http://e-flt.nus.edu.sg/

Electronic Journal of Foreign Language Teaching
2005, Vol. 2, No. 2, pp. 45-57
© Centre for Language Studies
National University of Singapore

What is New in the New TOEFL-iBT 2006 Test Format?

Alla Zareva
(Alla.Zareva@nau.edu)
Northern Arizona University, U.S.A.

Abstract

In recent years TOEFL has become one of the most popular high-stakes tests affecting not only what and how English language teachers teach but also what and how students learn (e.g. Johnson, Jordan, & Poehner, 2005; Alderson & Wall, 1993). The new 2006 TOEFL–iBT exam is on its way; yet, until now, information about the new test format and test preparation materials is scarce. Above and beyond interest in the test alone, the burning question is what demanded the revision of the current test, given that the latest computer-based TOEFL was introduced fairly recently worldwide. The paper elaborates on some of the major reasons that have promoted the current changes of the exam format emphasizing the realization of testing experts, researchers, ESL/EFL teachers, students, program administrators, and other end-users that to succeed in an academic environment in which English is the language of instruction, students need not only to understand English, but also to communicate effectively. Among these reasons is the growing awareness among all parties interested in the test results that if a test is aiming to be a test of English for academic purposes and a reliable instrument of language proficiency, it needs to capture the integrated nature of the use of skills in academic settings. The paper also discusses in greater detail the new revisions of the test format – that is, the inclusion of a new speaking section, the revision of the writing component, and the incorporation of note-taking throughout all sections of the test – in light of the theoretical considerations and research findings underpinning the modifications. The expected outcomes and implications of the test revisions are outlined with regard to a focus on communicative competence and the anticipated positive washback effect on the way English is taught in the future.

1 Introduction

The next generation 2006 Test of English as a Foreign Language Internet-Based Testing (TOEFL–iBT) is on its way. Educational Testing Service (ETS) recently announced that the transition to the new test worldwide will take place in a phase manner over the rest of 2005 and the following year. According to the ETS schedule, the test would be first administered in the United States in September, 2005, followed by its implementation in Canada, France, Germany and Italy in October the same year, while its introduction in the rest of the world will begin in 2006. However, until the new Internet-based test is implemented, ETS will continue to administer the current computer-based (CBT) and paper-based (PBT) versions, as well as Test of Spoken English (TSE) (as a stand-alone test) throughout the transition period. At this point of time, information about the new test format and test preparation materials is scarce – the only available source, to my knowledge, being ETS (consult <www.ets.org/toefl> for more information) – but the expectations of language testing experts, English as a second/foreign language (ESL/EFL) teachers, and program administrators seem to be optimistic (personal communication). Many of them feel that the next generation TOEFL will offer innovative elements in its new format by putting its main emphasis on test-takers' integrated skills and their ability to communicate in an academic setting – an idea that is generally neither new nor innovative in the language testing field (see Davies, 2003, for a review). The reason behind the test revision is the realization that to succeed in an academic environment in which English is the language of instruction, students need not only to understand English, but also to communicate effectively. Thus, the new test format is intended not only to help institutions make better decisions about prospective students' readiness for academic coursework, but also to give test-takers the confidence of knowing they have the skills to meet the demands of their areas of study.

2 What demanded the change?

The burning question that all interested parties have in mind is what demanded the revision of the current test, given that the latest TOEFL format (the CBT) was introduced fairly recently worldwide. By and large, this is not an easy question to answer considering the extraordinary power TOEFL has gained, particularly in U.S. colleges and universities, in making decisions about accepting students into programs, allocating funds to schools and programs, awarding scholarships, licences, or certification etc. In recent years TOEFL has become one of the most popular high-stakes tests affecting not only what and how English language teachers teach but also what and how students learn (e.g. Johnson, Jordan & Poehner, 2005; Messick, 1996; Alderson & Wall, 1993). In fact, compared to other proficiency tests used for selecting students for undergraduate and graduate programs, particularly in North American colleges and universities (e.g. the University of Michigan’s Michigan English Language Assessment Battery, the English Language Proficiency Test of the College Board, etc.), TOEFL is by far the most widely used primary instrument for making admissions decisions (Jamieson, Jones, Kirsch, Mosenthal & Taylor, 2000) and the most commonly taken international English language test worldwide (e.g. Pierce, 1992; Hamp-Lyons, 1998).

A brief survey of U.S. university admission requirements for international students shows that for most of them the threshold TOEFL score has been set at 550 (PBT) or 213 (CBT), reflecting a judgment on the part of these institutions of testees’ “readiness to learn” subject matter taught in English there (Tannenbaum & Wylie, 2004). Hence, students with TOEFL scores at or above the threshold score are considered to have demonstrated a sufficient level of English proficiency to study at these educational institutions, while those with test scores below the threshold are regarded as applicants who have not yet achieved a sufficient level of language proficiency to study in academic institutions, where English is the medium of instruction. However, there have been increasing concerns among administrators, scholars, language testing researchers and educators on several issues regarding the validity of the test as a proficiency measure and the real-life consequences it has for teachers, students, test designers, and material writers. To begin with, it has been frequently commented in literature that many international students admitted with TOEFL scores of 550 and above, in actuality, enter classes with insufficient writing and oral communication skills, which in turn severely hinders their ability to fully participate in academic programs. In fact, some studies indicate (e.g. Johnson et al., 2005) that there is a great deal of ambivalence, both on the part of test-takers and teachers, surrounding the status of TOEFL as an accurate measure of English language competence – at least the way it has been operationalized in comparison with the Common European Framework (CEF) (Tannenbaum & Wylie, 2004). According to a report based on the attempt of two panels of English language experts (representing 19 European countries) to provide guidelines for comparing existing levels of learners’ language competency as identified by TOEFL and CEF, the panel-recommended-TOEFL-cut score corresponding to the C1 level (or the entering point) of the “Proficient User” band (C1-C2) on the CEF was established at 560 (PBT) or 220 (CBT). In other words, in order for an international student to gain admittance into a university where the language of instruction is English (e.g. a North American or European university), typically he or she must achieve a certain standard score on the TOEFL (traditionally set at 550–560 [PBT]/213–220 [CBT]), which corresponds to the *minimum level* of English language competence necessary for the prospective student to function in academic milieu. Yet, in actuality, this minimum score does not correspond to a completely satisfactory academic performance, which has recently seriously challenged the validity of the TOEFL test as a measure of language proficiency.[1]

