

NOTE TO USERS

This reproduction is the best copy available.

UMI[®]



Université d'Ottawa • University of Ottawa



Université d'Ottawa - University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES

FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES

STOCKTON, Stephen

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

Ph.D (Biology)

GRADE - DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

A Preliminary Investigation into Discursive Models of Interpreting as a Means
of Enhancing Construct Validity in Interpreter Certification

Roda Roberts

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

G. Bastin

M. Simon

S. Lambert

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES
SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE
AND POSTDOCTORAL STUDIES

**A Preliminary Investigation into Discursive Models of
Interpreting as a Means of Enhancing Construct Validity in
Interpreter Certification**

Andrew Clifford

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the
Ph.D. degree in Translation Studies

School of Translation and Interpretation
Faculty of Arts
University of Ottawa

©Andrew Clifford, Ottawa, Canada, 2004



Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-494-01681-7

Our file Notre référence

ISBN: 0-494-01681-7

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.


Canada

Abstract

Interpreter certification has been largely ignored by the literature in Interpreting Studies. What little research that does exist on the topic is understandably elementary, discussing concerns that are general in nature or describing the development of individual certification instruments. No research has as yet examined the psychometric properties of certification instruments, despite the urgings of legal precedent, the weight of professional opinion, and the requirements of professional standards.

The present study has attempted to bridge this gap in two ways. First, to gain some sense of the current state of affairs in interpreter testing, a validation study of an existing certification test was conducted. Particular attention was paid to information about the test's constructs. Second, in a bid to improve construct validity, a new certification test was constructed and subsequently validated. Constructs in the new test were based on a discursive model of the interpreting process.

Validation of the two tests revealed some interesting findings. Constructs in the existing test bore only a weak relationship to the interpreting profession. Also, the constructs did not prove to be useful in predicting relationships among the existing test scores. It was expected that three separate constructs would emerge from the data; instead, empirical evidence suggested that the data were unidimensional in nature.

In contrast, constructs in the new test were strongly linked to the interpreting profession. They also proved to effectively predict relationships among the new test scores. It was expected that three separate constructs would emerge from the data, and empirical evidence suggested that this was, in fact, the case.

The results of the present study indicate that there is room for improvement in the current state of interpreter testing, and that a test instrument based on discursive constructs is likely to demonstrate acceptable levels of validity. There is obviously a need for a great deal of future research on the psychometric aspects of interpreter certification, and it is strongly recommended that this research adopt a discursive approach.

Acknowledgements

Since beginning my doctorate just over four years ago, my path has at times been barred a number of challenges.

The simplest of these was also the most daunting: summoning the courage to face the work each day.

I would therefore like to thank those whose example helped me to bring the work to its end.

Professor Simon, who agreed to lend a helping hand when others would not.

Professor Lambert, who took an interest early on and remained involved.

Professor Roberts, who read everything with a meticulous eye.

Michele, who blazed the trail and shared her experience.

And Steve, who simply remained constant.

I offer you all my thanks.

Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents	iv
List of Tables	vii
1. Introduction.....	1
1.1 Research Problem and Design of the Study.....	5
1.1.1 Scope of the Study	6
1.1.2 Components of the Study	8
1.2 Research Hypotheses	11
1.2.1 General Research Hypothesis	12
1.2.2 Specific Research Hypotheses	12
1.3 Conclusion	14
2. An Examination of Potential Research Models.....	17
2.1 Translation Quality Assessment Literature.....	18
2.1.1 Quantitative Models of TQA	18
2.1.2 Qualitative Models of TQA	24
2.2 Credentialing Literature	35
2.2.1 Research in the Health Care Professions	36
2.2.2 Research in Other Professions	43
2.3 Establishing a Model for the Present Study	54
2.3.1 Choice of a Model.....	55
2.3.2 Justification for the Choice of Model.....	57
2.4 Conclusion	63
3. The Identification of Interpreting Competencies.....	65
3.1 Interpretive Theory	70
3.1.1 The Interpreting Process	71
3.1.2 Linguistic Meaning and Meaning in Context.....	74
3.1.3 Units of Meaning.....	77
3.1.4 Simultaneity	79
3.2 Information Processing	83
3.2.1 Gerver.....	84
3.2.2 Moser.....	88
3.2.3 Cokely	93
3.2.4 Gile.....	96
3.2.5 Paradis	99
3.3 Discursive Models.....	106
3.3.1 Wadensjö.....	109
3.3.2 Roy	117

3.3.3 Setton.....	122
3.4 Conclusion	129
4. Validation Methodology	136
4.1 The Existing Test	136
4.1.1 Test Description	136
4.1.2 Assumptions about the Test Data.....	139
4.1.3 Sample.....	143
4.1.4 Procedure.....	148
4.2 The New Test	152
4.2.1 Test Description	155
4.2.2 Assumptions about Test Data.....	162
4.2.3 Sample.....	166
4.2.4 Procedure.....	167
4.3 Comparison of the Two Analyses.....	170
4.4 Conclusion	170
5. Validation Results	172
5.1 The Existing Test	172
5.1.1 Examination of the Constructs' Theoretical Origins	173
5.1.2 Estimation of Construct-Irrelevant Variance	178
5.1.3 Estimation of Construct Variance	185
5.1.4 Secondary Estimation of Construct Variance	190
5.2 The New Test.....	196
5.2.1 Examination of the Constructs' Theoretical Origins	197
5.2.2 Estimation of Construct-Irrelevant Variance	203
5.2.3 Estimation of Construct Variance	213
5.2.4 Secondary Estimation of Construct Variance	216
5.3 Conclusion	218
6. Interpretation	219
6.1 Comparison of Test Results	219
6.2 Hypothesis Testing.....	223
6.3 Conclusion	236
7. Conclusion.....	237
7.1 Future Research.....	238
References	243
Appendix I – Glossary	256
Section A	256
Section B.....	257
Section C.....	257
Section D	262
Section E.....	262
Section F	263

Section G	264
Section I	265
Section K	268
Section L	268
Section M	268
Section P	268
Section R	270
Section S	271
Section T	273
Section V	274
Appendix II – Existing Test Documentation	275
Rater Instructions	275
Intelligibility Rating Scale	276
Informativeness Rating Scale	277
Style Rating Scale	278
Rating Sheet	279
Appendix III – New Test Documentation	280
Transcript of the Ten-Minute Speech	280
Complete Test Instrument with Option Weighting	284
New Test Raw Scores with Indices	295
Distractor Analysis for New Test Raw Scores	296

