1 An Investigation of Speaking Test Reliability with Particular Reference to Examiner Attitude to the Speaking Test Format and Candidate/Examiner Discourse Produced

Brent Merrylees
LTC Language and Testing Consultants Pty Ltd

Abstract

This research project was designed to complement research being carried out at the time by the University of Cambridge Local Examinations Syndicate (UCLES) into candidate and examiner discourse produced in the Speaking Module of the test. The researchers felt that analysis of this kind was fundamental to having informed discussions on any possible changes to the test format and the debate would be further enhanced by consulting IELTS examiners, the practitioners who are actually required to apply the speaking test instrument. At the time there had been no large scale survey of IELTS examiners to establish their attitudes to either the speaking test format or to the band descriptors in their current form.

The research project investigated examiner attitude to the speaking test by carrying out a survey of IELTS examiners working at test centres in Australia, New Zealand, Malaysia, Thailand, Indonesia, Hong Kong, the Philippines and Taiwan. The survey was delivered in a two page questionnaire and was divided into the three broad sections of IELTS interview format, IELTS Band Descriptors and the different interview phases. The final sample size for the survey was 151 respondents. In addition to this survey, a dataset of 20 IELTS interview transcriptions was constructed and an analysis was carried out on examiner discourse and how it can affect the language produced by the candidate both in terms of quantity as well as quality. The dataset was also designed to provide a resource for more detailed analysis of the Speaking test if it were required in the future.

IELTS™
English for International Opportunity

Publishing details

International English Language Testing System (IELTS)

Research Reports 1999

Volume 2

Editor: Robyn Tulloh

IELTS Australia Pty Limited
ACN 008 664 766
Incorporated in the Australian Capital Territory
Web: www.ielts.org

© 1999 IELTS Australia.
This publication is copyright. Apart from any fair dealing for the purposes of private study, research or criticism or review, as permitted under the Copyright Act, no part may be reproduced by any process without written permission. Enquiries should be made to the publisher.

National Library of Australia
Cataloguing-in-Publication Data
1999 ed
IELTS Research Reports 1999 Volume 2
ISBN 0 86403 021 5

1.0 Background to the Research Project

The prime motivation for the design of this research project was to complement existing research being carried out at the time by the University of Cambridge Local Examinations Syndicate (UCLES). The UCLES commissioned research is concentrated on the speaking module through a linguistic analysis of the discourse produced. The researchers felt that analysis of this kind was fundamental to having informed discussions on any possible changes to the test format and the debate would be further enhanced by consulting IELTS examiners, the practitioners who are actually required to apply the speaking test instrument. At the time there had been no large scale survey of IELTS examiners to establish their attitudes to either the speaking test format or to the band descriptors in their current form and it was important that any investigations into these examiner attitudes be informed by input from both IELTS Australia and the British Council examiners. Given the fact that UCLES was currently carrying out research, the researchers thought it apposite to carry out a survey which would dovetail in with the UCLES research and produce a clearer picture of how the speaking module was performing.

Australia has more than 40% of the IELTS worldwide cohort and the LTC team had considerable experience, both through training of examiners and in the delivery of the test, of the issues/problems involved with the current speaking module.

The overall objectives were therefore:

- To establish examiner attitude to format, useability and perceived reliability of the speaking module
- To establish examiner attitudes to the speaking band descriptors focusing on the examiner ability to interpret the Band descriptors consistently and reliably
- To critically review the Band descriptors in order to provide data which will then be used to inform collaborative research with UCLES to investigate the effectiveness and reliability of the speaking module

2.0 Methodology

The initial phase of the project required a survey of a sample of examiners to investigate a number of issues:

- how the examiners feel about the format and the phases of the speaking module
- what changes, if any, to the speaking module format the examiners would like to see
- whether the examiners felt the current descriptors were easy to use
- how often the examiners refer to the descriptors when giving their rating of a candidate's performance in an interview
- whether there are areas, if any, of the descriptors examiners would like to see changed.

All examiners were surveyed on an anonymous basis, their permission having been first obtained for the research project. The survey was, wherever possible, given to examiners by the Test Administrator on the day that they were examining to ensure that the responses were based on fresh experience.

In the original project design, there was to be an analysis of the descriptors using a variety of linguistic tools, including a traditional approach and overall textual analysis focusing on continuity of assessment criteria contained in the band descriptors. As the project had been designed to include collaboration between the LTC team and UCLES, it was considered important that full and frank discussions were held between the two parties *before* the examiner survey and the analysis were carried out. The first of these meetings took place in November, 1996 in Bangkok between Clare McDowell and Nick Saville head of the Test Development and Validation Group at UCLES. During these discussions it became apparent that the initial plan for carrying out an analysis of the band descriptors would not be particularly beneficial and indeed could overlap with the research UCLES were involved with at the time. It was agreed, therefore, at these discussions that it would be more useful if some kind of discourse analyses were carried out on the actual language used both by the interviewer and the candidate in the speaking modules themselves.

Following the initial discussions in Bangkok, the LTC research team then reformulated the research design to reflect the changes as discussed. It was decided that the best way forward was to establish a dataset of 20 IELTS interviews which would be carefully constructed in order to control a number of key variables. The most important of these variables would be 1st language, final score allocated, gender and age of candidates. In addition to these variables it was also considered important to include only candidates who had "flat profiles" in the dataset; i.e. candidates whose scores in the other modules were within reasonable range of the speaking subtest scores. One other variable considered was to ensure that choice of topic in phase 2 and phase 3 was common to all interviews as a form of "anchor task". However, it soon became apparent that not all these requirements could be accommodated in the one dataset and so this preferred consideration was dropped. Another variable which came into play during the selection process was the quality of the interview recordings themselves to ensure they were sufficiently audible to be transcribed correctly.

The type of analysis to be carried out on the dataset was discussed by the LTC research team and confirmed in further communications with UCLES. It was decided that as the speaking subtest discourse analysis had not formed part of the original research proposal, it was important to keep it relatively straightforward to avoid going beyond the original scope of the project but it seemed important that the resultant transcripts could be used as a resource for both present and future research. A standard format for transcription was thus sought. Once the final form of the analysis had been agreed upon and completed, a final meeting took place in Cambridge between Brent Merrylees and Nick Saville to discuss the findings of both the examiner survey and the analysis of the interview discourse.

3.0 The Survey

3.1 Questionnaire design

When the questionnaire was designed, it was deemed relevant that it be precise, contain no ambiguities and that it could be completed in a time frame of under five minutes. The authors were keenly aware of the pitfalls of designing prompts in any questionnaire and paid particular attention to having prompt precision while not overly burdening the respondent with complex and ambiguous language. Once the design was finalised after a brief trial on a group of Sydney based examiners, the questionnaire (see Appendix 1.1) was able to be printed on two sides of one page, back to back for ease of distribution and handling, and ultimately contained 39 short questions. Respondents were asked to rate on a scale of 1- 4 a number of aspects of the IELTS speaking test. Some space was provided for examiners to include, in their own words, their views on certain aspects of the speaking test and to the speaking test as a whole.

The survey was divided into three broad sections

- IELTS Interview format
- IELTS Band Descriptors
- the interview phases

3.2 The Cohort

In January and February of 1997, all Australian IELTS Administration Centres were telephoned and asked to participate in the survey of IELTS examiners to ascertain examiner attitude to the Speaking module. Administrators were asked to seek the cooperation of their examiners, or part thereof, by having them complete an anonymous questionnaire. This instrument was, wherever possible, given to examiners on the day that they were examining, to ensure that the responses were based on fresh recollection of the exercise.

The Australian data was drawn from 113 respondents across the following centres: Sydney, Brisbane, Perth, Armidale, Launceston, Townsville, Wagga Wagga, Darwin, Melbourne, Gold Coast, Canberra, Newcastle and Wollongong. Following the "success" of the initial Australia based survey which had yielded both interesting and insightful results, and in accordance with the original research brief, further data was gathered during May and June at IELTS Australia Asia - Pacific centres to add to the preliminary findings. Centres in Malaysia, Thailand, Indonesia, New Zealand, Hong Kong, the Philippines and Taiwan were invited to participate and 38 responses were received from the first five countries; the findings were then analysed in the same manner as those received from the Australian centres. A bar graph was drawn up to illustrate the responses to each question for both cohorts and these can be viewed on the following pages. A comparison between the two sets of data follows on an item by item basis. The countries surveyed are not identified separately but rather the responses from the four countries involved have been collated together to produce the off-shore results attached. It would be possible to further identify the responses on a country by country basis but as the cohort is approximately a third the size of the Australian cohort it was felt this would not yield representative results if fragmented.

3.3. General Comments on the survey

It should be noted that the willingness with which the respondents were prepared to comment at the end of each section of the questionnaire is encouraging but also points to a perceived need to discuss such issues. Clearly respondents have jumped at the opportunity to express an opinion, albeit under cover of anonymity, on how they feel they manage an IELTS interview. The comments are both illuminating and helpful.

4.0 Analysis of the Data

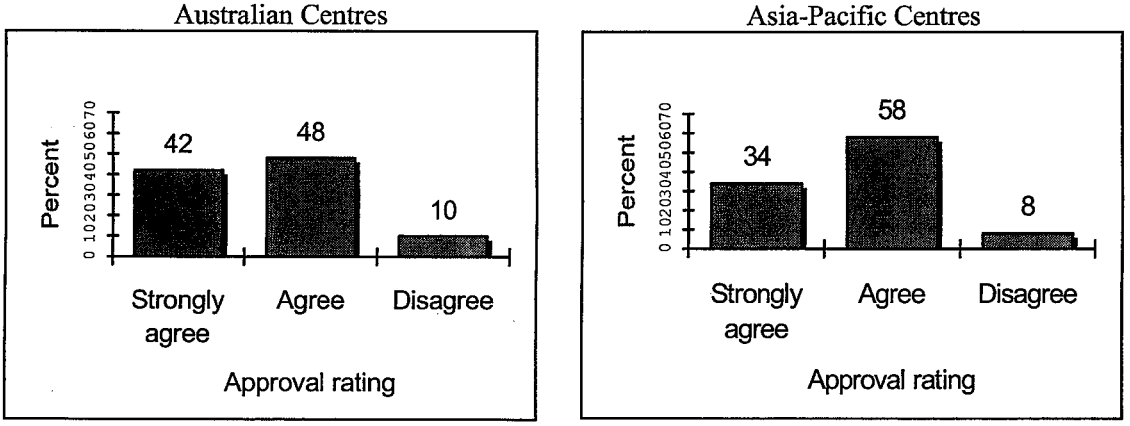
The information captured on the completed questionnaires was entered onto a database and then analysed to produce a statistical overview of the responses. The individual questions are produced below together with the responses so far received, presented in statistical form and accompanied by the researchers' interpretation of the data. In addition to the statistical data and the analysis, a summary of the respondents' individual comments is also attached at the end of each section.

5.0 The Data and Findings

5.1 Section One: The IELTS Interview format

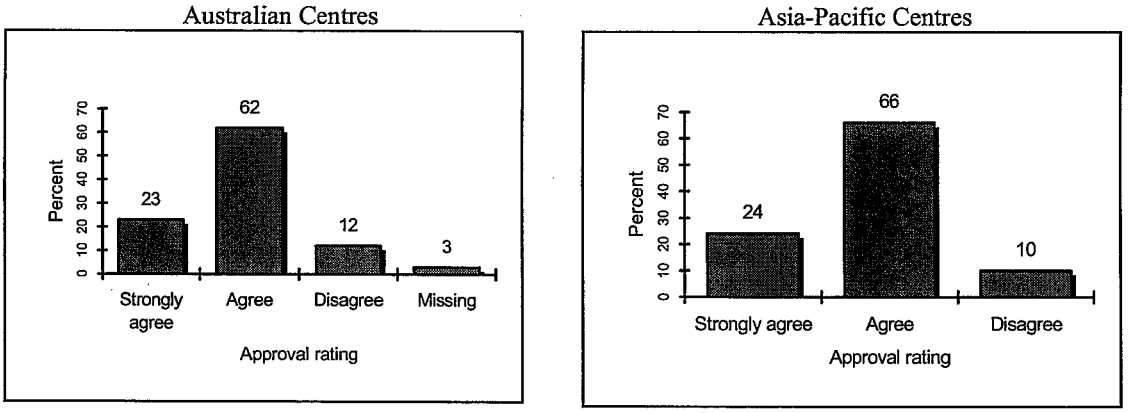
Respondents were asked to rate the following propositions from 1 to 4 with 1 being *"strongly agree"* and 4 being *"strongly disagree"*

Question 1 - The interview format is easy to manage



The vast majority of respondents from both cohorts agreed that this was true though 10% disagreed. We can assume from this that the format is generally acceptable to examiners.

