Language Assessment Quarterly
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hlaq20

Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination

Ardeshir Geranpayeh [a] & Antony John Kunnan [b]
[a] University of Cambridge ESOL Examinations ,
[b] California State University , Los Angeles, USA
Published online: 05 Dec 2007.

To cite this article: Ardeshir Geranpayeh & Antony John Kunnan (2007) Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination , Language Assessment Quarterly, 4:2, 190-222, DOI: 10.1080/15434300701375758

To link to this article: http://dx.doi.org/10.1080/15434300701375758

LANGUAGE ASSESSMENT QUARTERLY, *4*(2), 190–222

Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination

Ardeshir Geranpayeh
University of Cambridge ESOL Examinations

Antony John Kunnan
California State University, Los Angeles

When standardized English-language tests are administered to test takers worldwide, the test-taking population could be varied on a number of personal and educational characteristics such as age, gender, first language, and academic discipline. As test tasks and test items may not always be prepared keeping this diversity of characteristics in mind, it is essential for test developers to continuously monitor their tests in terms of whether all test takers are receiving a fair test. This study investigates whether the test items on the listening section of the Certificate in Advanced English examination functioned differently for test takers from three different age groups. The main results showed that although statistical and content analyses procedures detected differential item functioning in a few items, expert judges could not clearly identify the sources of differential item functioning for the items.

It is a truism to state that tests have to be fair to test takers so that test-score interpretations made by test users (e.g., for admissions officers and employers) are valid. Although test design, development, administration, and scoring procedures could take into consideration test fairness, many tests discover the problem too late in the test design-development-administration-scoring cycle. One approach to this problem has been to examine test scores from a pilot group or, if the test has already been launched, to examine test scores from a large sample of

Editor's note: The action editor for this manuscript was Associate Editor James Purpura.

Correspondence should be addressed to Ardeshir Geranpayeh, Principal Research & Validation Coordinator, University of Cambridge ESOL Examinations, 1 Hills Road, Cambridge CB1 2EU, United Kingdom. E-mail: geranpayeh.a@cambridgeesol.org

test takers and detect items that function differently for different test taking groups and to investigate the source of this difference. This approach is called differential item functioning (DIF), and it has been popular since the 1980s.[1] Recent codes and standards such as the Code of Fair Testing Practices in Education (Code; 1988, 2004) from the Joint Committee on Testing Practices in Washington, DC. and the Standards for Educational and Psychological Testing prepared by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999) have suggested DIF as a useful way of examining tests for their fairness. Researchers have also called for the use of DIF as an approach to address test fairness (Camilli & Shepard, 1994; Holland & Thayer, 1988; Kunnan, 1997, 2000, 2004).

CONTEMPORARY APPROACH

For more than two decades, the focus of DIF/test bias analysis was on the concept of *relative item difficulty* for different test-taking groups. The idea was to conduct a posttest administration (post hoc) studies to examine the performance of test takers with similar ability (as measured by the total score) from different subgroups with the expectation that there would be comparable individual item difficulty for the subgroups, as the test takers are matched in terms of overall ability. In cases where items performed or functioned differently for subgroups, such items were to be flagged and examined for potential content or response format bias. Considerable literature (e.g., Holland & Wainer, 1993) has developed around this concept and accompanying procedures. The major benefit from this procedure, as Camilli and Shepard (1994) put it, has been to help "clarify what a test is measuring and highlight the influence of irrelevant factors" (p. 16).

A more recent approach has focused on the concept that the general cause of DIF is the presence of *multidimensionality in items* displaying DIF (Ackerman, 1992; Shealy & Stout, 1993). Expanding on this, Roussos and Stout (2004) stated that "such items measure at least one secondary dimension in addition to the primary dimension that the item is intended to measure" (p. 108). Therefore, as DIF methods are based on comparable test takers matched with respect to the primary dimension or construct the test item is measuring, a large DIF value could mean the test item is measuring additional dimensions differently across the reference and the focal groups. The additional dimensions could be either intended secondary dimensions called an *auxiliary* or *benign* dimension or an unintended secondary dimension called an *adverse* or *nuisance* dimension that has crept into the test

[1]Until recently this type of research was popularly but somewhat erroneously called test bias.

item. This multidimensional paradigm of DIF offers researchers avenues to make judgments regarding these secondary dimensions.

Roussos and Stout (1996, 2004) call this approach a multidimensionality-based DIF analysis procedure that incorporates dimensionality analysis with DIF analysis. They suggest a two-step approach: development of DIF hypotheses and testing of DIF hypotheses. The development of DIF hypotheses can be made through referring test items and items that have already been flagged to item-writing specialists for their expert judgments. The testing of DIF hypotheses is to be based on the secondary dimensions that have been hypothesized in the development stage. Based on the sets of items that have secondary dimensions, test bundles can then be formed and differential bundle functioning (DBF) can then be estimated. Item bundles that have statistically significant DBF estimates can be said to have secondary dimension that exhibit strong evidence of causing DBF. Then, if a secondary dimension is identified as an auxiliary dimension, the DIF is labelled *benign*, whereas if the secondary dimension is identified as a nuisance dimension, the DIF is labelled as *adverse*. At this stage, the test developer or researcher should determine the action that needs to be taken with the items that have displayed DIF: If the DIF is statistically significant for items with an adverse secondary dimension, the items that are in question may need to be revised or dropped from the test and such items not included in further tests. If the DIF effect size is statistically significant but small, the test developer or researcher may alert test writers to the secondary dimension of the test items and their impact. If the DIF is statistically significant for items with a benign secondary dimension, test writers could be alerted to the nature of the secondary dimension (and as to how much of a role it should play).

PREVIOUS EMPIRICAL STUDIES

Following the popularity of DIF studies in educational testing, numerous DIF studies in language assessment have been conducted since 1985 as the main approach to examine tests for fairness. Table 1 presents the most widely discussed ones. The main focus was generally on whether test items exhibited DIF and not on the identification of the sources of DIF. The most popular studies focussed on whether test items exhibited DIF when test takers were from different native language backgrounds. This may be due to the findings from second-language learning studies that suggest when language learners study or test takers take a test in a second language, the main influence will be from their native languages. The influence of other personal test taker characteristics such as gender and academic major also were also explored. But none of the studies used a multidimensionality-based DIF detection method or tested any DIF hypotheses. Although we have benefited from these studies, no clear and definitive findings regarding DIF in tests based on test taker characteristics have emerged as yet.

