DESIGNING A TASK-BASED CRITICAL LISTENING CONSTRUCT FOR LISTENING ASSESSMENT

JONATHAN TRACE

University of Hawai'i at Mānoa

ABSTRACT

This study analyzed a task-based construct of critical listening in an academic listening test for placement purposes in a North American university English for Academic Purposes program. As the ability to listen critically in English is one of the outcomes for the program, it is necessary to utilize a placement instrument that can adequately measure this ability. Buck (2001) claims that one way of approaching this idea of critical listening in listening assessment is through the use of tasks that mirror the uses and functions an examinee will encounter in authentic situations. Using Rasch model analysis, this study first examined the current form of the test to identify how items are functioning and whether or not different, distinct constructs are present in the test. The test was revised using new pilot items based on a task-based model, and then analyzed again to determine the extent this construct was represented in the instrument. Based on these analyses, recommendations are made about the effectiveness of the test and the form further revisions of the test might take in future administrations.

The process of assessing listening comprehension for placement purposes is a challenging task. Level separation is understandably important in this context and it is, therefore, equally important to have a well-defined construct of what is being assessed for each independent level of the program. While Buck (2001, p. 114) provides a singular, inclusive construct of listening comprehension that includes processing, understanding the text, and basic inference, this model is ultimately too simple, as it fails to account for the context.

The purpose of this study is to explore how an appropriately contextualized construct of listening comprehension can be defined and operationalized for use in an English for Academic Purposes (EAP) program listening placement test. The test in question was developed as an

Second Language Studies, 32(1), Fall 2013, pp. 59-111.

academic listening test (ALT) for placement purposes within one of two levels of academic listening comprehension courses (Clark, 2007), and is aimed squarely at testing comprehension based on authentic, naturally occurring listening. While successful as a general model for comprehension, it appears to not take into account how the test actually places test takers into different proficiency levels in relation to their particular student learning outcomes (SLOs). Level separation between the two courses was initially designed according to Bloom's (1956) taxonomy of cognitive demand, with the intermediate course focusing on comprehension and application. This can be seen through SLOs that state that students will be able to, 'demonstrate good use of strategies for comprehending academic lectures in English' and 'demonstrate effective use of strategies for incorporating information from academic lectures into their overall studies'. The advanced course, likewise, incorporates processes of analysis and evaluation, as seen in SLOs such as students will be able to, 'demonstrate effective use of strategies for comprehending advanced lectures in English' and 'critically evaluate speakers' perspectives, techniques, and arguments'. (For a full listing of SLOs for each course, see Appendix A.)

A previous analysis of the ALT by Chun (2011) using classical testing theory attempted to understand how these differences in level and SLOs are represented in the construct of the ALT in terms of processing levels, though the conclusions did not report a clear distinction outside of general listening proficiency. Several studies in the field have addressed the separation of processing levels (Shohamy & Inbar, 1991; Rost, 2002; Hansen & Jensen, 1994), as well as created taxonomies of skills and strategies (Aitken, 1978; Buck & Tatsuoka, 1998; Richards, 1983), but there is still a question of how to effectively bring these ideas together into a defined construct for assessment, particularly in the field of listening critically. This study will look into these issues in an attempt to revise the ALT and integrate a suitable construct that can better attend to the separation of placement levels within the program.

LITERATURE REVIEW

Processing

Based on the SLOs for each course, and the differences between them, it is appropriate to begin with a discussion of processing levels and their relationship with language proficiency. Processing in terms of listening comprehension is primarily understood in terms of a variety of

different dichotomies, such as bottom-up and top-down (Vandergrift, 2007; Rost, 2002), local and global (Shohamy & Inbar, 1991), microprocessing and macroprocessing (Kintsch & Yarborough, 1982; Van Djik & Kintsch, 1983), as well as text-based and situational (Kintsch, 1998). Subtle differences exist between many of these terminologies, but what each has in common is a distinction between comprehension at the lexical, morphological, or syntactical level, and comprehension at the contextual, inferential, or interpretive level. Clark and Clark (1977) claim that comprehension is a construction process, where meaning and interpretation are formed by the listeners. This is accompanied by a simultaneous utilization process, where the interpretation is put to use for communication or other higher order operations.

In a sense, bottom-up (i.e., local, micro, or text-based) processing is more closely associated with explicit material in a lecture or passage, while top-down (i.e., global, macro, or situational) processing is linked to implicit knowledge that is either beyond the scope of the text, or spread out across different ideas within the same passage. Field (2008) talks about this in terms of decoding lexical knowledge versus meaning building through what he calls meaning enrichment or information handling. The former is based upon the listener drawing on background knowledge to make sense of what is heard, while the latter is the process of assigning importance and connecting ideas brought up in speech.

The temptation is strong among researchers to assume that differences in processing are equated to a well-defined hierarchy in terms of difficulty and proficiency level, with bottom-up processing occupying the lower end of the scale, and top-down on the opposite end. Indeed, several studies have reported on this very notion (Hansen & Jensen, 1994; Wagner, 2002), though the results have not borne out this theory as nicely as expected. Shohamy and Inbar, in a study on assessing listening comprehension, found that local items were easier for examinees than global items. In a separate study by Osada (2001), this difference is accounted for by the idea that listeners devote so much attention to decoding lexical forms that suitable resources for top-down processing are too few. However, Tsui, and Fullilove (1998), in a study of Hong Kong university students, found that questions related to bottom-up processing were better indicators of listening ability than top-down processing items. Studies in reading, also found that it was lower level learners who relied more on context and top-down processing as a crutch for bottom-up problems, such as unfamiliar text and vocabulary (Rost, 2002; Perfetti & Roth, 1981). Instead of bottom-up processing acting as a prerequisite for top-down processing, it is better understood

as an alternative route towards understanding at the level of context and schema (Kintsch, 1998; Olsen & Huckin, 1990).

One solution to this problem of separating out processing levels is to consider them as interactional, and working simultaneously towards the goal of comprehension (Kelly, 1991; Buck, 2001), and indeed most of the dichotomous models work under this assumption (Clark & Clark, 1977; Van Djik & Kintsch, 1983; Vandergrift, 2007). In this model, both top-down and bottom-up processing are activated by the learner as is needed for comprehension. The degree to which either top-down or bottom-up processing is utilized depends upon (a) the person, (b) the degree of accessible background knowledge, (c) the text or lecture, and (d) the purpose of the task (Vandergrift, 2007; Buck, 2001; Grabe, 1991). Proficiency is no longer tied to lexical knowledge, scheme familiarity, or any singular factor of processing, but is understood as the interplay between these factors and the available processing resources.

This still leaves a question of how to determine proficiency through an understanding of processing. While there may be no definitive answer, one interesting area of exploration is the relationship between metacognitive skill use and processing.

Metacognitive Skills

One way that learners approach comprehension in listening, as well as other language skills, is through metacognitive skills and strategies. It is first important to draw a distinction here between skills and strategies. In general, strategies differ from skills in that strategies are compensatory and self-conscious, whereas skills are unconscious and automatic (Field, 1998; Phakiti, 2003). Phakiti claims that where strategies exist between conscious incompetence and conscious competence, skills can only be understood as unconscious competence (2003, p. 683). In other words, when a learner can automatically process some piece of knowledge, be it cognitive (i.e., lexical understanding) or metacognitive (i.e., monitoring information), this is considered skillful competence. The main difficulty arises in that determining whether a learner is accessing strategic knowledge or skillful knowledge is impossible by observation (Hudson, 2011). While they operate on different levels of consciousness, the outcome of both skills and strategies seems the same, with the exception that one requires more processing resources than the other. This again brings us back to the question of processing.

It is also important to define what is meant by metacognition. Metacognition is that which

considers cognition from the point of view of having knowledge about the cognitive process itself (Zimmerman, 2008; Plonsky, 2011). Unlike cognitive strategies or skills, which facilitate comprehension and achievement of a certain task (e.g., understanding vocabulary, storing speech), metacognitive strategies or skills help provide knowledge about the learning process itself. Zimmerman (1986), in the field of cognitive psychology, proposed a list of metacognitive skills and strategies, including goal setting, reviewing, and organizing, though this is far from an exhausting list.

One issue that arises in relation to the definition of skills and strategies is that there is no agreement about the specific number or division of skills and sub skills (Song, 2008; Alderson, 2000; 1990; Buck & Tatsuoka, 1998). Flavel, Miller, and Miller (2002) suggest that metacognitive knowledge can be broken down into three distinct parts: persons, tasks, and strategies. Buck and Tatsuoka (1998) found 15 skills to be important for listening comprehension performance, though Buck (2001) also says that any determination of a set number will always be nowhere near perfect, and skills in this sense may be more related to language activities and not what is actually utilized by a learner. In the field of reading, too, different researchers suggest different breakdowns of skills and strategies (Alderson, 2000). These extensive taxonomies are further complicated by the fact that most of them are rooted solely in theoretical terms, without any real empirical foundation (Buck & Tatsuoka, 1998; Buck, 2001; Skehan, 1984).

The simplest division of skills typically comes down to a two-stage process (Carroll, 1972; Buck 2001; Clark & Clark, 1977), which again comes back to the original concepts of processing at a micro/macro or local/global level. In addition, several studies have linked metacognitive strategy use to top-down and global processing (Vandergrift, 2003a; 2003b; Chamot, 2005; Vandergrift, Goh, Mareschal, & Tafaghodatari, 2006; Graham, Santos, & Vanderplank, 2008). Vandergrift (2003a), in a study of awareness raising in listening comprehension tasks, found that learners used metacognitive strategies to facilitate top-down processing. However, Vandergrift admits that because data were gathered through reflection, it is still difficult to determine the actual extent of metacognitive use and top-down processing. In further studies by Vandergrift (2007, 2003b) he finds that learners that focus on bottom-up processing alone are unable to engage in metacognitive strategies at all, as decoding and translation become the only focus for comprehension. A study on vocabulary size and listening proficiency by Staehr (2009) finds that more vocabulary can lead to further automation of bottom-up processing, which in theory allows

for more access to top-down processing, though this is never confirmed in the research.

While the connection between metacognition and processing appears to be positive, there still exists a considerable amount of confusion as to just how to define particular skills or strategies. This becomes a real danger when test designers set out to assess a learner's use or knowledge of a particular skill or strategy. Without a clear definition, a construct that claims to be testing one skill might in fact end up testing several interacting skills (Alderson, 1990). Without a doubt, this makes the assessment of skills a very tricky undertaking, as there is little way to determine what is being tested with certainty (Anderson, Bachman, Perkins, & Cohen, 1991; Graham, Santos, & Vanderplank, 2011).

Graham et al. (2008) in a longitudinal study of two learners' strategy use found that metacognitive strategy use (and one assumes skills, as well) is a highly individualized process, and that different learners will activate different strategies depending on the text and the task at hand. Graham et al. (2011) later found that even when teachers gave instruction on particular strategies, students still employed different skills for comprehension. Similarly, in the field of reading, Alderson (2000) points out that skills do different things for different learners, in that one test-taker may have a certain set of skills or strategies available, and will in turn use those to resolve a difficult item, while at the same time another test-taker may have a completely different set of skills and yet still be able to utilize them to answer the same item. The amount of freedom in selecting and employing strategies is vast, and given the receptive nature of listening and reading, is no easy feat to pinpoint how an examinee approaches items on a test.

Task-Based Listening Construct

One solution to the problem of designing a listening construct for strategy or processing use is to focus first on the contextualized task itself and the actual target language use that is assessed. Rather than trying to begin with a specific processing level or strategy, which lack definition, a test designer is better served by considering the actual language use the test-taker is expected to display (Buck, 2001; Buck & Tatsuoka, 1998). As language use is built into task design, there is less need to define all the specific components of a strategy or skill (Brindley & Slayter, 2002; Bachman & Palmer, 1996). As long as the task engages the same skills as the target language, then there is a clear focus for creating a task-based listening construct (Buck, 2001, p. 106).

Rather than competence alone, this is a construct based on performance that takes into account the interaction of the task and the test-taker. This is essential for accounting for the individualism in strategy and skill use previously mentioned (Buck, 2001; Dunkel, Henning, & Chaudron, 1993; Jensen, Hansen, Green, & Akey, 1997). Buck (2001) claims that this interaction, "between the test-taker and the test task is similar to the interaction between the language-user and the task in the target-language use situation. In other words, they should require similar competencies" (p. 108). Even if two different test takers approach the same task in a different way with different skills or processing methods, if the task is matched to the target language, the appropriate construct should be operationalized. This allows the test itself to become generalizable to the authentic language situation, which "links authenticity to construct validity, since investigating the generalizability of score interpretations is an important part of construct validation" (Messick, 1996. p. 24). Authenticity is important for construct validity in that when a task on a test matches authentic use, then there is less of a chance that the task is testing invalid constructs.

There is still the need to be careful using tasks as the basis for construct validity, as even in authentic tasks there will always be some effect of the test on the task (Bachman & Palmer, 1996). This occurs simply because a test is not the same as an actual language situation, so authenticity can almost never be perfect. It is important, then, that task designers move carefully and account for possible variations within the task in terms of characteristics of input. This input can include the level of text in the lecture, topical knowledge, speed, accent, length, item type, item availability, and more (Buck, 2001). For the ALT, the lectures are already set, as well as the format of the test (e.g., item availability), which means that control over the input is limited here to item type alone, and that will be the focus of this study.

As this study is not creating an entirely new form of listening test, but rather only seeking to address the separation between intermediate and advanced level listening comprehension courses, the type of items being piloted will not cover the full scale of listening comprehension. Instead, it is the focus of this review to concentrate only on those tasks and target language uses that are applicable to the advanced listening course, and the subsequent listening-focused SLOs.

The current items on the ALT appear to address general comprehension successfully, based primarily on natural comprehension arriving out of natural production (Clark, 2007). When Clark designed the items, rather than working from the script alone, graduate students familiar with the

program took notes on the lectures. Information that was salient across all raters was used as the basis for information to be assessed. Indeed, this appears to relate to the SLOs of the intermediate course, based on comprehension and application, and possibly effective strategy or skill use at the level of note taking. However, it does not appear to address the SLOs or tasks associated with the advanced listening course, particularly in terms of critically evaluating speakers or tasks requiring more advanced strategies.

Despite the presence of two SLOs specifically tied to listening, as previously mentioned, only one of them seems to present an actual, definitive task at the advanced level that is wholly different from those at the intermediate level: critical listening. While the other applicable SLO mentions a difference between 'good strategy use' at the intermediate level and 'effective strategy use' at the advanced level, according to the 'Level Separation Chart' (Appendix B), the implication between this difference is that the former means making learners aware of possible strategies, while the latter means helping learners identify useful strategies and apply them to listening comprehension tasks. Since it has already been established that learners use strategies individually and unexpectedly (Graham, et al., 2011; Barnett, 1988), making this distinction at the level of a task within the test appears fruitless at this stage.

Critical listening, however, is a task that is identifiable as only being associated with the advanced listening course. According to the SLO, the task is defined as being able to 'critically evaluate speakers' perspectives, techniques, and arguments'. In the context of the EAP program, this means that listeners should be able to go beyond lexical comprehension alone and begin to question and analyze what is presented in lectures and other academic listening situations. This seems to require skills such as connecting ideas, determining the consistency of an argument, and identifying important information. However, before looking at these skills, it is again important to focus first on the task itself, and one of the primary characteristics of this task is that learners are to respond to what is heard, rather than just acknowledging it.

In an academic setting, the input from academic lectures is typically first responded to through the medium of notes. In many EAP courses, note-taking is one of the primary and most important skills identified by teachers and learners alike (Lynch, 2011; Ferris, 1998), though it is true that not all students utilize notes, and the degree to how much is recorded is very much dependent on the individual and the input received (Badger, White, Sutherland, & Haggis, 2001). In the ALT, as well, before students have access to the items for a particular lecture, they are

instructed to listen and take notes on what they hear.