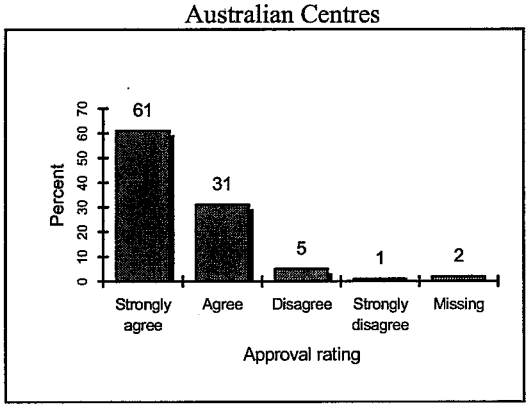
Question 2 - The interview format is effective in generating assessable discourse



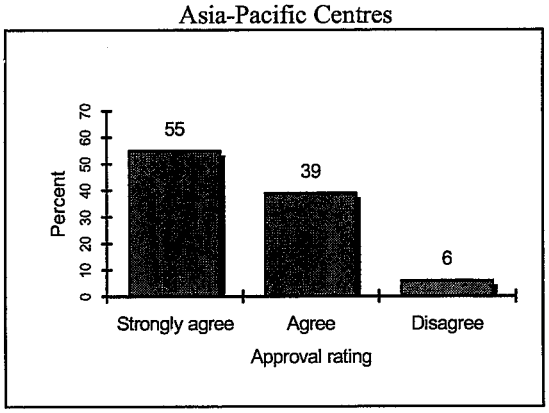
Over three quarters of the respondents felt this to be the case. In other words, the language produced by the candidates in response to the tasks, is adequate for an assessment to be made.

Question 3 - The interview is of a manageable length

Australian Centres



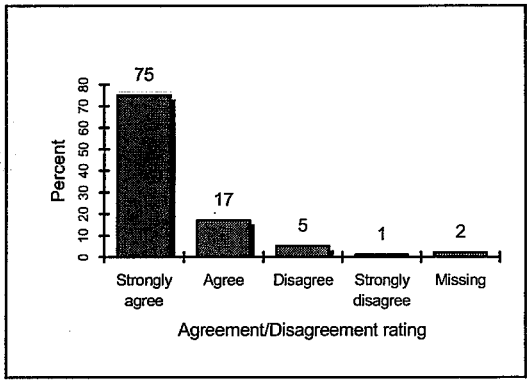
Asia-Pacific Centres



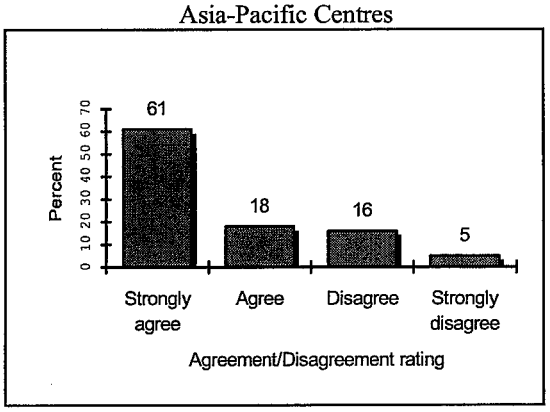
All but a handful of respondents felt the length was manageable.

Question 4 The taping of all interviews is a good idea

Australian Centres



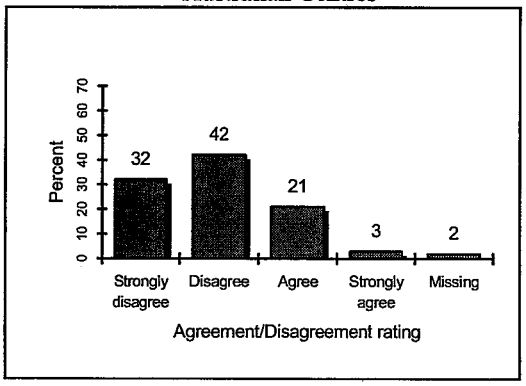
Asia-Pacific Centres



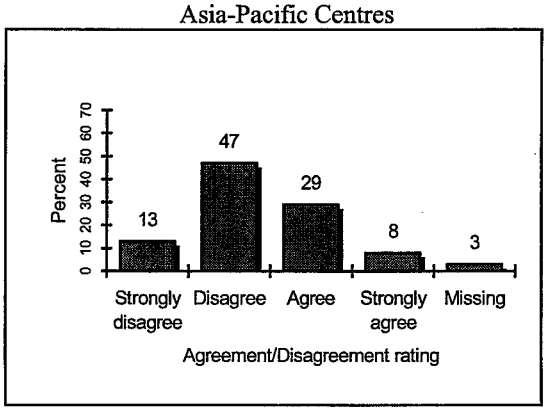
The vast majority of examiners had no problem with the taping of interviews though it can be seen that off shore responses reflect the fact that some examiners are not in favour of the taping. This may stem from a lack of understanding of the rationale for taping interviews.

Question 5 The interview should be less structured than it currently is

Australian Centres



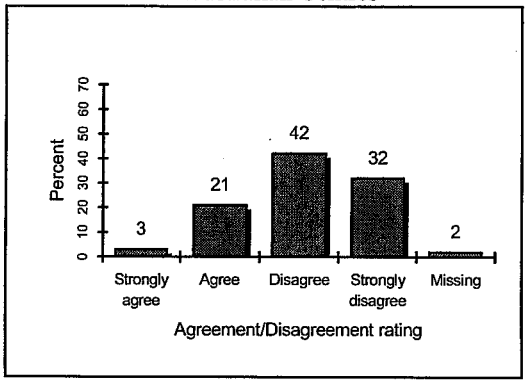
Asia-Pacific Centres



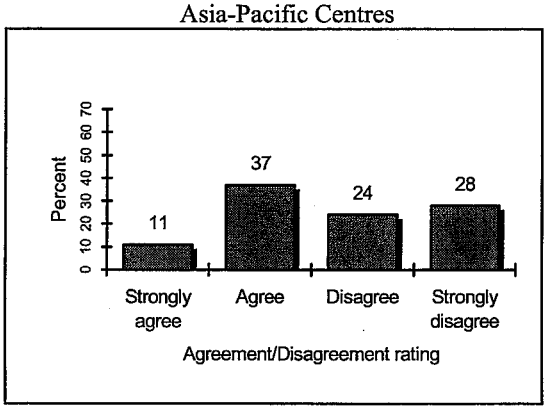
While 74% of the Australian respondents felt happy with the format, almost a quarter of them admitted that they would like it to be less structured than at present. This contrasts with the smaller number of 10% who answered in Q1 that it was not easy to manage. Those off shore respondents who would like a less structured format constituted a larger percentage of their cohort.

Question 6 The interview should include picture and/or photo prompts

Australian Centres



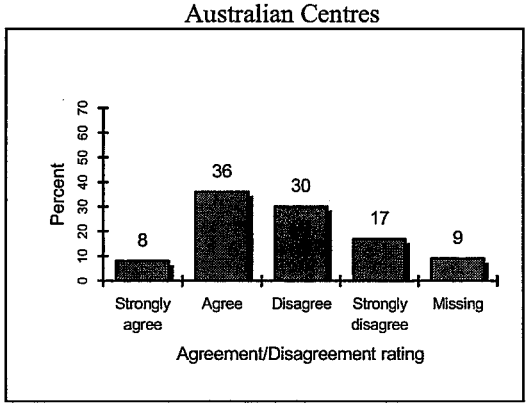
Asia-Pacific Centres



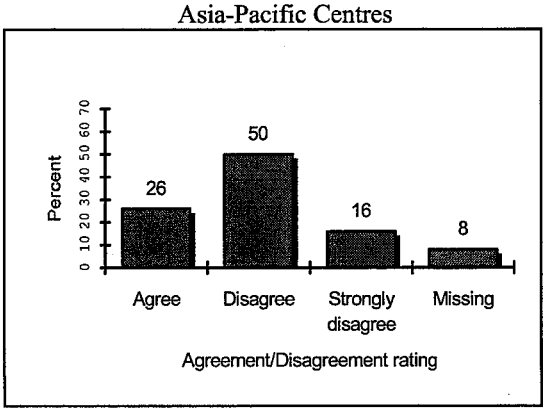
One quarter of the Australian respondents agreed with this proposition but the majority felt that pictures would not enhance the interview. Just under half of those who did not want pictures felt strongly about this. On the other hand, examiner comments included one remark which was strongly in favour of using pictures so this is an area of dispute. The off shore responses are noticeably different and have produced a favourable response to the idea of picture prompts from almost half the group, though the field is clearly divided here.

Question 8 The interview should include a negotiated task

Australian Centres



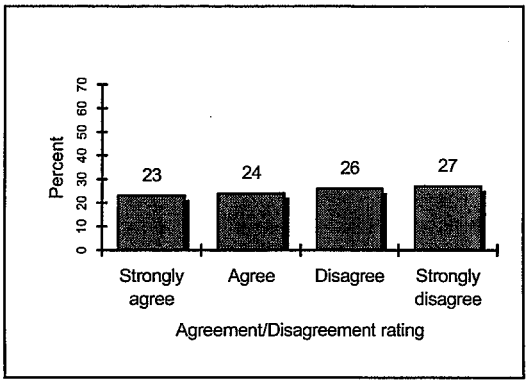
Asia-Pacific Centres



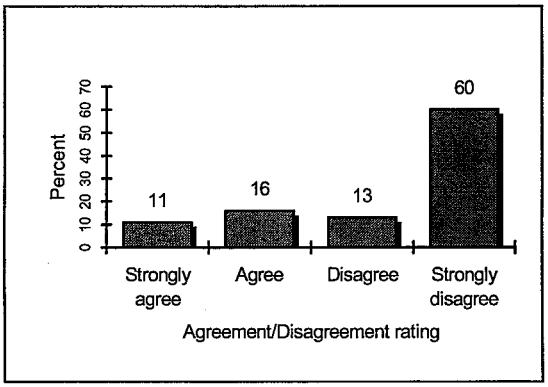
Here the Australian respondents were fairly evenly divided with 44% in favour and the rest against. However, those who were very in favour were almost as great in number as those who were very against the proposition, indicating that this is an area of contention. It seems that the overseas respondents were generally not in favour. It may be that some examiners misunderstood the proposition.

Question 9 There should be two examiners (one interlocutor and one assessor)

Australian Centres



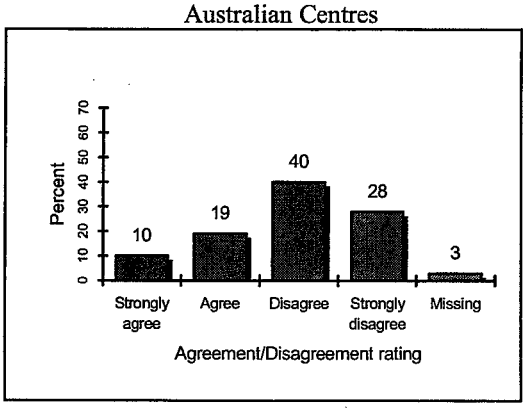
Asia-Pacific Centres



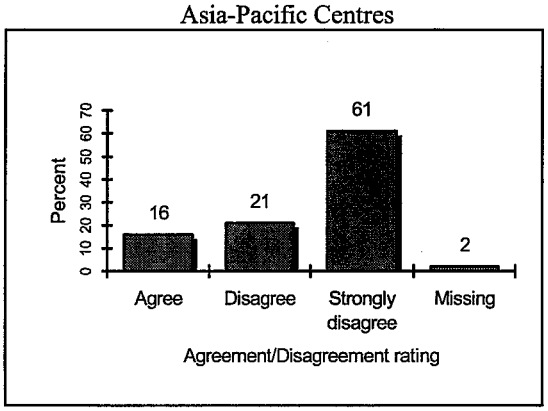
Again here we see a very evenly distributed field in Australia with almost half strongly in favour of the idea of having 2 examiners present while slightly over half were against. Those who strongly disagreed with the proposition were in the majority at 27% of the cohort. One respondent suggested that this approach was essential for new examiners. However, in the off shore centres, the idea was apparently not well received.