TABLE 1
Empirical Studies in Language Testing Focusing on Test Bias (1980–2005)

Author and Year of Study	*Specific Focus*
Swinton & Powers (1980)	Native language
Alderman & Holland (1981)	Native language
Shohamy (1984)	Test method
Alderson & Urquhart (1985a, 1985b)	Academic major
Chen & Henning (1985)	Native language
Zeidner (1986, 1987)	Gender, minorities
Hale (1988)	Major field and test content
Oltman, Stricker, & Barrows (1988)	Native language
Kunnan (1990, 1992)	Native language, gender
Sasaki (1991)	Native language
Shohamy & Inbar (1991)	Question type and listening
Ryan & Bachman (1992)	Gender
Kunnan (1995)	Native language
A. Brown (1993)	Tape-mediated test
Ginther & Stevens (1998)	Native language, ethnicity
Norton & Stein (1998)	Text content
J. D. Brown (1999)	Native language
Takala & Kaftandjieva (2000)	Native language
Lowenberg (2000)	Different Englishes
Kim (2001)	Native language
Pae (2004)	Academic major
Uiterwijk & Vallen (2005)	Native language

None of the studies cited examined the relationship between test performance and test takers' age. This is an important concern in the Cambridge English for Speakers of Other Languages (ESOL) Main Suite examinations, as there has been a shift in the traditional test population, where test takers of many age groups are taking these cognitively challenging tests. Anecdotal historical observations have indicated that if there were going to be any DIF in these examinations, it was likely to impact mainly listening items. Banks (1999) and Geranpayeh (2001) examined the country and age bias in examinations such as the First Certificate in English and Preliminary English Test and recommended further investigation into DIF of listening item types. As there had been no empirical research in this area with the Certificate in Advanced English (CAE) examination, it was decided to investigate whether listening test items would exhibit DIF across age groups. However, although we wanted to use the multidimensionality-based DIF analysis procedure and test DIF hypotheses, we were unable to do this because we could not obtain explicit statements of primary and secondary dimensions regarding the test items from test developers.

RESEARCH QUESTIONS

There were two research questions:

1. Do CAE listening paper test items exhibit DIF toward test taker groups in terms of age? And if so, to what extent?
2. Are CAE listening paper test items *biased* toward test taker groups in terms of age? And if so, to what extent?

To answer these questions, statistical analyses were performed first, items were then flagged, and content analyses were performed on these items as well as on the items that were not flagged.

METHOD

Data

Data in this study are based on 4,941 test takers who took the Cambridge CAE examination in December 2002. Test takers' background information was collected through electronic Candidate Information Sheets (CIS) completed before the test administration. CIS included information about each candidate's gender, age, first language, years of study in English, and previous exams taken.

There were three versions of the listening paper, but our study only reports the performance of those who took Version 1 of the CAE listening paper. Test takers were divided into three age groups: 17 and younger, 18 to 22, and 23 and older. It was assumed that 18- to 22-year-old test takers are the target test takers for a CAE examination, which represent a range of test takers finishing high school to those studying at college. Seventeen-year-old and younger test takers (called younger test takers) were assumed to be mainly high school students, whereas 23-year-old and older test takers were considered to be mature test takers. There were 83% of the test takers who fell into the first two age groups, which is also a typical representation of the CAE test takers.

The test takers from this pool of 4,941 test takers were randomly reduced to 1,000 by BILOG-MG (Scientific Software International, 2003) keeping in mind the proportion of test takers by age. This was done to make item estimation easier and to facilitate the interpretation of the significant differences found in any analysis. Large sample sizes tend to show statistical significant differences with minor variations in samples' performance, which in turn make the meaningfulness of the differences difficult to justify. Table 2 illustrates test takers' distribution by age of the sample.

TABLE 2
Distribution of Test Takers by Age

Age	*Original Pool*	*Random Sample*	*% of Total*
17 and younger	1,871	411	41.1
18 to 22	2,247	422	42.2
23 and older	823	167	16.7
Total	4,941	1,000	100.0

N = 1,000.

Instrument

The CAE listening paper contains four parts. Each part contains a recorded text or texts and corresponding comprehension tasks. The texts in Parts 1, 3, and 4 are heard twice; the text in Part 2 is heard once only. The recordings contain a variety of accents corresponding to standard variants of English native speaker accent and to English nonnative speaker accents that approximate to the norms of native speaker accents. Background sounds are included before speaking begins to provide contextual information. Subdued reaction from an audience to talks and speeches is also included. For all parts of the paper, test takers write their answers on an answer sheet; each question in the paper carries one mark. Items were scored dichotomously. More details of the topic, task type, and response format are provided in Appendixes A and B.

Analytical Approaches

Two complementary approaches were used in this study: statistical analysis and content analysis.

Statistical analysis. The item response theory (IRT)-based DIF detection procedure implemented in BILOG-MG, which compares latent trait item difficulty parameters across groups, was used for the study. To satisfy the IRT assumption of unidimensionality, an exploratory factor analysis was conducted. A one-factor solution was obtained with an Eigenvalue of 2.2 accounting for 55% of the total variance. The sample response data (n = 1,000) was read into BILOG-MG for the first analysis. The Marginal Maximum Likelihood Ratio Test (Thissen, Steinberg, & Wainer, 1993) was used to investigate DIF that is said to be present when the probabilities of success on a given item are invariant between two or more groups at the same ability level. We assumed that DIF does not extend to the item discriminating powers. In other words, the b_j parameters for the separate groups were estimated on the assumption that the slope parameters, a_j, were homogeneous across groups.

We ran two different models: the compact model and the augmented model. In the compact model no group differences were assumed, whereas in the augmented model, we assumed that the items being investigated would exhibit DIF. We first tested the compact model, analysing the data in a single group as though they came from the same population, and calibrated the items accordingly. We noted the marginal maximum log likelihood of the item parameters in the final Newton cycles (labelled *–2 log likelihood* in the output). We then analysed the data in separate groups using the augmented model, assuming the presence of DIF in the items and again noted the final -2 log likelihood. Using a chi-square test, we tested for a significant difference between the two final -2 log likelihood estimates.

Content analysis. Following Roussos and Stout (2004) in terms of identifying item dimensionality, the approach used here was to have subject officers[2] first examine the items that were identified as ones exhibiting DIF. Therefore, the items that showed DIF were submitted to the subject officer responsible for the CAE listening paper as well as five other content experts for further review. The subject examiner was the responsible officer for constructing the listening test items. She was asked to examine the items and decide if there was any evidence suggesting that the items favored any of the age groups in question. The five content experts included two experienced listening subject officers responsible for other Cambridge ESOL Main Suite exams, one experienced nonlistening subject officer, one relatively new nonlistening subject officer, and one newly appointed subject officer who had just been employed with at least 5 years of teaching experience. An additional analysis included the content experts examining the other test items (the ones that did not exhibit DIF) to identify items that might advantage a particular group of test takers even though statistical analysis did not identify them as displaying DIF. This broadened the scope of the content analysis.

The content experts were asked to rate the suitability of the test items for each age group using a questionnaire with a 5-point scale (see Appendix C for the questionnaire). The experts rated each item on a scale from 1 (*strongly advantage*) to 2 (advantage) to 3 (*neither advantage nor disadvantage*) to 4 (*disadvantage*) to 5 (*strongly disadvantage*). One of the researchers who conducted this study briefed the experts about the nature of the task and what was expected of them. In addition to the use of the rating scales, the experts were asked to fill in the comment field if they felt there was an area that was not captured in the questionnaire but was relevant to the performance of the age groups. The content experts were given a copy of the question paper with the test items, the listening

[2]Subject Officers in Cambridge ESOL are the experts responsible for the coordination of activities for constructing question papers. The expert in this case was responsible for the CAE Listening Comprehension paper in question and was engaged closely in selecting the original items for test construction.

transcripts, an audiotape of the listening paper, and the rating scale (see Appendix B for question paper with test items and the listening transcript). The expectation of the content analysis was that if there was evidence that the items that showed DIF were advantaging a particular group of test takers, then it might be possible to conclude that such items may be biased.

