Online ELIPT: Assessment Needs and Intent

Roderick Gammon

Under mandate from the Director of the University of Hawaii at Manoa’s English Language Institute (ELI), a project is actively investigating creation of a test administration system with the ability to serve users both in centralized and geographically distributed contexts, and that is upgradeable in the type and format of test that is given. This document supports that project by articulating general testing issues and reviewing the current state of the ELI Placement Test (ELIPT) system.

Introduction

The first section is a literature review intended as an introduction to project concerns including relevant development methodologies, common information system features and technologies, common testing concepts and types, computer based testing, and applicable legal and professional guidelines. The latter sections discuss an analysis of the current English Language Institute’s (ELI) Placement Test (ELIPT) and its context. Although presented first, the topics discussed in the literature stem from observations made during an initial needs analysis.

Although the project is focused on providing tests, it is useful to conceive of such tests as part of a comprehensive information management system. As the current system includes technology, and the project requirements suggest information technology (IT) usage, the proposed system may be considered an IT system. This is useful because such conception can ease future attempts to further digitize the ELI’s information systems, and because comparisons can be made with IT systems from other domains.

It should also be stated at the outset that this project is necessarily collaborative. Many different skills are needed, it is unlikely that all participants will be skilled in all areas, and thus participants will need to know how to communicate effectively. The following literature review will hopefully provide the foundation for such communication. Otherwise the project will suffer a relatively common error in the interface between IT and language learning, where one domain of expertise is represented while others are ignored; Doughty and Long (2003) review particular examples in the area of task-based language teaching.

The final introductory concern is that this is a draft rendering. Omissions and errors were unavoidable and comments are welcomed. For example, some published literature regarding the ELI
itself is provided in the last section of the literature review, however it is supposed that much is already known to the first round of readers and so was omitted in the interest of timeliness.

Basic Technologies

Godwin-Jones (2001) provides a review of the assorted technologies that establish the basic computer-based testing (CBT) implementation lexicon. This project will target a lowest common denominator in platform capabilities. For example, Lynch and Horton (1999) for example note that 800 pixels wide by 600 pixels high is a reasonable size to assume for target screens within which all display elements must appear, an important concern if the test is to be embedded into a web page or other framing component.

Often software is classed by the manner in which its source code is licensed. Source code is considered open if its licensing allows one to look at the code, while closed source does not allow one to access the fundamental code. This licensing difference affects all software users regardless of technical competency; open source tends to be less costly to maintain and more secure (Raymond, 2000).

Central to the ELI Director’s mandate is the coordination of testing activities among geographically distant areas. This requirement immediately suggests some sort of Internet distributed system, as coordination across distances is one clear advantage of the technology (Negroponte, 1995). More commonly, the technological selection will be made from the n-tier paradigm that conceives of an Internet application, usually worldwide web based, as composed of coordinated slices such as a database, a web page server, and scripting technologies for dynamic web page generation. Spainhour and Eckstein (2000) provides an excellent overview of these technologies and their features for the beginning implementation agent. In an open source paradigm using the Apple OS X (“ten”) operating system, the tiers are commonly composed of the MySQL database package, the Apache web server, and PHP-scripted dynamic pages. As Brown (1997, p. 45) notes: “Receptive-response items-including multiple-choice, true-false, and matching items—are fairly easy to adapt to the computer-assisted testing medium... Unfortunately, the more interesting types of language tasks (e.g., role plays, interviews, compositions, oral presentations) prove much more difficult to develop for computer-assisted testing.”
Documentation is also an essential element of an IT system. For example, Weigle (2000) reviews the Michigan English language assessment battery (MELAB), a paper, pencil, and audiotape based test. Despite various weak points, Weigle (2000, p. 453) still praises the system because of strong documentation, including both statistical reviews and operational manuals.

Kenyon and Malabonga (2001) discuss a computer implementation of the ACTFL Oral Proficiency Interview (COPI, OPI) that took great pains with documentation. The result, as Norris (2001, p. 100) describes, was that

"As reported in Kenyon and Malabonga's [(2001)] article, the COPI utilizes technology to address affective concerns by introducing examinees to the computerized test format with a hands-on tutorial... Thus, even though examinees generally evaluated the COPI and SOPI formats in equivalently positive ways on independent surveys, when they were asked to select between one of the two tests, they reported on average that they preferred the COPI because it (a) seemed less difficult overall [and] (d) had clearer directions..."

Locale Management

The goal of geographically distributed testing produces acute language management issues. In an IT system information that varies by geographical or cultural location is termed locale specific. Localization is the act of preparing an information system for a particular locale. Localized information includes many aspects of language including diction, formatting and punctuation standards (Lunde, 1998). Regarding non-Latin orthographies, character display technologies are governed by the complicated domain of character encodings. For Asian-Latin localization, the seminal resource has long been considered Lunde (1998).

Lokan and Fleming (2003, p. 167) discuss localizing a career guidance system from the American to Australian locales (they use the term "adaptation"). Even though this might seem to be a simple task in terms of language translation effort, the authors note (2003, p. 168):

"Although the same language, English, was involved, and in a superficial way one might think that..."
United States and Australian cultures are similar, there are substantial differences in education systems and educational pathways, and noticeable differences in occupational conditions and people's value systems.