Research has shown a connection between note taking and proficiency (Tsai, 2004; Lynch, 2011; William & Eggert, 2002). Organization in notes, in particular, has been found to be linked to higher processing abilities (Song, 2011), task performance (Jung, 2006; Tsai, 2004) and differentiating between important and trivial information (Cushing, 1991). Song (2011) compared the quality of notes across 257 test takers on an EAP placement test and found that when notes were written at a high quality it indicated a high level of proficiency. However, whether or not the opposite was true or not was difficult to determine, mostly because lower quality notes do not indicate a lack of understanding, but rather might be a sign of a high capacity for memory. She also mentions that while detailed notes are good for lexical understanding, organization in notes is more closely associated with higher functions, such as making inferences. Organization stands out, then, as one possible skill that makes up this task-based construct of critical listening.

Organization is closely tied to inference, as the ability to interpret a speaker's ideas or apply it to outside information requires the ability to recognize structure and patterns within a lecture or passage (Buck, 2001; Song, 2011; Hudson, 2011). Inference in general is a tricky topic to approach, as interpretations of this sort are oftentimes based in very individualized contexts, and what one learner may infer from a text could be completely different than the inference of another (Buck, 2001). It is also sometimes confused with students guessing at the meaning of a passage versus inferring the meaning because of a deficiency in bottom-up processing. For critical listening, however, inference might be better understood as the ability to connect ideas and establish relationships (Field, 2008). Once an inference goes too far beyond the explicit and implicit information of a task, then it is open to interpretation. Therefore inferencing must be tied to the task at hand (Hudson, 2011). By thinking of inference as understanding the larger relationships within a task or a lecture, then it becomes less about interpretation and guesswork, and more about understanding the speaker's intent.

Connecting ideas, then, is another possible skill that is closely related to the task of critical thinking. This also requires a recognition of discourse structure, as well as the ability to look beyond literal and explicit meanings alone towards more implicit ideas (Field, 2008; Alderson, 2000). Learners that are only decoding lexical information are made blind to the larger meanings across a passage, allowing no chance for deeper interpretations or connections, both in listening

(Field, 2008) as well as reading (Garner & Reis, 1986; Jimenez, 1996). This is also related to how learners are able to separate out important versus trivial information, in that learners that treat information as equal are likewise unable to determine which information is related to the main idea of arguments put forth by a speaker or text, or check for consistency (Goh, 2000; Field, 2008).

It seems possible to move forward with this task-based construct of listening comprehension for critical listening now that several possible skills have been identified: (a) organization of ideas, (b) connecting ideas, and (c) determining the importance of ideas. It remains a possibility that other skills or strategies might also account for a learner's ability to complete this task, but given the characteristics of the task and the close relationship between these skills and strategies, it is likely that a verifiable construct exists. To this end, this study attempts to answer the following research questions:

1. How does the current form of the ALT perform in terms of matching the ability and difficulty of the population of examinees, reliability of the scores on the test, and item and person fit?
2. What constructs are represented in the current form of the ALT?
3. How do items based on the skills of organization, connecting ideas, and determining the importance of ideas perform on the ALT in terms of matching the population of examinees, reliability of scores on the test, and item and person fit?
4. How do the piloted items compare to the items they replaced?

METHODS

Participants

The data for this study were collected in two stages, once for the current form of the ALT, and again for a revised version of the ALT containing five piloted items. Data for the current ALT were collected from 502 examinees of four past administrations[1] of the ALT between fall 2010 and spring 2012. In addition, data were collected from 141 examinees for the revised ALT, which was administered in fall 2012. The population of test-takers is made of all incoming

[1] The ALT is administered several times at the beginning of each semester. For the purposes of this study, data for one administration includes all tests given prior to a semester.

international students who do not speak English as their native language and who have paper-based TOEFL scores of between 500 and 600, or computer-based TOEFL scores of between 173 and 250. Participants include students at both the undergraduate and graduate level, as well as one- or two-semester international exchange students who are only entering the university on a temporary basis.

Materials

The ALT is one of two listening comprehension subtests that are part of a larger placement exam for an EAP support program within a North American university. Examinees are placed into one of two listening courses based on scores from the ALT, a dictation test, and their listening score from the TOEFL[2]. For the ALT, examinees that receive a score equal to or below the normed mean are placed into the intermediate level course, while examinees who receive a score between the mean and one positive standard deviation are placed into the advanced course. Scores above one standard deviation are qualified for exemption from academic listening support classes.

The test consists of 35 multiple-choice items based on five unscripted listening passages that are designed to resemble authentic academic lectures similar to those encountered by students in regular university classes (Clark, 2007). The first three passages are short lectures of between three and five minutes in length, while the final two lectures are longer, with runtimes of around eight minutes each. Examinees listen to a single lecture and are encouraged to take notes about what they hear, though they are not allowed to look at the items until the lecture is finished playing. Examinees are then given a limited time to answer items about the lecture based on what they remember and what information they wrote in their notes. The content of the lectures is based on a variety of different genres to account for the variation in background knowledge of the test-takers and prevent bias effects for certain fields of study. Examinees listen to each lecture only once.

Two versions of the revised ALT were developed by the researcher, with the intention of piloting five unique items on either form of the test, for a total of ten new items. The original listening passages were retained. The researcher selected five items from the current ALT as candidates for replacement based on Clark's initial analysis of the test. Clark identified four

[2] Final placement decisions are based on the examinee's highest score among these three measures.

items (A1, B1, C1, and E7) as not fitting the test using Rasch analysis. Item E5 was also identified by in the present study as a good candidate for replacement as it is one of three items currently not scored on the test (along with items B1 and E8), based on a separate analysis that showed these items to have a skewing effect on the placement results (Harsch, 2012, personal communication). Item E8 was kept based on findings from Clark's analysis of overall fit for this item, and with five more suitable items already identified for replacement, the researcher thought it too aggressive to remove any more items from an already short test ($k = 35$), fearing an effect on reliability and placement results.

The original numbering from the current ALT was preserved whenever possible, however, three of the five replaced items all shared a similar position in their respective question banks, and in order to avoid the chance of examinees deducing which items were scored from those being piloted, the numbering of the items was changed slightly. For ease of analysis, the numbering and scoring of both pilot tests were identical (i.e., pilot items for both tests were set as items A5, B4, C6, E7, and E8).

The fall 2012 ALT was administered four times over the course of two weeks. In order to capture an approximately equal sample size for each version of the test, one version (Pilot A) was used for the first two administrations ($N = 78$), and the second version (Pilot B) was used for the final two administrations ($N = 63$). Because the ALT is part of an active placement test, the revised items on each test were not scored for use in placement decisions.

Procedures

Revisions to the ALT for the fall 2012 administration involved the development of new items meant to represent a task-based construct of critical listening as operationalized by the skills of organization, connecting ideas, and determining the importance of ideas. To accomplish this, the researcher began by compiling available content from the lecture passages on the test. Following Clark's (2007) original design process for items on the ALT, the researcher acted as a model comprehender to draw out "actual instances of comprehension" (p. 145). This was done by listening to the lectures one time each while taking notes, then creating a summary of what was presented based on those notes and any other information from memory. The purpose for this was to more closely match the conditions of the actual test, where repeated listenings and highly detailed notes are not available. However, because the researcher had a specific goal in listening

for major themes in the passages rather than lexical details, there is a danger that this was not in line with how an actual test-taker might listen during the test. At the same time, since being able to listen for specific details as well as larger concepts should be skills possessed by advanced listeners, this discrepancy is assumed to not pose a real threat to the overall content validity of items developed in this way.

Items were next developed based on information available from both summaries and notes, which included information expressed both explicitly and implicitly in the lectures. In order to match content to skills for organization, connecting ideas, and determining importance, Field's (2008, p. 246) concept of information being integrated into discourse through (a) connecting information, (b) comparing information, and (c) constructing information was used as a guide. The researcher looked for areas in the content where several main ideas were spread out across different sections of the passage, or where important, related themes were separated by specific details. Other possible areas of interest were places where ideas were compared within a passage, or multiple, seemingly unrelated ideas were brought up in the same passage. In this way, connections represented these relationships among ideas, comparisons represented determining importance, and construction —which Field considers the formation of argument structure—is represented by organization of ideas. Using this framework, two new items were developed for each removed item for a total of ten pilot items. Each item was matched to one of three target skills, as shown in Table 1.

Table 1

Distribution of Item Type by Construct

Construct	Number of Items
Organization	2
Connecting ideas	6
Determining the importance of ideas	2

Organizational items were designed based on information within lectures where details for a singular idea were spread around the lecture in different areas and intermixed with other important details. Example (1) displays an organizational item for lecture A. For this type of item, the stem is asking the test-taker to reconstruct an entire idea from pieces laid out within the passage. It requires non-linear thinking, which in turn requires listeners to move beyond the level

of decoding (Field, 2008).

(1) Based on the lecture, what is the process of checks and balances in a law?

A. Congress creates a law, the president approves it, and the supreme court judges the law as constitutional.

B. Congress makes a law, the president approves it, and the supreme court enforces it.

C. The president writes a law, congress ratifies it, and the supreme court enforces it.

D. The president makes a law, congress vetoes it, and the supreme court judges the law as constitutional.

Items for connecting ideas were developed from instances where multiple main ideas were spread out across the length of a lecture, typically involving ideas that showed some relation to the topic of the lecture itself. Example (2) shows an item of this kind from Lecture E. In this case, the test-taker is asked to relate the two main topics from the lecture and ask about the relationship between them. Because of the length of this particular lecture (8:00+ minutes), it was easier to pinpoint distinct concepts from the passage that require examinees to consider how the details for each of these concepts creates a relationship between them that is not explicitly stated.

(2) How is the Drake equation connected to Fermi's paradox?

A. The Drake equation is necessary to answer Fermi's paradox.

B. The Drake equation proves the existence of extraterrestrial life.

C. Drake and Fermi worked together to develop the Drake equation.

D. Fermi's paradox proves the Drake equation to be true.

Lastly, items were developed that required examinees to discriminate between important and trivial details. These items were closely tied to explicit information within the lectures, and took the form of items that asked about locating support for an argument or idea presented by the speaker. Example (3) shows an item from Lecture B that asks test-takers to locate a supporting idea for a relationship mentioned by the speaker later in the lecture. The stem inquired about the relationship between two concepts, both of which are described separately. The examinee's task is to then identify the central idea of each outside of any other extraneous information provided.

(3) What best describes the relationship between branding and advertising?

A. Branding reinforces successful advertising

B. Advertising reinforces branding

C. Advertising uses publicity to create a brand

D. Branding begins by advertising alone

The final step in the development phase was to examine each item in terms of problems of bias, specifically in reference to objections made for the use of inference in items. All items needed to be answerable based on the content of the lectures alone, and not otherwise reliant upon outside or background information (Brown, 2005). In particular, items that asked about implicit connections still needed to be based off explicit information within the passage. To accomplish this, the researcher made transcriptions of each lecture, and then each item stem and correct response was successfully located within the lecture. It is thus assumed that no apparent bias of this type exists for these items.

A draft of the items, along with their paired skill-type, was sent to the director of the EAP program and an independent testing expert. Based on their combined feedback, the items were revised again by the researcher to improve clarity and precision, as well as remove any areas of ambiguity in distractors or stems. The final set of ten items as they appear in the pilot tests is included in Appendix C.

Analysis

Rasch model analysis was conducted on both the current version of the ALT, as well as the revised pilot versions of the test using the analysis program *Winsteps* (Linacre, 2010). Rasch analysis reports on the probability for a correct response on a test based on the relationship between an item's difficulty and the ability for an examinee to answer the question correctly (Bond & Fox, 2007). For example, a person has a 50% chance of answering an item correctly when both their ability level and the difficulty of the item are equal. Changes in either item difficulty or person ability will cause this probability to change accordingly (i.e., less ability or more difficulty will lead to a lower chance of success). Items and persons are arranged along an interval scale in terms of the expectancy of correct responses.

One of the main benefits of the Rasch model is that because it is based on probability, and items are defined along a fixed interval scale, it is possible to understand how items perform

independent of a single sample of examinees. Classical test theory (CTT) is limited to interpreting only a particular set of items by the given sample of examinees (Brown, 2012). Therefore, changes made to the test can also only be understood if the same sample is used again, making it impossible to generalize findings of reliability or item function (Henning, 1984). Rasch analysis makes it possible to revise items on the test and pilot them with a different sample of examinees and still understand how the items themselves are functioning.

Rasch analysis also provides information about model fit statistics in terms of item and person reliability, as well as response validity. CTT presents this information at a holistic level, but through Rasch, details for each individual item and person are revealed (Henning, 1984, 1985). Given that the current study is an investigation of different item types on a single test, this form of detailed, item-level information is invaluable for determining how these items are performing.

One limitation of the Rasch model is the requirement for unidimensionality, in that all items on a test should be explainable through a common factor (Bond & Fox, 2007). Rasch analysis is limited to understanding ability in one-dimensional terms, in that a person can only be measured in terms of having a degree of high or low ability in X, but not in Y, or X and Y at the same time. If the difference in their responses is caused by something other than ability in X, or a combination of effects, then the model will not work, and items or persons will be identified as misfitting.

Factor analysis was also run on the current version of the ALT using IBM SPSS 21. This form of analysis looks at correlations between items and matches those items that correlate together but independently from other items (Tabachnick & Fidell, 2013). This form of analysis is useful in testing as a way to interpret convergent-divergent construct validity for items on a test based on how items load together (Brown, 2010). A construct can be said to exist when certain items converge on a singular loading that all test the same kind of thing, while also diverging from other items and loadings that test different things. If a test is assumed to have multiple constructs, factor analysis can show whether or not items are actually representing those constructs in the test, thus satisfying concerns about validity.

Two primary forms of factor analysis are often used for understanding construct validity: principle components analysis (PCA) and exploratory factor analysis (EFA). PCA accounts for all associated variance of the solution, including unique and error variance. For this reason, PCA

is primarily used when there is no underlying theory about the number of expected factors in a solution (Brown, 2009). In contrast, EFA only accounts for common variance within the solution and is therefore used when the analysis is theory-driven.

RESULTS

Rasch Analysis

Before running Rasch analyses, data were examined for missing data or outliers, as the presence of either can result in a distortion of the analysis. Outliers consist of those examinees that received either a perfect score or a score of zero correct responses on the test. Only one individual in the group of examinees for the current ALT scored perfectly on the test. Data for this individual was not used, making the final number of participants for analysis of the current ALT 501. No other missing data or scores of zero correct responses were found on any form of the test. Due to an error on the part of the researcher, item E5 on Pilot A was found to be a different question from item E5 on Pilot B. Item E5 was a scored item, and all scored items on the ALT were intended to be identical across both tests. However, because of this error, and the fact that item E5 on Pilot A had a different correct answer than item E5 on Pilot B, responses for these items on both tests were removed from the overall analysis, making the final item count for the revised ALT 34.

Summary statistics of the Rasch analysis for the ALT between 2010 and 2012 show that there is a good match between person ability and item difficulty on the current version of the test. Table 2 shows person statistics along an interval logit scale, with a mean ability score of 0.57, a standard deviation of 0.77, and a range of -1.38 to 3.81 logits. Comparable findings are also found for items (Table 3), with a mean item difficulty of 0.00[3] a standard deviation of 0.80, and a difficulty range of -1.23 to 1.26. The person-separation reliability is also displayed for the model, with a value of 0.73. This is analogous to Cronbach alpha, and indicates that the test scores are 73% consistent.

[3] This value is set by default to 0.00 by the model.