Question 10 The interview should be in a paired format with two examiners

Australian Centres



Asia-Pacific Centres



The responses to this question varied considerably from the previous question. Only 29 % were in favour in Australia while almost 70% disagreed with the proposition. Off shore we see a similar pattern with the paired interview with two examiners being firmly rejected by the overseas examiners.

Since all UCLES main suite exams except CPE[1] and BEC[2] tests now prescribe this format for the oral component of the tests, it is interesting to note the response from the IELTS examiners. It may be that those in favour were already familiar with the UCLES model.

5.1.1 Summary of respondents comments : Questions 1 - 10

The general feeling from the examiners is that the format is good and quite manageable. Many comments related to the administrative difficulties that would be involved in changing the format to include more than one examiner, highlighting how organisational concerns often inform decisions. For a full listing of examiner comments please refer to Appendix 1.2 - *Comments on Format*:

5.2 Section Two: The IELTS Band Descriptors

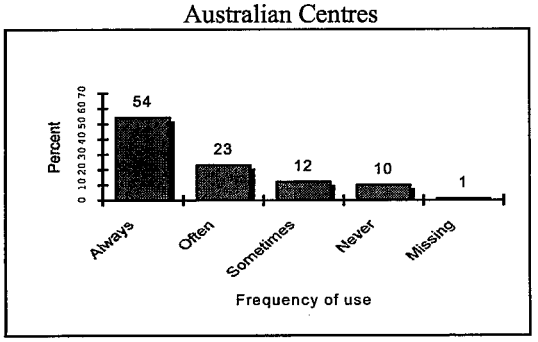
The questions in this section of the questionnaire were designed to probe examiner behaviour and the authors acknowledge this is a difficult area to deal with as often respondents give answers which they think is appropriate or expected. Nevertheless, it was considered important to investigate the issue.

[1] Certificate for Proficiency in English
[2] Business English Certificate

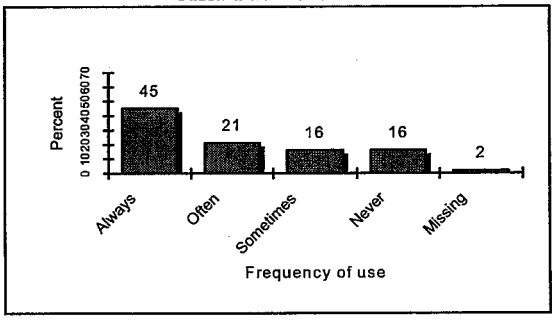
Respondents were asked to rate the following propositions from 1 to 4 with 1 being *"always"* and 4 being *"never"*

Question 11 - I refer to the descriptors before every examination session

Australian Centres



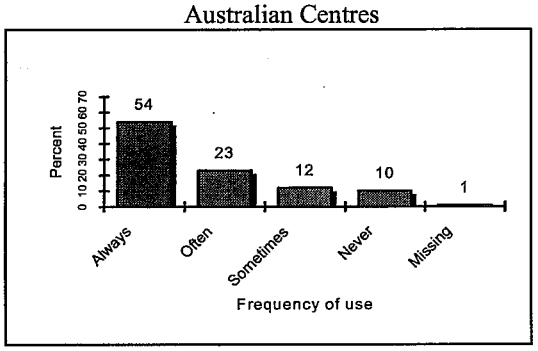
Asia-Pacific Centres



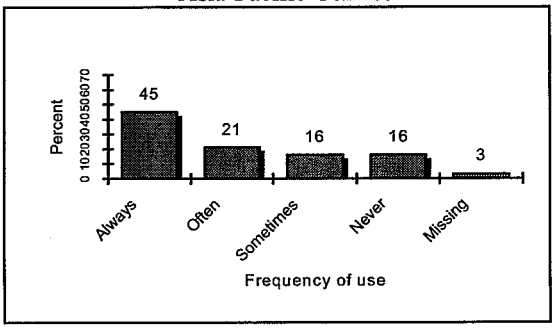
It was interesting to note that over half of Australian respondents were sure that they referred to the descriptors, though the fact that 10% admitted to never doing so is cause for concern. The overseas responses reflect the same pattern though the 16% who never refer is more alarming. Since a further 12% in Australia and 16% off shore replied that they only did this on some occasions, we can assume that approximately 25% of examiners are not referring regularly to the descriptors before an exam session.

Question 12 - I refer to the descriptors before every interview

Australian Centres



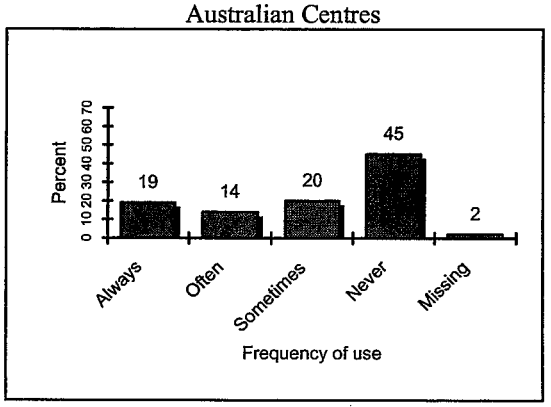
Asia-Pacific Centres



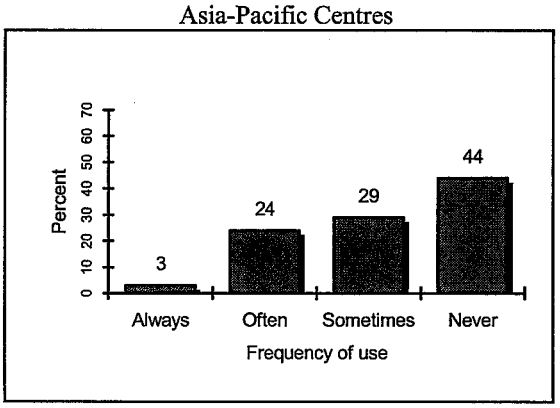
Here we find far fewer examiners admitting to this behaviour, presumably because the majority felt that it was sufficient to refer to them at the start of the session.

Question 13 - I refer to the descriptors during the interview

Australian Centres



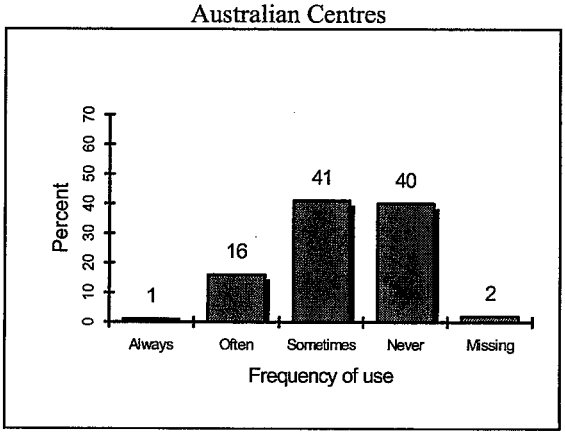
Asia-Pacific Centres



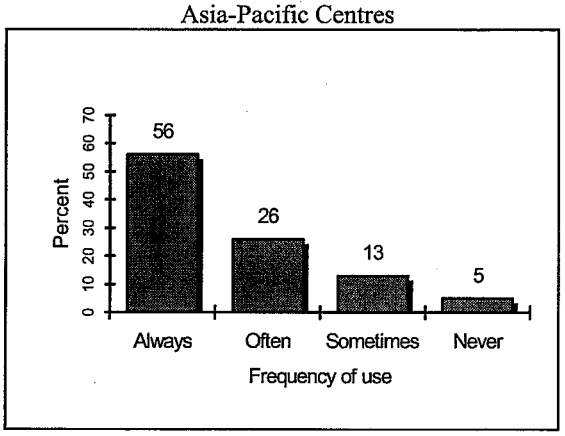
As expected fewer respondents indicated this pattern and, in fact, almost half pointedly registered that they do not do so, possibly because they would consider this to be intrusive examiner behaviour during the interview.

Question 14 - I refer to the descriptors after every interview when rating

Australian Centres



Asia-Pacific Centres

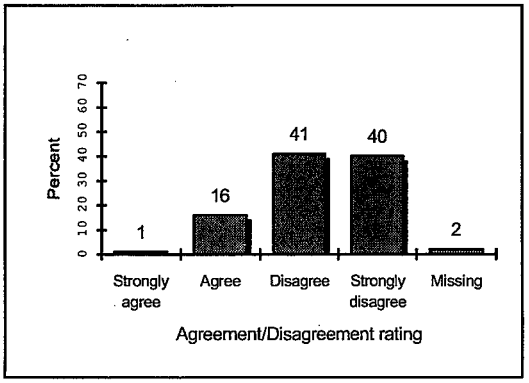


The overwhelming majority of respondents in Australian centres claim that they do not always refer to the descriptors after an interview. Since half have responded that they do not refer to them at the start, either, this response is disturbing. The overseas examiners, on the other hand, appear to be far more likely to refer to them at the end.

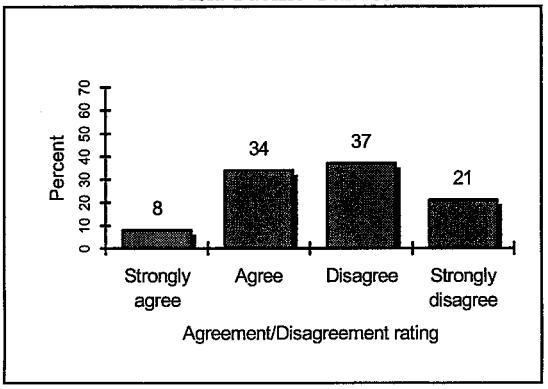
Respondents were asked to rate the following propositions from 1 to 4 with 1 being *"strongly agree"* and 4 being *"strongly disagree"*

Question 15 - I am thoroughly familiar with the descriptors and rarely refer to them

Australian Centres



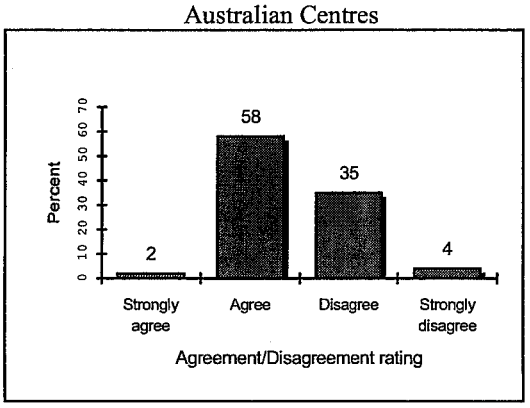
Asia-Pacific Centres



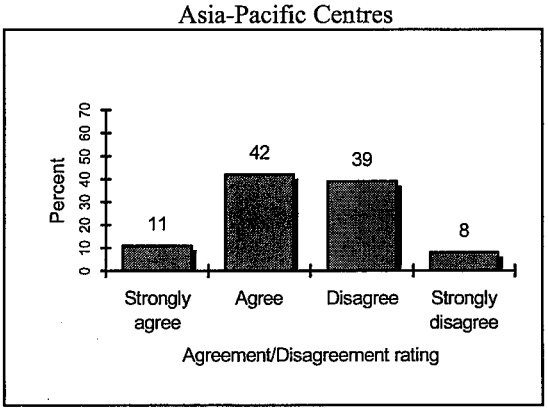
This question could have been thought of as a "trick" question which no one wished to get caught by. Most responded that they often refer to them which is in apparent contrast to the responses to the previous three questions.

Question 16 - I find the descriptors easy to interpret/apply

Australian Centres



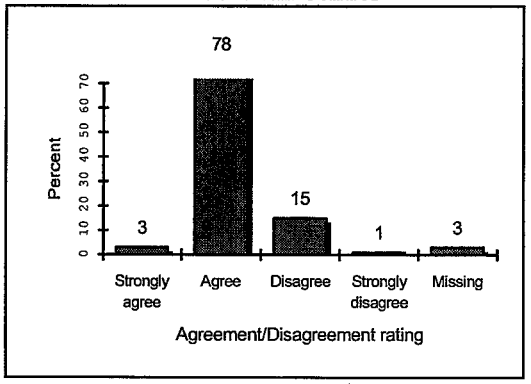
Asia-Pacific Centres



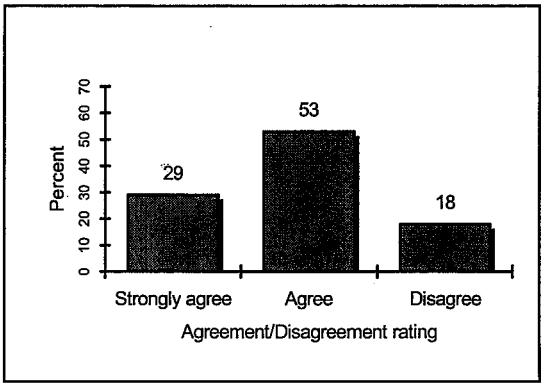
Over half of the respondents indicated that the descriptors were easy to apply but a large number in both groups (40-45%) did not agree and admitted to having difficulty using them. This is a disturbingly high proportion since the application of the rating scale is the key to reliable marking.

