's assessment of the examinee's performance on test items. Test items may be presented in the form of natural communication or role-plays. If the examinee does not feel comfortable about a certain topic, he or she can request a topic change. The interviewer controls the test-taking time on the basis of his or her perception about the examinee's proficiency level. Like the Simulated Oral Proficiency Interview in Cell 9, the examinee's performance may be rated using either the ACTFL proficiency guidelines or the 11-point ILR scale^{vii}, ranging from Novice (0+) to Superior (3 and above).

For Cell 12, one can imagine a grammar test with adaptive claims determined by the

examiner and observations selected by the examinee. Such a hypothetical cognitive grammar test (CGT) is an extension of the computerized cognitive test (CCT) illustrated by Zhang (this volume) in which examinees select cues that help them in responding to test questions. In the hypothetical CGT, the examiner may adjust the inferential targets—though still with a primary focus on grammar knowledge—depending on the cues examinees choose to solve the language problems. For example, if an examinee does much better in jumbled-word test items when they get help from metalinguistic cues—such cues identify the stem sentence (subject, verb, and object) in a complicated sentence structure – than when they skip such help options, the examiner may decide to make claims about the examinee’s metalinguistic competence in addition to his or her grammar knowledge. Such claims may be that the examinee displays high competency of metalinguistic knowledge but low grammar knowledge. Likewise, if the examinee has to rely on cues of key word definitions to assemble jumbled words into meaningful sentences, the examiner will then include vocabulary knowledge into the test claims. In this case, the test claims may be that the examinee shows high level of grammar knowledge but low level of vocabulary knowledge. The observations of the CGT are also adaptive but are subject to examinees’ control. While the cues are provided by the examiner, examinees may choose either types of cues or decide not to use any cues during test taking.

Examinee-controlled claims

The adaptive claims of a language test might also be selected by examinees, who choose the claims that they wish to be able to make about their language ability. Examples of such tests appear in Cells 13, 15, and 16. In Cell 13, the Free-English-Test.Net (FETN)^{viii} is a low-stakes English test website for English as second language (ESL) learners to assess their English proficiency in various specific aspects of language ability at three different difficulty levels. It has adaptive, examinee controlled claims and fixed, examiner controlled observations. The claims made by FETN are adaptive based on the examinee’s choices of five major sections, including three levels of grammar, synonyms, business English, usage, and idiomatic expressions. Under each section is a large number of sub-tests, each assessing one specific language aspect and each categorized into one of the three difficulty levels: elementary, intermediate, and advanced. Upon entrance to the test website, an examinee has the freedom to choose any sub-test under a certain section and at a certain difficulty level to receive claims about a specific language area and level. The FETN test website in its current form is limited to item-level feedback. However, if such a test could provide summative evaluation for examinees on each specific aspect of language, the score would better serve as an overall claim. As an internet-based test, FETN presents the same test items in each sub-test to all examinees and consistently uses the multiple-choice question format.

A high stakes example of a test in Cell 13, the Standard Test of Proficiency in Korean (S-TOPIK)^{ix} assesses Korean as a foreign language learners’ general Korean proficiency primarily for admission and hiring purposes. It has adaptive, examinee controlled claims and fixed, examiner controlled observations. The S-TOPIK makes claims that are

adaptive based on examinees' choices among three proficiency levels, i.e., beginner, intermediate, and advanced, at the beginning of the test. Scores are then contingent upon examinees' initial choices of proficiency levels. Each level of the S-TOPIK consists of four sections: vocabulary and grammar, writing, listening, and reading. The score of each level is reported in terms of a standardized total score, which corresponds to a lower or upper band of the level. The Korean proficiency of an examinee thus can be interpreted by looking up the descriptions of the band in which his or her score falls. Once one of the three levels has been chosen by the examinee, the examiner controls a fixed procedure of obtaining observations. The test presents the same task items in a predetermined order to examinees who select the same level of the test. Test items in all sections except writing are given in a multiple-choice format, each presenting 30 questions. Different limits regarding the length of the composition are posed to examinees based on their target proficiency level. An examinee is required to complete the writing section in an hour and each of the other sections in 30 minutes. Thus, the entire test lasts for three hours.