List of Tables

Table 1: Candidate Information	143
Table 2: Multitrait-Multimethod Matrix of Correlations for Constructs within Item Scores	179
Table 3: Descriptive Statistics for the Existing Test Items	183
Table 4: Correlations among the Existing Test Items	184
Table 5: Comparison of Correlated Coefficients for the Existing Test Items.....	185
Table 6: Descriptive Statistics for the Existing Test Constructs.....	186
Table 7: Correlations among the Existing Test Constructs.....	187
Table 8: Comparison of Correlated Coefficients for the Existing Test Constructs	189
Table 9: Principal Components Analysis of Existing Test Constructs	193
Table 10: Coefficient Alpha for Existing Test Constructs.....	195
Table 11: Raw Scores for Items in the Sample with Indices and Subtotals.....	206
Table 12: Sample Items Comparing Discrimination based on Totals and Subtotals.....	209
Table 13: Distractor Analysis for Items in the Sample based on Total Score	210
Table 14: Distractor Analysis for Items in the Sample based on Subtotals	212
Table 15: Descriptive Statistics for the New Test Constructs.....	214
Table 16: Correlations among the New Test Constructs in the Sample	215
Table 17: Comparison of Correlated Coefficients for the Existing Test Constructs	216
Table 18: Coefficient Alpha for New Test Constructs.....	217
Table 19: Comparison of Unrelated Coefficients for Constructs on Both Tests	222

1. Introduction

In 1980, after failing the US Federal Court Interpreter Certification Examination twice, Norma Seltzer and Margarita Torres-Cartagena sought an injunction against the Director of the Administrative Office of the United States Courts, in order to prevent him from continuing to use the exam. The plaintiffs claimed that the exam was inappropriate, because it consisted in part of a written language proficiency *test*¹ that was unrelated to actual on-the-job behaviour (*Seltzer v. Foley*, 1980).

In the ensuing court case, the defence lawyer placed a great deal of emphasis on the role that *psychometrics*² had played in the development of the test, by underscoring several points. First, he noted that the cut-off score (the score that differentiates between passing and failing grades) was determined by empirical means. Second, he pointed out that one of the *test developers*^{*} was uniquely qualified, not only because she was an experienced interpreter and interpreter trainer, but also because she was a Ph.D. student in educational *measurement*^{*}. Third, the defence lawyer stressed the point that the entire development process had been reviewed by an external psychometrician.

The focus on psychometrics was obvious in other aspects of the case. The defence lawyer spent a great deal of time establishing the psychometric property of *validity*^{*}: in other

¹ Throughout the present study, we have used a number of technical terms that may not be familiar to all readers. We have therefore included a glossary of terms in *Appendix I*. Terms that have an entry in the glossary are identified by italics and by an asterisk at their first appearance in the text.

² Psychometrics was defined in the case as the “psychology of measurement”, and it was further noted that measurement in this sense is mathematically based.

words, he argued that accumulated evidence supported the interpretations made from the test scores. The evidence in question, a study of the interpreting profession conducted by the test developers, showed that the test content was consistent with the realities of the profession. For example, the study indicated that *court interpreters** regularly encounter challenging technical vocabulary and language registers – most notably in the form of “legalese” – in the course of their work. It was therefore appropriate to use an advanced language proficiency test as part of the overall process to certify court interpreters.

After considering the arguments of both sides, the district court judge ruled in favour of the defendants, and the use of the exam was upheld. The judge noted in his findings that “[t]he said tests and/or the criteria used by the Director bore a rational and proper relation to skills appropriate, necessary and required for requisite precision interpretation by bilingual interpreters in courtroom settings and were valid and reliable in purpose and effect (p. 608).” The defence’s strategy was successful. By arguing generally that the exam was psychometrically sound, and specifically that decisions made from the exam scores were valid, the defence had satisfied the courts.

The court case had also set a precedent, in two ways. First, if the appropriateness of a *certification** test is called into question, the case shows that test developers must be prepared to show evidence of the test’s psychometric properties. For example, developers must demonstrate validity, as the defence lawyer did in the case above, or *reliability**, which would involve showing that the test scores are stable and dependable measurements. Second,

the focus on psychometric evidence is important, not only for professions like medicine and education, where it is already common practice³, but for interpreting as well.

Yet more than 20 later, it appears that little attention has been paid to the points raised by the court case. In the interpreting profession, tests continue to be used, but there is very little discussion of their psychometric properties. The rarity of such discussion has led Sawyer (2000) to call for greater awareness of the basic principles of measurement and *evaluation** within interpreting. He claims that interpreter testing is poorly done, and he cites a number of widespread problems – the arbitrary selection of test content, the lack of consistent test administration practices, the failure to establish and respect objective scoring criteria, and the inappropriate use of test results. He characterizes the state of affairs by saying that “...current testing practices are grossly inadequate in many programs for spoken language interpreters, professional associations of translators and interpreters, and the marketplace” (p. 33).

In light of the problems that Sawyer enumerates, it seems reasonable to hypothesize that existing tests are not measuring interpreting skill accurately. However, Sawyer offers no hard evidence – studies of reliability or validity for example – to support his claim. Neither is it possible to back up his contention with other research on professional interpreter testing. What little work does exist on the topic (Arjona, 1983; Schweda-Nicholson, 1986; Roberts, 1995; Roberts, 1998) has generally been limited to discussing the development and administration of particular instruments, and has not included a thorough evaluation of those

³ As the short literature review in *Section 2.2* will attest, it is common practice to publish studies of

instruments' psychometric properties. Without the results of a qualitative and quantitative examination of an actual test, Sawyer's claim about testing quality, as compelling as it is, remains little more than an educated guess.

Yet let us suppose for a moment that Sawyer is correct. What then? How might we go about improving interpreter testing? In what ways might we combine greater awareness of measurement and evaluation, as Sawyer suggests, with what we know about the interpreting process? For, at present, there seems to be a very wide disconnect between interpreting, on the one hand, and testing, on the other. Anyone who turns to Interpreting Studies for help in constructing an interpreting test with reliable measures and valid interpretations will find only half an answer⁴. Interpreting Studies can shed light on the world and mind of the interpreter, but it cannot yet inform our understanding of measurement principles or demonstrate the ways in which those principles may be applied. Similarly, a foray into the literature on testing available in other areas – such as psychology, education, and the professions that use *credentialing** – will only provide a partial solution. These fields can tell us a great deal about test theory and testing practices, but they can contribute little to our understanding of the act of interpreting and the context in which it is performed.

credentialing tests in the health care professions and in teaching that estimate levels of reliability and validity.

⁴ Some might argue that research on Translation Quality Assessment (TQA) should be used to orient the development of interpreter tests. We disagree with this viewpoint. As *Section 2.1* will demonstrate, TQA has been characterized by a concern for general quality in interpreting and translation, rather than by a concern for the sound development and maintenance of test instruments. As a result, the TQA research is able to offer a great deal of insight into the features of a “good” translation and a “good” interpretation according to a variety of schools of thought, but it cannot offer us insight into the issues that Sawyer raises.