Question 17 - I feel confident that my ratings are accurate when applying the scale

Australian Centres



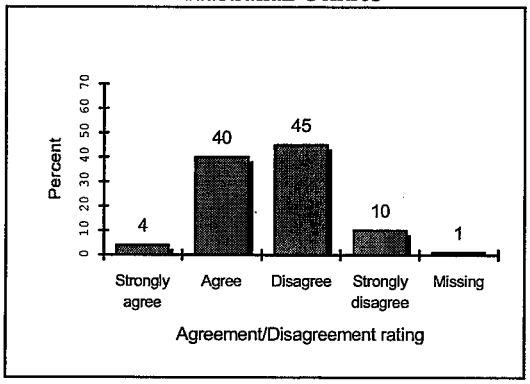
Asia-Pacific Centres



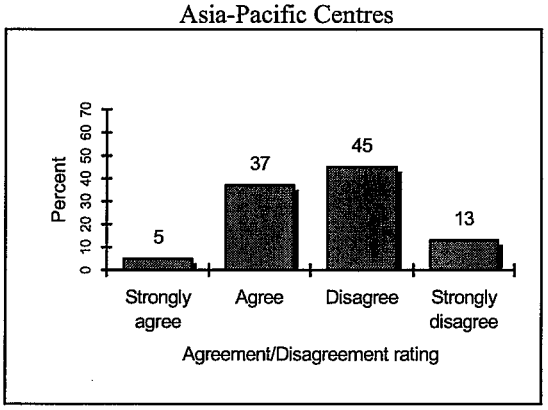
Well over three-quarters of the examiners felt confident about their own marking. This is encouraging and what one would expect. Nevertheless we find just under 20% who are not confident.

Question 18 - The descriptors discriminate clearly between the levels of proficiency

Australian Centres



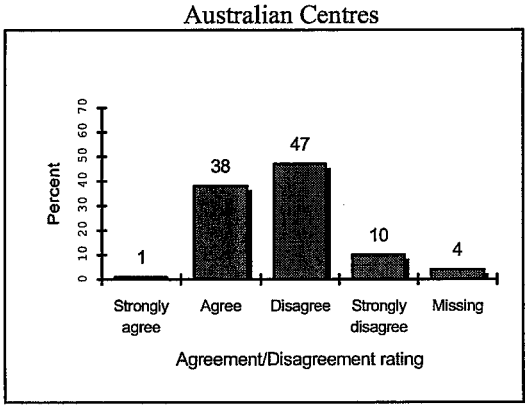
Asia-Pacific Centres



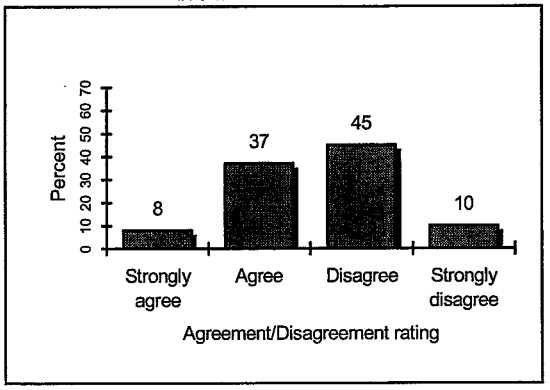
Here the field was clearly divided with just under half agreeing with the proposition and slightly over half disagreeing in both cohorts. The responses here are significant and point to the need to review the descriptors as these findings would indicate that examiners are having difficulty applying them with reference to the bands.

Question 19 - The descriptors are adequate for all phases of the interview

Australian Centres



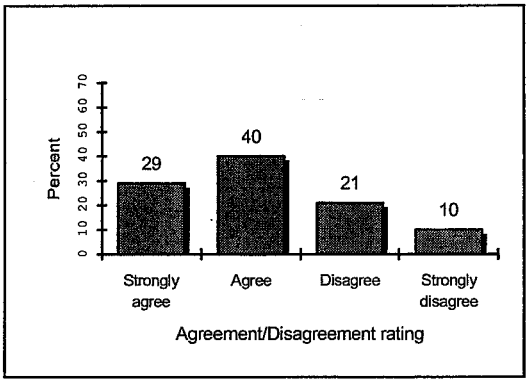
Asia-Pacific Centres



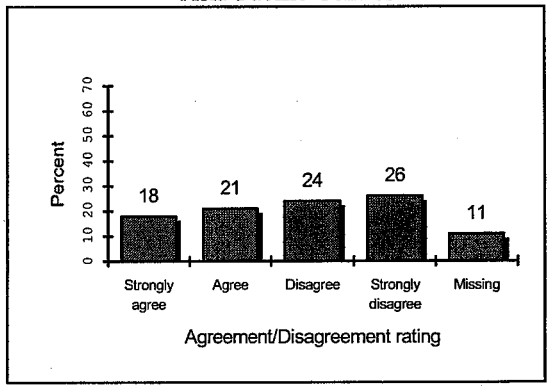
Over half of the respondents felt that the descriptors were inadequate for all phases. Both groups produced very similar split responses.

Question 20 - I would like to use a profile scale, as with writing, where individual aspects of performance are assessed (eg pronunciation, structure, fluency etc)

Australian Centres



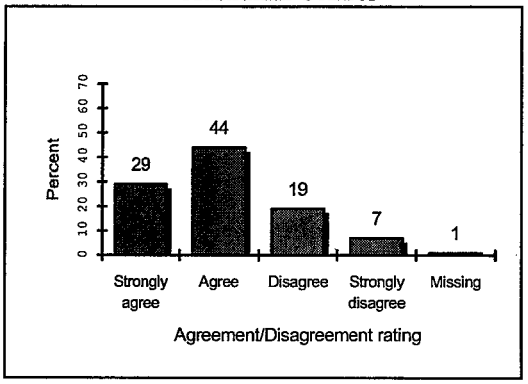
Asia-Pacific Centres



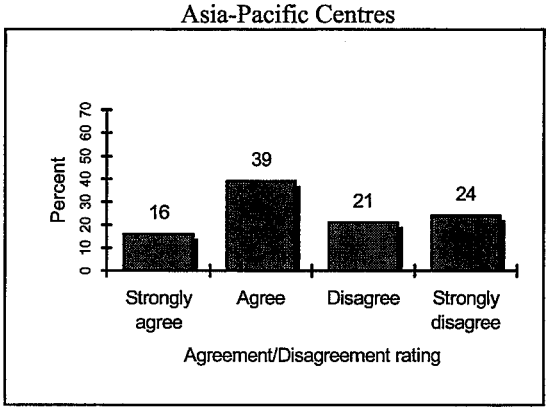
The responses to this proposition ranged across the spectrum with well over half of the Australian examiners (69%) indicating a preference for a profile scale but those not in favour also being split. Some 10% strongly disagreed with the proposition. However, the fact that nearly three-quarters of the respondents would welcome such a scale is significant. The off shore examiners, however, would appear not to be in favour.

Question 21 - I would like to use a combination of global and profile descriptors

Australian Centres



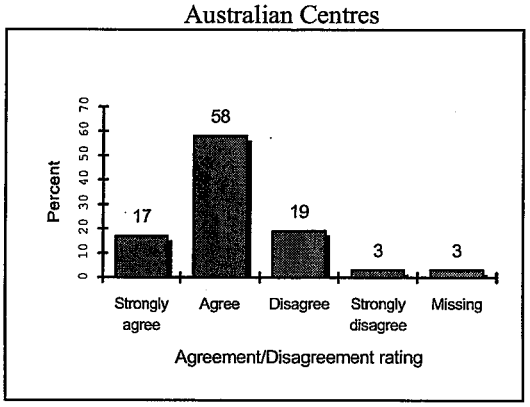
Asia-Pacific Centres



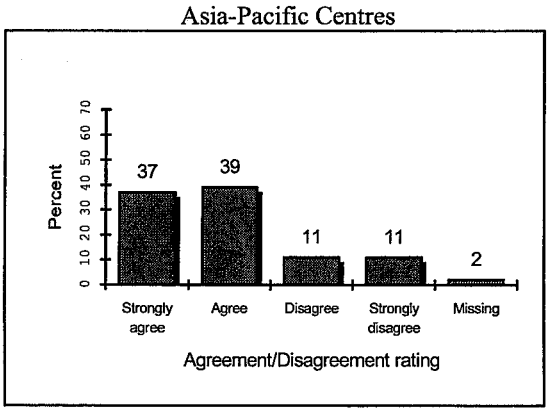
The responses to this were, as expected, very similar to the previous question though slightly more examiners were in favour of this arrangement than simply a profile approach. Significantly, the proportion of Australian examiners strongly opposed to the profile approach was slightly less when the opportunity to combine it with a global score was given. The off shore examiners were divided on this issue.

Question 22 - I am quite comfortable using the global descriptors

Australian Centres



Asia-Pacific Centres



It is significant that only 20% said they were not comfortable with the global descriptors when 75% of the Australian group had claimed that they would like to see a profile approach adopted. This would indicate that while respondents showed a preference for the profile approach, they were also quite able to use the global descriptors.

5.2.1 Summary of respondents comments: Questions 11 - 21

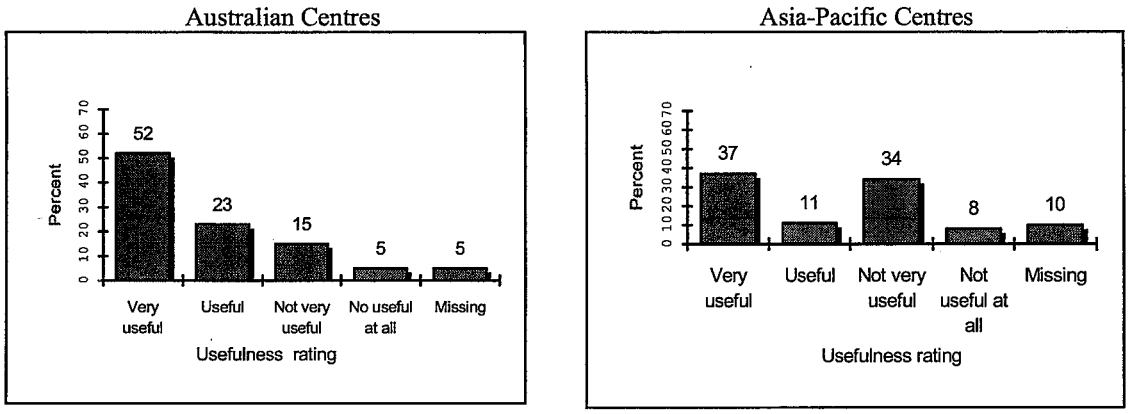
The comments were wide ranging and illuminating. Many respondents commented that they found it hard to differentiate between bands 5 and 6 as far as the descriptors were concerned and that clearer indicators were needed to guide the examiners in this area. Some people offered strong views about how profile descriptors would help enormously, particularly in areas such as pronunciation which is ignored in the descriptors for bands 5, 6 and 7. Others felt profiling would be time consuming. Many respondents made reference to the vagueness of the descriptors and the difficulties of interpretation which therefore arose. Terms such as 'fairly' and 'usually' were deemed unhelpful. For a full listing of examiner comments please refer to Appendix 1.3 - *Comments on Descriptors*

5.3 Section Three: The Interview Phases

Phase 1 - Introduction

Respondents were asked to rate the following propositions from 1 to 4 with 1 being *"very.."* and 4 being *"not at all"*

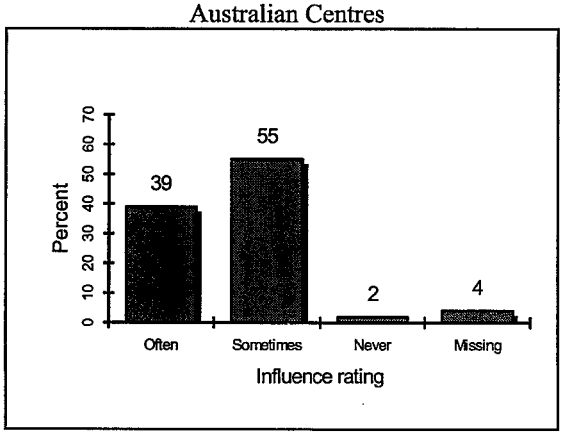
Question 23 - How useful is the candidate's CV/application form in Phase 1



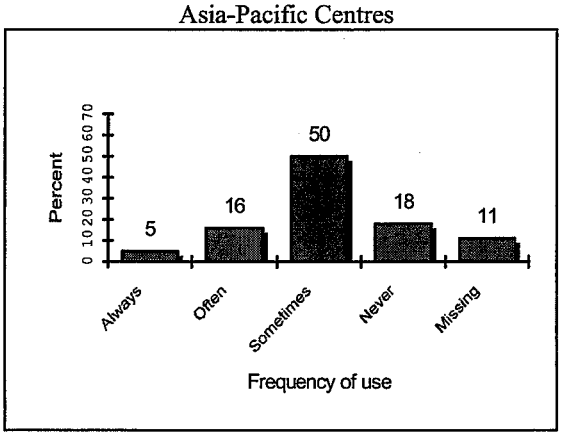
Three-quarters of Australian respondents advised that they found this useful with the majority of that group saying that it was very useful. This would indicate that the CV acts as a crutch or at least as a safety net for both the examiner and candidate. Only 20% felt that it was of little use or no use at all. Unlike the Australian-based examiners who were in favour of the CV, the overseas cohort was equally divided on this question. As it is standard practice in the Australian centres to use a CV but not so at British Council centres, this question may not have been viewed equally by all respondents.