Managing locales requires review and therefore can become part of piloting as in Lokan and Fleming (2003, 176). They also note that the act of doing the localization itself opened opportunities to correct minor errors from the original, for example naming errors. Lokan and Fleming (2003, p. 173) describe the experience:

"At the most basic level, the spelling, vocabulary and phraseology of every part of the program had to be checked for correctness and appropriateness. ...There is also the important 'resistance factor' among teachers and students that materials developed for another country could not possibly be fully appropriate for use in Australia... Although this part of the adaptation was very tedious and time consuming, it was in fact the simplest part of the work."

Methodologies

This section reviews methodologies for human systems review.

The unified process (UP) is a software development methodology whose strengths are that it focuses on human uses of IT systems and also seeks to unite traditional best practices in the IT industry (Jacobson, Bocch, and Rumbaugh, 1999). The UP characterizes IT systems using a paradigm based on user roles and actions. For example technology maintenance could be considered as a role encompassing the activities of infrastructure configuration and maintenance; roles tend to be nouns while activities tend to be verbs. Usage based consideration of a system, by keeping focus on human benefits, prevents the system from becoming an entity with circular justification.

Modular design is also encouraged by the UP, as well as several other systems approaches (Gamma, Helm, Johnson & Vlissides, 1995; Fowler, 1999). Condensing these sources, modular design is essentially the development of an IT system whose parts interact through unchanging, well-specified protocols. Such protocols are termed interfaces and represent the services a component provides, rather
than how those services are provided. This allows components to interact without reliance on internal structures, which tend to change over time. Additionally, modular design tends to ease the addition and elimination of features.

Another constant concern in IT projects is control of resources and ensuring that measurable progress is made. Much has been written on these topics, but in addition to Jacobson et al. (1999), respected resources include Brooks (1995) and Pressman (2004). Brooks (1995) clearly establishes that software projects are successful when they clearly establish their specifications and resources before implementation and then keep those elements from being modified before development phase completion. Additionally Brooks (1995) demonstrates that adding software developers in mid-development tends to increase development cost and time. Pressman (2004), writing in a textbook for software project managers, clearly establishes that development must be further coordinated by a focus on the development of tangible artifacts as milestones, and espouses the use of ever-more functional prototypes to coordinate transition among development phases.

Bachman and Palmer (1996, pp. 29 – 35) consider a clear understanding of intended score uses as central to successful test development. This concept is central to several language test development methodology, most recently Norris' (draft) Specification of Intended Test Use (SITU) that gives “the contexts, reasons, and expectations for a given assessment activity” and which serves as a guide to test developers. Norris (draft) organizes the SITU under the following questions: test score use, test-based interpretations, test users, test takers, instructional context, likely consequences of test use (including stakeholders), likely constraints on test use.

Test use is also a central element in establishing test validity. Chappelle et al. (2003, p. 411) note that validity is an argument accompanied by multiple information sources. (P. 412) start validity argument from "intended use". (p. 412) "Read and Chappelle (2001) elaborated components of test purpose as consisting of the inference that is to be made about the learner’s ability and/or subsequent potential for performance, the use of the test scores, and the impact that the test is intended to have beyond the specifics of measurement."

Finally, analysis requires knowledge of test subjects. Gorsuch (2004) details one project’s use of retrospective subject interviews to investigate sources of measurement error in CBT listening.
comprehension tests. Interview topics were derived in part from observations of test sessions (Gorsuch, 2004, pp. 349-350).

Knowledge of a subject's background is also essential for proper test calibration (AERA et al., 1999, pp. 51-55), including the features of gender, native language, and educational status. In addition, Kondo-Brown (forthcoming) has developed a background questionnaire for students of Japanese that has proved particularly useful when the population includes heritage learners. Brown and Iwashita (1998, p. 196), in the context of a Japanese second-language, grammar CAT, note that subject language background affected item difficulty. In particular a native language was found to affect difficulty according to the principle of zero contrast; target features are more difficult for those subjects whose native language lacks a similar feature (Brown and Iwashita, 1998, p. 197).

A later section discusses the ethical treatment of subjects.

Testing

Tests are often referred to in the literature by their uses and by their structure. Keeping with the UP methodology, uses will be discussed first.

Tests are commonly divided into two families, norm-referenced and criterion referenced (Brown, 1996; Brown & Hudson 2002). There are many features that distinguish these two types, but one succinct description is that the families are intended to produce different total score distributions. Norm-referenced tests are desired when one desires a normal distribution, or bell-curve (Brown & Hudson, 2002, p. 2). In contrast, criterion-referenced tests tend to produce skewed distributions (Brown & Hudson, 2002, pp. 103-105). NRT are useful for comparing students to each other. Brown and Hudson (2002, p. 4) define a CRT’s purpose as gauging “the performances of examinees in terms of the amount that they know of a specific domain of knowledge or set of objectives”.