1.1 Research Problem and Design of the Study

What the foregoing suggests is that there are two important *lacunae* that must be addressed. First, because there is a paucity of research on the psychometric properties of actual interpreting test instruments, there is a need for studies that determine whether existing tests are compliant with generally accepted testing standards. Second, if it should be revealed that Sawyer's claim is accurate, and that testing practices do need to be improved, there is a need for work that demonstrates how knowledge from Interpreting Studies may be combined with an understanding of the principles of measurement and evaluation to create technically sound interpreter certification test instruments.

We set out to conduct the research in the present study as an attempt to fill these gaps. To address the first need, and to gather concrete evidence of the current state of affairs in interpreter certification, we made a preliminary examination of three existing tests that are currently used to certify professional interpreters. Early information available on the tests suggested that they had not been prepared in accordance with accepted testing standards, which in turn cast doubt on the reliability of their measures, and on the validity of their interpretations. Consequently, we also decided to address the second need by attempting to use research from the discipline of Interpreting Studies to construct a test instrument, and to demonstrate that its psychometric properties were largely acceptable.

1.1.1 Scope of the Study

However, looking at all aspects of reliability and validity connected to a variety of test instruments would clearly have been a monumental task, and one that would have exceeded the limits of a single study. To narrow the scope of the research, we restricted its focus in two ways.

First, in the examination of existing certification tests, most of our attention was placed on one particular test. Although it might initially have seemed more productive to conduct a wider survey of actual instruments, the fact remains that wide-scale interpreter certification tests worldwide are relatively few in number⁵. It would therefore have been extremely difficult to identify and obtain data on the number of tests required for an empirical survey⁶. Also, the kind of case study conducted here is quite common in the testing literature in other disciplines. It appears that the practice in the health care professions and in education – areas that have a tradition of conducting psychometric research – is to publish validation studies of single credentialing tests. (For examples, see Violato, Salami &

⁵ Our repeated literature reviews and on-line searches revealed the existence of only a handful of national-scale interpreter certification tests: the US Federal Court Interpreter Certification Examination (FCICE) cited above; the National Judiciary Interpreter and Translator Certification exam offered by the National Association of Judiciary Interpreters and Translators (NAJIT); the test which is the focus of the present study; the Canadian Evaluation System administered by the Association of Visual Language Interpreters of Canada (AVLIC); the National Testing System offered by the US Registry of Interpreters for the Deaf (RID); the National Interpreter Certification formerly offered by the US National Association for the Deaf (NAD); and the interpreting tests run by Australia's National Accreditation Authority for Translators and Interpreters (NAATI).

⁶ By empirical, we mean a survey based on a random sample large enough to allow inferences about the entire population of tests. While some may argue that the present study could have examined more than one interpreter certification test, strictly speaking, unless our study was based on a statistically sound sample, we would still not be in a position to legitimately make inferences about the general state of interpreter testing based on our findings. Gathering such a sample appeared to us not only to be a monumental task, but likely an impossible one as well: it would have involved collecting and analyzing data on perhaps as many as 30 different tests, and we had no reason to believe that there are that many national-level interpreter certification tests currently in existence. Since an empirical analysis was out of the question, we chose instead to conduct a

Muiznieks, 2002; Price & Wilkins, 2001; Cobb, Shaw, Millard & Bomotti, 1999; Daniel & Siders, 1994; Norcini, Hancock, Webster, Grosso & Shea, 1988; and McGhan, Stimmel, Gilman & Segal, 1982. All are reviewed in *Section 2.2.*)

The test that was consequently chosen for the study is one used by a professional association (hereafter referred to as “Association 1”) to certify *conference interpreters**. This particular test was selected for a number of reasons. To begin with, data from several administrations of the test were readily available. In addition, information on the development of the test – which traces its roots to research conducted in the field of machine translation – was well documented. Finally, the test provided us with the opportunity to examine conference interpreting, which had been previously ignored by published scholarly research on the topic of certification⁷.

As a second means of limiting the study, we chose to conduct a validation study, and to focus specifically on the test *constructs**, that is, on the concepts or characteristics purportedly measured by the test. This decision was made because of the central place that *construct validity** occupies in testing standards. Traditionally, it has been commonplace to conduct a validation study in part by showing that evidence from a test supports beliefs about the test constructs. Also, in their most recent version of the *Standards for Educational and Psychological Testing*, the American Educational Research Council, the American Psychological Association, and the National Council on Measurement in Education (AERA,

case study of a single test. This choice had the added benefit of allowing us to analyze the test in question in great detail.

APA & NCME, 1999) argue that all scores are measures of some construct, and that all validation is, in effect, construct validation.

In addition, concentrating on test constructs seemed the most obvious way of combining knowledge from Interpreting Studies with the principles of measurement and evaluation. Since the early 1960s, researchers in Interpreting Studies have been concerned with the construction of theoretical models to explain and describe the interpreting process, and it seemed logical to use these models to define the psychological constructs that are the foundation of interpreting, and which should in turn be used to build the foundation of a certification test. We paid particular attention to a recent discursive model (Setton, 1999), because it is arguably one of the most comprehensive in the discipline (Pöchhacker & Shlesinger, 2002), combining aspects of previous schools of thought with concerns about *discourse**.

1.1.2 Components of the Study

While it was straightforward enough to determine a means of restricting the scope of our research, it was not immediately apparent how we should shape the study. To make this decision, we returned to the *Standards for Educational and Psychological Testing*, and we made note of two standards that seemed to be of particular relevance.

⁷ Arjona (1983) and Schweda-Nicholson (1986) both looked at court interpreting. Roberts (1995 and 1998) examined community interpreting.

Standard 1.1

A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation (AERA, APA & NCME, 1999, p. 17).

This standard is the first to be presented in the volume, and it expresses what should perhaps be considered the most basic expectation in testing of any kind – that those who construct a test (the test developers) and those who are responsible for its use (the test users) should evaluate it in a formal manner. According to the *Standards*, this entails making a statement about the way scores from the test will be used (for example, to certify that candidates have the skills necessary for practice in a given profession), and about the way they will be interpreted (which scores will indicate candidates have the requisite skills, and which scores indicate they do not). Conducting a formal evaluation of a test also involves supporting the test rationale in two ways: with a logical analysis of applicable theory, and with empirical evidence.

Standard 1.8

If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided (AERA, APA & NCME, 1999, p. 19).

This second standard has been included here because it describes situations where theoretical models are chosen to derive test constructs, and it suggests that test users and developers should explain these choices and provide evidence to support them. This may entail citing theoretical evidence to explain the model's relevance to the testing situation, or it may involve providing empirical evidence that demonstrates that the constructs behave in a manner predicted by the theory.

After a careful reading of these standards, it was decided that the study presented here should attempt to meet both. In other words, in examining the existing test used by Association 1, it should strive to uncover theoretical and empirical evidence that supported the test rationale. In using a cognitive model to derive the constructs on which to base a test, the study should endeavour to present theoretical and empirical evidence that confirms the applicability and predictions of the model.