Question 24 - How much does candidate performance in Phase 1 influence your final score?

Australian Centres



Asia-Pacific Centres

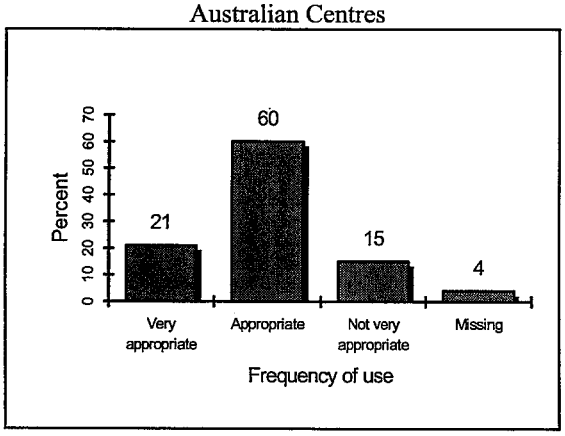


In Australia the vast majority 'admitted' to being influenced by the first impressions gleaned in phase 1 of the interview when technically no assessment should be taking place. The overseas examiners were not so revealing and quite a few chose not to answer this question.

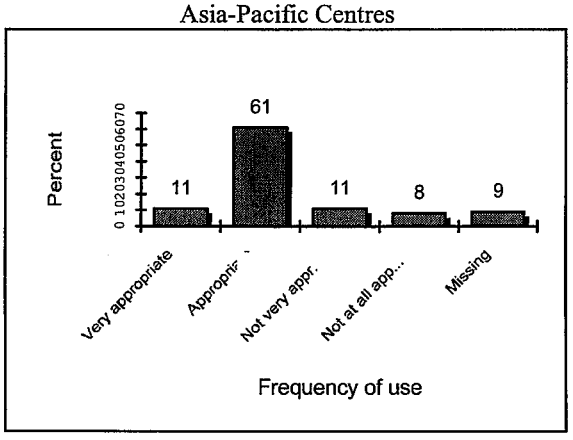
Phase 2 - Extended discourse

Question 25 - How appropriate is the choice of topics?

Australian Centres

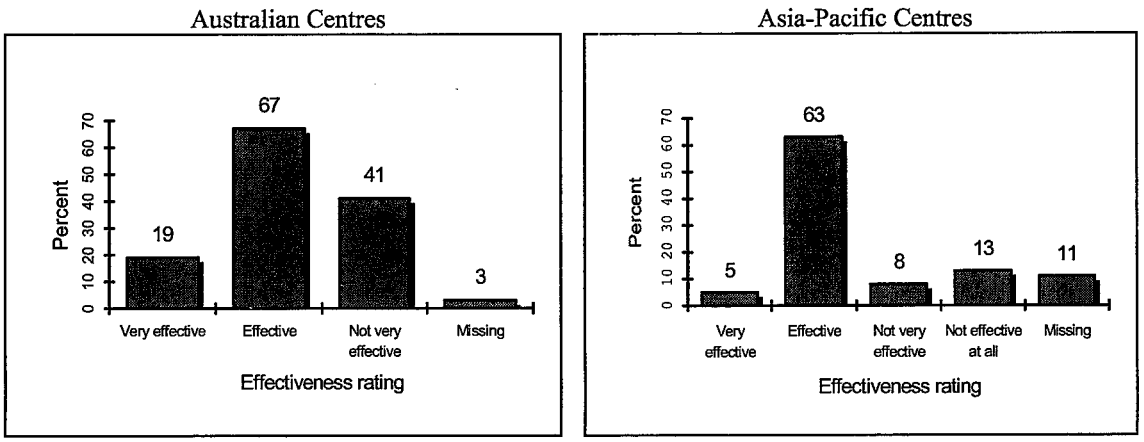


Asia-Pacific Centres



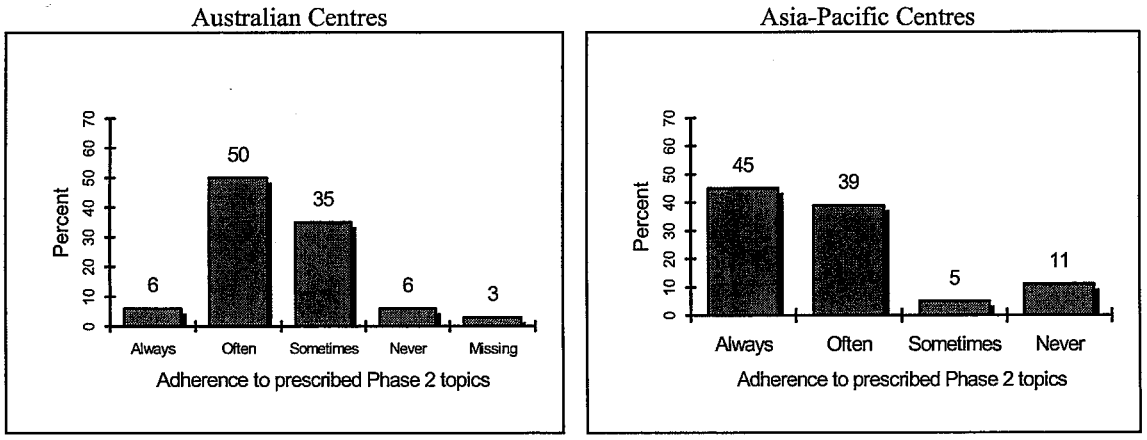
Over 80% felt that the choice of topics in Phase 2 was appropriate. This is significant as it now appears that these topics are 'public knowledge' and therefore can theoretically be practised in advance.

Question 26 - How effective are Phase 2 topics at producing assessable discourse?



Most respondents felt that they were effective.

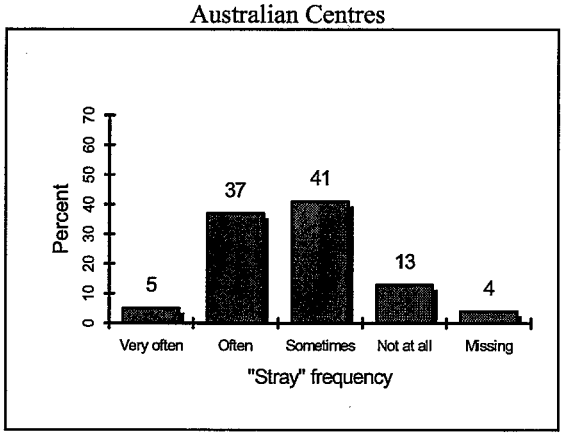
Question 27 - How rigidly do you stick to the prescribed Phase 2 topics?



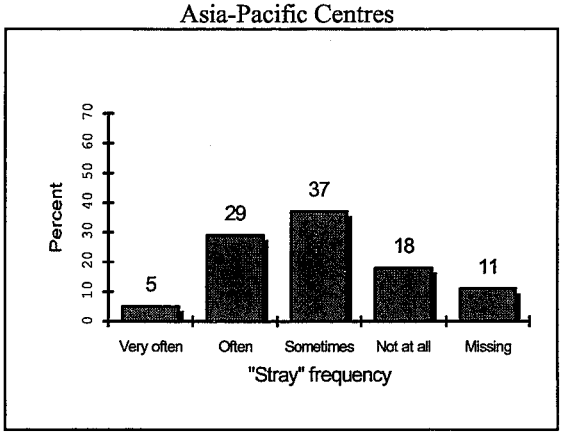
A significantly small number claim to use only the prescribed topics.

Question 28 - How often do you "stray" into Phase 4 topics in Phase 2?

Australian Centres



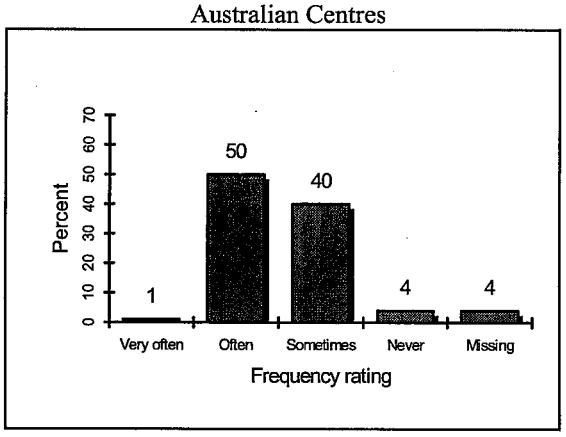
Asia-Pacific Centres



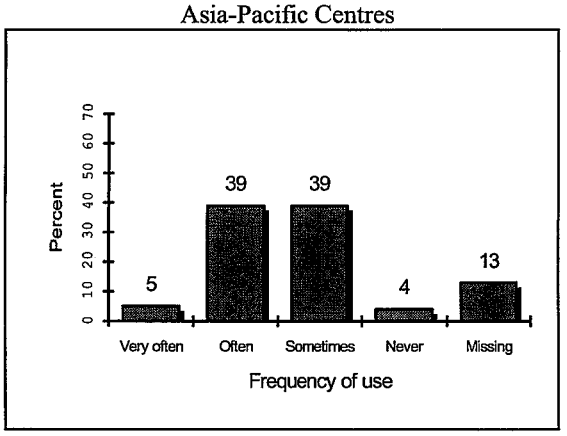
The pattern of responses is very similar from both groups. Again, the response indicates that examiners may touch on topics such as academic plans in Phase 2 which would tend to skew the format of the interview as the Phase 4 topics have then been used. A survey such as this is revealing but also allows us to remind examiners of the way in which they should be proceeding.

Question 29 - How often do candidates reach their linguistic ceiling in Phase 2?

Australian Centres



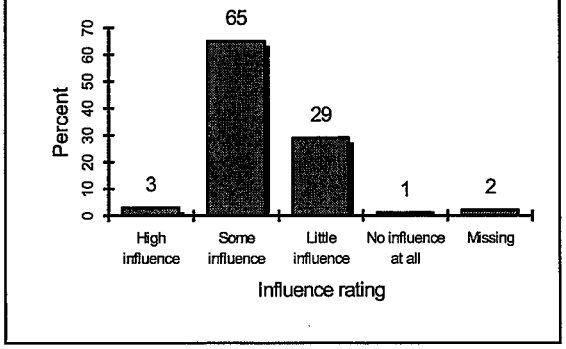
Asia-Pacific Centres



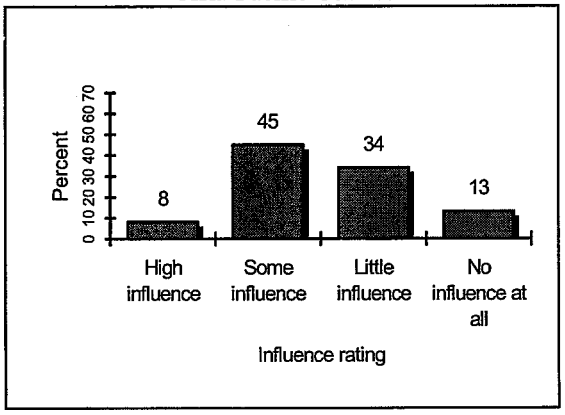
The response to this question would indicate that examiners feel that candidates often show their best performance by the end of Phase 2. This could be interpreted to mean that they are not sufficiently pushed in the latter part of the interview to show a higher level, or that indeed many candidates reach a performance plateau early in the interview.