Test uses are commonly analyzed into four types: placement, diagnostic, achievement, and proficiency (Brown and Hudson, 2002). These four types mirror decisions commonly made in curriculum management (Brown, 1995, pp. 108-112). ELIPT, as the name suggests, is a placement test.
Test structure is often classed by medium, into either paper and pencil or computer-based (CBT). However test structure may be more productively interpreted according to two features, item and unit storage, analysis of which yields four common types of test:

<table>
<thead>
<tr>
<th>Table 1. Common test structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Storage</td>
</tr>
<tr>
<td>1 Static</td>
</tr>
<tr>
<td>2 Item bank</td>
</tr>
<tr>
<td>3 Item bank</td>
</tr>
<tr>
<td>4 Generated</td>
</tr>
</tbody>
</table>

Often these forms are associated with a particular medium, either pen and paper or computer based. However a bit of creative thinking can quickly indicate that this need not be the case. Form one, where items are stored in a static repository and unit structure also remains constant, is the common paper and pencil test. However static storage is simply some unchanging container, both a piece of paper with the items on it and HTML pages loaded to a server and Adobe PDF files are example of static, digital stores.

The second form is in current use by ELI whose tests are distributed as static, on paper, but who manages test items using a database. As is evidenced by ELI's case, this method supports the production of parallel test forms. (Brown, 1997, p. 44) "Item banking covers any procedures that are used to create, pilot, analyze, store, manage, and select test items so that multiple test forms can be created from subsets of the total "bank" of items."

Form three facilitates the well-known computer-adaptive testing (CAT) paradigm. In this case the items are not selected until the test is begun, and the particular set of items used on a test are not known until the test is finished. Commonly this approach involves more complicated types of software, since the test needs to be generated at runtime rather than simply displayed. This approach also involves advanced training by its implementers, for example Brown (1997, p. 44) has pointed out the indispensability of item-response theory (IRT), an advanced statistical method.

Finally another emerging paradigm in item storage is runtime generation of test items. Nearly all published research in this area emerges from Educational Testing Services and is not focused on language testing per se; the seminal reference is *Item Generation for Test Development* (2002). Gammon
(2004, draft) has also experimented with the method for Mandarin listening comprehension. A particular benefit of item generative approaches is that implementing them requires a detailed comprehension of desired item types, content, and constructs; such development fosters a critical understanding of test needs that can also aid interpretation of test results later on.

Adaptation also need not be confined to test item selection. Murugesan and Ramanathan (2001), discussing IT systems in a more general context, note that the uses of personalization include the modification of interface affordances such as navigation opportunities, of presented content, and of content presentation. Brusilovsky’s (2001, p. 97) review collapses the distinction into interface and content adaptation. Especially useful in that latter review is a chart outlining adaptation activities within each domain such as hiding/revealing information and navigation (Brusilovsky, 2001, p. 100). Adaptive technology’s application to education has formed a prominent trend in personalization technology (Brusilovsky, 2001, p. 90). Goderis et al. (2001, p. 62) use the term dynamic determination, defined as “when personalization is different for different types of users”.

In general, CBT benefits are usually relevant to providing more convenient testing and also immediate feedback (Chappelle et al., 2003, p. 409). Brown (1997, p. 46) more objectively notes simply that CAT have different administrative concerns from paper and pencil tests. Wainer (2002, p. 301), drawing from the concept that over time computer based administration is more flexible and less expensive than pen and paper, notes that CAT are particularly suitable for classroom diagnostic tests where frequent testing is closely integrated to classroom content. Brown (1997, p. 47) lists more flexible test administration, greater standardization of test conditions, test individualization to examinee ability, and the ability to record multiple aspects of examinee behavior.

In a more practical vein, Kenyon and Malabonga (2001, p. 64) report that the COPI is shorter than the tape based OPI. Brown (1997, p. 47) reports on studies indicating that many fewer items are necessary in administering computer-adaptive language tests than are necessary in pencil-and-paper tests and that the testing time is correspondingly shorter. For example, the CALT in Madsen (1991) used an average of 22.8 items to adequately test the students in an average of 27.2 minutes. The comparable conventional reading test used in the study required 60 items and 40 minutes.” However Hudson and Park (XXXX) have learned that WBT require many items, because of concerns about item disclosure.
Piloting CBTs is more complicated because, as Brown (1997, p. 50) notes, CAT piloting breaks classical test theory assumption of a group taking the same form. Chappelle et al. (2003, p. 422) describe a typical pilot activity:

“Materials consisted of a brief paper questionnaire in addition to the forms of TYE that had been configured for the pilot study. The questionnaire asked participants to provide their TOEFL score if they had one, and to self-evaluate their English ability. The tests were developed using Macromedia Dreamweaver 3.0 and Macromedia CourseBuilder for Dreamweaver, and customized JavaScript was added in several places where CourseBuilder did not support the functions that were needed... 25 G4 Power Macintosh computers with 19-inch Sony monitors was the testing site...The screen resolution was set to 1024 x 768 pixels.”

A review cycle similar to the ELIPT’s is given by Chappelle et al. (2003, p. 423): “After the last question was answered, three actions occurred: (1) data in the text fields were copied to the top left frame; (2) these data were sent via email to two computer accounts of testing research assistants; and (3) data in the tracking frame were added to the Microsoft Access database on the server and then imported into Microsoft Excel spreadsheets. Steps (1) and (2) were back-up procedures in case the database connection failed; fortunately, no such problems occurred.”