With the scope limited to an examination of test constructs and the analysis of one existing test, and with the shape of the study defined by the standards cited above, it was possible to put forward a more concrete research design. It was decided that the study should have three components, each of which is briefly described below.

1. Analysis of an Existing Test

We decided to examine the certification test used by Association 1, which is based on three constructs first introduced in machine translation:

intelligibility, informativeness, and style. We undertook to validate the test,

by examining theoretical and empirical evidence related to the test constructs.

2. Construction and Analysis of a New Test

We set out to develop a new test for conference interpreters, which was inspired by a discursive model of the interpreting process (Setton, 1999), and which is based on three constructs: linguistic parsing, the assignment of reference, and the interpretation of inference. We undertook to validate the test, by examining theoretical and empirical evidence related to the test constructs.

3. Comparison of the Two Analyses

We decided to compare the validity evidence associated with the two tests. It was hoped that the comparison would show if the use of one set of constructs in connection with conference interpreting was more appropriate than the use of the other.

1.2 Research Hypotheses

To further develop the three components of the study, we constructed a series of research hypotheses. A general hypothesis describes our expectations with regard to the comparison of the analyses of the two certification tests, while a series of specific hypotheses describes our expectation with regard to each of the tests separately. All the hypotheses are listed below.

1.2.1 General Research Hypothesis

There will be greater evidence of validity of score interpretations in an interpreter certification test when the test constructs are based on a discursive model of the interpretation process than when they are based on research in machine translation.

We therefore expect to be able to make a stronger validity argument about the newly developed test, where the constructs are derived from a discursive model, than about the existing test, where the constructs are derived from a model of machine translation. We feel that this is likely because the machine-translation constructs were not designed with the realities of conference interpreting in mind, while the discursive constructs come from a model of the interpreting process that combines elements of other established models in a novel way.

1.2.2 Specific Research Hypotheses

The evidence referred to in the general hypothesis will come from four sources: 1) the origins of the test constructs, which may indicate their relatedness to the profession of conference interpreting; 2) variance in the test scores, other than that accounted for by the test constructs, which may provide an estimation of construct-irrelevant variance; 3) an initial measure of construct variance, which may show that test scores reflect beliefs about the constructs; and 4) a secondary measure of construct variance, which may confirm our findings. With these sources of evidence in mind, we were able to formulate four hypotheses specific to both the existing test and the newly developed test. They are listed below.

Hypotheses Regarding the Existing Test

1. *An examination of the origins of the constructs of the existing test will reveal a weak link to the interpreting profession;*
2. *An examination of construct-irrelevant variance will suggest that factors other than the constructs may explain patterns in the test scores;*
3. *An examination of construct variance will suggest that the constructs do not adequately explain patterns in the test scores; and*
4. *Secondary estimation of construct variance will also suggest that the constructs do not adequately explain patterns in the test scores.*

Hypotheses Regarding the New Test

1. *An examination of the origins of the constructs of the new test will reveal a strong link to the interpreting profession;*
2. *An examination of construct-irrelevant variance will suggest that other factors do not explain patterns in the test scores;*
3. *An examination of construct variance will suggest that the constructs adequately explain patterns in the test scores; and*
4. *Secondary estimation of construct variance will also suggest that the constructs adequately explain patterns in the test scores.*

In general, we expect to find that the evidence related to the new test will make a more convincing validity argument than the evidence related to the existing test. This in turn will support the general research hypothesis.

1.3 Conclusion

The pages that follow outline our attempts to test these hypotheses in the present study. Yet readers may note that we have been obliged to deviate somewhat from the expected format. In a traditional dissertation, for example, the literature review traces the development of an important theory or principle. It begins by examining the work of the original proponent (Pyrzczak & Bruce, 2003); it outlines the reactions of subsequent researchers; and it leads to the conception of a study that offers a novel contribution to the field in question. The literature review therefore provides a historical context: it shows how the dissertation is the logical extension of chain of thought on a given topic (Pyrzczak & Bruce, 2003) and how science in general evolves in an incremental and cumulative manner (Rudestam & Newton, 2001).

In the present study, it was not possible to structure the literature review in this fashion. We do discuss three currents of research – the study of *translation quality assessment** (TQA), research on professional credentialing tests, and a series of theoretical models of the interpreting process – yet we were not able to position the present study as a direct contribution to any of them, for a number of reasons. With regard to TQA, we do not share the same aims as the TQA research, nor do we use the same methods. With regard to

the credentialing literature, we cannot reasonably expect to wade into the debates we found there – which seem often to revolve around technical psychometric issues – given the state of measurement research in our own discipline. Finally, with regard to theoretical modelling, we do not offer a new account of the interpreting process, nor can we offer any direct insight into an existing one.

Instead, we reference this literature because, simply put, there is little else that can serve as a foundation for the present study. Scholarly attention to interpreter certification testing is rare, and empirical, psychometric research on the topic is virtually non-existent. In the absence of literature with a direct bearing on the present study, we have elected to cite peripheral literature (Pyrzack & Bruce, 2003).

Because we are relying on peripheral literature to provide us with a theoretical and methodological foundation, the present study is organized in the following manner. *Chapter 2* presents the TQA research and the credentialing literature, and it outlines our reasons for choosing to loosely base the methodology of the present study on the latter. In *Chapter 3*, we explain our conception of *competency**, and we review a series of models of interpreting as a means for identifying the competencies we targeted in our newly developed test. *Chapter 4* presents the validation methodology that was used to conduct the present study, and *Chapter 5* lists the results we obtained in the validation study. In *Chapter 6*, we explain our interpretation of the results, before making a series of comments about possible ways to orient future research in *Chapter 7*.

We believe that the potential scientific contribution of this research is threefold. First, by investigating a discursive model (Setton, 1999) of the interpreting process in an empirical manner, the present study may offer evidence to validate the model. Second, it may encourage a new trend of measurement-based research in Translation Studies and Interpreting Studies. Third, it may also suggest a means of measuring communicative competencies in context (as opposed to decontextualized linguistic competencies), thereby enriching the field of measurement and evaluation.

In terms of its social contribution, the present study may allow professional associations to improve their efforts in the development of certification exams. If the new test instrument demonstrates greater evidence of construct validity, it would suggest that there is merit in conducting more research on the use of discursive constructs in interpreter certification in the future. This future research may allow associations to better define and test for discursive competencies. It may also allow them to better defend themselves against criticism and litigation.

2. An Examination of Potential Research Models

On the surface, it might appear to some that a study of interpreter certification would logically be an extension of Translation Quality Assessment (TQA), a trend of research conducted to a certain extent in Interpreting Studies, and to a much wider degree in Translation Studies. After all, one of the goals of TQA is to devise methods for evaluating translations and interpretations by defining the features that determine quality. However, the present study was designed to address a need for measurement-based research in interpreter certification, a need that has been expressed in the profession (Sawyer, 2000) and in the courtroom (*Seltzer v. Foley*, 1980).