RESULTS AND DISCUSSION

Descriptive Test-Level and Item-Level Statistics

Table 3 reports mean scores, standard deviation, skewness, kurtosis, and mean correct and mean point biserial correlation for each age group and overall. Mean scores indicate that the groups are performing similarly on the test. However, the performance of the 17 and younger age group appears to be slightly lower than the other two groups. Although the difference is only two and a half raw marks, it may indicate some adverse impact of the test for this particular group. Skewness and kurtosis values show that the distribution of the data can be considered normal, the mean correct shows the mean item difficulty, and the mean point biserial correlation shows mean item discrimination.

Table 4 shows the item difficulty, Pearson (point biserial), and biserial correlation values for all 32 items in the test. From these indexes, it can be observed that there is a range of item difficulty and item discrimination values: Item 3 is the most difficult or least easy (0.37 on percent correct and 0.55 on logit) with relatively low item discrimination (point biserial coefficient = .32); Item 10 on the other hand is easily the least difficult or easiest item (0.91 on percent correct and –2.36 on logit) but has lower item discrimination (point biserial coefficient = .22).

TABLE 3
Descriptive Statistics of Raw Scores by Age

	Age Groups			
	17 & Younger	*18 to 22*	*23 & Older*	*Average*
M	18.56	20.56	21.00	19.88
Mdn	19.00	21.00	21.00	20.00
SD	5.45	5.51	5.07	5.52
Skewness	–0.06	–0.46	–0.29	–0.28
Kurtosis	–0.49	–0.21	–0.66	–0.46
Mean correct	58%	64%	66%	62%
Mean point biserial correlation	.36	.38	.35	.37

N = 1,000.

TABLE 4
Item Statistics Output From BILOG-MG Using 1-Parameter Logistic Calibration Model

	Item Difficulty[a]			
Item	*% Correct*	*LOGIT*	*Item Discrimination*[b]	*Biserial*
1	0.60	–0.41	0.35	0.44
2	0.51	–0.02	0.33	0.41
3	0.37	0.55	0.32	0.40
4	0.56	–0.23	0.29	0.37
5	0.47	0.10	0.39	0.50
6	0.83	–1.61	0.29	0.43
7	0.52	–0.06	0.41	0.52
8	0.63	–0.52	0.43	0.55
9	0.78	–1.28	0.08	0.12
10	0.91	–2.36	0.22	0.39
11	0.44	0.22	0.37	0.46
12	0.60	–0.42	0.34	0.43
13	0.73	–1.00	0.24	0.32
14	0.78	–1.28	0.25	0.35
15	0.51	–0.05	0.29	0.36
16	0.44	0.26	0.34	0.43
17	0.44	0.26	0.22	0.28
18	0.50	0.00	0.16	0.20
19	0.50	0.00	0.18	0.23
20	0.75	–1.11	0.40	0.55
21	0.60	–0.42	0.34	0.43
22	0.68	–0.77	0.14	0.18
23	0.83	–1.59	0.35	0.52
24	0.71	–0.91	0.44	0.58
25	0.70	–0.85	0.22	0.29
26	0.73	–1.02	0.35	0.47
27	0.61	–0.44	0.42	0.54
28	0.89	–2.04	0.31	0.51
29	0.64	–0.59	0.42	0.55
30	0.45	0.20	0.09	0.11
31	0.55	–0.21	0.40	0.50
32	0.73	–1.00	0.40	0.54

Note. $N = 1{,}000$.
[a]% correct is based on Classical Test Theory and logit estimated by item response theory. [b]Item discrimination index is based on the point-biserial Pearson correlations.

DIF Analysis

Table 5 shows that the difference between the –2 log likelihood values for the two models was significant, indicating that the augmented model better explained the data, that is, there was evidence that DIF was present.

TABLE 5
Chi-Square Test Results for Age Comparisons

	–2 Log Likelihood
Compact	37,095.96
Augmented	36,886.08
G2	209.88*

*$p < .05$, $df = 64$.

The next step was to identify which items showed DIF. The software BILOG-MG produces a table for the group threshold differences of item difficulty for each item in the augmented model for the three age groups. Group 2 (age 18–22) was chosen to be the reference group, as it comprises the biggest proportion of the CAE test takers and is considered to be the target age group, and Groups 1 and 3 were chosen to be the focal groups. As seen in Table 6, information is presented for all 32 items. The first row of information for each item is the Group Threshold Differences between Groups 1 and 2 and Groups 2 and 3. The second row of information for each item is the standard error estimates for the threshold differences. If the group threshold difference was bigger than two standard errors, we considered the threshold difference between the two groups to be significant at $p < .05$.

There are 64 DIF comparisons, and simply by chance alone, 5% or at least three significant results may be found. Table 6 shows that six items exhibit DIF: Items 4, 11, 18, 20, 21 and 27. Of these six items, only Item 4 shows DIF in both group comparisons (Groups 1 and 2 and Groups 2 and 3). A discussion of Item 4 follows.

Item 4. This item is in Part 1 (Environmental Adviser), and the subskill targeted (or primary dimension, if you like) is the ability to recall explicitly mentioned detail. The correct answer is "conservation group," but an analysis of the Common Wrong Answers[3] reveals that the three most frequent wrong answers are "conversation group," "consultation group," and "conciliation group." Were the younger group writing "conversation group" because it was familiar to them from conversation classes in school? Were the older group writing "consultation group" and "conciliation group," which are more sophisticated answers but

[3]Common Wrong Answer analysis is an exercise carried out in Cambridge ESOL tests with productive tasks in listening comprehension items, where a sample of approximately 500 live responses are captured for any possible acceptable response that might not have been conjectured at the time of test construction.

TABLE 6
Group Threshold Differences

Item	Group Differences: Between Group 1 and 2	Group Differences: Between Group 2 and 3	Item	Group Differences: Between Group 1 and 2	Group Differences: Between Group 2 and 3
1	–0.17	0.07	17	–0.42	0.30
	0.18	0.24		0.17	0.23
2	–0.09	–0.05	18	–0.01	**0.57**
	0.18	0.24		0.17	0.22
3	0.06	–0.20	19	–0.31	0.38
	0.18	0.23		0.17	0.22
4	**–0.46**	**–0.61**	20	**–0.55**	–0.07
	0.18	0.23		0.21	0.30
5	0.11	–0.15	21	0.03	**0.61**
	0.18	0.24		0.18	0.24
6	0.10	–0.32	22	–0.53	0.37
	0.23	0.34		0.18	0.23
7	0.23	–0.17	23	–0.00	–0.09
	0.18	0.24		0.23	0.33
8	0.13	–0.11	24	–0.02	–0.40
	0.19	0.27		0.20	0.29
9	–0.17	0.13	25	–0.17	0.02
	0.20	0.27		0.18	0.25
10	–0.14	0.14	26	0.35	0.07
	0.30	0.42		0.20	0.28
11	**0.48**	0.30	27	0.23	**–0.55**
	0.18	0.24		0.19	0.27
12	0.18	0.05	28	0.12	–0.18
	0.18	0.24		0.27	0.40
13	0.14	0.38	29	0.29	0.27
	0.19	0.26		0.19	0.25
14	–0.25	–0.19	30	–0.37	–0.15
	0.20	0.29		0.16	0.22
15	–0.12	0.07	31	0.30	–0.09
	0.17	0.23		0.18	0.25
16	–0.33	0.08	32	0.27	–0.46
	0.18	0.25		0.20	0.31