Question 30 - How much does candidate performance in Phase 2 influence your final score?

Australian Centres



Asia-Pacific Centres

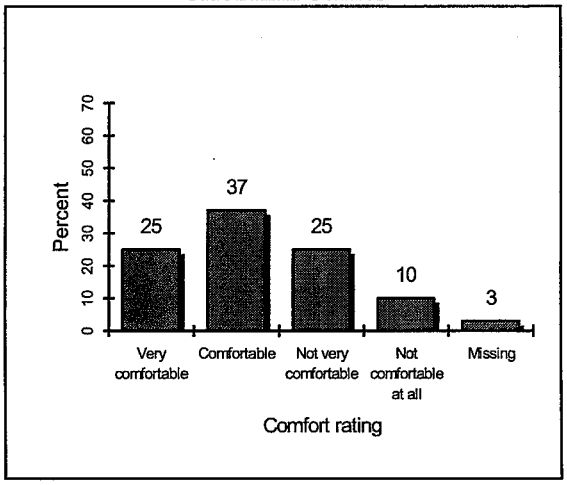


It would appear from the response to this question that many examiners effectively make up their mind about the rating by the end of Phase 2.

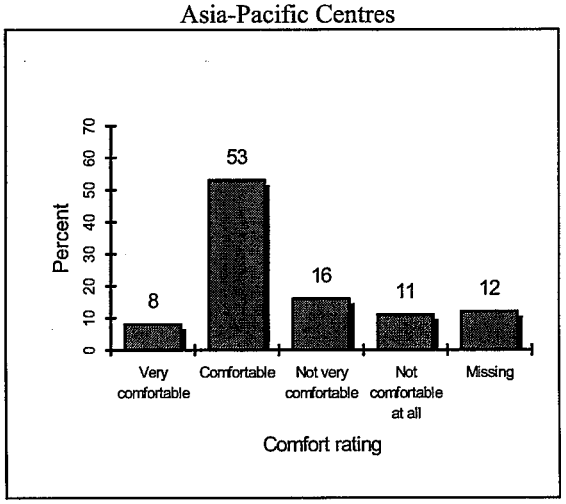
Phase 3 - Elicitation based on tasks

Question 31 - How comfortable do you feel about the interaction in Phase 3?

Australian Centres

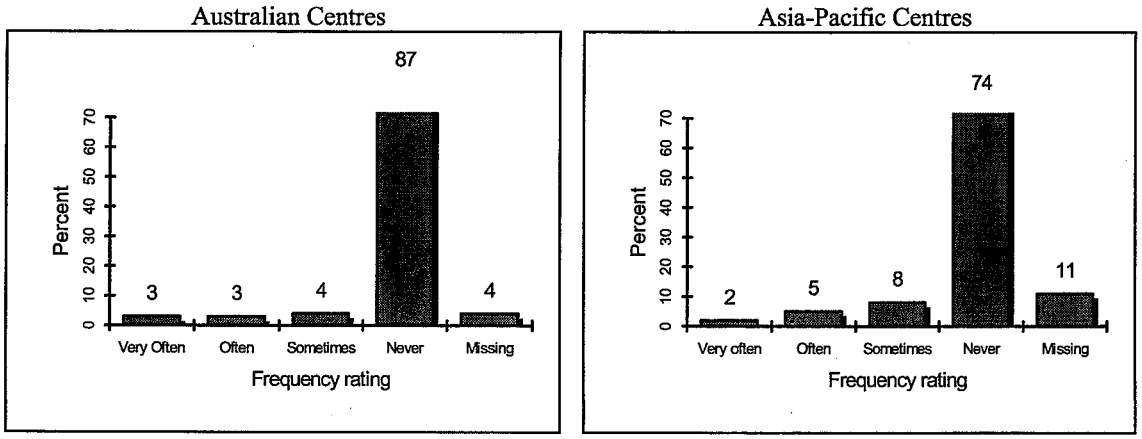


Asia-Pacific Centres



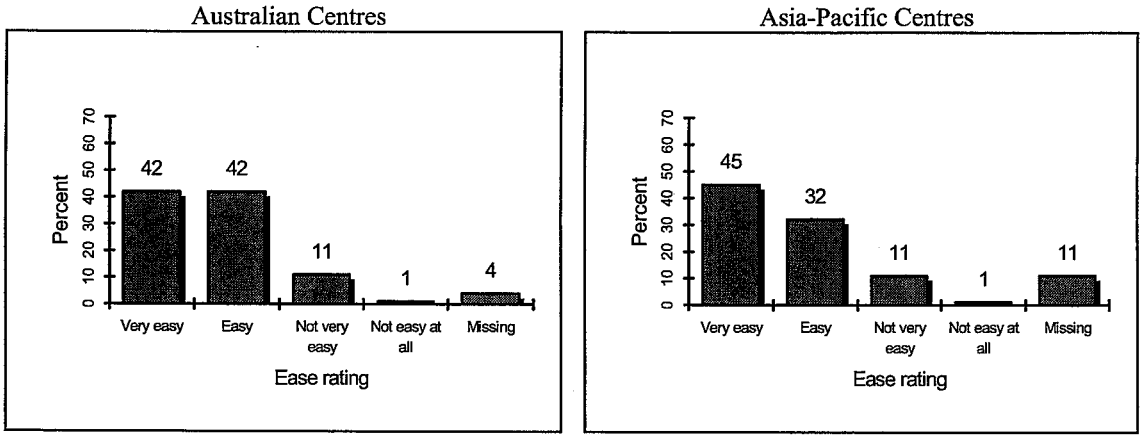
The response here was varied. Exactly 25% of Australian respondents felt very comfortable with the interaction, while an equal number expressed the view that they felt uncomfortable with it. The remaining 50% was mostly happy with the Phase 3 interaction though some 10% expressed a very negative view. The responses from the overseas examiners were similar though less extreme.

Question 32 - How often do you skip Phase 3 if candidate is struggling in Phase 2?



Since it is officially 'forbidden' to skip any of the phases of the IELTS interview, and examiners are trained not to do so, the responses here are revealing. While only a few respondents admitted to missing it out the fact that only 86% in Australia and 74% overseas answered that they never do so confirms some administrators' suspicions and is worrying, more from a point of view of procedural standardisation and thus face validity than anything else.

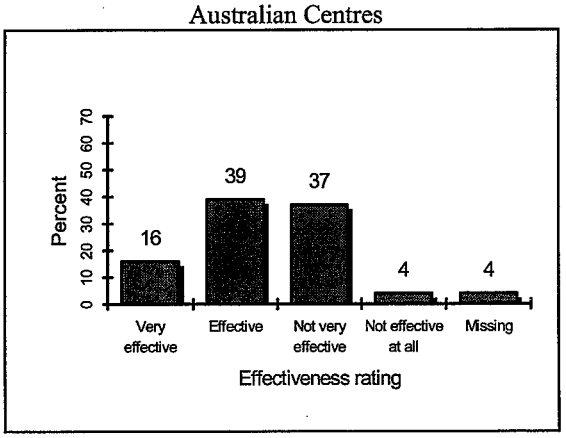
Question 33 - How easy do you find it to play the prescribed roles?



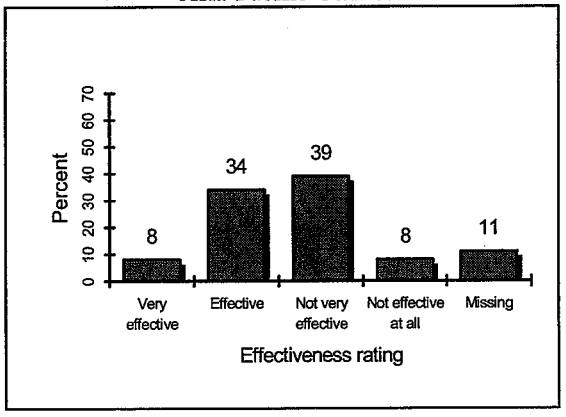
The responses from both groups are very similar. It seems that most examiners have little trouble playing the role that is expected of them in the Phase 3 elicitation phase. Slightly over 10% are not happy with the "role playing" aspect of the elicitation phase.

Question 34 - How effective are the tasks at producing assessable discourse?

Australian Centres



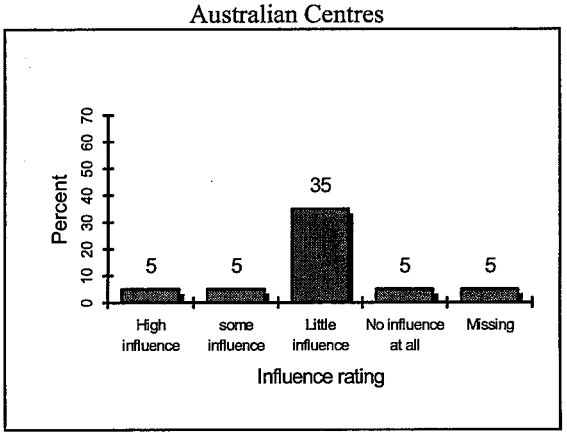
Asia-Pacific Centres



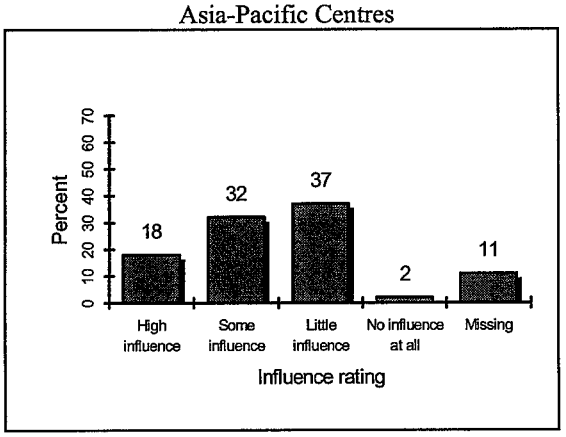
Here the field was evenly divided between those who find them effective and those who clearly do not. In other words opinions about the merit of Phase 3 cover the full range. This is revealing as it demonstrates that examiner attitude to this part of the test varies enormously.

Question 35 - How much does candidate performance in Phase 3 influence your final score?

Australian Centres

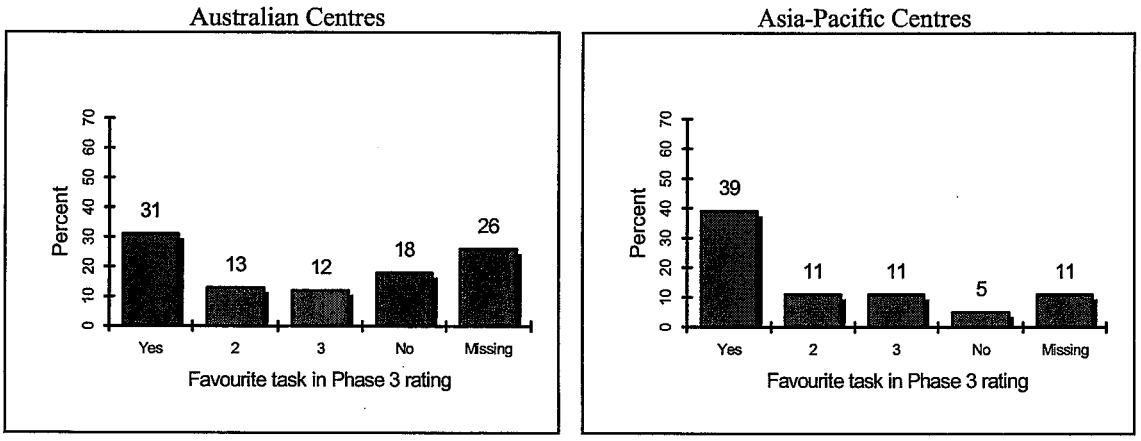


Asia-Pacific Centres



Again, the field is very divided here. We find that slightly over half of the examiners are influenced by the Phase 3 interaction and approximately 40% are not. These findings apply to both groups. This effectively means that the exercise is wasted in many cases. A significant number of respondents off shore did not offer a response.