Given the resources invested in the ELIPT, described later, and also that the ELIPT will need piloting, CBT and pen and paper (PNP) test forms will coexist for some duration. This yields the concern of cross-media equivalence. For example Choi et al. (2004, p. 299) claimed that CAT oral tests differ because visual data (the monitor) is explicitly included. In Choi et al.’s study (2004, p. 316), reading comprehension had the least comparability to paper and pencil forms while listening comprehension had the most.

Ethical Subject Treatment

Main aspects of subject treatment are fostering accessibility, preventing bias, and protection of confidentiality.
Brown (1997, p. 49) reviews literature establishing the following CBT disadvantage: "Differences in the degree to which students are familiar with using computers or typewriter keyboards may lead to discrepancies in their performances on computer-assisted or computer-adaptive tests."

Good documentation can also help with accessibility, for example Kenyon and Malabonga (2001, p. 65): "When taking the COPI, the examinee goes through nine phases: welcome, information on the purpose and structure of the COPI, input and correction of personal information, self-assessment of proficiency level, listening to an adequate response to a sample task(s), practice with the same sample task(s), responding to performance tasks (the actual test), feedback about the levels of the tasks that the examinee took, and closing. A photograph of a friendly virtual female "guide" accompanies the audio and written directions presented on the computer screens."

From Choi et al. (2003, p. 298): *Kirsch et al. (1998) and Taylor et al. (1998) address issues about equity and bias of test-takers' performances on TOEFL CBT. The study looked at approximately 90 000 TOEFL examinees in terms of their experience with computers and a number of examinee background characteristics. Their research findings reveal that more than 80% of the TOEFL population have at least moderate familiarity with computers, and that there was no evidence of adverse effects on the computer-based TOEFL performance due to lack of prior computer experience...."

Norris (2001, p. 101): *Although its innovations may offer certain advantages, the extent to which the COPI will serve as an adequate platform for assessing oral proficiency according to the ACTFL Guidelines (1999) as well as a model for the development of other CBTs of complex performance abilities, depends in large part on whether research reveals that changes in the testing format influence both the quality of speaking performances and the accuracy of proficiency ratings.* Norris (2001, p. 101) goes on to note that when subjects self select tasks or content domains there is an opportunity for bias to occur that needs quantified analysis.

The Americans with Disabilities Act requires testing agencies to accommodate the disabled (CASAS, 2002, p. 15). However test subjects needing accommodation are required to notify their
institution at registration time, although institutions can remain proactive by providing a standard "Request for Accommodation" form (CASAS, 2002, p. 16). Chisholm et al.'s (1999) federal guidelines on accessibility describe specific activities websites should engage in to enhance ADA compliance.

Additionally Sprague (1999) lists methods of website design that comply with American federal regulations regarding accessibility by the disabled. Nielsen's (2000) reference also provides a good introduction.

Regarding protecting subjects, Murugesan and Ramanathan (2001, p. 68) note that while data collection is central concern, users need to be informed as to when it will occur. Researchers working within the United States should design systems that comply with the Federal Policy for the Protection of Human Subjects (1991) from the very beginning. It should be noted that a participant's age is a critical factor in defining the ability to consent, under those (1991) regulations.

Norris and Ortega (2003, p. 719) typify SLA data with the universal definition of "repeated observations of particular patterns in behaviors... condensed into scores." Such data is governed by governmental and institutional regulations of informed consent.


Method

The remainder of this paper reviews the current state of ELIPT use and then discusses concerns arising from that usage as they apply to introduction of a web-based ELIPT. Information was collected through personal and email-based interviews, and through participatory observation of two ELIPT administrations and two essay grading sessions. The Director, Assistant Director, Curriculum Coordinator, current and past testing staff, and current lead teachers were interviewed in person. All ELI instructional staff were further interviewed through an opt-in email questionnaire. Additionally, "on the spot questions" were posed to participants in the ELIPT activities that were observed. When needed, information sources are referred to by their position. The Director, Assistant Director, and Curriculum
Coordinator are referred to throughout as senior administrative staff. The ELI Intranet, a document collection fostered and maintained by the Director, was also an important source.

There are also logistical staff who serve both ELI and SLS. They and actual test takers, or subjects, weren't contacted.

Instructional faculty were queried through short answer questions disseminated by email. These questions focused on satisfaction with the ELIPT and perceived relevance to curriculum. Responses are presented in the results section. The questionnaire structure is given in the appendix.

There were six respondents to the questionnaire. Respondents were alphabetically coded in order of response arrival. The codes and the amount of ELI instructional experience is given below:

A - 2 semesters, reading and writing
B - 2 semesters, unspecified areas
C - 2 semesters, unspecified areas
D - 4 semesters, reading and writing
E - 3 semesters, unspecified areas
F - 4 semesters, speaking

Results

Although this section is referred to as results, the fact is that the material in this section is also an analysis. Its real divergence from the following discussion section is that this section restricts itself to direct observations of the ELIPT testing process.

The ELIPT system uses heterogeneous media: paper-based, online, telephone, and in-person communications all occur. Tests and their results are the central artifacts; all roles and activities have these items as their central focus. In this sense, ELI testing is defined by the ELIPT; the ELIPT's purpose is to decide if a student is exempt, and if not then what courses are needed.