The question that must therefore be asked at this point is the following: is it possible to combine psychometric principles with the concerns that have normally characterized TQA? This chapter attempts to answer this question by taking the four following steps. First, we review the seminal research on TQA, being sure to include a number of articles that addressed the issue of interpreting quality. Second, we examine a sample of articles on credentialing tests in other professions as a means of determining how measurement and evaluation principles typically “translate” into applied research. Third, we identify the similarities and differences between the two research trends in order to determine their compatibility and their pertinence for the present study. Finally, we turn to a short series of articles within Interpreting Studies that were also useful in providing some context.

2.1 Translation Quality Assessment Literature

There are a number of thorough reviews of the TQA literature (see Waddington, 1999; Lauscher, 2000; House, 2001a; and Williams, 2001a), so the discussion here is not extensive. Instead, it merely points out some general tendencies in the research. One of these is the tendency to distinguish between quantitative TQA models, which provide numerical scores in order to give a general summary of the quality of a given translation or interpretation, and qualitative models, which are used to make a holistic statement about translation quality. We have used this distinction to shape our review of the research in this area.

2.1.1 Quantitative Models of TQA

A prime example of a quantitative TQA model is the Canadian Translation Bureau's *Système canadien d'appréciation de la qualité linguistique* (Sical). The Sical model was originally developed for use as an examination tool within the Bureau, and as a means of assessing the quality of the Bureau's translations (Williams, 2001a, p. 330). It was first implemented by the Bureau in 1976 (Larose, 1998, p. 166), and it underwent several revisions, before being retired in 1994¹ (Mossop, 2003).

¹ Williams (2001b) notes that the Translation Bureau introduced a "total quality" approach to TQA in 1994. From that point onward, no translation that contained any errors was delivered to a client. Around the same time, the Bureau stopped using Sical, or any other centralized grading system. Free-lance translations are now rated as "acceptable" or "unacceptable" by senior translators in individual sections of the Bureau, using a variety of criteria (Mossop, 2003).

The version of the system that is described in most detail in the literature (Larose, 1998; Williams, 1989; Williams, 2001a; and Williams 2001b) is Sical III, which was introduced in 1986. Sical III was based on a list of four types of errors: minor linguistic errors, major linguistic errors, minor translation errors, and major translation errors. A translated text was rated as superior, fully acceptable, revisable, or unacceptable, depending on the number of major and minor errors contained in a 400-word passage (Williams, 2001a, p. 330).

The first two ratings (superior and fully acceptable) were considered satisfactory; a translation of this quality was, for the most part², accepted as a passing score in an examination situation, or as an indication that a contractor's work met the Bureau's standards. The final two ratings (revisable and unacceptable) were considered unsatisfactory; a translation of this quality was considered a failing score in an examination situation, or an indication that a contractor's work was substandard (Williams, 2001b, p. 23). The distinction between satisfactory and unsatisfactory translation was made on the number of major errors found in the 400-word sample. Superior and fully acceptable translations contained no major errors, revisable translations contained one major error, and unacceptable translations contained more than one major error (Williams, 2001b, p. 23).

The concept of major error was pivotal in assessing a translation, and Sical III consequently defined major error in a manner that was more precise than those of earlier versions (Williams, 2001a, p. 330). A major translation error was described as a "complete

² Williams (2001b) indicates that individual operations managers and hiring managers had some discretion in

failure to render the meaning of a word or passage that contains an essential element of the message” or as a “mistranslation resulting in a contradiction of or significant departure from the meaning of an essential element of the message” (Williams, 1989, p. 26). A major linguistic error was described as an instance of “incomprehensible, grossly incorrect language or rudimentary error in an essential part of the text” (Williams, 1989, p. 26). The notions of “essential element” and “essential part” were apparently not defined, but Williams suggests that “essentialness” was determined by considering the consequences of the error for the client (Williams, 2001a, p. 330).

On the whole, Sical is noteworthy for two reasons. First, it was highly influential. It was adopted or adapted by a number of translation organizations outside the federal government (Williams, 2001a, p. 331). Second, it offers a clear illustration of the quantitative approach to quality assessment. Of all the TQA models, Sical offers what is arguably the most straightforward means of quantifying its definition of error.

A more complex approach to quantification can be seen in Gouadec (1989). The author examines the widespread use of hierarchical lists of errors to evaluate translations – lists that he considers vestiges of traditional *thème* and *version* exercises for language learning – and points out that such evaluations have three drawbacks: 1) they do not define error objectively; 2) they deduct points for errors in an arbitrary fashion; and 3) they prevent a better understanding of (student) translator behaviour.

setting their requirements.

To combat the first drawback, Gouadec proposes a functional definition of error. He suggests that an error, in translation, is any unjustifiable distortion between the source text and the target text, and he further elaborates on two specific types of errors. *Absolute errors* result from a failure to respect the general rules of communication. These rules may govern “cultural grammar” (logic, concept formation, concept organization), “linguistic grammar” (syntax, orthography, etc.), or the “use of general or specific stereotypes” (text type, terminology, etc.). *Relative errors* result from the failure to communicate or respect aspects of a given *translation project*. These aspects include such considerations as the target audience, the context of use of the translation, its subject matter, its themes, etc. (p. 38-39).

To combat the second drawback, he proposes a new method for weighting translation errors. The new method still makes use of the familiar hierarchical lists, but it multiplies the penalties for each type of error by a series of coefficients that are assigned to other features also worth considering. For example, Gouadec explains that the final weight of an error is the product of an initial value, multiplied by 1) a coefficient representing the extent of the error’s effect; 2) a coefficient representing the role of the error in a segment of text; and 3) a coefficient representing the level of the segment (sentence, paragraph or entire text) affected by the error. Using this system, the translator trainer is able to come up with a vast range of penalty values for errors in the translated text (p. 45).

Gouadec combats the third drawback by asking translator trainers to implement measures to achieve three goals. First, trainers should try to eliminate the risk of translator error by offering their trainees strategies to help them better understand source texts and to

help them better conduct documentation research. Second, trainers should carefully analyze the behaviour of their trainees. Third, trainers should attempt to limit the effect of trainee errors and their corresponding *cone of distortion* (p. 51).

The model proposed by Gouadec traces a complicated path between error and the scoring of error. It seems like no easy feat to follow the author's directions and arrive at a numerical score for a translation, but all of the ideas the author puts forward are clearly intended to shed light on student learning.

Strong & Rudser (1985) develop another quantitative model, one intended for use in assessing *sign language** interpreting. The authors suggest that there is an added level of complication in sign language interpreting that is not present in spoken language interpreting; sign language interpreters may be called upon to use *ASL**, *Signed English** or a form of communication somewhere in between³.