Secondly, there is a commonly shared perception, especially among teachers, that the test is not testing the variety of English "used in day-to-day interactions with native speakers of English" (Johnson et al., 2005, p. 71). While it is true that TOEFL does not reflect the everyday use of English, it should be pointed out that the philosophy behind the task design is to make the test suit the purposes it serves. Put differently, the primary purpose of TOEFL is not to test the day-to-day use of English but to capture test-takers' abilities to function successfully in academic context, which as a genre is linguistically distinct from everyday language use. Consequently, the TOEFL task design is intended to closely mirror the primary purpose of the test. For instance, as described by Jamieson et al. (2000), (1) the task participants are students, faculty, and staff of different gender, ethnicity, and age; (2) the subject matter of the tasks is focusing primarily on academic, class-related or extracurricular content; (3) the setting where the language acts occur is most often the instructional (e.g. lecture halls, labs, seminar rooms, classrooms) or the academic milieu (e.g. a study room in a dormitory, the library, the bookstore, a writing or a computer center), and less often the non-academic milieu (e.g. a business office, the health center, dormitory rooms etc.); (4) the level of formality spreads over several registers such as formal (e.g. lectures, class presentations, term papers), consultative (e.g. business letters), and less often informal (e.g. casual social interactions). Thus, on the one hand, the expectation is that including academically contextualized language use would allow for systematically evaluating how this aspect of a language task contributes to task difficulty (Jamieson et al., 2000). On the other hand, the hope is that the overall academic emphasis in all of the components of the test would create a positive washback effect on ESL and English for Academic Purposes (EAP) programs preparing students for the exam.

In this regard, a significant part of the new TOEFL format development efforts have been directed to accommodating college students' academic needs in the test format. Overall, reviews of the native language (L1) and second language (L2) literature on the academic needs of students in university settings (e.g. Benson, 1991; Ginther & Grant, 1996; Waters, 1996) show that many attempts have been made to derive the underlying competencies college students need from real-world academic tasks; yet, these attempts have met with only limited success. Nonetheless, there is ample evidence for the view that English for Academic Purposes (EAP) needs are both linguistic and cultural. A number of studies (e.g. Micheau & Billmyer, 1987; Olsen & Huckin, 1990) consistently point toward the close connection between language use in academic settings, and awareness of academic cultural norms and expectations for the study in the U.S. higher education system, for example. Along the same lines, in discussions of academic discourse and academic genres (e.g. Swales, 1990), researchers often emphasize that both L1 and L2 speakers of English must adapt to the demands of the academy through a process of enculturation to the oral and written "forms of talk" of the academy (Berkenkotter & Huckin, 1993). Berkenkotter and Huckin (1993) further explain that "genre knowledge is a form of *situated cognition*, that is, knowledge that is indexical, inextricably a product of the activity and situations in which it is produced" (p. 485). This knowledge, consequently, informs students' discourse-level skills, which appear to be very important, at least from the point of view of instructional faculty (e.g. Bridgeman & Carlson, 1983), for students' academic success.

Most of the EAP needs analysis research has primarily focused on identifying students' needs with respect to the four skills areas (reading, writing, listening, speaking) and it is logical to expect that the distribution of these needs will differ according to the point of view researched (e.g. students' or instructional staff's), level of study (undergraduate or graduate), and even area of specialization. Christison and Krahnke (1986), for example, used individual interviews in order to probe into student perceptions of EAP needs. Their subjects were 80 international students studying at five U.S. universities, who identified their perceived frequency of skill use as follows: Listening (50%), followed by reading (30%), speaking (10%), and writing (10%). In terms of perceived level of difficulty related to the four skills, speaking was identified as the most difficult (35%), followed by listening (32%), reading (9%) and writing (6%). So, it can be easily noticed that the students identified as significantly more difficult the skills where the pace of language delivery was not fully in their control (i.e. speaking and listening), whereas the ones where they could set up their own pace (i.e. reading and writing) were perceived as considerably less difficult. On the other hand, research into the perception of faculty about students' academic needs suggests that instructors regard all skills to be important in academic settings, though they agree that their distribution depends primarily on the educational level of study (i.e. graduate or undergraduate). In any event, as Waters (1996) has pointed out, a major weakness of EAP needs analysis research is that it largely ignores the fact that a good deal of EAP is carried out within an integrated-skill framework. That is, quite frequently, lectures involve not only listening but also note-taking or possible subsequent incorporation of the information into some form of writing (e.g. an essay, summary etc.), as well as a certain amount of reading from different sources (e.g. blackboard, lecture handouts etc.). Similarly, writing an academic paper involves not only a good deal of reading but also a good deal of listening (e.g. lectures, discussions, expressed points of view etc.), and, finally, speaking in an academic context has increasingly become an integrated skill based on reading, listening, or writing. Therefore, there is growing awareness among all parties involved in the TOEFL test that if a test is aiming to be a test of EAP, it needs to capture the integrated nature of the use of skills in academic settings in order to be considered a reliable tool of establishing language proficiency.