Table 2

Summary of 501 Measured Persons for the 2010-2012 ALT

	Total Score	Measure	Model Error	Infit		Outfit	
				MNSQ	ZSTD	MNSQ	ZSTD
Mean	21.40	.57	.39	1.00	.00	1.00	.00
SD	5.00	.77	.05	.13	.90	.20	1.00
Max	34.00	3.81	1.02	1.41	2.80	2.06	2.80
Min	8.00	-1.38	.36	.64	-3.10	.53	-3.00

Note. Winsteps v3.70.0.1
Person Reliability = .73

Table 3

Summary of 35 Measured Items for the 2010-2012 ALT

	Total Score	Measure	Model Error	Infit		Outfit	
				MNSQ	ZSTD	MNSQ	ZSTD
Mean	305.70	.00	.10	1.00	.00	1.00	.10
SD	79.60	.80	.01	.06	1.80	.11	1.90
Max	418.00	1.26	.12	1.13	4.30	1.26	4.40
Min	175.00	-1.23	.09	.87	-3.40	.85	-3.20

Note. Winsteps v3.70.0.1

Another way of viewing the results of the Rasch analysis is through a vertical ruler (Figure 1), which provides a more visual understanding of the relationship between ability and difficulty. Both person ability and item difficulty are placed onto the same logit scale, with an average difficulty standardized at 0.00 logits. Persons are shown on the left, as represented by pound (#) signs to indicate three people or dots (.) to represent one or two people. Items are displayed on the right, and are listed by listening passage (A-E), followed by the corresponding number within the passage (1-9). Increasing ability and difficulty levels are represented as moving up the scale from negative to positive logit values.

Both the summary statistics and Figure 1 show that person ability and item difficulty are well distributed and well matched overall. However, it appears that a mismatch is occurring between 1.25 and 2.00 logits, as there are persons with this ability level but no comparably difficult items. This might be problematic in that the current ALT is partially incapable of accurately measuring students with advanced listening abilities. However, the number of examinees with ability scores about 1.25 is rather few overall, and given that this is a placement test, it is not necessary to measure people of abilities that are clearly higher than the aims of the program.

Item and person fit to the model were also analyzed to identify instances of model misfit or

overfit. Misfit occurs when items or persons are responding in an unexpected manner based on the probability model of the analysis. In other words, when persons of higher ability repeatedly respond incorrectly on an item of lower difficulty, this item is understood as functioning in a way that cannot be adequately predicted by the model, and thus does not fit into the model. In contrast, overfitting items or persons fit the model too well (i.e, the odds are right every time).

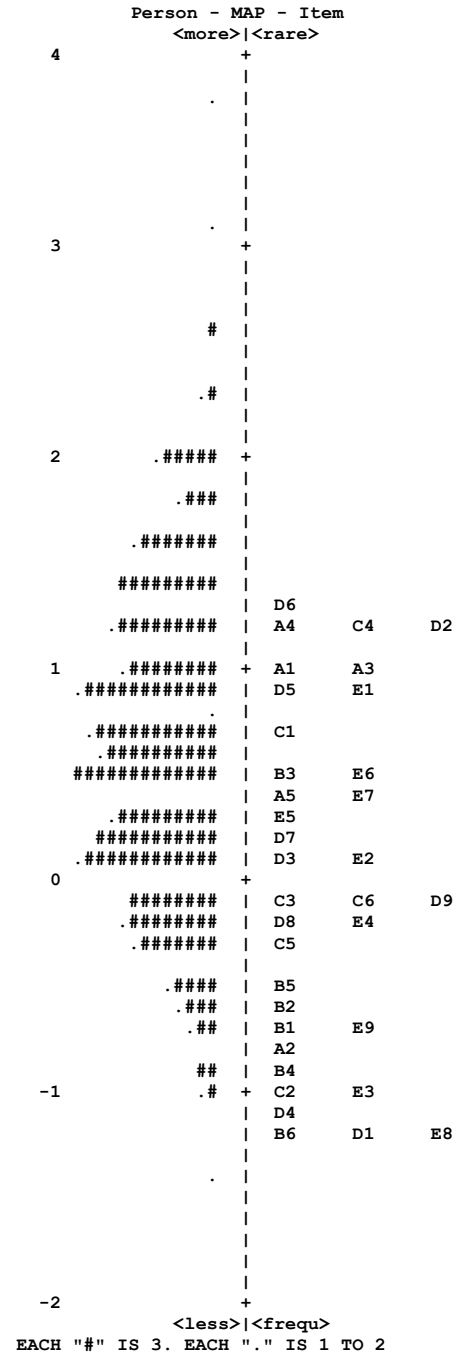

Figure 1. Item map for the 2010-2012 ALT

Item fit statistics are displayed in Table 4. Measure in the table indicates the logit difficulty value for each item, followed by infit and outfit statistics represented as mean-square (MNSQ) values and standardized *z*-scores (ZSTD)[4]. Infit is derived from instances when a person of a particular ability gets an item of matching or lower difficulty incorrect, while outfit indicates difficult items that persons lower ability are getting correct. Estimates to whether or not an item is overfitting or misfitting are done by taking twice the standard deviation of the infit *MNSQ* value (*SD* = 0.06) and adding or subtracting it from the mean infit *MNSQ* (*M* = 1.00). Using this formula, items that are misfitting have an infit *MNSQ* of 1.12 and higher, and items that are overfitting have an infit *MNSQ* of 0.88 or lower. According to Table 4, items E7 and C1 (*MNSQ* = 1.13 and 1.12 respectively) are seen as misfitting the model, while item A4 (*MNSQ* = 0.87) is the only overfitting item. Similar analysis of person fit statistics (Appendix D) show an upper bound *MNSQ* of 1.26 and a lower bound *MNSQ* of 0.74 (*M* = 1.00, *SD* = 0.13). Based on these values, 11 people are misfitting the model, and 10 people are overfitting the model. Misfitting items such as E7 and C1 are problematic, as they indicate that examinees are not answering these items in ways that would be expected based on the probability estimates of the model. This could mean the items are testing a different construct, and thus violating the assumption of unidimensionality of the test, or there is a guessing factor associated with the items. Persons classified as misfitting the model are unfortunately more difficult to explain based on analysis alone, and as such there is little interpretation that can be gained from these particular findings. Were the ratio of misfitting persons higher in regards to the entire population, it might indicate a problem with the test and how well it is actually deemed suitable for the population. However, 11 out of 501 examinees does not seem to be cause for alarm. Overfitting items and persons are interesting only in that they are working too well in the model.

[4] Point-measure correlation (PT-MEASURE CORR) is analogous to item discrimination in classical testing. Rasch analysis is not concerned with this statistic for the purposes of model fit, however.

Table 4

Item Fit Statistics for the 2010-2012 ALT

Person: REAL SEP.: 1.60 REL.: .72 ... Item: REAL SEP.: 7.52 REL.: .98

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
6	382	501	-.74	.11	1.10	1.8	1.26	2.9	A .12	.28	75.8	76.6	B1
16	337	501	-.25	.10	1.08	1.9	1.19	3.1	B .19	.31	69.1	69.4	C5
31	283	501	.26	.10	1.08	2.7	1.19	4.4	C .21	.33	60.5	64.7	E5
33	267	501	.41	.10	1.13	4.3	1.15	3.6	D .17	.33	56.5	64.2	E7
12	232	501	.73	.10	1.12	3.9	1.14	3.6	E .19	.34	56.7	64.5	C1
1	203	501	.99	.10	1.10	2.9	1.12	2.7	F .21	.34	60.7	66.3	A1
14	318	501	-.06	.10	1.07	1.8	1.11	2.1	G .23	.32	62.1	67.1	C3
18	418	501	-1.23	.12	1.01	.2	1.07	.6	H .22	.24	83.8	83.5	D1
20	295	500	.15	.10	1.06	1.9	1.06	1.3	I .25	.33	61.6	65.3	D3
17	324	501	-.12	.10	1.01	.4	1.06	1.2	J .29	.32	68.9	67.8	C6
5	264	501	.44	.10	1.06	1.9	1.05	1.3	K .27	.33	59.1	64.1	A5
35	380	501	-.72	.11	1.04	.7	1.05	.6	L .23	.28	76.2	76.2	E9
10	365	501	-.55	.11	1.02	.4	1.02	.3	M .27	.29	73.1	73.6	B5
30	332	501	-.20	.10	.97	-.7	1.01	.3	N .33	.31	71.3	68.7	E4
22	208	501	.95	.10	1.01	.4	1.01	.2	O .32	.34	64.9	65.9	D5
32	258	501	.49	.09	1.00	.0	.98	-.4	P .34	.34	62.9	64.1	E6
24	291	501	.19	.10	.99	-.4	.99	-.3	Q .34	.33	64.3	65.0	D7
25	330	501	-.18	.10	.99	-.3	.99	-.2	R .33	.31	68.9	68.5	D8
13	403	501	-1.01	.12	.99	-.2	.96	-.4	q .28	.26	80.8	80.5	C2
26	324	501	-.12	.10	.98	-.6	.98	-.4	p .34	.32	68.1	67.8	D9
15	180	501	1.21	.10	.96	-1.1	.97	-.5	o .38	.33	70.3	68.4	C4
34	413	501	-1.15	.12	.97	-.3	.88	-1.1	n .30	.25	82.8	82.5	E8
9	394	501	-.89	.11	.97	-.5	.91	-1.0	m .32	.27	79.0	78.8	B4
3	207	501	.95	.10	.96	-1.2	.95	-1.3	l .39	.34	66.7	66.0	A3
29	405	501	-1.04	.12	.96	-.6	.90	-1.0	k .33	.26	80.8	80.9	E3
7	370	501	-.60	.11	.95	-.9	.92	-1.1	j .35	.29	75.6	74.5	B2
11	418	501	-1.23	.12	.95	-.6	.86	-1.2	i .32	.24	83.8	83.5	B6
2	389	501	-.83	.11	.95	-.8	.91	-1.0	h .34	.27	78.4	77.9	A2
27	213	501	.90	.10	.95	-1.5	.93	-1.7	g .40	.34	68.3	65.5	E1
23	175	501	1.26	.10	.95	-1.3	.94	-1.1	f .40	.33	70.9	69.0	D6
21	407	501	-1.07	.12	.94	-.9	.85	-1.4	e .35	.26	81.6	81.3	D4
28	302	501	.09	.10	.93	-2.1	.91	-2.1	d .41	.32	69.7	65.7	E2
19	178	501	1.23	.10	.91	-2.5	.87	-2.6	c .45	.33	71.9	68.7	D2
8	256	501	.51	.09	.90	-3.4	.88	-3.2	b .46	.34	71.3	64.1	B3
4	177	501	1.24	.10	.87	-3.4	.86	-2.9	a .49	.33	74.5	68.8	A4
MEAN	305.7	501.0	.00	.10	1.00	.0	1.00	.1			70.6	70.8	
S.D.	79.6	.0	.80	.01	.06	1.8	.11	1.9			7.8	6.5	

Note. Winsteps v3.70.0.1

Rasch analyses were also run for each version of the pilot test. Rather than analyzing these tests independently from the findings of the current ALT, item anchoring was used for all common items between the current form of the ALT and both pilot versions. Anchoring allows the model to set the difficulty measure for common items on the pilot test to be based on those difficulties of the current ALT. With these items anchored in place, it is possible to understand how the piloted items are performing in the context of the working test, and not the pilot data alone.

Summary statistics of persons for the Rasch analysis of Pilot A are presented in Table 5, and for Pilot B in Table 6. Person ability scores were mostly consistent across both forms of the test, with logit means for Pilot A of 0.87 (*SD* = 1.00) and 0.69 (*SD* = 0.87) for Pilot B. The logit range

of abilities were almost identical across both tests, with Pilot A examinees between -1.25 and 3.69 logits of ability, and Pilot B examinees also between -1.22 and 3.77 logits of ability.

Items statistics are displayed in Tables 7 and 8. Item difficulty also appeared consistent across both tests. Item difficulty and person ability seem well matched, with a mean of -0.09 (*SD* = 0.78) for Pilot A, and -0.04 (*SD* = 0.85) for Pilot B. Because the most difficult item on the test is an anchor item (*M* = 1.26), there is still a discrepancy between person ability and item difficulty at the higher end of the logit spectrum. This appears to indicate that the piloted items had no apparent effect on the overall difficulty of the test. However, in terms of reliability, both This was expected, as examinees of both tests were similarly representative of the population. This appears to indicate that the piloted items had no apparent effect on the overall difficulty of the test. However, in terms of reliability, both tests were encouragingly improved over the current version of the ALT. Pilot A reported a person-separation reliability of .79, while Pilot B had a value of .76.

Table 5

Summary of 78 Measured Persons for the ALT Pilot A

	Total Score	Measure	Model Error	Infit		Outfit	
				MNSQ	ZSTD	MNSQ	ZSTD
Mean	22.90	.87	.44	.99	-.10	.99	-.10
SD	5.50	1.00	.12	.13	.90	.21	.90
Max	33.00	3.69	1.02	1.28	2.10	1.61	2.00
Min	9.00	-1.25	.37	.68	-2.40	.60	-2.30

Note. Winsteps v3.70.0.1
Person Reliability = .79

Table 6

Summary of 63 Measured Persons for the ALT Pilot B

	Total Score	Measure	Model Error	Infit		Outfit	
				MNSQ	ZSTD	MNSQ	ZSTD
Mean	21.60	.69	.42	1.02	.10	1.00	.10
SD	5.10	.87	.09	.15	1.00	.22	1.00
Max	33.00	3.77	1.02	1.44	2.90	1.58	2.80
Min	9.00	-1.22	.37	.67	-2.40	.43	-2.20

Note. Winsteps v3.70.0.1
Person Reliability = .76

Table 7

Summary of 34 Measured Items for the ALT Pilot A

	Total Score	Measure	Model Error	Infit		Outfit	
				MNSQ	ZSTD	MNSQ	ZSTD
Mean	52.60	-.09	.28	.98	-.20	.99	-.10
SD	12.20	.78	.03	.16	1.20	.33	1.30
Max	72.00	1.26	.34	1.35	2.80	2.37	4.70
Min	28.00	-1.23	.25	.60	-2.5	.44	-2.40

Note. Winsteps v3.70.0.1

Table 8

Summary of 34 Measured Items for the ALT Pilot B

	Total Score	Measure	Model Error	Infit		Outfit	
				MNSQ	ZSTD	MNSQ	ZSTD
Mean	40.10	-.04	.30	1.01	.10	1.00	.10
SD	10.30	.85	.04	.15	1.00	.22	1.00
Max	58.00	1.26	.41	1.30	2.20	1.43	2.60
Min	21.00	-1.64	.27	.59	-2.20	.46	-1.90

Note. Winsteps v3.70.0.1

Turning to the function of specific items on the test, fit statistics for all 34 items on each test are available in Tables 9 (Pilot A) and 10 (Pilot B). Anchor items are displayed by values with the letter "A" in the measure column. As before, item fit is determined using the model mean plus or minus twice the infit *MNSQ* standard deviation. Misfitting items have infit *MNSQ*s of 1.30 or higher for Pilot A (M = 0.99, SD = 0.16) and 1.31 or higher for Pilot B (M = 1.01, SD = 0.15). For Pilot A, only item D8 was found to be misfitting (*MNSQ* = 1.35), though this finding in itself was not altogether interesting as it can be explained by a problem during the administration of one of the tests for this item. During the second administration of Pilot A, the sound on the recording cut out where information pertaining to this item was found in the lecture. That this item was answered in unexpected ways is ironically expected, and if nothing else serves as a refreshing example of the Rasch model in action. One overfitting item was found for each version of the revised ALT, calculated using lower-bound *MNSQ* values of 0.66 and 0.71 for Pilot A and B respectively. Interestingly, a different item was overfitting for each version of the test despite the fact that both were anchored items from the current ALT. Item E6 (*MNSQ* = 0.60) was overfitting the model for Pilot A, while item D4 (*MNSQ* = 0.59) was

overfitting the model for Pilot B.