Compared with S-TOPIK, DIALANG's online diagnostic language testing system (DIALANG)^x in Cell 15 has the same claim status but different observation status. This low-stakes test intended for informing language learners about their proficiency levels as well as providing tips for language learning has adaptive, examinee controlled claims and adaptive, examiner controlled observations. Similar to S-TOPIK, DIALANG makes claims that are adaptive based on examinees' choices of languages and language aspects to be assessed. The online testing system offers assessments of fourteen languages in five language aspects, including listening, writing, reading, structures, and vocabulary but the examinees make decisions on what will be assessed when entering the test. DIALANG reports examinees' ability in each language aspect in six levels ranging from beginner (A1) to very advanced (C2) based on the Common European Framework. In addition, the test provides detailed score descriptions and suggestions for each level of learners. In contrast to S-TOPIK, the examiner controls the observations which are adaptive to examinee's responses during the assessment. Based on examinees' performance in a placement in which they are asked to distinguish between real words and pseudo-words and their responses to an optional self-report of language ability, the examiner routes the examinees to the appropriate levels of test items. Such routing activity always continues in the testing process. Thus, examinees of different levels of language proficiency encounter different test items. The task types used in the test include multiple-choice, gap filling, sentence completion, sentence insertion, error recognition, and word formation. The DIALANG test is not timed, and examinees are allowed to spend as much time as they want to on the test.

An example for Cell 16 would be a hypothetical online site^{xi} a language test system whose claims and observations are adaptive and controlled by the examinees. Imagine a website which has a variety of language tests in its database. A single test is categorized in regard to the target language abilities and difficulty level. To take a test, examinees visiting the website would type in the search box the name of a language skill that they want to be assessed (say, vocabulary). The search engine would retrieve all the tests that

make claims about vocabulary ability from its database and list them on the screen. The retrieved tests may be sorted by the difficulty level, sub-constructs (such as parts of speech, collocations, and idioms), or topics (such as hospital, cooking, school, travel, culture, etc.). Examinees then read through the list of vocabulary tests and select what they are interested in. They may choose to take more than one vocabulary test in various orders. Examinees may also quit the test in mid course and switch to another test in the list. In this scenario, claims are controlled by examinees in an adaptive manner perhaps informed by feedback they receive. As there is no restriction in selecting the target construct in addition to difficulty levels, examinees may enjoy freedom in searching for the language tests they wish to take on the website.

OBSERVATION STATUS

In the traditional CAT, the observation status is the dimension that is adaptive, and this dimension is controlled by the examiner through the selection of items based on an algorithm that considers the examinee's prior performance. We saw from the examples in cells beyond Cell 4 that adaptivity does not have to refer to adaptive observations, but can also refer to adaptive claims. The SOPI in Cell 9 and the FETN and S-TOPIK in Cell 13, for example, make adaptive claims but have fixed observations. The claims made by these three tests about examinees are adaptive and controlled by either the examiner's or examinees' choices of specific language abilities to be assessed. The adaptive claims made by SOPI are subject to the examiner's decision. Through a level-checking process of the test, the examiner places examinees into appropriate proficiency levels, each of which makes claims about different aspects of language ability. In contrast, the adaptivity of claims in FETN and S-TOPIK are controlled by examinees. In these two tests, examinees are entitled to decide the proficiency level or aspects of language ability to be assessed and, accordingly, the examiner makes corresponding claims. With these adaptive claims, the three tests have fixed observables which are controlled by the examiner. Once the scope of claims (or a "claim space" in Levy et al.'s terms) is determined, the examiner then presents a set of prearranged test items to examinees regardless of their test performance.

Although tests with adaptive claims and fixed observations exist for language assessment, Levy et al. point out that fixed observations are in many cases insufficient for assessment with adaptive claims because the predetermined fixed observables maybe optimal for assessing one part of the claim space yet inadequate for the other parts. In other words, an assessment having fixed observations may have limited flexibility to adjust its focus in the claim space. Such a drawback can easily be detected in the SOPI in Cell 9. Though the test has adaptive claims, its adjustment of focus in the claim space can take place only once and only at the beginning of the assessment before many observations have been gathered. Such a limiting adaptivity of test claims is largely attributed to the fixed test format which prohibits the assessment from calling for optimal observables to make specific claims and thus restrains the assessment from moving around the claim space freely in the testing process.

In contrast, language assessments having adaptive claims as well as an adaptive observation status can adjust the claims multiple times while flexible observables about the examinees are collected. Taking the hypothetical extension of CCT in Cell 12 as an example, the test is able to shift its focus in the claim space onto any of the three aspects of the examinee's ability, including grammar knowledge, metalinguistic competence, and vocabulary knowledge, based on the examinee's responses and uses of cues. Thus, the adaptivity of claims in language assessment can typically be better operationalized when the observables are adaptive as well.