To conduct their research, the authors use data from an existing study (Quigley, Brasel & Montanelli, 1973), in which 15 male and 15 female "interpreters" with varying degrees of ability were asked to carry out three tasks. The first task required the participants to interpret a lecture from English to a form of Signed English. The second involved

³ As Roberts (1987) points out, sign language interpreting generally involves a change in mode of communication, but does not always involve a change in language (p. 295). For example, a sign language interpreter may interpret a spoken English source message into Signed English (a change in mode, but not language) or into ASL (a change in both mode and language). The form of signed communication used in an interpreted encounter depends on a number of different factors, such as the language in which the Deaf consumer is most comfortable, the language favoured by the interpreter, the degree of formality of the encounter, and the complexity of the material to be interpreted (Strong & Rudser, 1985, p. 345).

interpreting a lecture and a story into ASL. The third and final task had participants interpret stories told in ASL and in Signed English into spoken English.

The authors then selected four raters from the pool of potential subjects (excluding those that had participated as subjects), and had them evaluate the data. The evaluation took place in two stages: first, raters divided the signed interpretations into their component propositions; and second, they answered a series of five questions for each proposition. The first asks the evaluator to judge the accuracy of the interpreter's proposition as accurate, inaccurate, modified or missing. The second asks the evaluator to identify the number of errors – defined as additions, omissions, substitutions, or problems with articulation – in the proposition. (It is at this point that a score is calculated; each proposition is initially worth 10 points, and a point is deducted for each error.) The third question asks if the interpreter has made any cultural adjustments. The fourth requires the evaluators to identify the “target language” (i.e., ASL, Signed English, or spoken English⁴). In the last question, evaluators must indicate the number of *fingerspelled** items in the proposition.

Strong & Rudser also make an attempt to evaluate *rater reliability** and *criterion validity** by comparing scores the interpreters obtained under their system with those arrived at in the study by Quigley, Brasel & Montanelli. This attempt to study reliability and validity quantitatively makes the authors stand out among their peers in both Translation Studies and Interpreting Studies.

⁴ While the target language of the spoken/signed interpretation might seem self-evident, raters were presumably asked to indicate the target language on the scoring sheets so that the authors could categorize the ratings without referring back to the recordings of the interpreters' performances.

2.1.2 Qualitative Models of TQA

One of the earliest qualitative models of translation assessment is the one advanced by House, which was first published in 1977, but revisited in 1997. House's model is an application of "pragmatics"⁵ to translation criticism. It allows the evaluator of a translation to do three things: 1) describe the linguistic and pragmatic features of the source text; 2) compare the source text and the target text; and 3) assess the quality of the translation by indicating how well the two texts match.

To conduct the assessment, evaluators must address two dimensions of concern. First, they must consider the dimensions of the language user, which consist of the regional dialect and social class dialect used in the text, and the features that mark the text as belonging to a particular time. Second, they must consider the dimensions of language use, such as the medium (speech or writing), participation (monologue or dialogue), social role relationship (between addresser and addressee), social attitude (the register used in the text), and province (occupational or professional activities – for instance, science or advertising – that may be at issue in the text). Using the model, evaluators will assess a translation positively if it preserves the semantic, pragmatic, and textual meaning of the original.

⁵ House explains that "research inside pragmatics deals with the rules and principles governing people's use of language, it tries to make statements about what makes some uses of language more appropriate than others in certain situations" (House, 2001b, p. 58). Her definition of *pragmatics* is much wider than the (Gricean) definition we have used in the present study. For more details see *Section 3.3*.

Like many other proponents of TQA, House devotes considerable attention to the definition of error. She distinguishes between two error types. "Overtly erroneous errors" are the result of a mismatch between the denotative meanings of source-text and target-text elements (such as additions, omissions, and substitutions), or of a breach of the target-language system (ungrammaticality, dubious acceptability). "Covertly erroneous errors" are the result of a mismatch along any of the language user or language use dimensions outlined above (House, 1997, p. 45).

House's model allows its users to make a holistic statement about the quality of a translation, a statement which is largely based on the notion of equivalence. In her eyes, judging a translation is largely a matter of observing the degree of correspondence between the source and target texts.

In stark contrast with House's approach are the ones adopted by Kussmaul (1995) and Nord (1996). Both authors dispense with source- and target-text comparisons, and instead base their views of the assessment process on Reiss & Vermeer's (1984) *Skopostheorie*. The theory argues that a translation should not be defined by its relation to the source text, but rather by its purpose (or *skopos*). Assessing a translation, according to the theory, becomes a matter of evaluating how well it conforms to the way it will be used in the target culture. The theory is common to both Kussmaul's and Nord's models of assessment, although there are differences in how the two authors apply it.

Kussmaul begins his discussion of translation assessment by introducing two views of translation teaching. The first has its origins in foreign-language instruction, and it sees the student translator as a learner of a second language. It therefore focuses on the student's knowledge of language, as evidenced by his or her ability to handle small linguistic units, such as words and phrases (1995, p. 128). The second view, Kussmaul maintains, is more consistent with the realities of professional translation. It sees the student translator as someone who must communicate an entire text to a recipient, for a particular purpose. It consequently focuses on larger "pragmatic"⁶ concerns, such as cultural, situational and text-typological adequacy (p. 128).

Because the professional translator's view tends to focus attention on larger, pragmatic units, while the foreign-language instructor's view is often concentrated on smaller, word-related ones, Kussmaul points out that his readers may assume that pragmatic errors should be penalized more severely than word errors (1995, p. 130). Yet he cautions against this assumption. He argues that to adopt a more appropriate approach to translation assessment, teachers should stop to consider the effect a particular error has on the reader of the target text, and on the text's use in the target culture. In some cases, a simple error in word meaning, Kussmaul explains, can distort an entire text.

To illustrate this approach, he examines five types of error – including examples of both the pragmatic and word-based varieties – and shows how each one can have far

⁶ Kussmaul defines *pragmatics* as "the study of the relationship between an utterance or text and its user, i.e. its speaker/writer and its hearer/reader. Language users do not exist in a vacuum, but operate in specific situations" (1995, p. 55). His use of the term roughly corresponds to our definition of *discourse* (see Section 3.3).

reaching effects. Discussing *cultural adequacy*, the author gives the example of an English novel translated into German where the place names and the names of the characters are translated. This violates a general tendency in German literature to preserve foreign settings and creates a text that lacks coherence. To illustrate *situational adequacy*, he underlines instances in the target text where the translator has left out sentences that were part of the original dialogue. If included, the omitted passages would have allowed the translator to more effectively display the social role relationship between characters in the novel, thereby helping the readers to better understand them. In his explanation of *speech acts*⁷, he explains how omitted lines of dialogue also deprive the translator of the opportunity to use illocution to give readers a better sense of the characters' personalities. The author points out how the *meaning of words* chosen in the target text can be problematic, particularly when word connotations impair the overall cohesiveness of the text. Kussmaul also underlines how *language errors* in the translation, such as faulty collocations, cause readers to assume the original author lacks linguistic competence (1995, p. 130-145).