Finally, the washback effect of high-stakes or institutional language tests, such as TOEFL, on teachers' instructional practices has been widely commented in the literature (e.g. Messick, 1996; Cheng, 2000; Chen, 2002; Spratt, 2005; Qi, 2005 etc.). As Swain (1985) succinctly puts it: "It has frequently been noted that teachers will teach to a test: that is, if they know the content of a test and/or the format of a test, they will teach their students accordingly." (p. 43) Bachman and Palmer (1996), however, note that washback is a more complex phenomenon than simply the effect of a test on teaching. They further argue that the impact of a test should also be evaluated with reference to societal goals and values, the educational system in which the test is used, and the potential outcomes of its use. Moreover, the nature of the washback effect – positive or negative – should be discussed with regard to, at least, three factors: the participants, the process and the product in teaching and learning (Hughes, 1993, cited in Bailey, 1999).

There is no doubt that the participants who are directly influenced by the TOEFL washback effect are the language learners, the teachers, the test developers, the materials developers and the publishers. However, their roles are so closely intertwined that it is hardly possible to look at the washback effects on one without considering the washback impact on the others. Along these lines, Andrews, Fullilove and Wong (2002) rightly point out that washback at the classroom level is largely indirect and unpredictable – indirect, because it depends on the mediation of teachers, publishers and textbook writers; unpredictable, because of individual differences among teachers and students. By and large, research findings indicate that language testing washback does influence teaching to a certain extent (e.g. the content of teaching, the time allocated to a particular skill [Andrews et al., 2002], the methodology of teaching [e.g. Alderson & Hamp-Lyons, 1996; Watanabe, 1996; Andrews et al., 2002]); yet, what we seem to be missing is sufficient evidence revealing how washback influences language learning. Research in this area is scarce and the findings do not seem to be encouraging. Andrews et al. (2002), for example, reported that the most apparent sort of washback from the changes in the Hong Kong Advanced Supplementary 'Use of English' oral exam (one of the high-stakes examinations in Hong Kong, which is a prerequisite for admission to university) appeared to be present at a very superficial level of learning (e.g. familiarisation with the exam format and the rote learning of exam-specific strategies and formulaic phrases); yet, the inappropriate use of these phrases by some of the testees, seemed to indicate mere memorisation rather than meaningful learning. Cheng's (1998) Hong Kong study also came up with negative conclusions regarding the limited washback effect on language learning since certain learning outcomes, such as students' perceptions of their motivation to learn English and their learning strategies, for example, remained largely unchanged.

As far as washback effects related to TOEFL are concerned, evidence is also slim, which opens an enormous gap to be filled in by future research. Alderson and Hamp-Lyons' (1996) study on the issue of washback effects linked to TOEFL preparation adds further support to the general conclusion of other researchers (e.g. Cheng, 1998; Chen, 2002) that while it may be relatively easy to use tests to bring about change in the content of teaching, it is much more difficult to achieve changes in teachers' methods of teaching. Some of the washback effects the researchers noted were quite superficial (e.g. allotting extra time to TOEFL classes in some institutions), while others would seem to play a more important role in determining the choice of methods used to teach exam classes (e.g. teacher attitude towards the exam, their use of exam-oriented materials, teachers' willingness to change their teaching methodology, the imbalance in the classroom interaction created by the greater dominance of the teacher and the lesser participation of the students etc.). One interesting finding from their study was the observation that the participating teachers believed it was the students who drove their test-preparation methodology by insisting on doing mostly practice tests and TOEFL-like items, which, as pointed out by Spratt (2005) raises the interesting possibility that one reason why some teachers tend to rely more heavily on exam preparation materials might be because they try to fulfil student expectations. In any event, given the dearth of research on this issue, one cannot but conclude that much more research is needed on the possible washback effects of TOEFL, especially in light of its new developments.

Another influential factor in promoting a washback effect are the TOEFL test preparation materials, particularly past papers and exam-related textbooks of different types of content (ranging from textbooks that are highly exam-technique oriented to ones that attempt to develop relevant language skills and language above and beyond the immediate content of the exam). There seems to be little question that if an exam is considered to be a sufficiently high stakes exam, it powerfully generates the publication of exam-related materials and, to a large extent, is in control of the time teachers spent working with them (e.g. Andrews et al., 2002; Spratt, 2005). It will be interesting to note here, though, that teachers and students approach with differing feelings and attitudes the use of exam preparation materials. Lumley and Stoneman (2000), for instance, found some mismatch between the attitudes of the teachers towards the contents of a learning package for a newly introduced test at tertiary level in Hong Kong and those of the students. The teachers clearly saw the potential of the materials as a teaching package in that it contained relevant and worthwhile teaching activities related to the test preparation but also going beyond it, while the students were more concerned with familiarising themselves with the test format and seemed to be less concerned with the broader suggestions for improving their language performance. In general, they showed little interest in the potential of using test preparation materials as an opportunity for language learning – a finding that sounds somewhat disturbing. Alderson and Hamp-Lyons (1996) also brought up the issue of differences in the perception of the usefulness of test preparation materials. On the one hand, the teachers felt that their methodology of using exclusively the test preparation materials and doing extensive practice with tests and TOEFL-like items was student-driven. On the other hand, some of the students felt that practice outside the classroom, such as having American friends, going to the movies, reading and generally using English outside class would be as beneficial for their test preparation as practicing test items. To find out more about the nature of the TOEFL preparation textbooks Hamp-Lyons (1998) looked at the content of exam preparation materials in a small-scale study of five TOEFL preparation textbooks. Not surprisingly, she found that the books used in her study "promote skills that relate quite exactly to the item types and item content found on the actual test rather than to any EFL/ESL curriculum or syllabus or to any model of language in use" (p. 332).