Note. Group 1 = (17 & younger), Group 2 = (18–22), Group 3 = (23 & older); first row for each item is the threshold difference, second row for each item is the standard error of the difference. Significant threshold differences are shown in bold.

equally incorrect? It is evident that all groups were having difficulty with this item and may have been trying to find an answer from their world experience (from Table 4, % correct is 0.56). In addition, in reviewing Table 4, we notice that Item 3 is the most difficult item on the test (% correct is 0.37). The correct

answer for Item 3 is "Wildlife Information Board." Because this is a listening test and the test takers had no control over the sequence of items, it is quite possible that the test takers were still trying to respond to Item 3 when they needed to respond to Item 4, and as a result they might have not listened to Item 4 in the most efficient way and hence might have missed part of the prompt clue. If that were the case, those who might have been affected by this factor would have certainly guessed the answer and their response should have little resemblance to the context of the response. In the light of this discussion, it is possible to conclude that the significant difference in the threshold estimates for Item 4 was confounded with guessing and test takers' world experience and it may not have related to the exhibition of DIF in this item. It was hoped that content expert analysis could provide support for this hypothesis.

Other items. It is difficult to interpret multiple group DIF analysis when the presence of DIF is only present in one of the comparisons. Items 11 and 20 showed DIF in Group 1 and 2 comparisons, whereas Items 18, 21, and 27 only exhibited DIF in Group 2 and 3 comparisons. Do we consider the items that showed partial DIF, that is, in only one comparison, as suspect? The DIF literature does not offer any consensus as how to deal with multiple group comparisons. Reise (personal communication, November 25, 2003) suggested combining the two groups that showed DIF in their comparison and evaluating them against the third group. This cannot be meaningfully applied to our data because we would end up having two different threshold differences for the same items. If we applied such a methodology to our data, we would lose the concept of the reference group. It would then become difficult to define what we meant by the "younger" (17 and younger) and "older" (23 and older) groups.

To summarize, the DIF analysis only pointed to one suspect item (Item 4) as showing DIF in comparisons with both the older group and the younger group. It was argued that the presence of DIF in this item could have been attributed to the impact of the most immediate preceding item (Item 3), which happened to be the most difficult test item. Also, we could not determine a meaningful way of investigating the source of DIF for items where DIF appeared only in one group comparison.

Content Analysis

Table 7 shows the average content ratings of the experts for all the items. The overall average ratings (rounded up to one decimal point) for the whole test show that there was consensus about the general suitability of the items for the reference age group 18 to 22 with an average rating of 3.0 (i.e., neither advantage nor disadvantage). For the younger age group, the items on average were rated 3.2, meaning that the items on average slightly disadvantaged the group. For the older

TABLE 7
Average Expert Content Ratings by Age Group

Items	*Group 17 and Younger*	*Group 18 to 22*	*Group 23 and Older*
1	4	3	2
2	3	3	3
3	3	3	3
4[b]	4	3	3
5	4	3	2
6	3	3	3
7	3	3	3
8	3	3	3
9	3	3	3
10	3	3	3
11[a]	3	3	3
12	3	3	3
13	3	3	3
14	3	3	3
15	3	3	3
16	3	3	3
17	4	3	3
18[a]	4	3	3
19	4	3	3
20[a]	3	3	3
21[a]	4	3	3
22	3	3	3
23	3	3	3
24	3	3	3
25	3	3	3
26	3	3	3
27[a]	3	3	3
28	3	3	3
29	3	3	3
30	3	3	3
31	3	3	3
32	3	3	3
Average	3.2	3.0	2.9

Note. 1 = strongly advantage; 2 = advantage; 3 = neither advantage nor disadvantage; 4 = disadvantage; 5 = strongly disadvantage. Values are rounded up to whole numbers.

[a]Item displayed DIF for at least one set of groups (Groups 1 and 2 or Groups 2 and 3). [b]Items displayed DIF for both sets of groups (Groups 1 and 2 and Groups 2 and 3).

age group, the items on average were rated 2.9, meaning that the items on average slightly advantaged the older group. But the differences were negligible, and generally speaking, the experts' ratings show that the items were suitable for both age groups.

At the item level, the items with the most variation in the content expert ratings were Items 1 and 5 (see Appendix B for test items and transcript for details). The ratings indicate that these items disadvantaged the younger age group or advantaged the older age group. However, these items were not flagged as DIF items in the DIF analysis.

As for the younger age group, the experts indicated in the comments that the context of Items 1, 4, 5, 17, 18, 19, and 21 could have disadvantaged this age group. This was not supported by DIF analysis except for Item 4. Of these items only Items 11 and 20 exhibited DIF for this age group but in the opposite direction, that is, the 17 & younger age group performed better than the reference group on these items. The experts did not indicate on the comments that the content of Items 18, 21, and 27 advantaged this age group over the reference group, as indicated by DIF analysis. These items are discussed next, except Item 4, which was discussed earlier.

Item 1. This item is from Part 1 (Environmental Adviser), and the subskill or primary dimension target is the ability to recall explicitly mentioned detail. The item was rated by the content experts as disadvantaging Group 1 or advantaging Group 3. The answer to the question is "rural studies," and this is information is given along with many other subject areas that the speaker considered studying. Item difficulty (0.60 % correct) and item discrimination (0.35) statistics from Table 4 indicate the item was not very easy. The topic of "taking a diploma" is probably something Group 1 test takers have less understanding of when compared to Group 3 members, who are older and perhaps already enrolled in such courses. Therefore, the subject matter of the question could have been the reason for the content experts' ratings.

Item 5. This item is also from Part 1 (Environmental Adviser), and the subskill or primary dimension target is the ability to recall explicitly mentioned detail. The item was rated by the content experts as disadvantaging Group 1 or advantaging Group 3. The answer to the question is "soil beneath," but the word *mirror* in the spoken text has to be matched with *reflected* in the question for the test taker to arrive at the correct answer. Item difficulty (0.47 % correct) and item discrimination (0.39) statistics from Table 4 indicate the item to be quite difficult. The content experts were perhaps indicating through their ratings that the item was difficult for Group 1 due to the level of the vocabulary.

Item 11. This item is from Part 2 (Tractors), and the subskill or primary dimension targeted is the ability to recall explicitly mentioned detail. This item exhibits DIF in only one direction between Groups 1 and 2. The average expert content ratings for each group on this item was 3 (see Table 7), which indicates the experts did not believe this item would advantage or disadvantage any age groups: "Jason's vintage tractor was found in a . . . behind his house," the correct

answer being "shed." Item difficulty (0.44% correct) and item discrimination (0.37) statistics from Table 4 indicate the item to be quite difficult, but there was no reason to believe the item was biased toward any age group.

Item 18. This item is from Part 3 (Tom Davies), and the subskill or primary dimension targeted is the ability to correctly identify the inference (from multiple-choice options) from implicitly mentioned ideas. This item exhibits DIF in only one direction between Groups 2 and 3. The expert content ratings on this item indicate no advantage/disadvantage for the aforementioned two groups. The experts, however, indicated a slight disadvantage for the first age group, which was not supported by DIF analysis. Item difficulty (0.50 % correct) and item discrimination (0.16) statistics from Table 4 indicate the item to be quite difficult, and the low item discrimination value could indicate that this could confound the DIF estimates on this particular item.