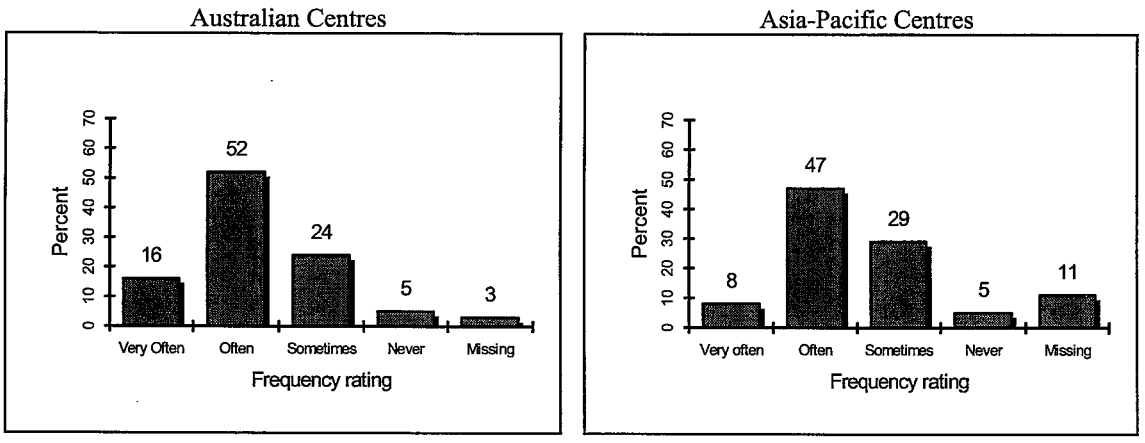
Question 36 - Do you have favourite tasks for phase 3?



Here the responses were varied with less than 40% claiming to have favourite tasks. Those cited by examiners again covered the full range. Many examiners mentioned the "Visiting a friend" task because of its authenticity as with the "wedding" and "evening course."
(See Appendix 1.4 for full overview)

Phase 4 - Speculation and Attitudes

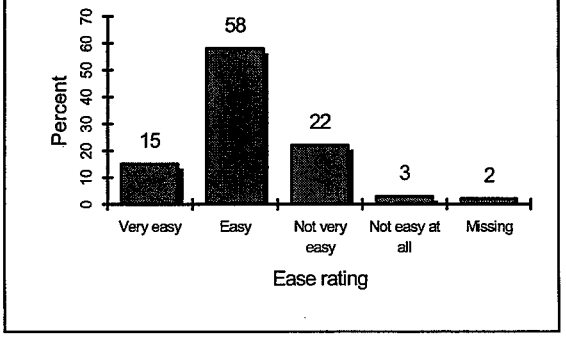
Question 37 - How often do you return to Phase 2 topics to generate Phase 4 discourse?



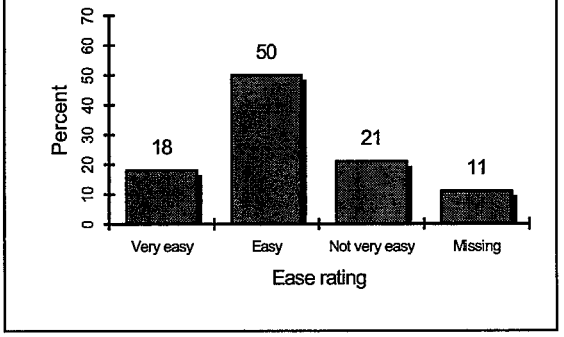
This question was intended to probe the issue of how examiners get back on track after the Phase 3 interaction and how best to prepare examiners for this in the training situation. The responses were varied enough to warrant looking at this issue further. Both groups gave very similar responses.

Question 38 - How easy do you find it to establish a useful Phase 4 topic?

Australian Centres



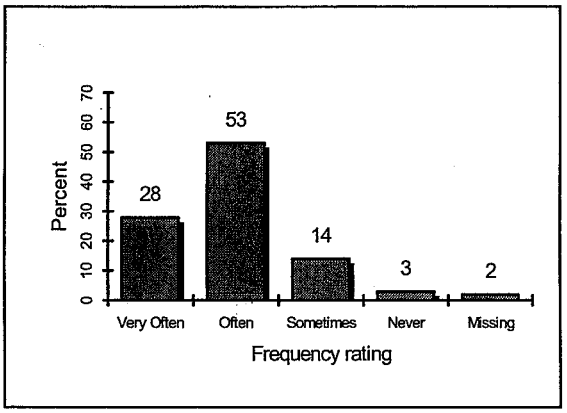
Asia-Pacific Centres



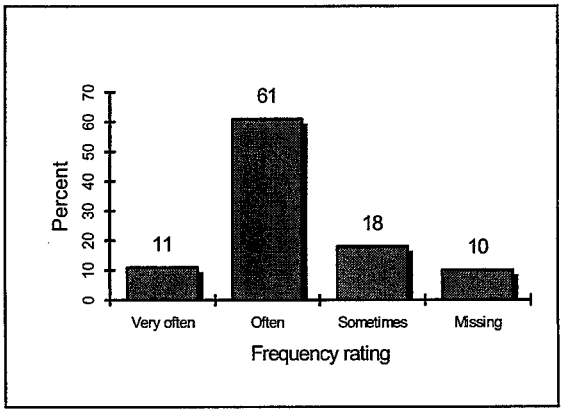
Approximately 70% of examiners claimed to have little or no trouble finding a suitable topic in Phase 4. This is encouraging and possibly what one would expect from experienced interviewers. The fact that 22% admitted to having difficulty would point to a need to include these techniques in the training procedure to provide a framework. Obviously some candidates are more difficult to interview than others, regardless of level, and examiners need to be able to deal with the more reticent ones.

Question 39 - How often do candidates reach their linguistic ceiling in Phase 4?

Australian Centres

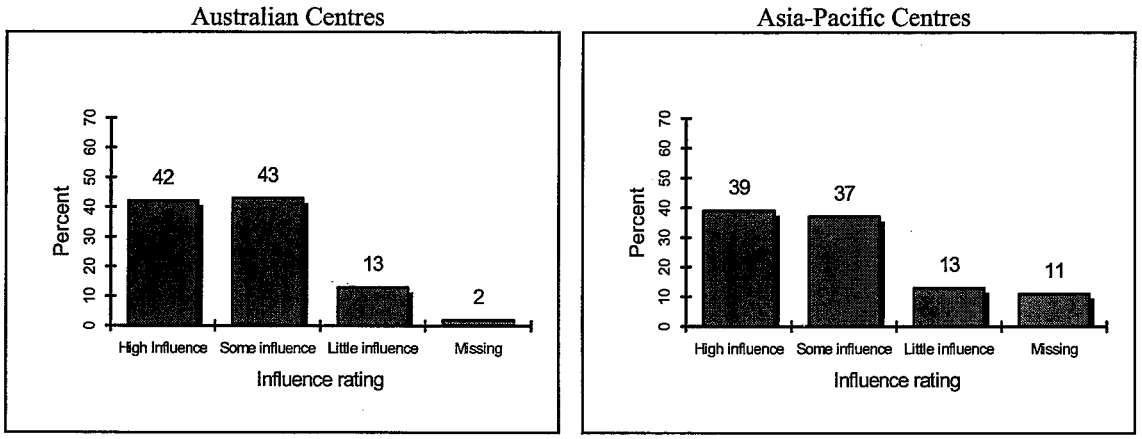


Asia-Pacific Centres



This question was included to find out how many examiners felt that they were able to take a candidate to his or her linguistic ceiling in the fourth phase. A very high proportion felt that they could do this which is encouraging and should point to the view that Phase 4 is working. However, another interpretation of these responses would be to say that ideally fewer respondents should be so sure that they are achieving this and so it may in fact point to the fact that they are unaware of their shortcomings as interviewers! The researchers suspect that this may be the case quite often.

Question 40 - How much does candidate performance in Phase 4 influence your final score?



The responses to this question reflect those of the previous probe in that 85% of the examiners surveyed felt that the candidate's performance in Phase 4 was significant in arriving at their final rating. Nevertheless the 13% of the Australian group who felt that it was not useful is significant enough to cause some concern, as is the fact that 11% chose not to answer in the overseas centres.

5.3.1 Summary of respondents' comments: Questions 31- 40

The comments vary enormously from those who are happy with the *status quo* to those who would eagerly accept change. This is inevitable and hardly surprising. However, the number of respondents who have commented on the non-academic nature of the interview is worth noting. Several respondents commented that candidates are now presenting for the test extremely test-wise having rehearsed nearly all the topics and Phase 3 tasks. Many examiners were prepared to volunteer the information that they found Phase 3 false and that it did little to focus their rating. Others however, felt that it provided a break between phases and maybe it should be seen as such.

Some of the comments were revealing in that they highlighted the individuality of approach. One respondent offered the information that s/he did not agree with basing a rating on the candidate's peak performance but rather relied on the "whole performance". Some expressed a strong desire to see profile type descriptors while others are obviously opposed to this approach.

6.0 Overall Comments on Survey Findings

The responses appear to have been supplied in a very open manner and the overall feeling of the research team was that they are honest and authentic. It is therefore felt that they should be taken seriously with regard to possible improvements to the test since these respondents are indeed the people implementing the instrument.

Several related issues were probed and now need attention.

1. There appears to be some divergence from the examiner guidelines with examiners taking liberties at times both with the format and the rating procedure. This can be addressed in refresher training for current examiners and also in training sessions for new examiners. It might be useful to produce a "Reminder Checklist" for examiners which is circulated by administrators at regular intervals.

2. It is evident that examiners would welcome amendments to the descriptors to provide clearer demarcation between a band 5 and 6 which are, to all intents and purposes, the critical levels.

3. There are clear differences of opinion about the merits of profile as opposed to global descriptors which stem possibly from experience in the field and from an adherence to a linguistic philosophy. The point was made, however, by a number of respondents, that it is difficult to operate a profile approach if one is also playing the role of interlocutor as well as the assessor.

   Individual comments from off-shore examiners reflected a scepticism about Phase 3 in terms of its actual usefulness and one or two commented that it interrupts the flow. These comments echo those of the Australian examiners. The question which prompted them to think about how often they embark on Phase 4 topics in Phase 2 has also pin pointed an inherent problem and needs to be addressed in training.

   The research team would like to suggest that the examiners working in off-shore centres may be intimidated by the idea of the two examiner system because, firstly it is an unfamiliar approach and also because it poses potential constraints on their interview style with a "watch dog" implication built in. Even if it is not adopted, there is certainly a need to monitor examiners more often than is currently occurring.

   Since all tests are a balance between what is practical in terms of reliability and also what is best practice, and for reasons of administrative ease and expediency, an approach to the IELTS speaking test which allowed for ease of delivery was adopted. The rationale for this decision must not be overlooked. IELTS is available practically on demand on a world wide basis and must therefore be easily administered. On the other hand, we do not want to lose sight of the importance of maintaining a reliable instrument simply in order to keep administrative arrangements to a minimum.

7.0 Analysis of Candidate/Examiner Discourse

Following discussions between the LTC research team and UCLES, it was decided to construct a dataset of 20 IELTS interviews, transcribe all the interviews and then use the transcripts as the basis for the analysis. The analysis would aim to ascertain to what extent, if any, examiner discourse measured in terms of the amount of "talk time" affects the language produced by the candidate both in terms of quantity as well as quality. This would be investigated through a number of avenues.

7.1 Construction of the Dataset

The interviews chosen for the dataset were selected from the bank of taped IELTS interviews held at International Programs, University of Technology, Sydney (UTS). UTS keeps all IELTS interviews for at least 4 months and as test administrations of 140 candidates are the norm, the research team felt there would be a sufficient range of materials to suit the particular requirements of the dataset.

An important constraint in the construction of the dataset was to use interviews of candidates who had relatively "flat profiles" i.e. whose scores on the reading, writing, speaking and listening subtests were within 1.5 bands of each other. To ensure this type of flat profiling, each interview was re-assessed twice by the Sydney senior examiners to ensure the original assessment given to the candidate was reliable. Where videoed interviews made for the training materials were used, this was not necessary as they had already been double marked by the senior examiners.