In Kussmaul's discussion of translation error, no one type is regarded as serious in and of itself. Instead, the author maintains that a given error should be evaluated in terms of its impact on the translation's function in the target culture. In translation assessment, the key is to determine the effect that each error has on the text's overall *skopos*.

⁷ In a preceding chapter, Kussmaul demonstrates how certain speech acts are associated with particular text types in a language culture. He compares a small sample of leaflets written in German and English, and he notes that the former tend to use the passive voice as an "illocutionary force indicator" for an instruction, while the latter tend to use the imperative for the same effect. Consequently, a translator who encounters a sentence like "Balneum Hermal wird in das Badewasser gegeben (nicht umgekehrt) und wird untergemischt" should avoid using a passive construction ("Balneum Hermal is added to the bath water (and not the other way around) and stirred") and opt instead for a more conventional English sentence ("Add Balneum Hermal to the bath water and stir well until dissolved").

Like Kussmaul, Nord also begins her article with a redefinition of error – as any feature of the translation that is not in keeping with the purpose and function of the target language text. However, she parts company with Kussmaul in her description of error, which she divides into three types.

The first type, pragmatic error, occurs when there is a failure to respect the purpose of the translation as indicated in the pragmatic instructions of the task at hand (i.e., to transmit the relevant information to the target audience, recognizing that there may be differences in knowledge and expectations between the source and target cultures). Nord gives the example of a German pamphlet, translated into Spanish, which announces the celebration of the University of Heidelberg's 600th anniversary, and in which there are numerous references in the Spanish text to "Ruperto Carolo University"⁸. She suggests that this is evidence the translator has failed to recognize that target readers will not understand that this is a reference to the University of Heidelberg, and that the text will be incoherent for them as a result (1996, p. 98).

The second type, cultural error, arises when a text violates general tendencies in the language and in the text type for communicating information. In German, Nord argues, it is common to use lengthy compound words ("*palabras compuestas*") so readers of the German pamphlet have no adverse reaction to seeing references to a "*Rechnernetz zur intelligenten*

⁸ Nord explains to her readers that German universities are often given sobriquets constructed from the Latinized names of their founders. The University of Heidelberg was founded by Prince Ruprecht Karl, a fact which explains its designation as "Ruperto Carolo University".

Informationsverarbeitung” (a computer network for the intelligent processing of data).

However, a Spanish phrase of similar length would be unusually technical in what is ostensibly a brochure for tourists (1996, p. 98).

The third type, linguistic error, can be seen when the norms of grammar, lexical choice, spelling, and punctuation are broken. Nord argues that linguistic errors should rarely be seen in professional translation, but that they are frequently sighted in student work. However, she maintains that the incidence of linguistic errors can be reduced in the classroom by giving the students a realistic translation task and clearly defining it (1996, p. 99).

Discussing the order of severity of the three types of error, Nord makes a distinction between the professional world and the classroom. For professionals, the order of severity is fixed (pragmatic errors are the most serious, linguistic ones, the least). For students, things are more flexible. Teachers may decide to alter the severity of the errors to suit individual pedagogical exercises (e.g., teachers may view linguistic errors as more serious when they are teaching the norms of target-language grammar).

Nord’s approach to translation assessment, like Kussmaul’s, is clearly intended for use in the classroom. In keeping with this *skopos*, she argues that assessment should be more than a simple tally of errors – it should instead offer students the motivation to find solutions to the problems they encounter in their work.

Schjoldager (1995) also designed a quality assessment model for use in the classroom. The teacher of a course on interpreting, she has created a *checklist** of the undesirable features of a spoken interpretation that is the counterpart of the lists of errors the previous authors have discussed with regard to written translation.

The questions on the checklist are broken down into four categories that are designed to discover whether 1) the listener can understand and bear to listen to the interpreter; 2) the interpreter's use of language is appropriate; 3) the interpreter's rendition is coherent and plausible; and 4) the interpreter is a loyal communicator (i.e. he or she has not added comments or changed the producer's emphasis). The questions in the first three categories are intended to assess interpreting from the listener's perspective, while those in the fourth are designed to assess interpreting from the speaker's. Scholdager suggests that while the four categories are equally important in principle, it is pointless to fulfill the expectations of the fourth category if those of categories one through three are not met first.

A mechanism for attributing a score could easily be added to Scholdager's checklist because of its breakdown of discrete interpreting errors. However, she indicates that she uses the checklist primarily as a means of providing formative assessment, that is, as a way of giving students feedback to help them learn and improve.

One final TQA model we feel is noteworthy is proposed by Williams (2001a)⁹. The author presents a full-text, argumentation-centred approach that uses an analysis and comparison of source-text and target-text argument macrostructures as a means of assessing translation quality. Williams bases his model on an account of argument macrostructure offered by philosopher Stephen Toulmin, which comprises four essential elements.

1. Claims (C)

A claim is the conclusion of the argument, or the main point toward which all other elements of the argument converge (e.g., the recommendation that a government introduce a new health program).

2. Grounds (G)

A ground is a piece of information that supports the claim (e.g., the fact that there is overcrowding in emergency departments of hospitals).

3. Warrants (W)

A warrant is a statement that indicates how the grounds are connected to the claim (e.g., the need for rapid response in an emergency department setting links overcrowding to the recommendation).

⁹ Williams (2001a) presents one component of a much larger model outlined in its entirety in Williams (2001b). We elected to review the single component in the interest of brevity; our goal in this chapter is to trace a representative portrait of the TQA literature, and not to produce an exhaustive literature review.

4. Backings (B)

A backing is an overarching principle, law, or standard governing the issue at hand (e.g., the principle of the universality of health care governs the health program).

Toulmin further identifies two elements that may also be part of argument macrostructure.

1. Qualifiers (Q)

A qualifier is a statement or phrase that enhances or mitigates the force of the claim (e.g., a government official may announce that the new health program may definitely, certainly, probably, or possibly be introduced).

2. Rebuttals (R)

A rebuttal is a statement of extraordinary or exceptional circumstances that is introduced for the sake of caution or modesty (e.g., an official may state that the program will be introduced “unless the government’s fiscal situation worsens”) (Williams, 2001a, p. 337-338).

Williams uses these elements to form a framework for his TQA model. To evaluate a translation, one of the evaluator’s primary tasks will be to determine whether the basic argumentative elements (B, W, G, C, Q, R) are accurately rendered in the target text if they are present in the source text. Williams suggests that this argumentation-based approach to quality assessment provides a means for clearly defining major error, something that was lacking in other models (see the review of Sical in *Section 2.1.1*). Failure to render one of the

argumentative elements presented above constitutes a “major/critical defect”. Conversely, if a translator successfully reproduces the argument macrostructure, he or she will have conveyed the “core message of the text” (Williams, 2001a, p. 342).

On the basis of this short review, it is possible to underline several points about the TQA literature. First, all the models examined above are concerned with the identification of errors. All of them require evaluators to examine a translation or interpretation in a fashion that is highly analytical – at times even linear – and to look for different types of discrete flaws in the work. The definition of error may vary from model to model, but the concept of error itself is always key.