Table 9

Item Fit Statistics for the 2012 ALT Pilot A

Person: REAL SEP.: 1.88 REL.: .78 ... Item: REAL SEP.: 2.52 REL.: .86

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	DISPLACE	Item
25	50	78	-.18A	.27	1.35	2.8	2.37	4.7	A .05	.35	64.1	72.5	.34	D8
32	*53*	*78*	*-.05*	*.26*	*1.12*	*1.2*	*1.63*	*2.7*	*B .19*	*.36*	*74.4*	*70.9*	*.00*	*E7Pilot*
26	53	78	-.12A	.26	1.17	1.5	1.39	1.7	C .18	.35	67.9	71.7	.07	D9
11	63	78	-1.23A	.34	1.26	1.2	1.02	.2	D .35	.26	79.5	85.8	.41	B6
10	56	78	-.55A	.28	1.23	1.6	1.23	.9	E .27	.32	71.8	77.6	.30	B5
20	50	78	.15A	.26	1.15	1.5	1.10	.6	F .24	.37	60.3	69.2	.00	D3
5	*46*	*78*	*.41*	*.25*	*1.07*	*.8*	*1.15*	*1.0*	*G .31*	*.39*	*67.9*	*67.4*	*.00*	*A5Pilot*
28	51	78	.09A	.26	1.07	.8	1.14	.8	H .29	.37	66.7	69.7	.00	E2
16	54	78	-.12A	.26	1.00	.1	1.14	.7	I .33	.35	74.4	71.7	.00	C5
8	67	78	-.89A	.31	.93	-.3	1.13	.5	J .13	.29	88.5	81.9	-.36	B3
17	*62*	*78*	*-.74*	*.30*	*1.11*	*.7*	*1.13*	*.5*	*K .20*	*.31*	*76.9*	*80.0*	*.00*	*C6Pilot*
29	61	78	-1.04A	.32	1.08	.5	.88	-.2	L .47	.28	79.5	83.7	.40	E3
27	46	78	.90A	.25	1.07	.8	1.08	.6	M .35	.41	66.7	67.2	-.50	E1
18	66	78	-1.23A	.34	1.03	.2	.96	.0	N .30	.26	85.9	85.8	.11	D1
13	50	78	-.06A	.26	.99	-.1	.88	-.5	O .46	.36	70.5	71.1	.22	C2
1	62	78	-.83A	.30	.99	.0	.82	-.5	P .39	.30	79.5	81.1	.10	A1
15	60	78	-.25A	.27	.94	-.5	.98	.0	Q .27	.35	80.8	73.5	-.33	C4
7	40	78	.51A	.25	.97	-.3	.97	-.1	q .45	.39	73.1	67.0	.27	B2
4	49	78	.44A	.25	.97	-.3	.95	-.3	p .38	.39	69.2	67.2	-.23	A4
12	66	78	-1.01A	.32	.94	-.2	.97	.1	o .26	.28	83.3	83.4	-.12	C1
33	*60*	*78*	*-.57*	*.29*	*.97*	*-.2*	*.86*	*-.4*	*n .36*	*.32*	*79.5*	*77.9*	*.00*	*E8Pilot*
24	48	78	.19A	.25	.95	-.5	.86	-.8	m .46	.37	67.9	68.9	.09	D7
30	58	78	-.20A	.27	.83	-1.5	.94	-.2	l .42	.35	76.9	72.8	-.21	E4
6	63	78	-.60A	.29	.94	-.4	.86	-.4	k .28	.32	80.8	78.3	-.23	B1
14	30	78	1.21A	.25	.92	-.8	.87	-.9	j .47	.41	70.5	68.6	.20	C3
22	34	78	.95A	.25	.90	-1.1	.87	-1.0	i .50	.41	71.8	67.5	.20	D5
2	40	78	.95A	.25	.87	-1.5	.81	-1.5	h .54	.41	71.8	67.5	-.17	A2
9	*51*	*78*	*.09*	*.26*	*.87*	*-1.3*	*.83*	*-.9*	*g .49*	*.37*	*76.9*	*69.7*	*.00*	*B4Pilot*
23	31	78	1.26A	.25	.84	-1.7	.78	-1.7	f .56	.41	78.2	68.9	.08	D6
34	68	78	-.72A	.30	.79	-1.4	.64	-1.3	e .24	.31	84.6	79.8	-.66	E9
19	29	78	1.23A	.25	.78	-2.4	.72	-2.3	d .60	.41	76.9	68.7	.24	D2
3	28	78	1.24A	.25	.77	-2.5	.71	-2.4	c .60	.41	78.2	68.8	.30	A3
21	70	78	-1.07A	.32	.72	-1.5	.68	-.9	b .23	.28	91.0	84.1	-.58	D4
31	72	78	-1.15A	.33	.60	-2.2	.44	-1.7	a .27	.27	93.6	85.0	-.85	E6
MEAN	52.6	78.0	-.09	.28	.98	-.2	.99	-.1			75.9	74.3		
S.D.	12.2	.0	.78	.03	.16	1.2	.33	1.3			7.6	6.4		

Note. Winsteps v3.70.0.1

Table 10

Item Fit Statistics for the 2012 ALT Pilot B

Person: REAL SEP.: 1.72 REL.: .75 ... Item: REAL SEP.: 2.52 REL.: .86

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	DISPLACE	Item
6	40	63	-.60A	.31	1.30	1.8	1.43	1.6	A .35	.30	68.3	76.3	.64	B1
9	*25*	*63*	*1.16*	*.28*	*1.23*	*2.2*	*1.41*	*2.6*	*B .09*	*.36*	*63.5*	*67.3*	*.00*	*B4Pilot*
14	31	63	1.21A	.28	1.18	1.7	1.28	1.9	C .27	.36	60.3	67.9	-.49	C3
8	47	63	-.89A	.33	1.16	.9	1.27	.9	D .32	.28	74.6	80.2	.34	B3
12	46	63	-1.01A	.34	1.25	1.2	1.26	.8	E .40	.27	74.6	81.8	.56	C1
13	33	63	-.06A	.28	1.22	2.0	1.25	1.4	F .31	.33	57.1	69.0	.63	C2
34	52	63	-.72A	.32	.90	-.6	1.21	.8	G .14	.29	84.1	77.9	-.36	E9
17	*38*	*63*	*.19*	*.28*	*1.16*	*1.6*	*1.19*	*1.2*	*H .17*	*.35*	*57.1*	*66.6*	*.00*	*C6Pilot*
18	50	63	-1.23A	.36	1.14	.7	.94	.0	I .42	.25	79.4	84.5	.39	D1
23	30	63	1.26A	.28	1.09	.9	1.11	.8	J .39	.36	58.7	68.4	-.47	D6
19	23	63	1.23A	.28	1.09	.9	1.10	.7	K .25	.36	63.5	68.1	.08	D2
1	49	63	-.83A	.33	1.10	.6	1.07	.3	L .24	.29	77.8	79.4	.09	A1
10	46	63	-.55A	.31	1.09	.7	1.04	.3	M .26	.30	69.8	75.6	.09	B5
28	33	63	.09A	.28	1.08	.8	1.01	.1	N .41	.34	60.3	67.4	.48	E2
20	41	63	.15A	.28	1.07	.8	1.05	.4	O .23	.34	65.1	66.9	-.19	D3
15	51	63	-.25A	.29	.97	-.2	1.07	.4	P .07	.32	71.4	71.4	-.73	C4
25	41	63	-.18A	.29	1.01	.1	1.06	.4	Q .36	.33	68.3	70.4	.14	D8
22	31	63	.95A	.27	1.02	.3	1.04	.4	q .36	.36	63.5	65.8	-.24	D5
27	29	63	.90A	.27	1.03	.3	1.02	.2	p .34	.36	63.5	65.6	-.04	E1
7	37	63	.51A	.27	.97	-.3	1.00	.1	o .36	.36	69.8	65.3	-.24	B2
24	33	63	.19A	.28	.99	-.1	.92	-.5	n .46	.35	61.9	66.6	.38	D7
16	44	63	-.12A	.28	.92	-.7	.97	-.1	m .35	.33	74.6	69.7	-.17	C5
30	43	63	-.20A	.29	.97	-.2	.88	-.6	l .38	.33	65.1	70.7	.00	E4
32	*33*	*63*	*.57*	*.27*	*.95*	*-.6*	*.89*	*-.9*	*k .43*	*.36*	*58.7*	*65.3*	*.00*	*E7Pilot*
5	*33*	*63*	*.57*	*.27*	*.94*	*-.6*	*.92*	*-.6*	*j .42*	*.36*	*71.4*	*65.3*	*.00*	*A5Pilot*
26	39	63	-.12A	.28	.93	-.6	.84	-.8	i .50	.33	73.0	69.7	.24	D9
2	29	63	.95A	.27	.92	-.9	.92	-.6	h .46	.36	73.0	65.8	-.09	A2
11	55	63	-1.23A	.36	.91	-.3	.86	-.3	g .18	.25	87.3	84.5	-.26	B6
4	33	63	.44A	.27	.89	-1.2	.84	-1.2	f .49	.35	69.8	65.5	.13	A4
3	21	63	1.24A	.28	.86	-1.4	.82	-1.3	e .47	.36	76.2	68.2	.24	A3
33	*56*	*63*	*-1.64*	*.41*	*.83*	*-.5*	*.57*	*-.9*	*d .43*	*.22*	*88.9*	*88.9*	*.00*	*E8Pilot*
31	56	63	-1.15A	.35	.81	-.9	.74	-.7	c .17	.26	87.3	83.6	-.51	E6
29	56	63	-1.04A	.34	.74	-1.3	.65	-1.1	b .22	.27	87.3	82.2	-.62	E3
21	58	63	-1.07A	.35	.59	-2.2	.46	-1.9	a .25	.27	90.5	82.6	-1.00	D4
MEAN	40.1	63.0	-.04	.30	1.01	.1	1.00	.1			71.1	72.5		
S.D.	10.3	.0	.85	.04	.15	1.0	.22	1.0			9.7	7.1		

Note. Winsteps v3.70.0.1

The vertical ruler for Pilot A is presented in Figure 2. Newly piloted items are marked accordingly alongside anchored items. All items seem centered within one logit of the zero mark on the scale, indicating that none of the items are particularly easy or difficult in comparison with the test as a whole. Measure statistics from Table 9 confirm this, showing a range for pilot items between -0.74 and 0.41 logits. In terms of difficulty alone, these items show no apparent differences with any other items on the test. Items B4 and E7 are grouped with several other items of the same difficulty, which corresponds to a larger grouping of examines in terms of ability on the left side of the scale. Items C6 and E8 are similarly grouped together with other items, though there are comparatively less examinees of equal ability to be found at this level, possibly indicating redundancy for these particular items.

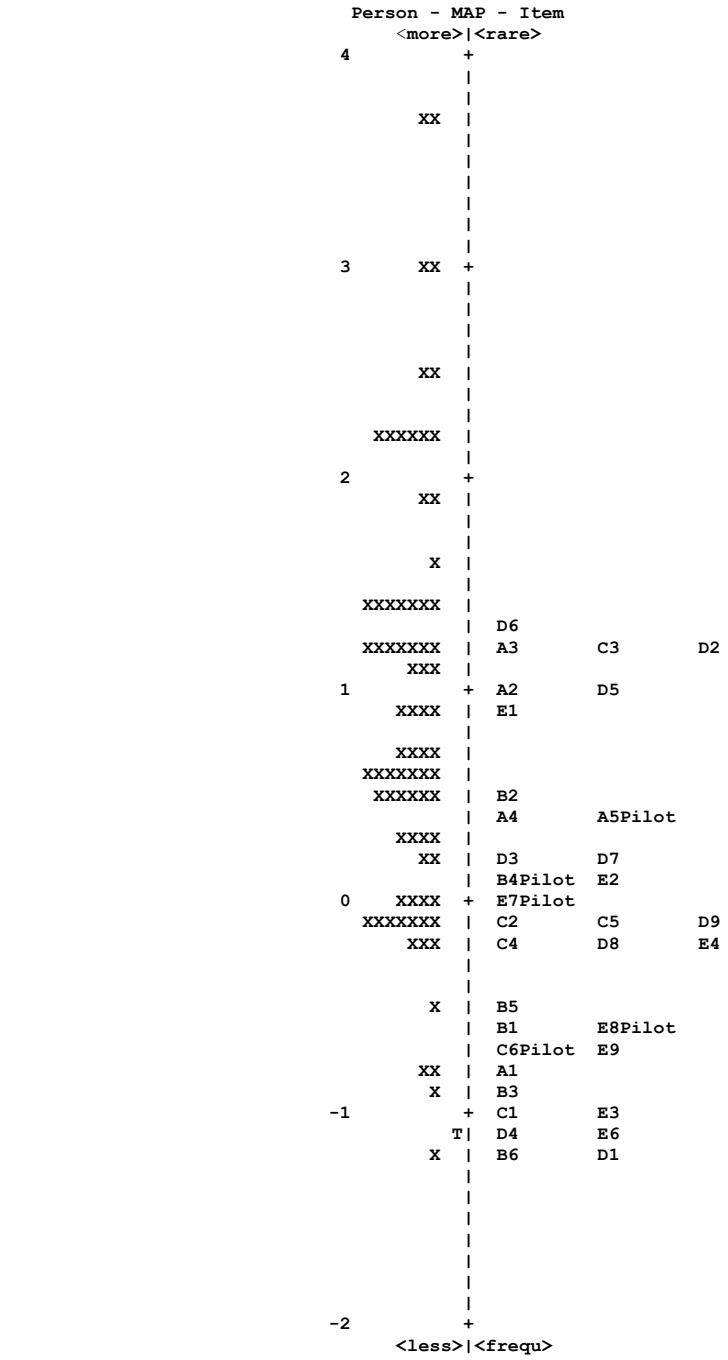

Figure 2. Item map for the 2012 ALT Pilot A

Pilot B is represented as a vertical ruler in Figure 3. Again, pilot items are marked in comparison to anchored items, but overall there appears to be more of a distribution of piloted items across the difficulty scale, with a range between -1.64 to 1.16 logits. Item E8 is far too easy

compared to the rest of the test, while item B4 appears to be acting at a more difficult level than most other items. Pilot items A5 and E7 occupy a place on the scale where there are few other items of comparable difficulty, which might indicate a need to keep or develop similar items of this kind.

Person misfit statistics are also presented in Table 11 for Pilot A and Table 12 for Pilot B. The same calculation as before is used in determining misfitting and overfitting persons for either test, with upper bound values of 1.25 and 1.32 logits and lower bound values of 0.73 and 0.72 logits for Pilot A and B respectively. Analysis of Pilot A showed no misfitting persons, though Pilot B had two misfitting persons. As this represents only 3.1% of the examinees on the test, these findings do not seem to warrant much concern for the test as a whole.