LOCUS OF CONTROL

The locus of control for adaptivity in the traditional computer-adaptive language test is the examiner, who sets the algorithm for item selection. Even though the algorithm includes information about the examinee's performance, the examinee does not have any explicit choice in the selection of observations. However, from Table 1, we saw that the locus of control is a third dimension that intersects not only the observation status but the claim status as well. Thus, the locus of control with regard to claims is another variable that distinguishes language assessment. For example, both FETN in Cell 13 and SOPI in Cell 9 have adaptive claims and fixed, examiner-controlled observations. The difference between the two tests lies in that the former allows examinees to select the language aspects to be assessed—examinee-controlled claims—but the latter reserves such a job for the examiner—examiner-controlled claims. Similarly, the choice of observations to be gathered during the assessment can be made either by the examiner or examinees. For example, although a traditional computer-adaptive test such as the ACT EPT in Cell 3 and the hypothetical grammar test in Cell 4 both have fixed, examiner-controlled claims and adaptive observations, the former empowers the computer (examiner) to select test items for examinees based on their performance—examiner-controlled observations—while the latter endows the examinees the freedom to decide the difficulty level of upcoming test items based on the feedback generated by the examiner—examinee-controlled observations.

Although the dimensions of claims, observations, and locus of control can theoretically intersect with each other, some combinations of these three dimensions have not been explored (or are not applicable) yet in the context of language assessment. For example, we did not find any existing language test for the Cells 5-8 which have fixed, examinee-controlled claims. According to Levy et al., such types of assessments are nonsensical because examinees do not have any control over the claims when the claim space (intended construct) is already fixed. Further research or further reflection may reveal assessments that do fit into these cells nevertheless. For the same reason, we did not find any examples for Cells 10 and 14 in which observations are fixed yet controlled by examinees. However, the Transparent Language Test in Cell 2 is an exception. In such a test, although all examinees encounter the same form of test items selected by the examiner, they are not required to follow the given testing procedure or complete all test items. Thus, different examinees may provide different observations (complete different

number of items and in a different order) to the assessment. In this case, the examiner and examinees share the right to make choices of observations. We categorized this test as a type with fixed, examinee-controlled observations because examinees have certain freedom on choosing observables. From this example, we see that cooperation between examiner and examinees in selecting test items is also possible, particularly in this low stakes test.

Table 1 reveals that the locus of control is a dimension that can be related to the stakes of language assessment. Specifically, tests under the examiner-controlled categories are of higher-stakes than those under examinee-controlled categories. Such distinctions can be easily detected by comparing pairs of language tests in two neighboring cells, such as the CET-4 vs. Transparent Language Test, SOPI vs. FETN, and OPI vs. DIALANG. The two tests in these pairs only differ in the locus of control for one dimension but they have very different stakes. In these cells, the tests more controlled by the examiner, such as the CET-4, SOPI, and OPI are widely used high-stakes test while those controlled by the examinees are low-stakes free online assessments.

Another such pair of examples is the TOEFL Internet-based Speaking test and its counterpart, the Online Practice Speaking Test (See Xi in this volume). The two tests, although having different test constructs and purposes, share the same test format. While doing the practice test to prepare for the real test, examinees may choose an untimed testing mode in which they are allowed to read and listen to testing prompts multiple times, prepare for their speech for as long as they want to, or quit and continue the test at any time. These options of examinee control are not available in the real test. Both tests have fixed observations as they present the same test items across examinees. However, the observations are controlled by the examiner in the real test and by the examinees in the practice test. Compared with those taking the real test, the examinees doing the practice test enjoy the freedom of proceeding through the test at their own pace. From the examples above, we saw that language assessments having examinee-controlled claims or observations are often used for self-assessment or practice purposes. Thus, one of the future directions for computer-adaptive language assessment will be to develop examinee-controlled tests to prepare learners for high-stakes test purposes. The Hypothetical Online Language Site in Cell 16 is a model for this type of test. In such a test site, examinees are not only permitted to choose the construct to be measured but select test items (observables) based on interest as well.

However, if the responsibility of deciding the constructs of a language test is completely put in examinees' hands, the examinees accustomed to examiner-controlled tests may not feel empowered but bewildered instead because of their lack of knowledge about their own language ability. Suppose that a learner of Chinese hopes to see claims about his Chinese proficiency. Given the privilege to select the language skills to be assessed—possibly including the abilities to spell *pinyin* for Chinese characters, to recognize the meaning of Chinese characters, to pronounce tones correctly for given words, to order the strokes for writing a Chinese character, etc.—he will probably feel at a loss. Thus, in many cases, it might be a good idea for the examiner and examinees to share the job of

selecting the language abilities to be assessed. For example, examinees may receive feedback and suggestions from the examiner about what should be tested based on their performance and then make decisions accordingly. In other words, the examiner will guide the examinees through the test yet the examinees will still have the control over the target inferences of the test.

Tests with adaptive, examinee-controlled claims will very likely need the examiner to provide examinees with individualized feedback based on their performance. For the computer to offer such feedback, it has to be equipped with the ability to diagnose responses, such as examinees' constructed responses (see further discussion in Cotos and Pendar, this volume). In addition, the computer examiner must rely on student models to make decisions or provide suggestions on the language aspects to be assessed. Such models define the important variables related to the language ability of the examiner or the examinees' interest (Mislevy, Steinburg, Almond, & Lucas, 2006). In addition, student models are the key to providing useful feedback to examinees as they are informed by the analysis of learner text (see further discussion in Schulze, this volume).