Item 20. This item is from Part 3 (Tom Davies), and the subskill or primary dimension targeted is the ability to correctly identify the inference (from multiple-choice options) from explicitly mentioned ideas. This item exhibits DIF in only one direction, between Groups 1 and 2. The average expert content ratings for each group on this item was 3 (from Table 7), which indicates the experts did not believe this item advantaged or disadvantaged any age group. Item difficulty (0.75 % correct) and item discrimination (0.40) statistics from Table 4 indicate the item to be relatively easy and discriminating quite well. Therefore there was no reason to believe that the item was biased toward any age group.

Item 21. This item is also from Part 3 (Tom Davies), and the subskill or primary dimension targeted is the ability to correctly identify the inference (from multiple-choice options) from implicitly mentioned ideas. This item exhibits DIF in only one direction, between Groups 2 and 3. The expert content ratings on this item indicate no advantage or disadvantage for these two groups. On the other hand, the experts indicated a slight disadvantage for age Group 1, which was not supported by the DIF analysis. Item difficulty (0.60 % correct) and item discrimination (0.34) statistics from Table 4 indicate the item to be quite easy and discriminating quite well. Therefore there was no reason to believe that the item was biased toward any age group.

Item 27. This item is from Part 4 Task 1 (Communication), and the subskill or primary dimension targeted is the ability to correctly identify the main idea of five short monologues in a statement -matching response format. This item exhibits DIF in only one direction, between Groups 2 and 3. The average expert content ratings for each group on this item was 3 (Table 7), which indicates that the experts did not believe this item would advantage or disadvantage any age

groups. Item difficulty (0.61 % correct) and item discrimination (0.42) statistics from Table 4 indicate the item to be relatively easy and discriminating quite well. Therefore there was no reason to believe that the item was biased toward any age group.

Summary and Discussion of Analyses

The DIF analysis identified six items as exhibiting DIF: Items 4, 11, 18, 20, 21, and 27. Of these items, only Item 4 exhibited DIF for both group comparisons. The content analysis identified seven items that could have disadvantaged one group over another group: Items 1, 4, 5, 18, 20, 21, and 27. Of these, the most variant ratings were on Items 1 and 5, which were not identified as items exhibiting DIF in the DIF analysis. Therefore, the common items that exhibited DIF and were rated by content experts as advantaging one group were 4, 11, 18, 20, 21, and 27. As the content expert ratings of these items only showed a 1-point rating difference, it is possible that the items are not clearly biased toward the age groups under investigation. We also hypothesized earlier that the source of DIF in Item 4 could be related to the impact of test takers' difficulty in responding to Item 3. We did not, however, have an explanation why the rest of the items exhibited DIF.

Second, the reason for test taker responses variability across age may also arise from differences in test taker processes: for example, in Items 1, 4, 5, 11, and 21, the ability to recall information or the ability to use memory strategies may be critical and different age groups might use these processes differently. Further, in Item 27, test takers are expected to listen to five monologues, and once again test taker responses variability across age may be due to the differences in the ability to recall information or the ability to use memory strategies across age groups. As this study did not collect data on test-taking strategies across age groups, we could not examine this possible explanation for items that displayed DIF.

Yet another possible explanation is the multidimensional nature of CAE listening items. We have already discussed the views of researchers (Ackerman, Roussos and Stout, and Shealy and Stout) who claim that one general cause of DIF is the presence of multidimensionality in items displaying DIF. Geranpayeh's (2005a, 2005b) recent studies on the CAE examination have shown that the CAE listening items have moderate to high correlations with items that test reading, writing, and speaking skills in addition to having high correlation with items that test grammatical ability. That is to say, the CAE listening items measure multiple dimensions to some extent. The large DIF values observed on some of the items are probably due to measuring those additional dimensions differently across the reference and the focal groups. In the communicative approach to testing listening skills, on which the CAE is based, measuring secondary dimension is not only possible but also desired, and an intended part of CAE's focus. It is this aspect of the test that has probably caused the items to function differentially across the groups. Unfortunately we could not pursue

this further, because we could not obtain formal item dimensionality analysis from the content experts, as they were not trained to do this.

Implications for Test Development

The implications of this study for test development are not obvious for various reasons: First, this is one of the first studies focusing on DIF in terms of age, and second, the findings of the study did not clearly show item bias toward any of the age groups examined. However, it is possible that the younger age group may have found the test topics somewhat less attractive as evidenced through their general lower performance. More systematic analysis of DIF by topic and age need to be conducted before a definitive recommendation is offered. But very generally speaking, it is clear that attention should be paid to choice of test topics so that as far as possible no age group is clearly disadvantaged when the test taking population is as varied as for the CAE.

LIMITATIONS AND FURTHER RESEARCH

One of the limitations of the study was that we did not have explicit statements of primary and secondary dimensions regarding the test items from test developers. If this were available, it could have helped us hypothesize DIF in items that had multidimensionality and then we could have tested the hypotheses fully implementing Roussos and Stout's (2004) multidimensionality-based DIF analysis procedure in addition to using Stout's SIBTEST DIF detection method.

Another limitation was that the study focused on item-level functioning but did not consider the four parts of the test as testlets or bundles to investigate differential bundle functioning (see Lee, 2000). This avenue might be useful, as Part 1 of the test had three items (1, 4, and 5) and Part 3 of the test had three items (18, 20, and 21) that were identified by either the DIF or the content analyses as items that needed to be examined further. It is possible that the topics of the parts of the test are the main cause of DIF and could be potential contributors to test bias, particularly toward Group 1 test takers (17 and younger).

A third limitation was that the study did not have data regarding the differences in test-taking strategy use across age groups, as this might be a possible explanation for the DIF.

CONCLUSION

In this study, we examined the CAE examination for DIF attributable to age. A two-step approach was used: First, the test items were examined for DIF, and

second the items that were flagged (as well as the ones that were not flagged) were subjected to content analysis by expert judges. Six items exhibited DIF in the statistical analysis; five of these items were identified as possible items that advantaged one age or the other, but no clear pattern emerged in terms of why particular items were identified as having DIF. Further, the expert judges could not clearly identify the sources of DIF for the items. As a result, it can be concluded that the CAE listening test (Version 1 of the December 2002 administration) that was the object of this investigation is probably not biased against the test taker age groups included in this study. But further analyses along the lines mentioned in the previous section are necessary to rule out bias with certainty.

To conclude, as Roussos and Stout (2004) stated,

> The general purpose of conducting a DIF analysis is to help ensure test equity or fairness. The statistical flagging of items that exhibit evidence of DIF represents an essential contribution toward the achievement of this objective. Because tests are inherently multidimensional and multidimensionality is the basic cause of DIF, increased understanding of test dimensionality and the effects of these dimensions on DIF hold the potential for a more accurate interpretation of the test score, more control over the influence of relevant auxiliary dimensions, and the reduction of influence by unintended and irrelevant nuisance dimensions. (p. 114)

We hope this article has made an initial contribution to the issue of test fairness to all test takers, irrespective of their ages using the DIF detection approach.