7.2 The Dataset

Band 7

Candidate Number/ Interview number	Gender	Age	First Language	Phase 3 task	Assessments			Original Examiner	
					Original	Second	Third	Gender	Age
V 17/ Interview 20	M	20's	German	withdrawn Phase 3 task	7	-	-	F	40s

Band 6

Candidate Number/ Interview number	Gender	Age	First Language	Phase 3 task	Assessments			Original Examiner	
					Original	Second	Third	Gender	Age
0950/ Interview 3	M	27	Korean	4	7	6	6	F	42
1686/ Interview 4	F	28	Burmese	5	6	6	6	F	46
V 9/ Interview 17	M	30's	Thai	-	6	6	-	M	40's
V 10/ Interview 18	F	20's	Serbian	-	6	6	-	F	45
V 14/ Interview 21	F	20's	Portuguese	-	6	6	6	M	40s
1474	M	31	Italian	1	6	6	6	F	

Band 5

Candidate Number/ Interview number	Gender	Age	First Language	Phase 3 task	Assessments			Original Examiner	
					Original	Second	Third	Gender	Age
1669	F	23	Indonesian	16	6	5	5	M	61
1503	F	27	Thai	15	7	5	5	F	
0918	M	24	Korean	8	5	5	5	F	44
1745	F	25	French	2	6	5	5	M	28
1817	F	24	Vietnamese	15	5	5	5	M	36
V 1/ Interview 15	F	20	Chinese	-	5	5	-	M	35
V 6/ Interview 16	F	20	Korean	-	5	5	-	M	
0949	F	29	Korean	6	4	5	5	F	44

Band 4

Candidate Number/ Interview number	Gender	Age	First Language	Phase 3 task	Assessments			Original Examiner	
					Original	Second	Third	Gender	Age
0983	F	19	Chinese	4	5	4	4	F	35
1529	M	21	Indonesian	2	4	4	4	F	
V 13/ Interview 19	M	20's	Korean	-	4	-	-	F	40s
1508	F	18	Japanese	4	4	4	4	F	39

Band 3

Candidate Number/ Interview number	Gender	Age	First Language	Phase 3 task	Assessments			Original Examiner	
					Original	Second	Third	Gender	Age
1790/ Interview 2	M	21	Thai	3	3	3	3	F	45
1000/ Interview 9	M	15	Thai	9	4	3	3	M	

7.3 Analysis of the dataset

The research team felt it would be useful in the analysis to investigate the interaction in terms of length of turn between the examiner and the candidate as this could possibly affect candidate performance and hence the final assessment. One area of particular interest was the possible dominance of the interviewer in the different phases of the interview. There are various methods for measuring the turns and it was decided to use a word count of each turn. All the interview transcripts were annotated so that the number of words in each turn was placed next to the turn number. The data were then processed to produce a "map" for each interview which tracked each turn and highlighted the boundaries of phase 3 and phase 4. A small sample of these maps can be seen in Appendix 1.5 and provide a useful overall impression of the interviews. As a further measure of examiner dominance, Nick Saville of UCLES suggested calculating the average turn length for phases 1&2, 3 and 4&5 for both candidate and interviewer to see how these match the information displayed on the graphs.

The table below shows average turn lengths in number of words:

Interview number	Interview participant	Phases 1&2	Phase 3	Phases 4&5	Average number of words per turn	Band
1	Examiner	3.4	5.5	3.7	4.0	5
	Candidate	5.4	2.8	5.8	5.0	
2	Examiner	5.4	15.1	5.1	6.8	3
	Candidate	7.2	3.2	11.4	7.8	
3	Examiner	3.7	5.4	3.6	4.1	6
	Candidate	11.2	8.4	16.0	12.0	
4	Examiner	5.5	11.2	7.1	7.5	6
	Candidate	6.3	3.7	8.4	6.5	
5	Examiner	4.9	6.9	6.2	5.7	5
	Candidate	7.5	4.4	7.7	6.9	
6	Examiner	8.0	8.5	13.7	9.5	4
	Candidate	4.8	6.7	9.1	6.1	
7	Examiner	5.0	5.5	4.1	4.9	4
	Candidate	4.1	3.1	4.4	3.9	
8	Examiner	6.1	9.2	9.8	8.3	5
	Candidate	28.6	12.4	10.3	17.3	
9	Examiner	5.0	7.0	6.8	6.0	3
	Candidate	3.5	4.1	3.7	3.7	
10	Examiner	3.2	7.0	3.7	1.8	5
	Candidate	4.6	2.8	4.8	4.4	
11	Examiner	4.7	7.7	5.2	5.2	4
	Candidate	4.5	3.6	7.5	5.1	
12	Examiner	2.7	5.5	3.9	3.9	6
	Candidate	8.5	4.1	5.0	6.1	
13	Examiner	4.2	8.5	5.4	5.8	5
	Candidate	7.8	5.9	8.2	7.3	

14	Examiner	3.7	6.1	3.8	4.2	5
	Candidate	6.2	5.5	7.8	6.7	
15	Examiner	6.0	6.9	6.7	6.4	5
	Candidate	11.1	4.6	9.3	9.1	
16	Examiner	4.4	14.3	8.3	7.7	5
	Candidate	6.9	6.4	7.0	6.9	
17	Examiner	5.1	8.7	4.9	6.0	6
	Candidate	6.3	7.1	7.9	7.3	
18	Examiner	4.8	12.0	7.6	7.4	6
	Candidate	6.8	6.7	7.6	7.0	
19	Examiner	2.8	6.4	3.9	4.3	4
	Candidate	6.9	4.9	7.9	6.4	
20	Examiner	6.6	11.3	4.8	6.6	7
	Candidate	12.7	7.0	9.2	10.2	
21	Examiner	8.0	31.1	8.1	10.9	6
	Candidate	11.6	7.3	10.6	10.6	

It can be seen from the data above that Phases 1 and 2 are producing longer average turns from the candidate than the examiner, which is obviously desirable though there appears to be no correlation between the lower band speakers and the higher band speakers in terms of turn length. Candidate 8 (Band 5) has produced by far the most language in these phases but is exceptional. Since the early phases of the interview perform the function of establishing identity, settling the candidates down and then allowing them to speak on a familiar topic at a level with which they feel comfortable, one would not expect an enormous difference in average turn length as the examiner has to initially lead the discussion. Clearly it is unsatisfactory, however, for the examiner to be speaking more than the candidate.

However, when we look at the turn lengths for Phase 3 we find in all but one case that the examiner is producing far more discourse than the candidate. Again there does not appear to be any particular correlation between examiner input and candidate level; in fact the opposite since candidates 20 and 21 (Band 7 and Band 6) have produced significantly less language than their examiners.

The phenomenon of examiner over-input may not be a direct fault of the test design but rather of test delivery. In other words, examiners could well be accused of simply not following the test guidelines which stipulate that very short, almost unhelpful responses need to be given in answer to the candidate's Phase 3 questions in order to force them to produce more questions. However, since it appears to be such an all pervasive difficulty which occurs at all band levels

and across a broad range of examiners, one is bound to consider that the problem may stem from the test format itself rather than simply the execution of Phase 3. [3]

Phases 4 and 5 show a more desirable result with the majority of candidates producing significantly more discourse than their examiners. The Band 3 candidate (Interview 9) and Band 4 candidate (Interview 6) are notable but explicable exceptions while the Band 7 candidate is able to demonstrate his proficiency level by taking obvious control of the interview. The Band 5 and 6 candidates are also sufficiently comfortable with their language to be able to take the apparent initiative in this latter part of the test.

Nevertheless, a number of the Band 6 candidates appear not to have produced much more language than their examiners which could mean that they had not been pushed sufficiently towards their linguistic ceiling, since the examiner is still producing a good deal of the discourse. This of course begs the question of whether they are true Band 6 candidates. In other words, had they been pushed harder in a less supportive manner with the examiner relying less on the scaffolding technique, would they still have been deemed Band 6? It is not within the scope of this research project to pursue this line but there is certainly room for further investigation of the data.

Another area of interest to the research team was the area of standardised delivery of the interview and one way to investigate this was to measure the exact lengths of the different phases of the interview and the overall length of the interviews.

[3] Discussions held by the research team with experienced examiners in refresher workshops as well as with trainers and trainees have invariably shown that examiners find it very unnatural to reply in an "unhelpful" staccato manner to questions posed by candidates. There is therefore an almost irrepressible desire to provide further information. This can be "trained out" of examiners for a period of time but apparently not for ever.

The table below shows the timings of phases 1&2, phase 3, phase 4&5 and the total length of the interview in minutes and seconds.

Interview Number	Candidate Number	Length of Phase 1&2	Length of Phase 3	Length of Phase 4&5	Total length of interview	Band Level
1	0949	5.14	3.02	9.28	17.44	5
2	1790	5.55	3.55	3.15	13.05	3
3	950	5.21	3.54	5.20	14.35	6
4	1686	3.35	3.55	5.45	13.15	6
5	1745	5.45	3.05	3.47	12.37	5
6	1529	5.32	1.48	3.14	10.34	4
7	1508	5.07	3.53	2.41	11.41	4
8	1817	5.19	3.57	3.07	12.23	5
9	1000	3.15	4.50	1.11	9.16	3
10	918	6.07	3.46	6.43	16.36	5
11	983	5.33	2.27	2.41	10.41	4
12	1474	4.46	3.10	2.47	10.43	6
13	1503	5.29	4.41	2.44	12.54	5
14	1669	5.32	2.45	5.52	14.09	5
15	Video 1	5.55	2.10	5.13	13.18	5
16	Video 6	3.20	2.30	5.25	11.15	5
17	Video 9	2.55	4.00	5.35	12.30	6
18	Video10	3.40	3.50	2.53	10.23	6
19	Video 13	4.30	4.32	3.29	12.31	4
20	Video 17	6.00	2.17	5.08	13.25	7
21	Video 14	3.55	3.09	5.26	12.30	6

As expected the length of the interview with weaker candidates is invariably shorter but most of the interviews appeared to fall within the guidelines of 12 to 15 minutes with two exceptions which went overlength. The internal timings, however, are quite varied and would indicate that there is a lack of standardisation and that this is an area which should be addressed both in the examiner training as well as the refresher courses to ensure that candidates receive the same version of the test.

8.0 Overall Findings and Conclusions

The analysis of the candidate/examiner discourse revealed a number of important issues.

1. The data has provided empirical evidence that examiners are speaking as much as and in some cases more than their candidates. It shows that Phase 2 is for the most part being conducted in a standard manner, but that in Phase 3 there is far too much "examiner talk" occurring and in Phase 4, there is insufficient candidate discourse being elicited. Since the object of the IELTS interview is to elicit assessable discourse from the candidates but not to assess their listening comprehension skills, this is clearly a problem.

2. The data has shown that in many cases the length of the different phases varies considerably from the specified norm and that in some cases, phases are cut very short. This is possibly acceptable when very low level candidates are examined but it is not acceptable at the higher levels.

3. There seems to be a correlation between the number of candidates assessed as being Band 6 but whose examiners are continuing to provide a good deal of scaffolding in Phase 4. Candidates in the group who were rated as Band 5, on the other hand, have in many cases been given a more obvious opportunity to demonstrate their weaknesses in Phase 4 as the data shows less "examiner talk". These interviews may in fact be better examples of a "good" IELTS interview. While we have no question that the bands awarded are accurate, these findings call into question the reliability of the test in the hands of examiners who choose not to extend their candidates fully. There appears to be a tendency among some examiners to use phase 4 to engage in general discourse rather than taking the candidates to their linguistic limits.

4. The average number of words per turn produced by the candidates appears to be low.

5. A reading of the transcripts reveals that a great many closed questions are being posed by examiners, leading to short responses which provide little assistance to the assessor/examiner.

9.0 Further Research

There are clear grounds for further research to be undertaken in the two broad areas of interest covered by this study building on the original data gathered for this research.

Firstly in the area of examiner attitude, it would be useful to undertake trials with a set of profile descriptors which differentiate more clearly between the critical levels of Bands 5, 6 and 7.

Secondly, there is an apparent need to monitor examiner performance with regard to standards of test delivery, both in the area of timing and also in the requirement to bring candidates to their linguistic ceiling in Phase 4. At the present time examiner monitoring takes the form of checking to see whether the assessment is within acceptable levels of accuracy and does not effectively address whether correct IELTS interview procedures are being followed.

Thirdly, there is scope for investigating whether more prescriptive examiner language could be introduced into the interview format. Introducing more tightly controlled examiner language or even an interlocutor frame in the form of a finite list of specific phrases or questions to be used by the examiner could be one method for standardising examiner discourse. In addition, examiners need to follow the guidelines with the aim of bringing the candidates to their true linguistic ceiling in phase 4.

Fourthly, there is a need to investigate whether prompts and/or pictures could assist in the process of eliciting longer turn discourse as the study has shown that candidates use few words in their responses and rarely produce a true stream of language. The discourse produced is often quite authentic within the 'conversational genre' but does not always provide sufficient evidence on which to base an accurate assessment.