Security is an important concern for artifact maintenance. For test security the ELI ensures that only valid subjects take the test, and also that the test is protected from early disclosure. Result security entails balancing consumer access to results while protecting subject confidentiality. Security often
focuses around protecting and validating artifacts such as student credentials, test booklets, and test forms.

ELI enjoys two general classes of resource, staff and stuff. Following the Unified Process methodology, the results are organized by actor roles to keep the human element in primary focus. The top-level roles are, alphabetically: administrators, authors, consumers, scorers, and subjects. These roles work to fulfill ELI’s two purposes of providing a bridge between ESL and EAP for students, and of providing a laboratory environment for SLS students.

Role sub-types and activities are described below. Additionally, a single person may fulfill multiple roles during the execution of their ELI duties. In fact multiple role fulfillment appears to be an adaptive response to the resource constraints faced by the ELI. At this section’s end, the roles and their actions will be mapped to a testing lifecycle, to illustrate their interaction sequences.

Because the emailed questionnaire focused on instructor satisfaction with the test and perceived relevance to curriculum, the majority of responses are reviewed in the section on consumers. However many responses discussed scoring aspects of placement, and comment on test administration; comments appear relevant sections. Responses may be elided but they have not been corrected from grammar, punctuation, or spelling.

Test Administration

Administrator roles are those involved in the actual carrying out of the test. This class of roles is bounded at one end by authors; administration roles are not concerned with revising or producing tests. At the other end of this hypothetical spectrum, administration is bound by consumers; administration roles do use test results. Test administration role types include managers who coordinate labor and resources; proctors who oversee subject-test interaction; and support personnel who ensure that proctors are properly outfitted and are interacting with the correct subjects. All administrators are alike in that their purpose coordinate the interaction of other roles with the test and test results.

Managers are identifiable by their decision making capability and their knowledge of extra-ELI channels. This makes the presence of one essential on testing days. On both observed days the testing
location was double-booked and a manager was needed. On one of these days the test had to moved
after it had already begun.

The Assistant Director is the primary logistical manager. For example the test cycle may be seen
as beginning when the Assistant Director authorizes a new test form, received from the ELI test authors.
Additionally the Assistant Director produces the reports utilized by consumers. In this sense, the
Assistant Director fulfills critical roles; this administrator is key to the creation of the central system
artifacts of test and test results.

However all senior administration is essential overall, as they all undertake important
management activities, including scheduling, online registration, test form printing, and other resource
allocations such as ensuring that pencils and paper are supplied. Ultimately responsibility falls on the
Director, and so the Director maintains close observation of the process. The Curriculum Coordinator is
central to fundamental ground-level activities, such as coordinating proctors and support staff during
testing.

Managers oversee the issuing of test taker IDs. Managers also set ELI protocol, which has
generally been defined to the benefit of subjects for example there is "up to last minute" registration. The
manager's duties increase with general admissions, which is when test registration begins. Proctors
prepare the test room, answer subject questions about the test and ELI (or defer them to more qualified
staff), ensure that subjects take the test properly by giving directions and policing the session, and collect
and submit materials for grading. Proctors have access to written protocols, including scripted directions
to read to subjects. Proctors follow a manual provided by the senior administrative staff. Although largely
confined to the start of semester test dates, proctoring will continue throughout the semester. This later
proctoring is usually embodied by the Assistant Director who proctors to late-comers throughout the
semester.

Proctors already face the difficult technology management requirements as it is both varied and
conducted under pressure (before a room full of subjects). IT confusion existed on both observed
days—once because a new room was needed and its hardware configuration differed, on the other day
because personnel were newer. Even superficially simple technologies involve complex interactions. For
example audio prompts on CD media have subtle complexities: there are pauses recorded on cd,
proctors need to pause the cd between dictation and AL sections, and proctors manage track transitions in some cases. In addition to being observed by this author, this difficulty has also been noted by the Curriculum Coordinator, who oversees much of the proctoring.

Although such an instance did not occur during the observed sessions, proctors also need to manage instances of special needs. To some extent the testing artifact supports this by noting that alternate formats are available and by giving a telephone number to call. However final responsibility lay with the human agents. Proctors also answer, or are at least asked, questions regarding test results and ELI policy.

Finally proctors are essential components in test security procedures. They are responsible for ensuring that cheating does not occur during the test, and that it is protected against disclosure outside of the test. To this end they are responsible for collecting test materials and notes taken during test, to prevent their dissemination among future candidates. The proctors are greatly aided by a clear and detailed manual, which was in use at both observed sessions.

Support personnel perform onsite registration and subject verification, act as runners when resources (e.g. pencils), forms need restocking, and provide information to subjects about the test. On test administration day they check Picture IDs, carry materials from the ELI offices to the test site, and run errands between the site and ELI (e.g. scores and forgotten items). Support personnel are also critical to test security as they ensure that only valid subjects interact with the test (proctors count heads but do not check identities).

Overall, EUPT administration is an essential ELI activity that consumes time, resources, and labor and eclipses other ELI activities during its duration. In many ways the administration is extremely successful. For example, even in the face of unusual difficulties (Assistant Director) such as room double-booking, the testing time overrun was less than one hour, only 1 student was unable to complete the test, and standard operating procedures were still consistently applied despite changes in available technology. Support staff also realize the effort, some positively and others negatively, as in the questionnaire responses below (respondent, question):

B, 1 – "...very carefully administered"
E, 2 - "The test is pretty long. The administration of the test is hard on students, teachers and faculty."