ACKNOWLEDGMENTS

We thank content experts Beth Weighill, Dittany Rose, Andrew Balch, Glyn Hughes, Diane Reeve, and Jason Street; we would not have been able to complete the content analysis of the study without their expertise. We also thank Annie Yu and the participants in the language testing seminar at Tunghai University, Taiwan; two anonymous *Language Assessment Quarterly* reviewers; and Jim Purpura for their thoughtful constructive criticism on earlier drafts.

REFERENCES

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67–91.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Alderman, D., & Holland, P. (1981). *Item performance across native language groups on the TOEFL* (TOEFL Research Rep. No. 9). Princeton, NJ: Educational Testing Service.

Alderson, J. C., & Urquhart, A. (1985a). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*, 192–204.

Alderson, J. C., & Urquhart, A. (1985b). This test is unfair: I'm not an economist. In C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 15–24). Ottawa, Canada: University of Ottawa Press.

Banks, C. (1999). *An investigation into age bias in PET* (Cambridge ESOL Internal Research & Validation Rep. No. 22). Cambridge, England: University of Cambridge.

Brown, A. (1993). The role of test taker feedback in the test development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*, 277–304.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*, 217–238.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*, 155–163.

Geranpayeh, A. (2001). *Country bias in FCE listening comprehension* (Cambridge ESOL Internal Research & Validation Rep. No. 271). Cambridge, England: University of Cambridge.

Geranpayeh, A. (2005a). *Building the construct model for the CAE examination* (Cambridge ESOL Internal Research & Validation Rep. No. 698). Cambridge, England: University of Cambridge.

Geranpayeh, A. (2005b, November). *Language proficiency revisited*: Demystifying the CAE construct. Paper presented at the 12th Language Testing Forum, Cambridge, England.

Ginther, A., & Stevens, J. (1998). Language background and ethnicity, and the internal construct of the Advanced Placement Spanish Language Examination. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 169–194). Cambridge, England: Cambridge University Press.

Hale, G. (1988). Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing, 5*, 49–61.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenzel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). (Eds.). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*, 89–114.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 741–746.

Kunnan, A. J. (1992). The author responds to comments on Kunnan (1990). *TESOL Quarterly, 26*, 598–602.

Kunnan, A. J. (1995). *Test taker characteristics and test performance*. Cambridge, England: Cambridge University Press.

Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta, V. Kohonen, L. Kurki-Suomo, & S. Luona (Eds.), *Current developments and alternatives in language assessment* (pp. 85–105). Jyväylä, Finland: University of Jyväylä.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–13). Cambridge, England: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages Conference Papers*, Barcelona, Spain (pp. 27–48). Cambridge, England: Cambridge University Press.

Lee, Y.-W. (2000). Identifying suspect item bundles for the detection of DBF in an EFL reading comprehension test: A preliminary study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 105–127). Cambridge, England: Cambridge University Press.

Lowenberg, P. (2000). Non-native varieties and issues of fairness in testing English as a world language. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 43–59). Cambridge, England: Cambridge University Press.

Norton, B., & Stein, P. (1998). Why the "Monkeys Passage" bombed: Tests, genres, and teaching. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 231–249). Cambridge, England: Cambridge University Press.

Oltman, P., Stricker, L., & Barrows, T. (1988). *Native language, English proficiency and the structure of the TOEFL for several language groups* (TOEFL Research Rep. No. 27). Princeton, NJ: Educational Testing Service.

Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21*, 53–73.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Measurement in Education*, 20, 355–371.

Roussos, L., & Stout, W. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook for social sciences* (pp. 107–115). Newbury Park, CA: Sage.

Ryan, K., & Bachman, L. (1992). DIF on two tests of EFL proficiency. *Language Testing, 9*, 12–29.

Sasaki, M. (1991). A comparison of two methods for detecting DIF in an ESL placement test. *Language Testing, 8*, 95–111.

Scientific Software International. (2003). BILOG-MG manual. *IRT from SSI*. Lincolnwood, IL: SSI Inc.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159–194.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*, 147–170.

Swinton, S., & Powers, D. (1980). *Factor analysis of the TOEFL for several language groups* (TOEFL Research Rep. No. 6). Princeton, NJ: Educational Testing Service.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analyses of an L2 vocabulary test. *Language Testing, 17*, 323–340.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second-generation immigrants in Dutch tests. *Language Testing, 22*, 211–234.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing, 3*, 80–95.

Zeidner, M. (1987). A comparison of ethnic, sex, and age biases in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing, 4*, 55–71.

APPENDIX A

TABLE A1
General Format of the Listening Paper

Part and Topic	*Task Type and Focus*	*No. of Questions*	*Text Type*
1 Environmental Adviser	Sentence completion, note taking Understanding specific information	8	A monologue of approximately 2 min, heard twice, from the following range of text types: announcements, radio broadcasts, telephone messages, speeches, talks and lectures.
2 Tractors	Sentence completion, note taking Understanding specific information	8	A monologue of approximately 2 min, heard once only, from the range of text types above.
3 Tom Davies	Sentence completion, multiple choice Understanding specific information, gist and attitude	6	A conversation between 2 or 3 speakers, of approximately 4 min, heard twice, from the following ext types; interviews, discussions.
4 Communication	Multiple matching, multiple choice Identifying speakers and topics, interpreting context, recognizing function and attitude	10	A serious of five themed monologues of approximately 30 sec each; the whole sequence is heard twice. In the multiple-matching format there are two tasks; the questions require selection of the correct option from a list of eight. In the multiple-choice format there are ten questions with two questions for each speaker. The questions require selection of the correct option from a choice of three.

APPENDIX B

Centre Number | Candidate Number

Candidate Name _____

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE

Examinations in English as a Foreign Language

CERTIFICATE IN ADVANCED ENGLISH **Test 1**

PAPER 4 Listening

DECEMBER 2002 Approx. 45 minutes

Additional materials:
Answer sheet

TIME Approx. 45 minutes

INSTRUCTIONS TO CANDIDATES

Do not open this booklet until you are told to do so.

Write your name, Centre number and candidate number in the spaces at the top of this page and on the answer sheet unless this has already been done for you.

Answer **all** questions.

You should write your answers in the spaces provided on the question paper. You will have ten minutes at the end to **transfer them to the separate answer sheet**.

At the end of the examination, you should hand in both the question paper and the answer sheet.

INFORMATION FOR CANDIDATES

This paper requires you to listen to a selection of recorded material and answer the accompanying questions.

There are four parts to the test. You will hear Part 2 **once** only. All the other parts of the test will be heard twice.

There will be a pause before each part to allow you to look through the questions, and other pauses to let you think about your answers.

This question paper consists of 5 printed pages.

SP (SM) S23408/2

Part 1

You will hear a woman called Margaret Shelley talking about her job as an environmental adviser. For questions **1-8**, complete the sentences.

You will hear the recording twice.

ENVIRONMENTAL ADVISER

Margaret Shelley took a diploma course called _____ **1**

She didn't feel that the _____ **2** section of her course was relevant.

She says that a talk given by the head of the

_____ **3** influenced her greatly.

She feels proud of the _____ **4** she established.

In her first job, she found it difficult to analyse how trees and grasses reflected the

_____ **5**

She was angry that the nature reserve where she now works showed evidence of

_____ **6**

She feels glad that the number of orchids seems to have more than

_____ **7** in the last five years.

She's looking forward to the return of the _____ **8** to the reserve.