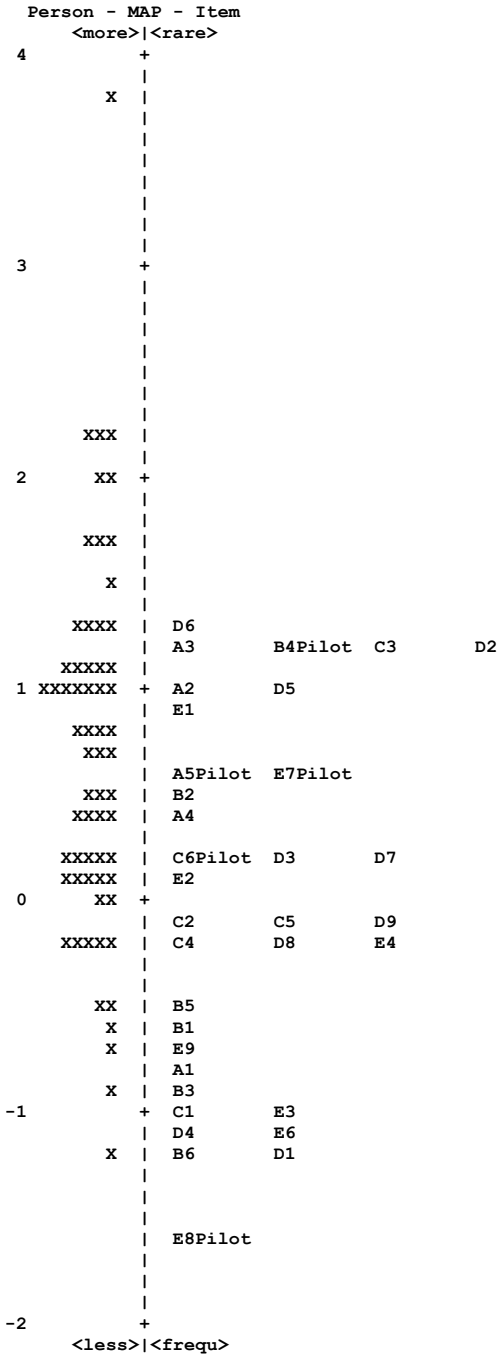

Figure 3. Item map for the 2012 ALT Pilot B

Table 11

Person Fit Statistics for the 2012 ALT Pilot A

Person: REAL SEP.: 1.88 REL.: .78 ... Item: REAL SEP.: 2.52 REL.: .86

Person STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
31	29	34	1.88	.50	1.19	.7	1.61	1.3	A-.05	.27	85.3	85.3	12144a
32	33	34	3.69	1.02	1.06	.4	1.45	.7	B-.02	.13	97.1	97.1	12125a
48	32	34	2.95	.74	1.11	.4	1.42	.7	C-.03	.18	94.1	94.1	12236a
51	30	34	2.15	.55	1.11	.4	1.41	.8	D .05	.24	88.2	88.2	12230a
12	30	34	2.15	.55	1.20	.6	1.38	.8	E-.04	.24	88.2	88.2	12115a
71	18	34	.04	.37	1.28	2.1	1.31	2.0	F .02	.36	52.9	65.3	12239a
2	33	34	3.69	1.02	1.05	.4	1.27	.6	G .01	.13	97.1	97.1	12133a
62	31	34	2.49	.62	1.14	.5	1.25	.6	H .01	.22	91.2	91.2	12232a
24	22	34	.60	.38	1.12	.8	1.25	1.2	I .17	.35	67.6	69.8	12117a
40	23	34	.75	.39	1.09	.6	1.24	1.1	J .19	.34	70.6	71.3	12122a
55	27	34	1.43	.45	1.11	.5	1.24	.7	K .14	.30	79.4	79.4	12244a
64	27	34	1.43	.45	1.13	.6	1.23	.7	L .12	.30	79.4	79.4	12237a
29	22	34	.60	.38	1.16	1.0	1.23	1.2	M .14	.35	67.6	69.8	12141a
25	23	34	.75	.39	1.20	1.2	1.22	1.0	N .11	.34	58.8	71.3	12116a
69	27	34	1.43	.45	1.14	.6	1.22	.7	O .12	.30	79.4	79.4	12241a
47	27	34	1.43	.45	1.17	.7	1.21	.7	P .10	.30	79.4	79.4	12242a
19	27	34	1.43	.45	1.01	.1	1.17	.6	Q .25	.30	79.4	79.4	12110a
10	22	34	.60	.38	1.12	.8	1.17	.9	R .19	.35	67.6	69.8	12151a
23	18	34	.04	.37	1.16	1.3	1.16	1.1	S .17	.36	47.1	65.3	12140a
1	22	34	.60	.38	1.14	.9	1.16	.8	T .18	.35	61.8	69.8	12131a
30	21	34	.45	.38	1.10	.7	1.15	.9	U .22	.35	67.6	68.3	12142a
34	21	34	.45	.38	1.15	1.0	1.11	.7	V .19	.35	55.9	68.3	12149a
41	26	34	1.24	.43	1.09	.5	1.14	.5	W .20	.32	73.5	76.5	12238a
75	21	34	.45	.38	1.12	.9	1.14	.8	X .20	.35	61.8	68.3	12227a
59	22	34	.60	.38	1.04	.3	1.12	.7	Y .27	.35	73.5	69.8	12209a
42	11	34	-.93	.39	1.11	.7	1.09	.5	Z .20	.32	55.9	69.3	12215a
BETTER FITTING OMITTED													
68	32	34	2.95	.74	1.00	.2	.66	-.2	.24	.18	94.1	94.1	12204a
46	30	34	2.15	.55	.99	.1	.79	-.2	.29	.24	88.2	88.2	12206a
9	26	34	1.24	.43	.94	-.2	.82	-.5	z .41	.32	73.5	76.5	12114a
38	20	34	.31	.37	.94	-.4	.92	-.5	y .43	.35	70.6	67.2	12156a
14	22	34	.60	.38	.93	-.4	.90	-.5	x .44	.35	73.5	69.8	12146a
49	26	34	1.24	.43	.92	-.3	.84	-.4	w .42	.32	79.4	76.5	12226a
73	17	34	-.10	.37	.92	-.7	.91	-.6	v .46	.35	73.5	64.9	12225a
33	17	34	-.10	.37	.90	-.8	.88	-.8	u .48	.35	67.6	64.9	12143a
77	21	34	.45	.38	.90	-.7	.87	-.7	t .48	.35	67.6	68.3	12247a
43	12	34	-.79	.38	.90	-.7	.82	-.9	s .47	.33	70.6	67.9	12234a
20	14	34	-.50	.37	.89	-.9	.83	-1.0	r .49	.34	73.5	66.1	12107a
13	17	34	-.10	.37	.87	-1.0	.85	-1.1	q .51	.35	67.6	64.9	12118a
6	21	34	.45	.38	.85	-1.0	.87	-.7	p .52	.35	79.4	68.3	12128a
7	17	34	-.10	.37	.87	-1.1	.85	-1.1	o .52	.35	73.5	64.9	12145a
63	17	34	-.10	.37	.87	-1.1	.84	-1.2	n .52	.35	67.6	64.9	12233a
8	16	34	-.23	.37	.86	-1.2	.83	-1.2	m .53	.35	70.6	65.0	12154a
58	20	34	.31	.37	.85	-1.1	.82	-1.2	l .54	.35	76.5	67.2	12201a
67	18	34	.04	.37	.83	-1.4	.83	-1.2	k .55	.36	76.5	65.3	12223a
50	19	34	.17	.37	.83	-1.3	.81	-1.4	j .56	.36	76.5	66.1	12240a
15	25	34	1.07	.41	.81	-.9	.68	-1.2	i .58	.33	79.4	74.5	12135a
61	20	34	.31	.37	.80	-1.5	.77	-1.5	h .59	.35	70.6	67.2	12207a
35	24	34	.90	.40	.79	-1.2	.72	-1.2	g .60	.34	82.4	72.7	12155a
39	24	34	.90	.40	.78	-1.3	.73	-1.2	f .60	.34	82.4	72.7	12106a
56	21	34	.45	.38	.78	-1.6	.74	-1.6	e .62	.35	79.4	68.3	12221a
65	26	34	1.24	.43	.78	-1.0	.68	-1.0	d .59	.32	79.4	76.5	12203a
78	18	34	.04	.37	.75	-2.2	.73	-2.1	c .66	.36	88.2	65.3	12235a
5	20	34	.31	.37	.71	-2.4	.66	-2.3	b .71	.35	82.4	67.2	12113a
53	23	34	.75	.39	.68	-2.1	.60	-2.1	a .74	.34	82.4	71.3	12212a
MEAN	22.9	34.0	.87	.44	.99	-.1	.99	-.1			75.9	74.3	
S.D.	5.5	.0	1.00	.12	.13	.9	.21	.9			9.9	8.9	

Note. Winsteps v3.70.0.1

Table 12

Person Fit Statistics for the 2012 ALT Pilot B

Person: REAL SEP.: 1.72 REL.: .75 ... Item: REAL SEP.: 2.52 REL.: .86

Person STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
36	25	34	1.15	.41	1.22	1.1	1.58	1.8	A .01	.34	76.5	74.1	12341
43	16	34	-.18	.37	1.44	2.9	1.49	2.8	B-.10	.38	44.1	66.6	12416
23	30	34	2.24	.55	1.16	.5	1.41	.8	C .01	.24	88.2	88.2	12338
14	18	34	.10	.37	1.35	2.4	1.37	2.2	D .01	.38	44.1	66.7	12336
33	23	34	.82	.39	1.25	1.5	1.31	1.3	E .08	.36	61.8	71.0	12324
55	12	34	-.74	.39	1.22	1.4	1.28	1.3	F .12	.37	61.8	69.6	12431
6	24	34	.98	.40	1.18	1.0	1.27	1.0	G .13	.35	67.6	72.4	12340
24	11	34	-.90	.39	1.18	1.1	1.25	1.1	H .15	.36	58.8	70.7	12316
39	28	34	1.72	.47	1.17	.7	1.24	.7	I .08	.29	82.4	82.4	12401
42	18	34	.10	.37	1.18	1.3	1.22	1.3	J .18	.38	55.9	66.7	12414
11	21	34	.52	.38	1.10	.7	1.21	1.1	K .24	.37	61.8	68.6	12334
40	23	34	.82	.39	1.21	1.3	1.17	.8	L .14	.36	55.9	71.0	12426
9	14	34	-.45	.38	1.15	1.1	1.21	1.2	M .20	.38	64.7	68.0	12335
21	24	34	.98	.40	1.09	.6	1.18	.7	N .23	.35	67.6	72.4	12350
52	25	34	1.15	.41	1.18	1.0	1.11	.5	O .16	.34	64.7	74.1	12405
58	19	34	.24	.37	1.09	.7	1.17	1.1	P .26	.38	67.6	67.2	12428
10	24	34	.98	.40	1.15	.9	1.16	.7	Q .19	.35	67.6	72.4	12315
49	22	34	.67	.39	1.06	.5	1.14	.7	R .27	.37	70.6	69.8	12407
57	23	34	.82	.39	1.14	.9	1.05	.3	S .23	.36	55.9	71.0	12412
27	28	34	1.72	.47	1.05	.3	1.13	.4	T .22	.29	82.4	82.4	12327
47	24	34	.98	.40	1.03	.3	1.11	.5	U .29	.35	73.5	72.4	12432
16	19	34	.24	.37	1.09	.7	1.07	.5	V .29	.38	61.8	67.2	12322
29	16	34	-.18	.37	1.07	.6	1.09	.6	W .30	.38	61.8	66.6	12306
46	21	34	.52	.38	1.08	.6	1.02	.2	X .31	.37	61.8	68.6	12410
48	19	34	.24	.37	1.08	.6	1.07	.5	Y .30	.38	61.8	67.2	12422
28	18	34	.10	.37	1.00	.1	1.07	.5	Z .36	.38	67.6	66.7	12329
BETTER FITTING OMITTED													
25	26	34	1.32	.43	.99	.0	.91	-.2	z .35	.32	76.5	76.5	12325
37	19	34	.24	.37	.99	.0	.99	.0	y .39	.38	67.6	67.2	12345
30	21	34	.52	.38	.97	-.1	.89	-.5	x .42	.37	61.8	68.6	12337
44	22	34	.67	.39	.97	-.1	.90	-.4	w .42	.37	64.7	69.8	12411
2	33	34	3.77	1.02	.95	.3	.43	-.1	v .25	.13	97.1	97.1	12312
45	14	34	-.45	.38	.93	-.4	.92	-.4	u .45	.38	64.7	68.0	12419
12	29	34	1.96	.50	.93	-.1	.77	-.3	t .37	.27	85.3	85.3	12326
1	23	34	.82	.39	.93	-.4	.83	-.7	s .46	.36	67.6	71.0	12342
26	25	34	1.15	.41	.93	-.3	.81	-.6	r .44	.34	76.5	74.1	12333
59	26	34	1.32	.43	.92	-.3	.89	-.2	q .41	.32	76.5	76.5	12415
50	20	34	.38	.38	.91	-.6	.88	-.7	p .48	.38	70.6	67.9	12421
31	16	34	-.18	.37	.90	-.7	.86	-.9	o .50	.38	67.6	66.6	12307
63	20	34	.38	.38	.89	-.8	.89	-.6	n .49	.38	76.5	67.9	12420
60	16	34	-.18	.37	.88	-.8	.86	-.9	m .51	.38	73.5	66.6	12433
62	13	34	-.60	.38	.88	-.8	.88	-.6	l .50	.37	76.5	68.6	12427
15	24	34	.98	.40	.88	-.7	.73	-1.0	k .51	.35	73.5	72.4	12304
32	26	34	1.32	.43	.87	-.6	.76	-.7	j .49	.32	76.5	76.5	12339
19	18	34	.10	.37	.84	-1.2	.82	-1.1	i .55	.38	73.5	66.7	12343
5	24	34	.98	.40	.84	-.9	.72	-1.1	h .54	.35	79.4	72.4	12320
20	27	34	1.51	.45	.84	-.6	.74	-.6	g .49	.31	79.4	79.4	12314
4	17	34	-.04	.37	.83	-1.3	.81	-1.3	f .57	.38	73.5	66.5	12346
34	17	34	-.04	.37	.83	-1.3	.78	-1.5	e .58	.38	67.6	66.5	12309
8	24	34	.98	.40	.82	-1.0	.70	-1.2	d .56	.35	79.4	72.4	12330
61	20	34	.38	.38	.81	-1.4	.78	-1.3	c .58	.38	76.5	67.9	12429
17	20	34	.38	.38	.70	-2.3	.65	-2.2	b .70	.38	82.4	67.9	12308
54	22	34	.67	.39	.67	-2.4	.59	-2.2	a .73	.37	88.2	69.8	12408
MEAN	21.6	34.0	.69	.42	1.02	.1	1.00	.1			71.1	72.5	
S.D.	5.1	.0	.87	.09	.15	1.0	.22	1.0			10.4	6.8	

Note. Winsteps v3.70.0.1

Factor Analysis

In order to examine the construct of the current ALT, a principle components analysis was run using varimax rotation on all 35 items of the current ALT (N = 501). The choice to use PCA was due to the exploratory nature of this analysis, and the fact that the researcher had no prior expectations of relationships among the items.

Extraction was set to delineate component loadings with Eigen values greater than 1.00, resulting in 13 unique factors that accounted for only 52.9% of the total variance. An examination of the loadings for this solution failed to reveal any components of real interest, and given that several components only had one item loading within them, this solution seemed more likely the result of random groupings than any indication of specific constructs. This result was not unexpected, as the original design of the current ALT called for a single, unified construct for listening comprehension (Clark, 2007). This was purposefully done in order to meet the assumption of unidimensionality for the original Rasch model used in analyzing the test.

Factor analyses were also conducted on the revised versions of the ALT using EFA on the theoretical basis that a solution of more than one factor would be discovered. It was hoped that the additional construct task-based critical listening would be validated by the presence of a second factor. Results of these analyses are not included in this study, however, as the findings were not encouraging and did not support a clear multiple-factor solution.

DISCUSSION

RQ1: How does the current form of the ALT perform in terms of matching the population of examinees, reliability of the scores on the test, and item and person fit?

One of things that makes the Rasch model so valuable is its ability to predict how well items on a test match a population of examinees. For the current version of the ALT, results of the Rasch analysis showed a match between both person ability and item difficulty, as is apparent from the mostly even distributions of both along either side of the vertical ruler in Figure 1.