In a nutshell, the role of the test preparation materials as a source of washback has never been underestimated, though it is still largely understudied. Nonetheless, as Pierce (1992) justly noted, the mere existence of such materials is already indirect evidence of washback, though reports of TOEFL test washback "remain anecdotal" and its existence "can only be extrapolated from the vibrant industry in TOEFL preparation books" (p. 687). In Spratt's (2005) words, "We can already see clearly, however, that while the relationship between exams and washback is sometimes thought of as a simple one in which exams generate washback, these studies indicate that rather than there being a direct, automatic and blanket effect of exams, washback is more complex and elusive. It seems to be a phenomenon that does not exist automatically in its own right but is rather one that can be brought into existence through the agency of teachers, students or others involved in the test-taking process." (p. 21)

3 What is new in 2006 TOEFL–iBT format?

The new test, as its previous versions, is still organized around the four modalities (speaking, writing, listening, and reading) but in a way that allows for them to be tested both integratively and independently. That is, there are a number of tasks that assess each skill area independently as well as integratively by using combinations of prompts, such as reading a text and writing a summary, listening to a question and providing a spoken response, or reading an article, listening to a lecture, and comparing and contrasting information in an essay. Thus, when the information from the integrated tasks is combined with information from the independent tasks, the overall TOEFL will construct a fuller profile of the language abilities for each examinee.

Furthermore, as ETS explains (consult <www.ets.org/toefl> for more information), the new test format will offer a new approach to assessing the academic English language skills needed in higher education by including tasks that closely reflect the variety of academic discourse in colleges and universities. Therefore, the 2006 TOEFL–iBT exam is not just an updated version of the current CBT format but a test that includes several new components, i.e. new integrated *Writing* and *Speaking* tasks intended to measure test-takers' ability to combine information from more than one source and communicate about it. There is no longer a *Structure* section, which only means that grammar will be tested on questions and tasks integrated into each section rather than as a separate component. The new TOEFL–iBT *Reading* and *Listening* sections do not seem to be dramatically different from the ones on the current CBT format, though the lectures and the conversations in the *Listening* section are longer and the speech may include a variety of English accents (e.g. British, Australian, American). The good news is that *note-taking* is allowed. In fact, note-taking is allowed throughout the entire test, which officially assigns an academically meaningful status to the note-taking skill, which it did not enjoy previously.

For reasons of brevity, the following paragraphs will only touch upon the new revisions of the test format – that is, the inclusion of a new *Speaking* section, the revision of the *Writing* component, and the permission to *take notes* throughout all sections of the test. I will first discuss the changes in each of the components mentioned above; then, I will elaborate on the theoretical background and research findings that have promoted the modifications; and, finally, I will outline the expected outcomes and the implications of the test revisions.

3.1 The Speaking section

A new and important component of TOEFL–iBT is a multitask speaking measure. As envisioned earlier (e.g. Butler, Eignor, Jones, McNamara & Suomi, 2000), three types of speaking tasks will be included in the speaking section of the new test: *Independent speaking* (IS) tasks and two types of *integrated* tasks, combining listening, reading, and speaking (LRS) as well as listening and speaking (LS). The inclusion of integrated tasks has been advanced based on findings that with such tasks test-takers are less likely to be disadvantaged by insufficient information upon which to build their argument (Read, 1990) and, at the same time, the validity of the tasks will be enhanced by simulating real-life communication tasks in academic contexts (Wesche, 1987; Lee, 2005). In the main, the independent tasks require the test-takers to use their personal experiences or general knowledge in their responses, whereas the integrated tasks require the examinees to understand academic texts first (delivered in different modalities) and then to construct spoken responses that demonstrate understanding of the material.

The speaking tasks will be rated by judges; thus, there are at least two major concerns regarding the assessment of a component which consists of extended constructed responses rated by human judges: On the one hand, the extent to which the overall score on the speaking component may be lacking generalizability across task types (e.g. Miller & Linn, 2000) and, on the other hand, the extent to which ratings may be influenced by subjective rater judgments. Overall, current research on performance-based holistic assessments (e.g. Miller & Linn, 2000) indicates that rater variance is relatively small compared to examinee-by-task variance, which suggests that what distinguishes between the speaking performances of different test-takers is the quality of their responses to different task types rather than raters' bias. In terms of task types, some intriguing questions for teachers preparing students for the speaking component of the new TOEFL–iBT exam are whether performances on each of the three task types (given that each of them might be tapping a somewhat distinct aspect of speaking) is comparable across the speaking section as a whole and to what extent the speaking score might be negatively impacted by the heterogeneity of the tasks. Regrettably, there is insufficient empirical research to date to answer these questions, though some studies (e.g. Lee, 2005) have found that while the tasks are, on average, comparable in difficulty, they are not uniformly difficult for all examinees. Interestingly, Lee (2005) found that, in comparison to reading-speaking (RS) or IS tasks, the percentage of the testee variance was the greatest in the LS tasks, which implies that this task type is distinguishing between examinees' speaking performance better than the RS or IS sub-sections. This comes to tentatively suggest that for test preparation purposes, putting more emphasis on the listening-speaking skills may pay off well in increasing test-takers' chances of getting a higher score on the overall speaking component of the new TOEFL–iBT.