Part 2

You will hear a man talking about an event where you can see old-fashioned farm vehicles. For questions **9-16**, complete the sentences.

Listen very carefully as you will hear the recording ONCE only.

VINTAGE TRACTOR EVENT

The event will be held in Fordham on _____ **9**

Jason's vintage tractor originally belonged to his _____ **10**

Jason's vintage tractor was found in a _____ **11** behind his house.

Jason says that many owners of vintage tractors are not _____ **12** by profession.

There are special _____ **13** on sale for people interested in old tractors.

An old tractor costs at least £_____ **14** to buy.

The event will start in Fordham, at the _____ **15**

The best place to see the tractors is the _____ **16** car park in Fordham in lunchtime.

Part 3

You will hear a radio interview with the writer, Tom Davies. For questions **17-22**, choose the correct answer **A, B, C** or **D**.

You will hear the recording twice.

17 How does Tom feel now about being a writer?

A It is no longer as exciting as it was.
B He used to get more pleasure from it.
C He is still surprised when it goes well.
D It is less difficult to do these days.

18 How does Tom feel about the idea for a novel before he begins writing it?

A He lacks confidence in himself.
B He is very secretive about it.
C He likes to get reactions to it.
D He is uncertain how it will develop.

19 Tom's behaviour when beginning a new novel can best be described as

A determined.
B enthusiastic.
C impulsive.
D unpredictable.

20 What does Tom say happens to writers as they get older and better known?

A Their friends are more honest with them.
B Publishers are less likely to criticise them.
C They get less objective about their own work.
D They find it harder to accept criticism.

21 What does Tom admit about his novels?

A They are not completely imaginary.
B They are open to various interpretations.
C They do not reflect his personal views.
D They do not make very good films.

22 What did Tom feel about the first film he was involved in making?

A He enjoyed being part of a team.
B He found it much too stressful.
C He earned too little money from it.
D He was reassured by how easy it was.

Part 4

You will hear five short extracts in which different people are talking about communication.

TASK ONE

For questions **23-27**, match the extracts with the effects technology has had on communication, listed **A-H**.

TASK TWO

For questions **28-32**, match the extracts with the effects of communication on our lives in the future, listed **A-H**.

You will hear the recording twice. While you listen you must complete both tasks.

A Opportunities for showing how we feel have increased.

B We have more conversations about nothing.

C We find the idea of the internet exciting.

D Community life has begun to disappear.

E It's impossible to develop a caring relationship on the phone.

F We can recognise emotions without seeing people.

G Spoken language has become easier to control.

H We are free of people's value judgements.

Speaker 1 **23**

Speaker 2 **24**

Speaker 3 **25**

Speaker 4 **26**

Speaker 5 **27**

A We'll develop a sense of community spirit on the internet.

B We will all start to speak and behave in the same way.

C The art of chatty conversation will die.

D People will get to know their neighbours better.

E It's oral communication that will be remembered.

F Nobody will be disadvantaged in any way.

G We'll anticipate instantaneous replies when contacting people.

H The internet will result in a lack of communication.

Speaker 1 **28**

Speaker 2 **29**

Speaker 3 **30**

Speaker 4 **31**

Speaker 5 **32**

Dec02

PART ONE

Environmental Adviser

Well, I'm what's called an environmental adviser. When I left school, the future was far from clear for me . . . . All I did was hang around for the first few months. I thought I'd do a course, and took brochures from the library about everything from sports education to zoo-keeping, but finally settled on a diploma in Rural Studies, and off I went to college. It was good overall, I guess . . . The course covered a lot – perhaps too much. There was one section on working with farm machinery, obviously very practical, and then tourism was another and I have to say I never quite got the point of that . . . I think they just added it on . . . and aspects of ecology, animal management – which I loved . . . but I still felt I was drifting . . . Then one day, this guy turned up, and gave this brilliant presentation . . . he was the head of the Wildlife Information Board, and he basically said that the point was not all this farming in itself, but how farming could be made to fit in with the natural countryside instead of damaging it . . . everything clicked into place for me . . . I knew exactly what to do with my life. I set up what I called the conservation group to get together and discuss the issues, and go on visits to projects and so on. That was great, and I'm happy to say it's still running – in fact, I went back to give a talk to them the other week . . . that was my first real achievement, I think . . . meant more than what marks I got for assignments. So I was full of enthusiasm when I completed the course, ready to go to work. In my first job I had to do a botanical survey of a large area of woodland, but I didn't do it very well. I needed to look at how different tree types and grasses are a kind of mirror of the soil beneath. And it was a lot easier said than done. My present job is challenging too, but more rewarding. For the last few years, I've been looking after a local nature reserve. When I first took it on, I was also furious at the way you could see the effects of pollution there. But the hard work's starting to pay off. It's pretty healthy now. Five years ago, there were only about eighty orchids left . . . it was touch and go, really, but now they've doubled in number, at the very least, which isn't bad going. And there are five more species of bird visiting regularly, which is brilliant. It's a great sense . . . Giving nature back to nature . . . the blue butterflies that used to be here, I'm hoping we can get them back . . . the vegetation they feed on is back in place, so it should happen. Well, I can happily recommend environmental advising to anyone concerned about the……..(fade).

PART TWO

Tractors

Good morning. My name's Jason Turnbull and I've come to talk to you about the Vintage Tractor Event which is taking place at Fordham. Now, I drive a tractor for a living, but like many people up and down the country, I like nothing better in my free time than restoring old tractors to working condition and then displaying them at events like the one this Saturday in Fordham.

My particular interest in old tractors began when my parents took over the family farm from my grandparents when I was seven years old. At the back of the house, we found one of the first tractors that my grandfather had used there in the 1930s. It had been standing neglected at the back of the shed for many years and we had a lot of repair work to do before we could get it out of the shed and get it going. For me, that was the beginning of a life-long fascination.

Driving a vintage tractor is very different from driving its modern day equivalent. Nowadays tractors even have air-conditioning, but on these old machines, you're out in the open air, exposed to all weathers. Yet many people drive them as a hobby at the weekends, and strangely enough, they aren't all farmers either. We've got doctors, lawyers, people from all sorts of professions amongst the owners bringing their vehicles to the event. They have no connection with farming, but are just people who like playing about with old engines.

Today, if you want to buy one of these old vehicles, there are several magazines on the market which carry advertisements and regular sales are held in different parts of the country. The magazines are also a good place to advertise for spare parts and pick up other bits of information. To buy an old tractor, you have to pay anything from £5,000 upwards and, of course, particularly rare models, maybe where not many were actually made originally, can fetch as much as £50,000 at auction.

So, if you're interested in seeing some of these old tractors in action, we're meeting at the agricultural college in Fordham where there are excellent unloading facilities, because many of these old vehicles are being transported to the event by lorry. Then we are setting off from the college at around 10.30 am, driving slowly around the backroads of the area for a couple of hours, avoiding the traffic, and stopping for lunch at the Village Hall back in Fordham at around one o'clock. So the tractors will be stationary for about an hour in the car park at the Village Hall from about one o'clock onwards, and that's probably your best chance to have a good look…..(fade).