There does exist a slight gap at the upper range of the scale, between roughly 1.25 and 2.00 logits, where there are a number of examinees of this ability but no appropriately difficult items. This is of note because it means that the ALT cannot effectively measure examinees with this level of advanced proficiency. This might be simply characteristic of a placement exam, where

the goal is not to assign a proficiency score but to place people into one of two levels of academic support (Clark, 2007). Those people who score outside the limitations of the test are beyond the scope of the program.

Conversely, there are several items at the low end of the difficulty scale relative to the number of examinees with a corresponding ability level. This indicates that for a majority of the population, these items are too easy. This could have a skewing effect on placements if there are a plethora of easy items but only a few items around the cut point between intermediate and advanced level or advanced and exempt level placements. It seems possible that an examinee could have their score inflated by responses on easy items, then be placed on the basis of a few correct guesses. With a more even distribution of items, or more items towards the middle of the scale, there seems to be less of a chance of misplacement, as examinees would need to repeatedly display their ability at these levels.

Reliability of the scores on the ALT were found by the analysis to be mostly reliable, with a Cronbach alpha level of .73, meaning the test is 73% consistent. While this number is not especially low, it also means that the items are 27% inconsistent. This might be accounted for by the low number of total items on the ALT, as reliability tends to be lower for a test with relatively few items compared to a test with lots of items.

Results showed that the items as a whole appear to be fitting the general model of the test, as only two items out of 35 (C1 and E7) had infit scores on or slightly above the acceptable range, indicating unexpected responses by examinees on these particular items. Items A1 and B1, identified as misfitting in Clark's previous analysis, held the next highest infit values, but were not found to be misfitting in this analysis. This could mean the items are testing something different from other items on the test, or that some examinees of lower ability were able to guess the correct response despite, and thus beating the odds that these items were expected to be too difficult.

Looking more closely at the items themselves, it could be that these items were deemed as misfitting by the analysis based on how they were constructed in terms of wording or choice of distractors. Item C1 appears to have some ambiguity in the available distractors, as shown in (4) below.

(4) The purpose of the lecture was to:

A. Determine the best definition of freedom.

B. Illustrate principles of freedom.

C. Discussion the history of freedom.

D. Introduce different definitions of freedom.

Both options A and D are very similar in that they are both asking about definitions, and it seems possible that this could cause confusion for some examinees, or may be an effect of reading comprehension rather than due to anything related to listening. There is also a fair amount of redundancy in the distractors with the repetition of the phrase "of freedom," which again might be over-complicating the responses (Brown, 2005).

Item E7 also appears to be ambiguous in terms of its design and what it is asking of examinees. As seen in (5), while the distractors for the item appear soundly constructed, the question itself is rather confusing, and comparatively lengthy to other items within the ALT.

(5) In the lecture, two possible reasons were given for why extraterrestrials do not leave their home worlds. One is that they are not interested in interstellar travel. The other reason is that:

A. Visiting Earth is not very interesting.

B. Civilizations are not as common as we think.

C. Interstellar travel is not possible.

D. Advanced civilizations tend to destroy themselves.

The section of the passage this items is asking about is structured so that the speaker gives two supporting reasons for a particular claim, and within these reasons further states another dichotomous set of explanations. In other words, there is a set of information nested within another set of information, and item E7 is asking about the nested information only. While the ability to organize different hierarchies of information might indicate a high degree of listening comprehension, this item was not measured as being overly difficult (0.41 logits). A better explanation might be that because the format of the test requires examinees to listen and take notes before having access to the questions, the chance of misunderstanding what information this item is attempting to draw out seems high, and the margin for error rather small.

It is possible that item E7 falls into the category of a task-based item based on the skill of organizing content, and the misfit value is an indicator that this item does not fit within a unidimensional construct of the test as a whole. While the assumption of unidimensionality is not overly strict, items that measure a different dimension than the rest of the test would appear as misfitting the model (Bond & Fox, 2007).

Persons were also generally fitting the model based on the findings of the Rasch analysis. Only 11 of 501 examinees were identified as misfitting the model, which translates to a mere 2.2% percent of the population. Were this percentage higher there might be cause for concern, but such a low value might be more attributed to random error in the model than a problem with the test itself.

RQ2: What constructs are represented in the current form of the ALT?

Based on a PCA of the items on the ALT, the assumption of unidimensionality for the test seems to have been met (Clark, 2007). There were no indications of multiple constructs outside a single factor of listening comprehension. Because PCA requires a large sample size to effectively discriminate between components, multiple components might be observable with a larger number of examinees (Tabachnick & Fidell, 2013). However, given that 501 examinees already contribute a lot of data, there is little reason to suspect a different outcome than what was found in this analysis.

RQ3: How do items based on the skills of organization, connecting ideas, and determining the importance of ideas perform on the ALT in terms of matching the population of examinees, reliability of scores on the test, and item and person fit?

General findings for the revised versions of the ALT were comparable with the analysis of the current ALT. Items carried over from the original test were anchored to provide a more complete comparison between those items and newly developed items. Person ability and item difficulty again seem matched overall, despite some discrepancies at either end of the scale. New items on both versions of the test fell within the current range of difficulties on the test, with the exception of item E8 on Pilot B, which was exceptionally low in difficulty. This finding was not surprising, as despite the fact that new items were made to tap into a different listening construct, there was no assumption that this would have a direct effect on difficulty.

Revised items on Pilot A were mostly classified near the center of the difficulty scale. Pilot B showed a bit more spreading out of items in terms of difficulty, and with the exception of item E8, all items were above 0 logits. These were positive findings, as the previous analysis showed a lot of redundancy at lower difficulty levels, but few items towards the middle of the scale. As this is the area where items will likely have more effect on placement decisions, the addition of

more similar item types at this mean difficulty level should benefit the accuracy of these decisions. In addition, items such as B4 on Pilot A and A5 and E7 on Pilot B are located at difficulty ranges where there were previously no comparable items, which again only improves the ability of the test to measure across a variety of ability levels.

This is further validated by the effect of the pilot items on the reliability of the scores on both Pilot tests. Compared with a Cronbach alpha of .72 for the current ALT, revised versions of the test reported alpha levels of .79 and .76 for Pilot A and B respectively. While these values are not extraordinary, this is an encouraging finding as any increase in reliability indicates an increase in the effectiveness of the instrument. As analyses of the pilot tests only included 34 items rather than the full 35 items, and item number has an effect on reliability, it is possible that these scores might go even higher when a full analysis is run on future administrations of the test.

Item misfit for both pilot tests was also found to be low, with only item D8 on Pilot A misfitting the model, and recall that this was explained by an error in the recording that occurred in relation to this item. Overfit was found for items E6 in Pilot A and D4 in Pilot B, which means these items were in some way performing too well, and is not of significant concern. It is curious that the different items were overfitting on either test, but this is probably explained by minor differences in the population of test-takers and variance in the model itself.

As with the previous analysis, persons misfitting the model were also found to be low, with only a combined 1.4% of the population (n = 2) not fitting. This is within acceptable limits, and can also be explained by the fact that no model is perfect for every single person and some error is to be expected. With a relatively small population of examinees (N = 141), such a small percentage is actually quite encouraging as it helps to reaffirm the effectiveness of the test for the given population.

RQ4: How do the piloted items compare to the items they replaced?

The Rasch model is useful for examining the performance of items on a test, but it is also worthwhile to consider how different items compare across tests, in particular how newly piloted items stand up against those items that were removed. Such comparisons can better reveal just what is different about these two groups of items. In addition, because a factor analysis failed to distinguish a new construct present in the piloted items, a closer inspection of the items

themselves might serve as an alternative methods of gaging which items from either the current or revised versions of the ALT should be selected for future models of the test.

At a conceptual level, the differences between the original items on the test and the new items seem mostly clear. Outside of the two misfitting items (C1 and E7) explained above, the remainder of the replaced items from the original test (A1, B1, E5) all shared a similar construction in that they required listeners to define terminology as it was explained in the lecture. The answers for each item was explicitly stated in the lecture, and almost always followed a formulaic pattern with both the target of the item and the answer being mentioned in short succession by the speaker. These items all appear to access primarily bottom-up modes of processing based on individual, word-level comprehension (Vandergrift, 2007; Field, 2008).

Piloted items, as described in the procedures section above, were designed to tap into the skills of organizing, connecting ideas, and discriminating important from unimportant information. The majority of these items required examinees to construct meaning based on information spread out across different sections of an individual passage. Example (6) shows this in an item that asks listeners to not only draw on several different definitions and explanations throughout a lecture, but also consider it from negative perspective.

(6) A5 (Pilot B). Which is NOT a power of the President?

A. Ratify treaties.

B. Veto laws.

C. Negotiate treaties.

D. Enforce laws.

While these items still draw on explicitly stated information, the separation of the content into various sections of the lecture was intended to force listeners to engage in more than just skills for comprehension, but also the target skills associated with critical listening (Field, 2008). On the surface, it appears that these items are asking the examinees to listen differently than the original, replaced items.

The only item that did not match the rest of the piloted items was item E8 on Pilot B, which is given below in example (7). This item alone was more self-contained to a specific section of the passage, and both the question and correct response were very close to one another in the lecture.

(7) E8 (Pilot B). What example does the speaker give for extraterrestrials leaving us alone on purpose?

A. People's interest in extraterrestrials.

B. Nuclear weapons.

C. The Drake Equation.

D. Bacteria in Alaska.

While the intention was for this item to test examinees' ability to determine what information was important, looking back it seems that this item is more closely related to word-level comprehension alone, and resembles the original items on the ALT more than the pilot items. It should also be noted that this item scored exceedingly low in terms of difficulty (-1.64 logits), and was the easiest item on any version of the test. This is probably explained by the realization that in the context of the lecture, the only distractor that can really be considered an example is the last one, which is the correct answer. Therefore, this item might have simply been obvious regardless of whether or not examinees comprehended the associated listening content.

A comparison of items by model attributes also yields interesting results. All of the new items fit the model well, and this is reflected in the improved reliability for the revised versions of the ALT. However, when the original items and the piloted items are compared in terms of difficulty, it is not so apparent which items are working better. Pilot items for passages A and C on both tests were lower in difficulty than the items they replaced. However, when these items are considered from a positional standpoint (i.e., where they fall in comparison to other items along the logit scale), even though they are lower in difficulty than the items they are replacing, the positions they occupy on the scale are areas where there are slight gaps or a lack of equivalent items. While the original items tend to be surrounded by like items, the piloted items for passages A and C are both in more isolated positions, and thus possibly more able to uniquely measure ability. This is not true in all cases (e.g., C6 in Pilot B appears with several other similarly difficult items), and the tradeoff between difficulty and matching person ability is one that cannot be fully understood at this stage. Nevertheless, these items might contribute something unique to the model.

The picture is a bit clearer for items for passage B. In this case, both of the revised items (B4) were higher in difficulty than the original item (B1). In particular, item B4 on Pilot B was found to be one of the most difficult items on the entire test with a logit value of 1.16, while the item

from the original test was measured at only -0.74 logits, making it one of the easier items on the test. While item B4 on Pilot B shares the same spot on the difficulty scale as several other items, as the test is both lacking in difficult items and a bit overwhelmed with easy ones, it seems like there is more to be gained from the piloted item.

Comparisons for items in passage E also seem straightforward. Both versions of item E8 were disappointingly low in difficulty, and even item E7 in Pilot A is low (-0.05 logits) compared with the items it replaced in the original test (E5, 0.26 logits and E7, 0.41 logits). These items do not offer any value in terms of position, as both the original and revised items are alone in their respective positions on the scale. Only item E7 for Pilot B has a higher difficulty, though it also shares it position with another piloted item (A5 on Pilot B), and so the amount of information it is contributing to the model is also uncertain.

In the end, the choice of items to include in future revisions of the test is probably best done by looking all three aspects of fit, position, and difficulty in combination and making judgments based on what will best serve the test. A general model for decisions might entail first looking at how well an item fits the overall model, as this has a direct effect on reliability. Following this, considerations about which items provide unique information for the overall test is beneficial for tapping into a full range of abilities. Lastly, items with higher degrees of difficulty should be selected as the test itself is skewed towards having more easy items than difficult ones when compared to the abilities of the population. Table 13 presents an overall summary of comparisons between original and revised items, as well as recommendations and suggestions based on these findings as to which items to include on a revised version of the ALT.

Table 13

Item Comparisons and Recommendations

Passage	Item[a]	Measure	Infit *MNSQ*[b]	Recommendation
A	A1O	0.99	1.10	Though A1O was the most
	A5A	0.41	1.07	difficult item, A5B had a
	A5B	0.57	0.94	relatively good difficulty rating and also occupied a more unique position on the test.
B	B1O	-0.74	1.10	B4B both fit the model and was
	B4A	0.09	0.87	difficult, which indicates this
	B4B	1.16	1.23	item should be included in future revisions.
C	C1O	0.73	1.12*	Lack of fit for C1O and low
	C6A	-0.74	1.11	difficulty scores for C6A
	C6B	0.19	1.16	indicate C6B as the best choice for future revisions.
E	E5O	0.41	1.13*	Lack of fit for E5O and lower
	E7O	0.26	1.08	difficulty values for E7A, E8A
	E7A	-0.05	1.12	and E8B suggest that these items
	E7B	0.57	0.95	might be best left out of the test
	E8A	-0.57	0.97	in favor of E7O and E7A.
	E8B	-1.64	0.83	

Note. Measure = difficulty value; Infit *MNSQ* = infit mean square.

[a] O = items from the 2010-2012 administrations of the ALT; A = items from Pilot version A; B = items from Pilot version B.

[b] *MNSQ* < 1.12 on the 2010-2012 ALT is considered fitting the model. *MNSQ* < 1.30 on Pilot A is considered fitting the model. *MNSQ* < 1.31 on Pilot B is considered fitting the model.

* Misfitting the model

CONCLUSION

This study set out to accomplish two specific tasks in relation to the ALT, the first of which was to revise and develop new items for the test to improve the overall quality and usefulness of the instrument. The findings suggest that this process was successful, with all of the newly piloted items fitting the general model of the test and serving to improve the reliability of the scores. New items were able to create a better match between the population and the test by replacing redundant items and measuring a more complete spectrum of abilities. However, the ALT is still limited in the maximum range of listening comprehension abilities it can measure, though this is a common and unavoidable outcome of placement testing.

The second aim of this study was to test a suitable construct for better determining different

levels of placement using a task-based model of critical listening. Unfortunately, the results did not show evidence for such a construct at this stage. While several items were designed with this construct in mind, and on the surface these items appeared to be asking different things from the test-takers, the data did not support this hypothesis. This leaves the question of how well the test can effectively measure different components of listening comprehension unanswered.

There are, however, several explanations that might help account for this lack of findings. Outside of the need for a higher number of participant data for piloting, the most likely reason is due to the narrow range of proficiencies that the ALT is attempting to measure. The population of examinees is so similar that identifying clear differences in the way they use skills and strategies to listen is quite difficult. Were the population more varied across a broader range of proficiency levels, such differences might be more pronounced and discernable from a general construct of listening alone.

Another consideration is the design and structure of the listening passages themselves and the possibility that they simply do not lend themselves to many opportunities for critical listening. The lectures were created to be natural, unscripted, and authentic representations of actual academic lectures. However, the content of the lectures was rather simple, and the time allowed for each was constrained to between three and nine minutes. Lectures followed a straightforward model each time, with the speaker introducing the topic and then explaining each main idea with some details in succession. While this may be very representative of the first five minutes of a university class, lectures probably do not follow this pattern through to the end. There were no real arguments being constructed, or positions defended, or beliefs explained for the listeners to really have a chance to think critically.