There are several aspects in the assessment of the speaking component that the tasks are designed to measure but the ones immediately noticeable are comprehensibility, coherence, and the ability to combine the appropriate information from, at least, two sources in providing a complete answer. Also, the fact that content-wise all tasks are academically situated calls for specific attention to the use of language by college and university teachers and students in the prompts as a valuable source of information for constructing a response. Unfortunately, there are few studies (e.g. Cutting, 1999; Biber et al., 2004) describing the linguistic characteristics of different spoken registers common to university life, which in actuality would leave many EFL teachers and learners to rely mostly on their intuitions about how to approach the preparation for the speaking component. Based on a relatively large corpus (over 2.7 million words) representative of the range of spoken and written registers that students encounter at U.S. universities, Biber et al. (2004), for example, found that all university spoken registers are characterized by features like present tense verbs, first and second person pronouns, contractions, rare use of passive constructions etc., which notably distinguish them from the so-called informational formal registers. This finding reveals that classroom teaching, at least in U.S. colleges and universities, is much more interactive and less fully scripted (including formal lectures) than the prepared discourse many EFL teachers and students might be culturally used to. At the same time, the features indicating a lower level of formality of academic discourse *are* well-represented in the integrated speaking task prompts and, respectively, successful responses should closely mirror these features. This becomes particularly important in light of analyses of student discourse showing that most of the L2 students use a more formal style of spoken language in academic settings (usually informed by the more formal nature of academic discourse in their own culture), while most of the L1 students speak more conversationally. Therefore, it is evident that TOEFL test preparation should address the need for students to be able to deal with a reasonably general style of English in an academic context. Likewise, it is essential that we as teachers distinguish between EAP needs in the sense of academic English and English for communicating about academic topics because, as pointed out by Waters (1996), it is the latter, not just the former, that EAP involves.

A cursory look at the 2006 TOEFL–iBT sample prompts for the speaking tasks reveals that task 1 and 2 are IS tasks, asking test-takers to briefly describe (in 45 seconds) their own experience with an academic event (e.g. a favorite class) or express an opinion related to some aspects of academic life (e.g. dormitory life). Tasks 3 and 4 combine two prompts: a short reading passage (90–120 words) that contextualizes a dialogue or a lecture (150–200 words) on the same topic. Test-takers are given 45 seconds to read the text, which is then followed by a question prompting a 60-second response. It is interesting to note here that lexically, the frequency of the vocabulary used in the reading and listening prompts is close to the word frequency typical of academic reading and speaking contexts. For example, a word frequency analysis, carried out by using *The*

Compleat Lexical Tutor (v.4) (Cobb, n.d., available at <http://www.lextutor.ca/>) of the reading passages shows that 81% – 83% of the words there belong to the 2,000 most frequently used words (West, 1953), another 9% is vocabulary that can be found in the Academic Word List (Coxhead, 2000) (e.g. *facilities, technology, research, commitment, approximately, domesticate, indicator* etc.), and the rest of the vocabulary is more specialized and falls beyond these frequency bands (e.g. *renovating, upgrading, mammals, herd* etc.). The listening prompts, based on short dialogues taking place in academic contexts (e.g. library, lecture hall, office etc.), are very close in word frequency to everyday spoken language use, regardless of the fact that the prompts are academically situated. For example, the percentage of the 2000 most frequently used words increases to 90% – 92%, while the percentage of the academically frequent lexis decreases to 3% – 4% and limits itself to a few academic words specific to the topic of discussion (e.g. a history talk) rather than the register. At the same time, the number of conversational collocations and formulaic expressions identified by using *MonoConc Pro (v. 2.0)* software (e.g. *Well/ but I mean, you know, a bunch of, another thing that* etc.) increases, which evidences the more conversational nature of academic spoken discourse in U.S. universities. In support of this conclusion, Biber at al. (2004) found that students seem to encounter generally the same structural linguistic features, regardless of their level of study or subject matter, where the physical mode of production (i.e. spoken or written) seems to be by far the most important predictor of linguistic variation in academic discourse. Therefore, a linguistically relevant spoken response should not only be comprehensible and coherent but should also be relevant to the register it reflects.

As far as the level of difficulty of the questions following the prompts is concerned, it is usually determined by the type of information the test-takers are requested to provide. In previous large-scale assessments of adults' and children's literacy studies (e.g. Kirsch & Mosenthal, 1995), by using a 5-point scale to score the difficulty of the information variable, researchers identified that questions asking highly concrete information (e.g. to identify a person, animal, or thing) were the easiest to answer; hence, they were assigned the lowest value. Questions that required examinees to identify an unfamiliar term or phrase for which respondents had to give an interpretation or express an opinion were assigned the highest value, because they were judged to be the most abstract and difficult. Following the same 1 to 5 scale for evaluation of difficulty, it would be safe to say that the tasks are of the highest difficulty, since they require the examinees, for instance, to provide *evidence* that justifies a claim, to express an *opinion* reflecting the belief or perspective of a character in the prompts or the testee herself or himself, or give an *explanation* consisting of enumeration of causes or reasons associated with an identifiable effect, outcome, or condition.

In sum, communicative competence in oral academic language requires control of a wide range of phonological and syntactic features, vocabulary, oral genres and the knowledge of how to use them appropriately (Butler et al., 2000). At the same time, it is important to realize that success in spoken interaction is determined by at least three factors, i.e. the nature of the tasks the interaction involves, the conditions under which the participants are required to perform, and the resources individuals brings to the interaction (Butler et al., 2000). In the new TOEFL–iBT test, the examinees are asked to demonstrate their oral communication skills across a variety of academic genres, functions, and situations. The tasks focus on the middle to upper range of ESL/EFL proficiency and aim at simulating realistic communicative situations by including integrated tasks – for example, ones involving listening and speaking, or reading, listening and speaking. This portion of the test is considered to be a very important one for assessing test-takers' speaking proficiency because it is hardly possible to test speaking apart from the other skills. In general, test-takers are called upon to speak about topics they are somewhat knowledgeable about, but they will also be expected to talk about subjects they are just learning about. This would allow for their responses to be rated based on a number of features of performance, such as accomplishment of task (in terms of discursive requirements, coherence etc.), sufficiency of response in terms of length and complexity, comprehensibility (including control of phonological and prosodic features), adequacy of grammatical resources, range and precision of vocabulary, fluency, and cohesion (e.g. Bachman, Lynch & Mason 1995; Butler et al., 2000). Finally, there is an expectation that the introduction of an oral communicatiion component in the new TOEFL–iBT exam will have a positive effect on the ESL/EFL teaching and learning community. By using constructed-response items, which are less likely to be coachable, learners will be encouraged to learn to communicate orally (not to learn a skill simply to do well on a test) and teachers will be encouraged to teach skills integratively.