F, 1 - "It is well-organized, I think. But since I don't know any other placement tests, I cannot really compare the ELIPT to others and therefore there is no way for me to know if there is any better ways of placement tests. (By "better", I mean economical and efficient.)"

F, 2 - It is long, so the ELIPT duties also get long.

Test Authorship

Authorship roles are among the most expensive ELIPT roles in that eligible labor requires specialized training in a traditionally under populated ESL subfield. These roles are also long term, as the test item publication cycle appears to average about a year. On the other hand, ELI has a ready pool of talented researchers, SLS students and faculty, that can develop and analyze survey instruments and results.

There are three roles in this class: item authors, who write items and units; proofreaders, English native speakers who review item prompts and responses; and authenticators who ensure face and quantified validity. This role interacts with managers directly by passing forward completed tests, making revisions as ordered, and receiving scores as data for quality review. The author role also indirectly affects scorers and subjects.

The author roles are also very technology reliant. At their simplest, tests, results, and their drafts are examined with Microsoft Word and disseminated using email. There is also a set of specialized technology. For example a scantron machine is used for result tallying. The item bank is managed using the PAR suite of applications (Score and Test), a specialized testing application that includes statistical functions. For the more complex multivariate, inferential statistics that are a test author's particular domain, there are also a host of other, specialized applications that are essentially specialized calculators. Currently the reading comprehension, cloze, and listening sections have computer support.
The general authorship workflow is: data capture (e.g. scanning by the Scantron reader), analysis aided by PAR score, new item authoring, test form generation aided by PAR Test, and then ratification and revision as directed by senior management.

The PAR Test program is well-appreciated by the authoring team. It simplifies new test form creation; security is maintained through a login program; current and cumulative statistics are visible; and item formats include true/false, multiple choice, matching, blank completion, and essay.

However there are some problems. First, overall the PAR system does not exactly support what ELI is doing: items are banked but not media (e.g. audio); it does not directly support reading passages with items attached (the “instructions” block is used as a stopgap); PAR Test only supports work with item scores, not total and unit scores (such as needed for variance based statistics); and Par Score maintains information by “course”, which ELI instead uses for “test sessions”. Test output from the system is also not directly usable, much manual editing needs to be performed to achieve satisfactory page formatting. Finally, the statistical applications are designed so that separate tracking of pilot items is complicated. Overall poor ease of use was a consistent complaint with the system; it is particularly intolerant of human error and further lacks an undo feature. A critical problem with PAR Score is that it interacts poorly with Microsoft Excel which is an essential application throughout the testing world.

One problem the authors face is poor documentation. Item specifications are reverse engineered and spotty, produced by student projects.

Test authors have an important purpose in the ELI, they provide the research that validates the test. Therefore test authors operate as authorities within the ELI, in practice advising senior administration as experts and, in the ELI’s mental space, acting almost as totems appealed to with a special language and having particular abilities of control (remember, only managers are decision makers). Questionnaire responses indicate both that this latter aspect exists and that better communication between test authors and greater ELI is needed. For example the following ascribes planning capacity to authors:

A, 4 - “I don’t know much about the listening test... [but] it would be logistically difficult to test presentation and discussion skills in the context of a placement test (perhaps this is a challenge
you could take up later on in the semester?)"

This response utilizes a special lexicon:

D, 4 - "1, but I will qualify this by saying that I do not think the placement test should/need to address what I'll teach. It's by design an NRT, not a CRT."

By the same respondent, the following comment indicates that authorship activities need better communication with the greater ELI community:

D, 5 - "I think it's fine, but I would like to see some reliability and validity testing done, if it isn't already."

Finally, this comment indicates a perceived specialist status of authors:

F, 1 - "...since I don't know any other placement tests, I cannot really compare the ELIPT to others and therefore there is no way for me to know if there is any better ways of placement tests. (By "better", I mean economical and efficient.)"

Consumers

Consumers are those who use tests. This is the point at which test results meet greater-ELI administrative policy. One obvious class is ELI senior administration, who have the responsibility of communicating placement decisions to the greater UH community. Candidates for consumer roles include students, ELI instructors, ELI and greater UH administrators, and third parties (e.g. employers) needing language skill verification. Current consumer sub-types are advisors, students, and teachers.

Advisors, usually also managers, provide guidance to students and determine exemptions. Advisors also see themselves as advocates for non-native English speaking students: there is an "exempt until proven otherwise" multiple methods approach to placement decisions, a high exemption rate (A.
Dir.), students near cut point get special attention, 20-30% of students are exempt in 1 or more ELI areas, and a calculated standard error of measurement statistic allows exceptions around cut point.

Students are generally interested in outcomes only; are they exempt and, if not, what classes they need to take. It has been noted by ELI senior administration that this is just as well, under the current system it would be costly to give more detailed reports to students.