PART 3

Tom Davies

Interviewer: My guest today is Tom Davies. He has written a series of highly-acclaimed novels, as well as a play and two successful filmscripts. He has said 'I love the solitude, the sheer pleasure of writing, the secret excitement.' Tom, writing is a solitary business, but does it go on being exciting?

Tom Davies: Well, writing *is* an exciting process, although there are good days and bad days, obviously. I remember when I started, I used to sweat for so long over one sentence that it really wasn't much of a pleasure. But I got past that stage and yes, I do find that when things go well, when things are working out, it *is* very absorbing.

Interviewer: But surely less secret these days, now that you've won major prizes?

Tom Davies: Possibly. I recently read out a whole chunk of my work-in-progress at a literary festival because it's one way of trying these things out, whereas in the past I'd been too frightened that if I talked about what I was writing, I would somehow lose control of it. But I think generally I don't talk about what I am intending to write, because I'm still not entirely sure myself which way it's going to go. But once something is down in a first or second draft, then you can try it out and see how it sounds.

Interviewer: And you've said that at any one time there are as many as ten or fifteen ideas for novels floating around in your head. How do you choose which one to follow up?

Tom Davies: You've got to find the idea that's got the right kind of urgency and it's not a rational decision. It's patience and luck and turning up at your desk every morning even when nothing seems to be coming. If you're not there, then nothing is precisely what will happen. But once I get started, then a good day would be two or three hundred words.

Interviewer: And then do you hone it, do you go back over it?

Tom Davies: I go back all the time until I get to the stage when I won't look at it again because you need the distance of time to look back and see it from a different perspective.

Interviewer: And is there anyone who you can then give this manuscript to and say, 'Look, before I go any further, tell me what you think of this.'?

Tom Davies: I give the finished draft to certain old friends who're permitted to be as brutal as they like. That's very useful because I think there's a danger for writers as they get older, as their reputations get established, that publishers won't tell them if they've any serious doubts about a piece. So sceptical friends are very important to give you the benefit of a truthful opinion.

Interviewer: And you trust these friends?

Tom Davies: Absolutely. The first time I tried this, years ago, a friend of mine said 'Look, I think this novel's absolutely terrible, put it in a drawer and forget about it.' And I didn't speak to him for eighteen months. But after that I learnt that if you give someone your novel to read, you've got to allow them to say that kind of thing. These days I wouldn't take it so personally.

Interviewer: And although you've denied any suggestion that you write about yourself, there are actually all sorts of bits and pieces of you dotted all over your work, aren't there?

Tom Davies: Someone said that you can't write two hundred words in a novel without giving something of yourself away and I suppose that's true. Perhaps that's why I've always been a bit defensive about my work.

Interviewer: Now, despite those two successful filmscripts, you haven't, strangely, had a lot of luck translating your stories onto the big screen, have you? Why's that?

Tom Davies: Oh well, my first experience was of a low-budget English film. And because we had so little money to work with, it was wonderfully uncomplicated and I thought, 'Oh what a brilliant life. I could write novels and then in between each one, I could do a film.'

Interviewer: Because it's so much easier?

Tom Davies: Well, it was such fun being away on location surrounded by fabulously competent people, all taking fierce pride in their ability to do something so well and very quickly. The panic of the ticking clock, the things going wrong and then somehow being solved at the last minute, all that was marvellous for someone who usually spends his time locked up in an empty room.

Interviewer: So it's actually harder to write a good screenplay?

Tom Davies: No, I wouldn't say that. Indeed, I don't think a screenplay is a literary form in itself. It's more a set of instructions, a bit like a recipe. And you can fool yourself into thinking that you can see what's going to be on the screen, but actually too many people intervene in the finished product, you're just a part of the process, so it's quite unlike a novel where you're in sole charge, as it were.

Interviewer: Tom, there, unfortunately, we have to leave it. Thank you…….(fade)

PART 4

Communication

Speaker One

I like telephones. I think they're useful because they give out so much information about whoever's calling you. You can almost guess when a person's smiling at the other end of the line because a smile changes the voice slightly. I think the reason spoken language works so well – is so successful, is that it's difficult to control. I've heard people say that printed words are forever and that spoken ones don't last – but the memories we have don't back that up. When you look back and try to remember people you once knew, you remember the sound of their voices – not a few words they might've written on a piece of paper.

Speaker Two

The amazing thing about the internet is that it has cut out what you might call stereotyping. No-one can see what kind of person we are, or where we come from, or hear the way we speak, so we're working towards a society in which we're all equal! What the internet has done is provide us all with the first method of communication without any kind of barriers. Sometimes, when you hear someone speak you immediately put them into a kind of box, often educationally and socially. Sometimes even the fact that you are male or female can actually work against you. That won't happen any longer.

Speaker Three

Conversations about nothing in particular are the sort of conversations friends and acquaintances have – but, in my opinion, these conversations are beginning to disappear. To my mind, the whole thing really began with television and other forms of home entertainment, and with a faster pace of life. I read the other day that nearly a third of the population have never even spent an evening with a next-door neighbour! In fact, looking back, I think things really began to pick up speed when the internet came along. You see, the problem lies in how human beings communicate. We think of it as a product of the mind, but in fact, it's actually mostly done by body language.

Speaker Four

You can't blame the internet for lack of communication nowadays! We've been conducting loving, caring and trusting relationships over the telephone for more than a hundred years now. What the internet's done is simply develop new ways of telling people how you feel. Written emails are often better than a

conversation because we can always go back to them and read them – not like a phone call, which we can forget in a few minutes. I mean, take poetry, for instance. Hundreds of people send their loved ones romantic messages over the internet every day and get a response immediately. I'm sure this kind of thing will go on increasing.

Speaker Five

I think the internet's still a plaything to most people. We're still kids who press a button to see what happens and then are thrilled to bits when something does. The real problem with all this, however, is that we're still in the very early stages of communication of this kind, and at the moment all we can really do is enjoy it. What we haven't done is establish a sense of belonging to a group. So what I think will happen in the next decade, or even century, is that we'll learn to care for a community of people we contact by email – in other words, we'll feel responsible for others that we communicate with.

APPENDIX C

DIF Questionnaire

Examine the question paper you have been given and decide whether each item is likely to advantage/disadvantage test takers in the following groups. Use the scale below and write the numbers in the boxes provided. Please use the space below each question to explain your choices.

Scale

1 - Strongly advantage; 2 - Advantage; 3 - Neither advantage nor disadvantage
4 - Disadvantage; 5 - Strongly disadvantage

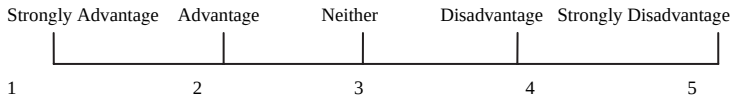
Groups

Group 1 = Test takers aged 17 & under; Group 2 = Test takers aged 18 – 22; Group 3 = Test takers aged 23 & above

Example:

Questions	Group 1 (17 & under)	Group 2 (18 – 22)	Group 3 (23 & above)
1	4	3	2
Comment	The topic was suitable for older test takers		

This shows that the item is likely to advantage 23 and above



Items	Group 1 (17 & under)	Group 2 (18 – 22)	Group 3 (23 & above)
1			
Comment			
2			
Comment			
3			
Comment			
4 . . . 32			
Comment			