3.2 The Writing component

Writing is the other component in the new TOEFL–iBT that has been significantly modified to measure test-takers' ability to use writing to communicate in an academic environment. There are two writing tasks in the new test format: an integrated task, requiring the examinees to read a short passage, listen to a brief lecture and then answer a question in writing based on what they have read and heard; and another task (very similar to the writing tasks in the Test of Written English [TWE] and TOEFL–CBT writing component) asking the test-takers to respond in writing to a question based on their own knowledge and experience. Students will be allowed to take notes during the reading and listening prompts and the time allotted to each task differs depending on the nature of the task – that is, 20 minutes response time for the integrated task and 30 minutes for the essay. Thus, by using a reading and listening text as content input to the integrated writing prompt, the new writing task is aiming at supporting test-takers in their response construction by providing them in the prompts not only with some information for their writing, but also with some vocabulary to lean upon and some genre conventions to model their response upon. With the argumentative essay, the challenge is to keep the breadth of the task balanced because some research (e.g. Hamp-Lyons & Mathias, 1994) suggests that when the task is very narrow, there is less room for a test-taker to reveal his or her unique qualities as a writer and more room to fail to write within the acceptable boundaries of academic writing. On the other hand, research evidence also shows that the more freedom there is in a writing task, the more a writer is thrown onto his/her own resources, and the more seriously he/she can be disadvantaged if those resources are limited (Hamp-Lyons & Kroll, 1997). As far as the quality of the writing responses is concerned, both tasks will be judged not so much on their length but on the completeness and accuracy of content, the development of ideas, the organization of the compositions, and the quality and accuracy of the language used to express meaningfully connected ideas. Hence, in this format, the writing component of the new TOEFL–iBT exam significantly differs from its predecessors (TWE or the writing component of the TOEFL–CBT), which is definitely a step forward in testing writing competence. The modifications also chime well with the fact that most of the academic practices in the colleges and universities of English-speaking countries require the production of written texts within an integrated-skill framework, in which a student should show competence both in the subject matter and the writing genre. In most colleges, a great deal of writing is expected of college students in a wide range of contexts – from the formal and analytic (e.g. writing research papers, theses, analytic responses to texts etc.) to the less formal and critical contexts (e.g. lab reports, creative writing etc.). Other types of writing also play an important role in students' academic life. For example, note-taking in lectures and from text materials, summarizing, keeping learning logs and journals etc. all place heavy demands on students' written competence, while at the same time critically contributing to their understanding of the ideas they encounter (Hamp-Lyons & Kroll, 1997).

In a nutshell, it has been frequently noted in the EAP literature that while good listening and reading skills are essential to students' reception of the knowledge the academy has to offer, speaking and writing are essential to students' knowledge integration and production. There have been numerous attempts to shed light on what writing skills students from a wide cross-section of academic disciplines need to successfully accomplish different academic writing tasks. In this regard, among other skills, Leki and Carson (1994) highlight the importance of students' ability to synthesise ideas from multiple sources as an input to writing, to constructing well-organized and coherent texts, and select with sufficient speed the most concise form of expression to use. Thus, by acknowledging that learning purposes calling for writing rarely exist completely separated from the purposes involving wider language functions, the new format of the writing component also acknowledges the need for integrating testing of writing with the other language skills. From this perspective, researchers and test designers unanimously agree that in the new TOEFL–iBT exam, it is imperative to ensure that the kinds of writing tested reflect as closely as possible the academic communicative demands of English-speaking colleges and universities across disciplines and levels of education. Finally, many arguments about possible positive washback effects in writing practices have been advanced in favor of modifying the writing section in the new TOEFL–iBT exam. Accordingly, the expectations are that writing will receive the attention it deserves in the ESL/EFL programs not only as a test preparation component but as a life-long skill that shapes a well-rounded language user.

3.3 Note-taking

Allowing note-taking across all sections of the new TOEFL-iBT exam is a major improvement of the test format, which takes it closer to reproducing authentic academic tasks college students perform on a regular basis. In the new TOEFL, note-taking is perceived as an important complement to all tasks but, probably, most useful in the integrated tasks where multiple sources of information need to be combined. Research shows that students intuitively view note-taking as the primary means of creating "a record of information" of the academic activities they are involved in. In a study of college students' perception of the primary purposes of note-taking, Van Meter, Yokoi and Pressley (1994), for instance, found that students assign a number of goals to it relating to attention (e.g. focus of attention to an information source), understanding (e.g. facilitation of comprehension and memory of material), organization (e.g. aid to a better connection of ideas), structure (e.g. holistic representation of information content), study and homework aid, and others. In investigating the perceived effect of note-taking on TOEFL listening comprehension tasks, other researchers (e.g. Carrell, Dunkel & Mollaun, 2002) enriched the list by adding several other advantages students attribute to note-taking. For instance, students have reported that under test conditions their level of comfort and ease with the listening tasks has significantly increased as a result of being allowed to take notes; that note-taking has aided their performance in answering questions about the lectures, and that their recall of information has been positively influenced by their notes. In addition, researchers believe that being allowed to take notes seems to aid students' processing of information, though test-takers generally admit that they have difficulty using their notes under time constraints. Nonetheless, the interesting question whether or not there is a match between students' feelings about the benefit of note-taking and the actual effect of note-taking on their test performance remains still unanswered and is yet to be researched.