Teachers consume test results indirectly, by having classes formed on the basis of test results. However there is also a general feeling within the ELI community that closer interaction between teachers and test results would be beneficial. ParScore data is available but seldom shared with teachers: the teachers need to request it, they may not know how to read the reports, and there is not a clear coordination between the ELIPT and ELI curriculum.

It should be noted that instructors are of different minds about the relevance of the test to their classroom, although there appears to be interest in rectifying the situation. Regarding question 4, which directly queried perceived relevance, the Likert answers ranged from 2-3.5. The Curriculum Coordinator echoed the general belief that writing and reading sections of the ELIPT are the most effective. The following short answer responses elaborate:

A, 4 — "The free response portion of the writing test gives some good clues about students' ability to handle either 100 or 83, versus 73."

B, 1 — "...more than one mode of test for certain areas (e.g., dictation + listening comprehension, etc.)"

B, 2 — "i) Dictation seems to be (intuitively) a poor measure of listening skills, and it seems to be strongly affected by other factors, such as working memory capacity, etc. ii) I know it's difficult, but it would be great if we could test speaking skills too"

B, 5 — "How about something like, listening to a lecture and summarize the content? That doesn't measure speaking, but still it measures listening comprehension and (written) production skills."
D, 1 - "...I'd like to see a vocab section added as well. For writing, the essay prompts are fine, though I would like to see them be on topics less common since it seems some students are able to mold them into essays that they have written in some form or fashion before." [note: This was also echoed in an exchange with the Curriculum Coordinator.]

E, 4 - "2. We mostly teach strategies and the test is based on performance, although I don't think this is such a bad thing."

F, 4 - "3--one major reason is that the ELIPT does not test speaking skill, which is one of the major components in ELI70/80."

Scorers

Scorers produce the test results that are one of the ELI testing system's central artifacts. Although many of those who perform administrative roles also perform scoring, the scoring roles are kept distinct on the basis that scoring activities do not involve direct interaction with test subjects.

Scorers interact with two general types of test response: listable responses and natural language responses. Listable responses are produced by multiple-choice sections (processed by a scantron reader) and by the cloze test (human graded using a list of acceptable responses, primarily at the testing check-in table). Natural language responses are essays graded by at least three raters in a more organic process. Scoring at all stages is guided by a reasonably well-defined set of protocols and rubrics. It should be noted that essay scoring requires dual roles as the scores are immediately converted into placement decisions, although those decisions still require ratification by other actors.

Scoring is hard work; by all reports, the Assistant Director is overwhelmed with a sudden deluge of work during the scoring activities. Essay scoring requires 3 human raters, with an additional 4th to break deadlocks.

The Assistant Director manages grading and performs grading. He has a Scantron for grading, uses StatView software for raw-scale score conversion. Issues with his scoring machinery include that
StatView doesn't allow for massaged changes based on human choice and that the Excel based form generation is clumsy and based on monolithic (rather than modularized) code.

No moderator manual is available for facilitating the rater grading process. There are substantial grading hallmarks for individual scorers though.

Regarding effectiveness of scoring and subsequent placement decisions, ELI teachers had quite a bit to say. Among the following responses are implicit criticism of the "place higher" policy:

A, 1 - "I've only graded writing placement tests, and it seems that the rubrics are quite good."

A, 2 - "From what I've seen in the classroom, it should be made very clear to the scorers of the written tests that the different writing needs of undergrads and grad students must be kept in mind when making placement recommendations. I've had grad students in my 83 course who belonged in 73 due to lack of awareness or training in basic writing. Much the same holds true for 72 vs. 82 in the area of academic reading skills."

A, 4 - "The reading test seems to be slightly easier than I would prefer to see- i.e., it's not too difficult to place higher in a course than one may actually be prepared for."

B, 3 - "Fine, although there were about three students in my class that needed to be promoted to a higher-level, that still seems minor."

D, 1 - "I think the cloze test and reading comp sections are adequate for determining level placement..."

E, 1 - "I have come to believe that it is pretty good at placing students, and this makes my job much easier."

Subjects
Subjects are those who take the test. The ELI placement test population Fall 2003 was approximately 220, Spring 2004 was approximately 150. Fall 2004 was 220. Performance in this role excludes performance of other roles, not least because of security protocols.

Not all students have identical testing experiences. For example undergraduates perform an essay section defined by the Manoa Writing Project, while graduate students perform one defined by the ELI.

Students also vary in that there are a variety of test taking strategies observable during test sessions. For example during reading comprehension, students like to skip ahead and check out questions before reading (also a Princeton Review test-taking strategy). Many students also enjoy having a visible timer. Although this document does not seek to catalogue these strategies, it should be noted that any CBT that obviated such strategies would likely also impact aggregate result patterns.

ELIPT lifecycle

Now that the primary roles and activities have been described, their sequencing, the ELIPT lifecycle, will be reviewed. First, a new test is produced by the authors; they collect reading passages, recycle prior forms, and fix obvious mistakes from the prior test. The resulting test form is sent to senior administration who validates it, possibly after requesting changes. A manager then oversees printing of the test and forms.

During the final stages of authorship, the managers are also registering students and setting the test dates and locations. On those days, support staff arrive with materials and begin checking in subjects. Proctors prepare the test room, receive the subjects, administer the test, and collect materials. Response materials are then taken to the support staff, who transmit them to scorers. In the case of Scantron compatible forms, they are taken to the Assistant Director. Cloze tests are graded at the check-in table. *Essays are saved for later, group-based rating.*

After a test session has been scored, placement results are made and posted online and in the ELI (Moore 5th floor). At this point students begin receiving their advisement.