CONCLUSION

The three dimensions of claim status, observation status, and locus of control elaborate the meaning of adaptivity beyond the one dimensional concept of a test capable of making a fixed examiner-controlled claim or claims based on an examiner-controlled set of observations. The three dimensions provide a space to recognize the adaptivity inherent in existing assessments that are used across purposes. In the examples, we saw not only the traditional linear and computer-adaptive test, but also the range of options provided by a rich concept of adaptivity. Such a rich concept allows test developers to consider a range of potential types of adaptivity for assessments that are used in proficiency and placement, as well as in achievement and diagnosis. Besides, it provides language for analyzing the ways that a test may be adaptive in some ways but not others, to meet the purpose of the test, and provides support for determining the measurement models and adaptation algorithms that best suit these purposes. The richer concept of adaptivity thus allows for assessments that are useful for the benefit of educators who wish to place and evaluate examinees as well as for learners who wish to better understand what they know and what they need to work. With such a range of assessments defined by these options for adaptivity, test developers can better consider the ways in which technology can be used to develop a new generation of language assessments.

REFERENCES

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, and assessment*. Cambridge: Cambridge University Press.

Levy, R., Behrens, J. T., & Mislevy, R. J. (2006). Variations in adaptive testing and their online leverage points. In D. D. Williams, S. L. Howell, & M. Hricko (Eds.), *Online assessment, measurement, and evaluation* (pp. 180-202). Hershey, PA: Information Science Publishing.

Mislevy, R. J., Steinburg, L. S., Almond, R. G. & Lucas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-47). Mahwah, N. J.: Lawrence Erlbaum Associates.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

ⁱ Though a well-recognized English test in China, CET-4 WT does not have its official website. Even so, many test preparation websites in China provide detailed information about the test. A sample test provided by *QQ CET-4 Test Preparation Center* can be found at <http://edu.qq.com/a/20071204/000124.htm>.

ⁱⁱ For more information about the *Transparent Language* English Proficiency Test, please visit its official website at <http://www.transparent.com/tlquiz/proftest/esl/tlesltest.htm>.

ⁱⁱⁱ For more information about the ACT ESL Placement Test, please visit its official website at <http://www.act.org/esl/overview.html>.

^{iv} Developed as an equivalent of the face-to-face Oral Proficiency Interview (OPI), SOPI is a semi-direct speaking test, administered in places where a limited number of trained raters are available for the test. More information can be found in the following documents on the website of the Center for Applied Linguistics: *Simulated Oral Proficiency Interviews: Recent Developments* (<http://www.cal.org/resources/digest/0014simulated.html>) and *Testing/Assessment: Simulated Oral Proficiency Interviews* (<http://www.cal.org/topics/ta/sopi.html>).

^v The ACTFL Guidelines for speaking can be found at www.sil.org/lingualinks/languagelearning/OtherResources/ACTFLProficiencyGuidelines/contents.htm

^{vi} Although it is a communicative oral proficiency test administered face-to-face, the OPI test has many drawbacks such as inefficient test administrations, requirement of a great number of trained interviewers, and lower reliability. To solve such problems, different types of semi-direct oral proficiency tests equivalent to the OPI have been developed, one of which is the Simulated Oral Proficiency Interview. Currently, the Computer-mediated Oral Proficiency Interview is being developed with the feature of examiner-controlled adaptive observations. For more information about the OPI development and its scoring rubrics, please visit the following two websites: Defense Language Institute (http://dlielc.org/testing/opi_test.html) and Center for Applied Linguistics (<http://www.cal.org/resources/digest/oralprof.html>).

^{vii} The ILR scale and oral proficiency descriptions can be found at <http://www.govtilr.org/ILRscale2.htm>.

^{viii} For more information about the FETN test website, please visit its homepage at <http://www.english-test.net/>.

^{ix} For more information about the Standard Test of Proficiency in Korean, please visit its official website at http://www.topik.or.kr/guide/topik_en_01_d.html.

^x For more information about DIALANG, please visit its official website at <http://www.dialang.org/intro.htm>.

^{xi} This hypothetical test is created based on Levy et al.'s explanation of a non-language test:

“Consider a simple case where a user’s query results in a list of documents, possibly structured by

some criterion such as perceived relevance. The user then selects some of the documents from the list for further consideration. A great deal of observable information can be collected from such a process. Which documents were viewed? In what order? How much time did the user spend reading each? These only scratch the surface of what data could possibly be collected. In these systems, the user is in control of the claim space, via the query, and the observables, via the actions taken with respect to the produced list of documents” (p. 29).