Most of the L2 research on note-taking has focused primarily on examining its relationship to listening comprehension (in particular to lecture comprehension) or overall learning outcomes. Yet, findings about its effects on these two variables are far from being conclusive, which further underscores the importance of looking at comprehension and learning as relatively independent outcomes. Studies investigating the link between note-taking and academic lecture comprehension (e.g. Olsen & Huckin, 1990) emphasize that effective lecture listening comprehension involves not just an understanding of the macro-markers but also detecting the overall schemata that the speaker is representing. As far as research regarding the relationship between note-taking and learning outcomes goes, inquiries into whether note-taking is facilitative, debilitative, or of no particular usefulness to information recall, are, to say the least, inconclusive. Some researchers (e.g. Gibbs, 1981) cite sources showing that taking notes is not necessarily associated with better learning outcomes than not taking notes; other researchers (e.g. Hartley & Davies, 1978) argue that there is conflicting evidence regarding the facilitative effects of note-taking on information recall. In any event, there is evidence suggesting that test achievement may not be so much related to the quantity but rather to the quality of note-taking. Dunkel (1988), for instance, investigated lecture note-taking among undergraduate L1 and L2 speakers of English, who took notes while viewing a video-taped lecture and then were given a test. Interestingly, the analysis of students' notes and their relationship to the test results indicated that test achievement was more closely linked to terseness and inclusion of main points in the process of note-taking than to the quantity of notes. Overall, the analysis revealed that effective L1 and L2 note-takers were those who compacted large amounts of spoken discourse into propositional-type information units, transcribed content words (e.g. names, dates, or statistics) using abbreviations, symbols and a few structure words. The L2 note-takers who did not perform well on the test wrote down numerous structure words (e.g. articles and prepositions) so that their notes contained fewer information units overall but a larger quantity of words or notations. These findings seem to suggest that good note-takers are essentially good summarisers, which highlights the importance of the skill of abstracting and reformulating the gist of information as a vital and widely applicable academic skill. Yet, I should hasten to add here that researchers frequently note that there is no single, unitary note-taking method which is effective for all groups of students, at all levels of education (Dunkel, 1988; Waters, 1996). However, what needs to be more seriously researched is the role of students' conceptualisation of the process of note-taking and how it can be reconstructed, if needed, to better serve their academic needs.

In sum, test-takers consistently report that note-taking enhances their level of comfort with the TOEFL tasks. This self-perceived comfort of being able to jot down notes throughout all test tasks may also allow examinees to demonstrate a higher level of performance since they will not have to rely so heavily on their memories to store all the information from the prompts while constructing their responses. Instead, they would be able to reference their notes to check information asked in the test questions. Furthermore, the face validity of the test will substantially improve because, in reality, academic life strongly encourages students to take notes and use them in their academic preparation. In the context of the new generation TOEFL test, the intuitive belief held by test designers is that allowing note-taking under exam conditions will encourage TOEFL examinees to place greater value on their ability to take notes across a wide variety of tasks and use them meaningfully in their performance.

4 Conclusion

A test that is aiming at being a test of proficency needs to be analytical, communicative, and integrative, among other things (Davies, 2003). Unfortunately, until the introduction of the TOEFL–CBT format in 2000, the test had remained almost unchanged for over 40 years, "having no truck with the communicative revolution" (Davies, 2003, p. 360), which makes the new revision of the format more than timely. It should be noted here that the new generation 2006 TOEFL–iBT exam is a result of the collective effort of testing specialists, researchers, material designers, and teachers. The new test format has evolved from the understanding that for the assessment of university-level English language skills we first need to fully understand the linguistic challenges faced by students in university contexts. There are obviously special demands presented by academic reading and writing, especially in relation to textbooks, research papers, and student essays and term papers. There are also special demands associated with academic listening and speaking, which are tightly linked to students' abilities to participate in classroom activities, such as discussions, presentations, team projects etc. The inclusion of a new *Speaking* section, the use of integrated tasks in the *Speaking* and *Writing* components, the incorporation of *note-taking* in the test format, and the emphasis on communicative competence are hoped to have impact on the way English is taught in the future. However, all interested parties in this high-stakes exam are yet to find out whether the new test format will increase their confidence that the test results adequately reflect test-takers' ability to communicate and be successful in their academic studies.

Author's Note

The author has never been involved with the Educational Testing Service in any capacity and the views presented in the article do not reflect the views of the organization. The author would also like to thank two anonymous reviewers for their insightful comments on an earlier version of the manuscript, which substantially improved the quality and the clarity of the discussed issues.

Correspondence concerning this article should be addressed to Alla Zareva, Faculty of Humanities, Bourgas Free University, 62 San Stefano St., Bourgas 8001, Bulgaria. E-mail: azareva@bfu.bg

Notes

[1] By and large, the investigation of the relationship between language tests and academic outcomes is not a futile line of inquiry (Fox, 2004). Interest in the relationship between test results and academic performance relates to a large extent to the predictive validity of a test that claims to be an EAP test. In other words, one of the most important questions in this regard is the extent to which inferences drawn from test performances may be deemed more or less valid, as Bachman (1990) points out, on the basis of the extent to which performance on the test is consistent with predictions we make on the basis of a theory of abilities. Because the TOEFL claims to test English use for academic purposes, it is of critical importance to its validity as a measure of language proficiency to know that the test accurately identifies the language threshold and usefully predicts the learning trajectory of the test-takers.