Additionally, the fact that the ALT is a multiple-choice test is a real limitation to the way in which the assessment can address questions of critical listening. Measurement involving critical listening is typically best accomplished by interactive assessments (Buck, 2001; Lynch, 2011). Multiple-choice tests are limited in the kind of responses they can ask about (Rodriguez, 2003; Traub, 1993), and lecture listening in particular is a very passive process, so the opportunity for creating interaction is almost impossible, despite the best efforts of the researcher when designing the items.

Future revisions of the ALT might consider not only item replacement, but also a closer look at the passages themselves, both in terms of content and presentation. One alternative might be to

present a listening passage that is a segment of a lecture in progress. This would require examinees to build the context from what they hear in the lecture without any clear introductory signposts, and this could take advantage of the kind of information that comes out in the later points of a classroom lecture. Interactions could also be included in the listening passage between multiple speakers (e.g., students asking a question of the teacher, a student commenting on another students response) to simulate what happens in an authentic classroom. Rarely is there just one speaker throughout the course of an entire lecture, and being able to process and comprehend comments, questions, and opposing opinions might be an excellent avenue of further research.

Lastly, there is the possibility of creating lecture passages that are not fully reliant on audible clues alone, but rather utilize video so that examinees can make use of verbal clues and other visual responses in connection to their listening. This, too, is more authentic to what students in an academic university will encounter, and it provides the opportunity to include different elements of critical thinking and comprehension, such as pragmatics, into a listening test.

In the end, the revisions made to the ALT are successful and should be implemented into future versions of the test in the short term, though continued piloting of the revised items is also suggested. The test in its current format is successful in measuring the population in terms of listening comprehension, but long-term revisions that continue to consider this idea of assessing critical listening could be a valuable process for the program as a whole.

REFERENCES

Aitken, K. G. (1978). Measuring listening comprehension. *English as a second language. TEAL Occasional Papers,* Vol. 2. Vancouver: British Colombia Association of Teachers of English as an Additional Language. ERIC Document No. ED 155 945.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Anderson, N., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing, 8*(1), 41-66.

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Badger, R., White, G., Sutherland, P., & Haggis, T. (2001). Note perfect: An investigation of how students view taking notes in lectures. *System, 29*, 405-417.

Barnett, M. A. (1988). Reading through context: How real and perceived strategy use affects L2 comprehension. *Modern Language Journal, 72*(2), 150-162.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of education goals; Handbook 1: Cognitive domain.* New York: Longmans, Green.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testings, 19*(4), 369-394.

Brown, G. (1995). Dimensions of difficulty in listening comprehension. In D. J. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening.* San Diego: Dominie Press.

Brown, J. D. (2005). *Testing in language programs.* New York: McGraw-Hill.

Brown, J. D. (2009). Principle components analysis and exploratory factor analysis – Definitions, differences, choices. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(1), 13-20.

Brown, J. D. (2010). How are PCA and EFA used in language test and questionnaire development? *Shiken: JALT Testing & Evaluation SIG Newsletter, 14*(2), 30-35.

Brown, J. D. (2012). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 303-315). New York: Routledge.

Brown, J. D., & Kondo-Brown, K. (2012). Rubric-based scoring of Japanese essays: The effects on

generalizability of numbers of raters and categories. In J. D. Brown, (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 169-182). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge. University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119-157.

Carroll, J. B. (1972). Defining language comprehension. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisitions of knowledge* (pp. 1-29). New York: John Wiley & Sons.

Chamot, A. U. (2005). Language learning strategy instruction: Current issues and research. *Annual Review of Applied Linguistics, 25*, 98-111.

Chun, J. Y. (2011). *The construct validation of ELI listening placement tests.* Unpublished manuscript, University of Hawaii at Manoa, Honolulu, HI.

Clark, H., & Clark, E. (1977). *Psychology and language: An introduction to psycholinguistics.* New York: Harcourt Brace Jovanovich.

Clark, M. (2007). *Listening placement test development and analysis from a Rasch perspective.* Unpublished PhD dissertation. Honolulu, HI: University of Hawai'i at Mānoa.

Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-491.

Cushing, S. T. (1991). *A qualitative approach to the study of note taking in UCLA's English as a second language placement examination*. Unpublished manuscript, University of California, Los Angeles.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal, 77*(2), 180-191.

Ferris, D. (1998). Students' views of academic aural/oral skills: A comparative needs analysis. *TESOL Quarterly, 32*, 289-318.

Field, J. (2008). *Listening in the language classroom.* Cambridge: Cambridge University Press.

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive development.* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Garner, R., & Reis, R. (1981). Monitoring and resolving comprehension obstacles: An investigation

of spontaneous text loopbacks among upper-grade good and poor readers' comprehension. *Reading Research Quarterly, 16,* 569-582.

Goh, C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28,* 55-75.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*(3), 375-406.

Graham, S., Santos, D., & Vanderplank, R. (2008). Listening comprehension and strategy use: A longitudinal exploration. *System, 36,* 52-68.

Graham, S., Santos, D., & Vanderplank, R. (2011). Exploring the relationship between listening development and strategy use. *Language Teaching Research, 15*(4), 435-456.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening* (pp. 241-268). New York: Cambridge University Press.

Hudson, T. (2011). *Teaching second language reading.* Oxford: Oxford University Press.

Jensen, C., Hansen, C., Green, S., & Akey, T. (1997). An investigation of item difficulty incorporating the structure of listening tests: A hierarchical linear modeling analysis. In A. Huhta, V. Kohonen, L. Kurki-Suinio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 151-164). Jyvaskyla: University of Jyvaskyla.

Jimenez, R. T. (1997). The strategic reading abilities and potential of five low-literacy Latina/o readers in middle school. *Reading Research Quarterly, 32*(3), 224-243.

Jung, E. H. (2006). Misunderstanding of academic monologues by nonnative speakers of English. *Journal of Pragmatics, 38,* 1928-1942.

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL, 29,* 135-149.

Kintsch, W. J. (1998). *Comprehension: A paradigm for cognition.* Cambridge: Cambridge University Press.

Kintsch, W. J., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Educational Psychology, 74*(6), 828-834

Linacre, J. M. (2010b) *User's guide to Winsteps: Rasch-model computer programs.* Chicago: Author.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes, 10,* 79-88.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 242-256.

Miyata, M. (2007). *A Rasch analysis of the ELI listening placement test.* Unpublished manuscript, University of Hawaii at Manoa, Honolulu, HI.

Osada, N. (2001). What strategy do less proficient learners employ in listening comprehension? A reappraisal of bottom-up and top-down processing. *Journal of the Pan-Pacific Association of Applied Linguistics, 5*, 73-90.

Olsen, L. A., & Huckin, T. N. (1990). Point-driven understanding in engineering lecture comprehension. *English for Specific Purposes, 9*, 33-47.

Perfetti, C., & Roth, S. (1981). Some of the interactive processes in reading and their role in the reading skill. In A. M. Lesgold & C. Perfetti (Eds.), *Interactive processes in reading* (pp. 269-297). Hillsdale, NJ: Erlbaum.

Phakti, A. (2003). A closer look at gender and strategy use in L2 reading. *Language Learning, 53*(4), 649-702.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61*(4), 993-1038.

Richards, J. (1983). Listening comprehension: approach design, procedure. *TESOL Quarterly, 17,* 219-240.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Rost, M. (2002). *Teaching and researching listening.* Harlow, England: Longman.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing, 1*(2), 202-220.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question-type. *Language Testing, 8*, 23-40.

Song, M.Y. (2008). Do divisible sub skills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing, 25*(4), 435-464

Song, M. Y. (2011). Note-taking quality and performance on an L2 academic listening test. *Language Testing, 29*(1), 67-89.

Staehr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *SSLA, 31*, 577-607.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.*)*. Upper Saddle River, NJ: Pearson.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed-response, performance testing, and portfolio assessment* (Electronic version on Questia database, pp. 29-44). Hilldale, NJ: Lawrence Erlbaum.

Tsai, Y. O. (2004). *The role notetaking plays in an English for academic purposes listening test*. Unpublished manuscript, University of California, Los Angeles.

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics, 19*(4), 432-451.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Vandergrift, L. (2003a). From prediction through reflection: Guiding students through the process of L2 listening. *The Canadian Modern Language Review, 59*(3), 425-440.

Vandergrift, L. (2003b). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning, 53*, 463-496.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40*, 191-210.

Vandergrift, L., Goh, C., Mareschal, C., & Tafaghodatari, M. H. (2006). The metacognitive awareness listening questionnaire (MALQ): Development and validation. *Language Learning, 56*, 431-462.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, 2*(1), 1-39.

William, R. L., & Eggert, A. C. (2002) Notetaking in college classes: Student patterns and instructional strategies. *The Journal of General Education, 51*(3), 173-199.

Zimmerman, B. J. (1986). Development of a structures interview for assessing students' use of self-regulated learning strategies. *American Educational Research Journal, 23*(1), 614-628.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166-183.

APPENDIX A
Student Learning Outcomes

Intermediate Level

By the end of the course students will be able to:

- Demonstrate good use of strategies for comprehending academic lectures in English.
- Demonstrate effective use of strategies for incorporating information formal academic lectures into their overall studies.
- Make short academic presentations with some accuracy and cultural and stylistic appropriacy.
- With guidance, lead academic discussions using academic English.
- Demonstrate good use of strategies for participation in academic discussions with other students who are second-language users of English.
- State a range of strategies for using listening/speaking opportunities to develop academic vocabulary (in English) and specify which they have an active command of in their repertoire.
- State a range of strategies for developing academic English and specify which they have an active command of in their repertoire.
- Self-assess their strengths in terms of listening/speaking abilities, as well as identify areas for continued development.

Advanced Level

By the end of the course students will be able to:

- Demonstrate effective use of strategies for comprehending advanced academic lectures in English.
- Critically evaluate speakers' perspectives, techniques, and arguments.
- Make academic presentations (individually or in group or panel contexts) with a high degree of formal accuracy and cultural and stylistic appropriacy.
- Autonomously lead academic discussions using academic English.
- Demonstrate excellent use of advanced strategies for participation in academic discussions with expert users of English.
- State a range of strategies for using listening/speaking opportunities to develop academic vocabulary (in English) and specify which they have an active command of in their repertoire.
- State a range of strategies for developing academic English and specify which they have an active command of in their repertoire.
- Self-assess their strengths in terms of listening/speaking abilities, as well as identify areas for continued development.

APPENDIX B
Level Separation Chart

	Listening strategies (Fluency in listening)	Critical listening	Extensive listening
ELI 70	**Required** • Introduce pre-listening (obtaining background information, having a discussion), during-listening (note-taking, getting main idea and details), and post-listening strategies (reviewing notes, having a group/class discussion). • Teach pronunciation as an aid of listening comprehension . Use intermediate-level academic listening materials.	**Optional** Focus more on general comprehension.	**Optional** (If time allows, we might want to require this for further practice of listening strategies.)
ELI 80	**Required** • Review pre-listening (obtaining background information, having a discussion), during-listening (note-taking), and post-listening strategies (reviewing notes, having a group/class discussion). • Use strategies (e.g., "Which strategies in particular work effectively for you?") more effectively. • Teach pronunciation as an aid of listening comprehension. Use advanced-level academic listening materials.	**Required** Introduce how to listen critically to the materials (not merely comprehending the materials, but responding to the materials).	**Optional** (If time allows, we might want to require this for further practice of listening strategies.)

APPENDIX C
Revised Items List for the ALT

Short Lecture A

A5A. Based on the lecture, what is the process of checks and balances in a law?

A. Congress creates a law, the president approves it, and the Supreme Court judges the law as constitutional.
B. Congress makes a law, the president approves it, and the Supreme Court enforces it.
C. The president writes a law, congress ratifies it, and the Supreme Court enforces it.
D. The president makes a law, congress vetoes it, and the Supreme Court judges the law as constitutional.

A5B. Which is NOT a power of the President?

A. Ratify treaties
B. Veto laws
C. Negotiate treaties
D. Enforce laws

Short Lecture B

B4A. What best describes the relationship between branding and advertising?

A. Branding reinforces successful advertising
B. Advertising reinforces branding
C. Advertising uses publicity to create a brand
D. Branding begins by advertising alone

B4B. What does the common-sense definition of marketing actually resemble?

A. Branding
B. Advertising
C. Demographics
D. Publicity

Short Lecture C

C6A. What implication can be made about freedom according to this lecture?

A. Only people with money are free
B. Freedom does not have a single definition
C. Freedom is the ability to think what we want
D. Freedom is impossible

Appendix C (continued)

C6B. Based on the lecture, why is it difficult to define freedom?

A. Freedom requires money
B. Freedom has many definitions
C. Freedom only exists in America
D. Freedom is just a feeling

Long Lecture E

E7A. How is the Drake equation connected to Fermi's Paradox?

A. Drake and Fermi worked together to develop the Drake equation
B. The Drake equation proves the existence of extraterrestrial life.
C. The Drake equation is necessary to answer Fermi's Paradox.
D. Fermi's Paradox proves the Drake equation to be true.

E7B. The speaker mentions that one reason extraterrestrials do not leave their homes is because they might have destroyed themselves. How is this inconsistent with the Drake equation?

A. The Drake equation only considers civilizations with interstellar travel
B. The Drake equation only considers planets in habitable zones.
C. The Drake equation only considers civilizations that still exist.
D. The Drake equation only considers planets in the Milky Way.

E8A. What implication can be made about the existence of extraterrestrial life based on the lecture?

A. Extraterrestrials do not exist in the universe.
B. Extraterrestrials existed at one time, but have since destroyed themselves.
C. Extraterrestrials have visited the earth, but only in secret.
D. It is possible that extraterrestrials exist, but so far no proof exists.

E8B. What example does the speaker give for extraterrestrials leaving us alone on purpose?