Some time after the ELI has “calmed down”, authors then begin the review phase that ends and starts a test lifecycle. Validity and reliability studies are conducted, change requests are received from
managers, and, in practical terms, these studies are conducted as part of each author's research program. A month before a new test date the cycle restarts.

Known Directions

Having described the activities that accompany ELI testing, this final results section will document some long term goals voiced by the ELI community. These fall into the general categories of enhancing the existing system and expanding the consumer base.

One example of expanding consumership is the Vietnam-located MBA (VEMBA) program offered by the UH College of Business. Such students need ELI clearance for UH study, and such clearance would be in the form of placement testing. Additionally SLS and ELI students have begun teaching ESL on location in Vietnam. State future activities have included diagnostic and achievement testing, distance education components, and a better accordance with the MBA curriculum that is more text-oriented (long narratives, case-reports, lecture comprehension) than ELI's current EAP focus.

Enhancing the existing system tends to focus around improving current research activities and also coordinating more closely with curriculum. One concern is system proliferation, some in ELI have voiced an interest with an system that integrated even with the greater MyUH IT system. Using the ELIPT to represent curriculum matches its original intent.

Clark (2002, 10) provided a detailed set of future research directions including revision and sound enhancement of the oral passages, improvement of item specifications, investigating the relationship between academic vocabulary size and ELIPT scores, examining alternatives to the dictation test (cloze also comes up frequently), and reviewing the ELIPT's relation to TOEFL distributions.

Discussion

The focus of this discussion is how current lessons from the paper and pencil ELIPT can inform development of an IT test-platform that will publish tests in addition to a web-based ELIPT (WELIPT). Realizing that practicality demands that a WELIPT needs much piloting, the WELIPT is further considered as concurrent parallel form to the paper and pencil ELIPT.
A web-based ELIPT would be a system delivered to a standard web browser and using an n-tier design. Such a system could vary from affordable (e.g. a Mac G3 with OS X and open source software) to very costly solutions. Whatever platform is decided, it is urged that the system be dedicated wholly to ELI testing, and ELI testing only. One problem with existing ELI infrastructure is that little of it is dedicated to ELI, proliferating chances for malfunction. Additionally, the ELI should plan on a system that allows simultaneous piloting and provision of tests, either through a feature-rich single system, or through two separate, perhaps less complicated, systems.

Having parallel forms using different media would result in different test experiences. Although this should not be treated lightly, it is worth remembering that the present ELIPT is not monolithic; there are different essay sections for example. Also, late-coming students necessarily have different testing experiences from those who take it en masse at the start of the semester.

There are three main challenges faced by the ELI: rapid staff turnover, inconsistent funding, and lack of dedicated physical resources. As a result, information processing and coherent long-term planning are complicated. A unified IT system could be beneficial.

Regarding information processing, a dedicated IT system could serve as a long-term data repository. For example it could hold documents produced by authors, for the benefit of consumer (i.e. comment D.5 above). Such support for instructors could in turn benefit authors—a clearly stated research plan and item specification set could be conducive to getting teachers to produce items, in turn easing burdens on existing authors and possibly making test items more relevant to classroom curriculum. Such a mechanism already exists in the ELI Intranet, which this report suggests expanding. A more integrated IT publishing system might also lesson costs of giving students more textured result information.

An immediate advantage of a web-based system would be that it could greatly reduce effort spent by the Assistant Director when proctoring the ELIPT for late coming students.

However an expanded, IT based testing system will also present new challenges. System maintenance will become a concern, and dedicated computer labs with trained proctors will be needed. Student IDs will still need to be checked and note paper will need to be collected.
Although subjects can presently be condensed into one group on the basis of their taking the same test, this is not the most beneficial view. First, the program's stated intent is to produce a variety of tests. Second, their subject variation sufficient to, even now, cause differences in the tests they take. Finally, allowing for differences among test result consumers, subject type is one trigger for selection among available result uses.

In conclusion a list of features desired in a web-based ELIPT include: Result browsing by person, group, and item; facilitation of scoring; the opportunity to expand into adaptive tests; localizable, adaptive, parallel forms; multimedia; adherence to ethical standards; provision of pilot and live modes; automatic generation of “research ready” data; and equivalence to a paper ELIPT.

However before decisions are made regarding that list, more discussion is needed. Particularly absent from this presentation were responses from office logistical staff and test subjects.

References


Appendix A

Email questionnaire for instructional staff

1. What do you like about the current ELI Placement Test (ELIPT)?

2. What do you not like about the ELIPT?

3. Are your students well-placed in class? (Yes or no)
4. On a scale of 1 (least relevant) to 5 (most relevant), does the ELIPT test what you teach?

5. Do you have any particular ideas for improving the ELIPT or its relation to our curriculum (particularly keeping in mind that we have to be able to test, score, and report results for 30-90 students within a 12-hour period)?

6. Counting this semester, how many semesters have you been teaching in the ELI? How many times have you taught the class(es) you are currently teaching?