References

Alderson, J.C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115-129.

Alderson, J.C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing, 13*, 280-297.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback – a case-study. *System, 30*, 207–223.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Bachman, L.F., Lynch, B.K, & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Bailey, K.M. (1999). *Washback in language testing* (TOEFL Monograph Series No. 15). Princeton, NJ: Educational Testing Service.

Benson, M.J. (1991). University ESL reading: a content analysis. *ESP Journal, 10*, 75-88.

Berkenkotter, C., & Huckin, T. (1993). Rethinking genre from a sociocognitive perspective. *Written Communication, 10*, 475-509.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. (TOEFL Monograph Series No. 25). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Carlson, S.B. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students*. (TOEFL Research Report No. 15). Princeton, NJ: Educational Testing Service.

Butler, F.A., Eignor, D., Jones, S., McNamara, T., & Suomi, B.K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph No. MS-20). Princeton, NJ: Educational Testing Service.

Carrell, P.L., Dunkel, P.A., & Mollaun, P. (2002). *The effects of notetaking, lecture length and topic on the listening component of TOEFL 2000*. (TOEFL Monograph Series No. 23). Princeton, NJ: Educational Testing Service.

Chen, L. (2002). *Washback of a public exam on English teaching*. (ERIC Document Reproduction Service No. ED 472167)

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation, 24*, 279-300.

Cheng, L. (2000). Washback or backwash. A review of the impact of testing on teaching and learning. (ERIC Document Reproduction Service No. ED 442280)

Christison, M.A., & Krahnke, K.J. (1986). Student perceptions of academic language study. *TESOL Quarterly, 20*, 61-81.

Cobb, T. (n.d.) The Compleat Lexical Tutor (v.4). Retrieved June 4, 2005, from http://www.lextutor.ca/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238.

Cutting, J. (1999). The grammar of the in-group code. *Applied Linguistics, 20*, 179-202.

Davies, A. (2003). Three heresies of language testing research. *Language Testing, 20*, 355–368.

Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and its relation to test performance. *TESOL Quarterly, 22*, 260-279.

Educational Testing Service (n.d.). Retrieved 4 June, 2005, from Educational Testing Service Website, http://www.ets.org/toefl

Fox, J. (2004) Test decisions over time: Tracking validity. *Language Testing, 21*, 437-465.

Gibbs, G. (1981). *Teaching students to learn.* Milton Keynes, England: Open University Press.

Ginther, A., & Grant, L. (1996). *A review of the academic needs of native English-speaking college students in the United States*. (TOEFL Monograph Series No. 1). Princeton, NJ: Educational Testing Service.

Hamp-Lyons, L. (1998). Ethical test preparation practice: the case of the TOEFL. *TESOL Quarterly, 32*, 329-337.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000-writing: Composition, community, and assessment.* (TOEFL Monograph Series No. 5). Princeton, NJ: Educational Testing Service.

Hamp-Lyons, L., & Mathias, S. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3,* 1.

Hartley, J., & Davies, I. K. (1978). Notetaking: A critical review. *Programmed Learning and Educational Technology, 15*, 207-224.

Hughes, A. (1993). *Backwash and TOEFL 2000.* Unpublished manuscript, University of Reading.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper*. (TOEFL Monograph Series No. 16). Princeton, NJ: Educational Testing Service.

Johnson, K.E., Jordan, S. R., & Poehner, M.E. (2005). The TOEFL trump card: An investigation of test impact in an ESL classroom. *Critical Inquiry in Language Studies: An International Journal, 2,* 71-94.

Kirsch, I.S., & Mosenthal, P.B. (1995). Interpreting the IEA reading literacy scales. In M. Binkley, K. Rust & M. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study* (pp. 135-192). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Lee, Y. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. (TOEFL Monograph Series No. 28). Princeton, NJ: Educational Testing Service.

Leki, I., & Carson, J. G. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly, 28*, 81-101.

Lumley, T., & Stoneman, B. (2000). Conflicting perspectives on the role of test preparation in relation to learning? *Hong Kong Journal of Applied Linguistics, 5*, 50-80.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241-256.

Micheau, C., & Billmyer, K. (1987). Discourse strategies for foreign business students: Preliminary research findings. *ESP Journal, 6*, 87-97.

Miller, M.D., & Linn, R.L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*, 367-378.

MonoConc Pro Version 2.0. [Computer Software]. Athelstan.

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

Pierce, B.N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly, 26,* 665-691.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9,* 109-121.

Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research, 9,* 5–29.

Swain, M. (1985). Large-scale communicative testing: A case study. In Y.P. Lee, A.C.Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 35-46). Oxford: Pergamon Press.

Swales, J. (1990). *Genre analysis*. Cambridge, England: Cambridge University Press.

Tannenbaum, R.J., & Wylie, E.C. (2004). *Mapping test scores onto the Common European Framework: Setting standards of language proficiency on the Test of English as a Foreign Language (TOEFL), the Test of Spoken English (TSE), the Test of Written English (TWE), and the Test of English for International Communication (TOEIC).* Princeton, NJ: Educational Testing Service.

Van Meter, P., Yokoi, L., & Pressley, M., (1994). College students' theory of note-taking derived from their perceptions of note-taking. *Journal of Educational Psychology, 86,* 323-338.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13,* 318-333.

Waters, A. (1996). *A review of research into needs in English for academic purposes of relevance to the North American higher education context*. (TOEFL Monograph Series No. 6). Princeton, NJ: Educational Testing Service.

Wesche, B. (1987). Second language performance testing: the Ontario test of ESL as an example. *Language Testing, 4,* 28-47.

West, M. (1953). *A general service list of English words.* London: Longman, Green.