A. People's interest in extraterrestrials
B. Nuclear weapons
C. The Drake equation
D. Bacteria in Alaska

APPENDIX D
Person Fit Statistics for the 2010-2012 ALT

Person: REAL SEP.: 1.60 REL.: .72 ... Item: REAL SEP.: 7.52 REL.: .98

Person STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
1	8	35	-1.38	.42	1.41	1.8	2.06	2.6	A-.30	.30	77.1	77.1	11216
419	28	35	1.57	.44	1.27	1.1	1.77	1.8	B-.13	.29	80.0	80.0	10359
244	28	35	1.57	.44	1.20	.8	1.68	1.7	C-.04	.29	80.0	80.0	80245
54	29	35	1.78	.47	1.11	.5	1.64	1.4	D .04	.27	82.9	82.9	11206
99	26	35	1.21	.41	1.21	1.1	1.59	1.8	E-.01	.32	74.3	74.4	80206
427	29	35	1.78	.47	1.23	.9	1.58	1.3	F-.06	.27	82.9	82.9	10342
450	30	35	2.01	.50	1.13	.5	1.55	1.1	G .04	.25	85.7	85.7	30104
341	29	35	1.78	.47	1.13	.5	1.51	1.2	H .05	.27	82.9	82.9	10124
343	24	35	.89	.39	1.21	1.3	1.46	1.8	I .03	.34	62.9	70.8	10126
320	24	35	.89	.39	1.16	1.0	1.45	1.8	J .07	.34	74.3	70.8	10147
248	13	35	-.61	.37	1.38	2.5	1.45	2.2	K-.11	.35	45.7	68.2	80239
393	25	35	1.05	.40	1.30	1.6	1.45	1.6	L-.05	.33	60.0	72.5	10363
480	18	35	.06	.36	1.38	2.8	1.45	2.8	M-.09	.37	48.6	65.9	30321
217	17	35	-.07	.36	1.39	2.8	1.44	2.8	N-.09	.36	42.9	65.9	80326
270	24	35	.89	.39	1.22	1.3	1.41	1.7	O .04	.34	62.9	70.8	10206
98	27	35	1.38	.42	1.19	.9	1.41	1.2	P .04	.31	77.1	77.1	80101
327	18	35	.06	.36	1.38	2.7	1.40	2.6	Q-.07	.37	42.9	65.9	10116
230	25	35	1.05	.40	1.27	1.5	1.40	1.5	R-.01	.33	65.7	72.5	80179
153	11	35	-.89	.39	1.23	1.4	1.37	1.5	S .04	.34	68.6	70.5	80117
402	19	35	.20	.36	1.33	2.4	1.36	2.3	T-.02	.36	48.6	66.2	10345
454	21	35	.47	.37	1.27	1.9	1.36	2.0	U .03	.36	51.4	67.4	30120
101	29	35	1.78	.47	1.21	.8	1.36	.9	V .00	.27	82.9	82.9	80429
483	24	35	.89	.39	1.22	1.3	1.33	1.4	W .06	.34	57.1	70.8	30205
225	14	35	-.47	.37	1.27	1.9	1.33	1.8	X .03	.36	54.3	67.5	80140
395	33	35	3.07	.74	1.09	.3	1.33	.6	Y .00	.17	94.3	94.3	10351
263	23	35	.75	.38	1.16	1.0	1.32	1.5	Z .12	.35	71.4	69.6	10221
354	28	35	1.57	.44	1.14	.6	1.32	.9	.08	.29	80.0	80.0	10133
283	21	35	.47	.37	1.25	1.7	1.31	1.8	.06	.36	51.4	67.4	10216
200	30	35	2.01	.50	1.07	.3	1.31	.7	.13	.25	85.7	85.7	80173
120	26	35	1.21	.41	1.26	1.3	1.30	1.1	.02	.32	62.9	74.4	80208
331	22	35	.60	.37	1.29	1.9	1.30	1.5	.02	.35	54.3	68.5	10117
223	28	35	1.57	.44	1.13	.6	1.30	.9	.11	.29	80.0	80.0	80325
312	16	35	-.20	.36	1.27	2.0	1.29	1.9	.05	.36	48.6	66.1	10149
134	22	35	.60	.37	1.17	1.2	1.29	1.5	.13	.35	65.7	68.5	80223
428	28	35	1.57	.44	1.15	.7	1.28	.8	.09	.29	80.0	80.0	10312
155	12	35	-.75	.38	1.17	1.1	1.28	1.3	.12	.34	60.0	69.2	80310
4	13	35	-.61	.37	1.22	1.5	1.28	1.4	.09	.35	57.1	68.2	11223
259	22	35	.60	.37	1.19	1.3	1.27	1.4	.11	.35	65.7	68.5	10202
473	16	35	-.20	.36	1.25	1.9	1.27	1.8	.07	.36	54.3	66.1	30310
440	19	35	.20	.36	1.25	1.8	1.27	1.8	.08	.36	54.3	66.2	10305
172	14	35	-.47	.37	1.20	1.4	1.27	1.5	.11	.36	60.0	67.5	80221
490	25	35	1.05	.40	1.10	.6	1.26	1.0	.18	.33	71.4	72.5	30255
80	12	35	-.75	.38	1.25	1.6	1.26	1.2	.05	.34	54.3	69.2	80224
310	18	35	.06	.36	1.24	1.8	1.26	1.7	.08	.37	54.3	65.9	10102
247	22	35	.60	.37	1.14	1.0	1.26	1.3	.16	.35	65.7	68.5	80138
42	26	35	1.21	.41	1.16	.9	1.26	.9	.11	.32	62.9	74.4	11301
109	24	35	.89	.39	1.25	1.5	1.24	1.0	.06	.34	57.1	70.8	80378
482	23	35	.75	.38	1.13	.9	1.24	1.2	.16	.35	65.7	69.6	30202
185	28	35	1.57	.44	1.14	.6	1.24	.7	.11	.29	80.0	80.0	80313
220	15	35	-.33	.37	1.20	1.5	1.24	1.5	.12	.36	57.1	66.8	80335
401	26	35	1.21	.41	1.15	.8	1.24	.9	.13	.32	68.6	74.4	10344
366	19	35	.20	.36	1.24	1.8	1.24	1.6	.09	.36	48.6	66.2	10395
26	21	35	.47	.37	1.22	1.6	1.23	1.4	.10	.36	51.4	67.4	11312
182	20	35	.33	.37	1.23	1.7	1.23	1.5	.10	.36	48.6	66.7	80420
56	16	35	-.20	.36	1.18	1.4	1.23	1.5	.14	.36	60.0	66.1	80205
3	19	35	.20	.36	1.19	1.4	1.23	1.5	.14	.36	54.3	66.2	11320
121	16	35	-.20	.36	1.22	1.7	1.23	1.5	.11	.36	48.6	66.1	80114
21	15	35	-.33	.37	1.17	1.3	1.22	1.4	.15	.36	57.1	66.8	11314
496	17	35	-.07	.36	1.19	1.5	1.22	1.5	.14	.36	54.3	65.9	30214
344	27	35	1.38	.42	1.14	.7	1.22	.8	.14	.31	77.1	77.1	10166
369	23	35	.75	.38	1.10	.7	1.22	1.1	.20	.35	71.4	69.6	10327
91	30	35	2.01	.50	1.07	.3	1.22	.6	.14	.25	85.7	85.7	80301
392	21	35	.47	.37	1.22	1.5	1.21	1.2	.11	.36	57.1	67.4	10362
72	22	35	.60	.37	1.19	1.3	1.21	1.1	.13	.35	60.0	68.5	80372
198	22	35	.60	.37	1.15	1.1	1.21	1.1	.16	.35	65.7	68.5	80232
399	26	35	1.21	.41	1.21	1.1	1.20	.8	.08	.32	68.6	74.4	10324
135	29	35	1.78	.47	1.08	.4	1.21	.6	.16	.27	82.9	82.9	80120
387	19	35	.20	.36	1.18	1.4	1.21	1.4	.15	.36	60.0	66.2	10330
288	30	35	2.01	.50	1.13	.5	1.21	.6	.09	.25	85.7	85.7	10406

Appendix D (continued)

329	27	35	1.38	.42	1.20	1.0	1.14	.5	.10	.31	77.1	77.1	10151
40	20	35	.33	.37	1.18	1.4	1.20	1.3	.15	.36	54.3	66.7	11306
335	23	35	.75	.38	1.20	1.3	1.20	1.0	.12	.35	60.0	69.6	10163
	BETTER FITTING OMITTED												
148	28	35	1.57	.44	.96	-.1	.77	-.6	.38	.29	80.0	80.0	80308
274	28	35	1.57	.44	.96	-.1	.73	-.7	.40	.29	80.0	80.0	10231
268	31	35	2.28	.55	.95	.0	.74	-.3	.32	.23	88.6	88.6	10228
266	25	35	1.05	.40	.94	-.3	.79	-.8	.44	.33	65.7	72.5	10224
309	28	35	1.57	.44	.94	-.2	.78	-.5	.39	.29	80.0	80.0	10101
23	28	35	1.57	.44	.93	-.2	.74	-.6	.42	.29	80.0	80.0	11222
60	28	35	1.57	.44	.92	-.2	.73	-.7	.42	.29	80.0	80.0	80430
73	31	35	2.28	.55	.92	-.1	.73	-.3	.36	.23	88.6	88.6	80304
31	26	35	1.21	.41	.91	-.4	.78	-.7	.45	.32	74.3	74.4	11310
286	25	35	1.05	.40	.91	-.4	.79	-.8	.46	.33	71.4	72.5	10404
459	32	35	2.62	.62	.91	.0	.53	-.6	.38	.20	91.4	91.4	30111
368	24	35	.89	.39	.91	-.5	.78	-.9	.48	.34	68.6	70.8	10371
235	23	35	.75	.38	.89	-.7	.79	-1.0	.50	.35	71.4	69.6	80151
92	27	35	1.38	.42	.89	-.5	.71	-.9	.48	.31	77.1	77.1	80352
163	25	35	1.05	.40	.89	-.6	.76	-.9	.49	.33	77.1	72.5	80123
141	24	35	.89	.39	.89	-.7	.80	-.9	.49	.34	68.6	70.8	80119
414	26	35	1.21	.41	.88	-.6	.79	-.7	.47	.32	74.3	74.4	10315
77	24	35	.89	.39	.88	-.7	.78	-.9	.50	.34	74.3	70.8	80163
358	26	35	1.21	.41	.88	-.6	.74	-.9	.49	.32	74.3	74.4	10159
105	24	35	.89	.39	.88	-.7	.75	-1.1	.52	.34	68.6	70.8	80155
74	27	35	1.38	.42	.87	-.5	.72	-.9	.49	.31	77.1	77.1	80302
208	13	35	-.61	.37	.87	-.9	.80	-1.1	.52	.35	68.6	68.2	80356
400	25	35	1.05	.40	.87	-.7	.74	-1.0	.51	.33	71.4	72.5	10367
115	30	35	2.01	.50	.87	-.3	.58	-.9	.46	.25	85.7	85.7	80157
231	11	35	-.89	.39	.86	-.8	.76	-1.1	.52	.34	74.3	70.5	80135
489	30	35	2.01	.50	.86	-.3	.58	-.9	.47	.25	85.7	85.7	30226
152	26	35	1.21	.41	.86	-.7	.73	-.9	.51	.32	74.3	74.4	80139
464	25	35	1.05	.40	.86	-.8	.74	-1.0	.52	.33	71.4	72.5	30121
255	26	35	1.21	.41	.86	-.7	.78	-.7	.50	.32	74.3	74.4	80162
86	22	35	.60	.37	.86	-1.0	.77	-1.3	.54	.35	71.4	68.5	80437
423	26	35	1.21	.41	.85	-.7	.71	-1.0	.52	.32	80.0	74.4	10340
157	15	35	-.33	.37	.84	-1.2	.80	-1.3	.55	.36	74.3	66.8	80247
346	25	35	1.05	.40	.84	-.9	.75	-1.0	.53	.33	71.4	72.5	10142
404	27	35	1.38	.42	.84	-.7	.71	-.9	.52	.31	77.1	77.1	10347
265	23	35	.75	.38	.84	-1.1	.74	-1.3	.56	.35	71.4	69.6	10223
149	19	35	.20	.36	.83	-1.3	.79	-1.5	.57	.36	71.4	66.2	80333
83	27	35	1.38	.42	.82	-.8	.70	-.9	.53	.31	77.1	77.1	80358
466	25	35	1.05	.40	.82	-1.0	.69	-1.3	.57	.33	71.4	72.5	30123
93	27	35	1.38	.42	.82	-.8	.66	-1.1	.55	.31	77.1	77.1	80324
371	18	35	.06	.36	.82	-1.5	.78	-1.6	.59	.37	71.4	65.9	10372
36	23	35	.75	.38	.81	-1.3	.79	-1.0	.57	.35	82.9	69.6	11220
363	22	35	.60	.37	.81	-1.4	.73	-1.6	.60	.35	77.1	68.5	10325
332	24	35	.89	.39	.81	-1.2	.71	-1.3	.58	.34	74.3	70.8	10119
403	27	35	1.38	.42	.80	-.9	.61	-1.3	.58	.31	77.1	77.1	10346
326	21	35	.47	.37	.80	-1.5	.79	-1.3	.59	.36	74.3	67.4	10115
39	14	35	-.47	.37	.80	-1.5	.74	-1.6	.60	.36	77.1	67.5	11316
484	12	35	-.75	.38	.80	-1.4	.79	-1.0	.57	.34	82.9	69.2	30206
340	26	35	1.21	.41	.80	-1.0	.66	-1.2	.58	.32	80.0	74.4	10143
90	19	35	.20	.36	.80	-1.6	.79	-1.5	.60	.36	77.1	66.2	80165
132	19	35	.20	.36	.79	-1.7	.79	-1.5	.60	.36	77.1	66.2	80428
433	21	35	.47	.37	.79	-1.6	.74	-1.6	.61	.36	80.0	67.4	10332
219	18	35	.06	.36	.79	-1.8	.75	-1.9	.62	.37	71.4	65.9	80148
11	20	35	.33	.37	.78	-1.7	.77	-1.6	.61	.36	77.1	66.7	11205
292	21	35	.47	.37	.78	-1.7	.73	-1.7	.62	.36	74.3	67.4	10412
276	20	35	.33	.37	.78	-1.8	.77	-1.6	.61	.36	82.9	66.7	10211
439	26	35	1.21	.41	.78	-1.1	.63	-1.4	.60	.32	74.3	74.4	10304
88	22	35	.60	.37	.78	-1.6	.71	-1.7	z .62	.35	77.1	68.5	80348
17	18	35	.06	.36	.78	-1.9	.75	-1.9	y .63	.37	82.9	65.9	11225
300	20	35	.33	.37	.78	-1.8	.75	-1.7	x .62	.36	82.9	66.7	10422
8	15	35	-.33	.37	.78	-1.8	.76	-1.7	w .62	.36	80.0	66.8	11322
250	16	35	-.20	.36	.77	-1.9	.74	-1.9	v .63	.36	82.9	66.1	80235
448	21	35	.47	.37	.77	-1.8	.71	-1.9	u .64	.36	80.0	67.4	30125
210	20	35	.33	.37	.77	-1.9	.74	-1.8	t .63	.36	82.9	66.7	80341
405	21	35	.47	.37	.76	-1.8	.71	-1.9	s .64	.36	85.7	67.4	10348
386	23	35	.75	.38	.76	-1.6	.70	-1.5	r .63	.35	82.9	69.6	10385
456	17	35	-.07	.36	.76	-2.0	.74	-2.0	q .64	.36	82.9	65.9	30108
443	19	35	.20	.36	.76	-2.0	.72	-2.1	p .65	.36	77.1	66.2	10308
195	16	35	-.20	.36	.75	-2.1	.75	-1.8	o .65	.36	77.1	66.1	80125
133	21	35	.47	.37	.75	-2.0	.70	-1.9	n .66	.36	80.0	67.4	80412
30	21	35	.47	.37	.75	-2.0	.71	-1.9	m .66	.36	80.0	67.4	11325
370	16	35	-.20	.36	.74	-2.2	.70	-2.3	l .67	.36	77.1	66.1	10328
262	18	35	.06	.36	.74	-2.2	.71	-2.2	k .67	.37	82.9	65.9	10205
243	18	35	.06	.36	.73	-2.3	.70	-2.3	j .68	.37	82.9	65.9	80997
410	18	35	.06	.36	.73	-2.3	.73	-2.1	i .68	.37	88.6	65.9	10314
81	15	35	-.33	.37	.73	-2.3	.69	-2.2	h .68	.36	85.7	66.8	80104

Appendix D (continued)

	130	23	35	.75	.38	.73	-2.0	.63	-2.0	g .69	.35	77.1	69.6	80374
	260	21	35	.47	.37	.72	-2.2	.68	-2.1	f .68	.36	85.7	67.4	10219
	213	18	35	.06	.36	.72	-2.4	.70	-2.3	e .69	.37	82.9	65.9	80379
	458	22	35	.60	.37	.69	-2.4	.62	-2.3	d .73	.35	82.9	68.5	30110
	362	19	35	.20	.36	.69	-2.7	.64	-2.7	c .74	.36	82.9	66.2	10141
	281	23	35	.75	.38	.68	-2.3	.61	-2.1	b .72	.35	88.6	69.6	10213
	426	20	35	.33	.37	.64	-3.1	.60	-3.0	a .79	.36	88.6	66.7	10311
MEAN		21.4	35.0	.57	.39	1.00	.0	1.00	.0			70.6	70.8	
S.D.		5.0	.0	.77	.05	.13	.9	.20	1.0			9.3	5.9	