Second, the models are also concerned with the identification of quality. However, since each one is intended for use in a specific context, the definition of quality varies from model to model. For instance, Sical was used with instrumental (non-literary) translations before they were delivered to clients of the Translation Bureau, and quality in this instance was freedom from linguistic and translation errors. Gouadec’s model helps instructors to identify the sources of student error and to enhance learning; its definition of quality is 1) freedom from distortions in grammar, rhetoric, and textual organization; and 2) the successful achievement of the *translation project*. House’s equivalence-based system can be applied with a variety of text types – from literary translations to the translation of advertising and administrative texts – and it argues that quality is an adequate match in the linguistic and pragmatic features of the source and target texts. Nord’s *Skopos*-inspired

model is intended for use inside and outside the classroom, and it suggests that quality is consistency with the purpose and function of the target-culture text.

Third, as a result of the preceding concerns about error and quality, the object under study in TQA is always the translation or interpretation, and never the person behind it. In the models examined above, error is always pivotal. As a result, evaluators are instructed to look for the absence of a particular feature (what the feature is depends on the model in question) in an inanimate object. None of the TQA literature we examined considered an alternative approach – searching for evidence of a human attribute. The discussion of quality is similar. All of the TQA models discussed quality as it related to things, and not to people. In some of the pedagogically oriented models (Gouadec's for example), the goal is to eventually understand and influence student learning, but that learning is not assessed directly, passing instead through the intermediary of error. The focus is not on the producer, but rather on the product.

The concern with error, quality, and product has allowed TQA to address a number of issues that are central in Translation Studies and Interpreting Studies. For instance, TQA has been used to define the features of “good translation” (House, 2001a; Waddington, 2001; and Williams, 2001a), to establish a form of translation criticism (House, 2001a; and Martínez Melis & Hurtado Albir, 2001), and to argue for a particular theory of translation (Waddington, 2001). However, as universal as the concepts of error, quality, and product appear to be in TQA, they are not the only ones that can orient research on evaluation.

2.2 Credentialing Literature

As a basis for comparison, we now turn our attention to a series of articles that illustrate how research on credentialing tests in other professions is typically conducted. A great deal of the available literature examines such testing in the health professions – perhaps because the consequences of improper testing in this area pose such a high potential risk to the public – and we have tried to preserve that focus here. However, we have also attempted to consult research on credentialing in a number of other professions in an effort to show that the concerns of the research are surprisingly uniform, despite the fact that it is carried out in several different disciplines.

As was the case in our discussion of the TQA literature, our aim is not to be exhaustive. If we were to write a thorough review of the credentialing literature – detailing the development of measurement-related concepts and methods from study to study, and positioning the present study at the end of the review – it might appear that we are attempting to make a direct contribution to the further development of those concepts and methods. However, given the state of psychometric research in Interpreting Studies, it would be unreasonable to assume that any study of interpreting could make such a contribution at the present time. Within Interpreting Studies, the current lack of familiarity with measurement theory and practice is simply too great to advance either theoretical or practical knowledge of psychometrics in a meaningful way.

Instead, the following sample of credentialing research has been included here as a means of hopefully overcoming our subdiscipline's apparent lack of familiarity with evaluation and measurement. By reviewing a number of studies of *licensure** and certification tests, we hope to gain a sense of what the hallmarks of those studies are, and to understand how they might be reproduced in current and future work. In other words, we have included a review of the credentialing literature at this juncture – and kept it brief – because we are considering it as a model to guide similar research in Interpreting Studies.

2.2.1 Research in the Health Care Professions

The article by Violato, Salami & Muiznieks (2002) presents a typical example of a certification exam and discusses the exam from a psychometric perspective. The authors focus their attention on the Alberta Registered Massage Therapists Society (ARMTS) certification examination, which tests candidates in three domains: professional knowledge; clinical judgement; and clinical skills. The first domain and second domains are evaluated with written, multiple-choice tests (a 140-item test for professional knowledge, and a 60-item test for clinical judgement). The third domain is evaluated with a performance test conducted on three clients in three sessions. Each session is devoted to one of the following: taking a case history; assessing a physical condition; and treating a physical condition.

In addition to explaining the make up of the exam, the authors also give a brief outline of its development. For example, the authors note that the exam was *pilot tested**, and that the performance test of clinical skills was scored with a detailed checklist. They also

mention that the raters were thoroughly trained in the use of the checklist before being asked to score actual candidates.

However, the majority of the article is devoted to a discussion of the exam's psychometric properties, and to the statistical tests that were used to determine them. To investigate *internal reliability**, the authors examined the scores in each of the exam's five components (two written tests and the three performance test sessions), and calculated a reliability coefficient for each, using *Cronbach's alpha**¹⁰. Values for the coefficients ranged from $\alpha = 0.60$ to $\alpha = 0.88$ (the highest coefficient values were obtained from the two written components). The authors consider that all the values are in the "adequate" to "good" range, and they conclude that this is satisfactory evidence of internal consistency within each component.

To examine criterion validity, the authors set up pairwise comparisons between each of the five test components (i.e., the five components were compared two at a time), and they use *Pearson's product-moment correlation**¹¹ to generate correlations for each pair. The test components correlate significantly ($p < 0.05$), which leads the authors to conclude that there

¹⁰ *Cronbach's alpha* measures the internal consistency of a test (Everitt & Wykes, 1999). It shows the degree to which scores for a series of test items are similar. Values for alpha (represented by the symbol α) typically range between 0 and +1, although negative values are also possible (George & Mallery, 2001). Cronbach's alpha is often used to provide some preliminary indication of unidimensionality in the test data (Cortina, 1993). The higher the value of α (the closer it is to +1), the more likely it is that variation in the data is due to a single underlying factor.

¹¹ Pearson's product-moment correlation measures the degree to which two continuous variables are related (George & Mallery, 2001, p. 359). Values for Pearson's (designated by a lower-case r) typically range between -1 and +1. A negative value indicates that as one variable increases, the other tends to decrease; the closer the value is to -1, the stronger is that tendency. A value of 0 indicates that there is no relationship between the variables. A positive value indicates that as one variable increases, the other also tends to increase. The closer the value is to +1, the stronger is that tendency (George & Mallery, 2001, p. 112-113).

is satisfactory evidence of validity. They further report that the significant correlations range from $r = 0.24$ to $r = 0.78$.

To examine construct validity, the authors used *factor analysis*^{*12} to extract a number of components from the data. They report that there were three *eigenvalues*^{*13} greater than 1, and that the corresponding components accounted for 70% of the total variance. These three factors roughly matched the three constructs around which the exam is based (professional knowledge, clinical judgement, and clinical skills). However, the factor analysis does suggest a minor reworking of the constructs is required to make them more accurate. (The authors suggest that the new constructs be named information processing, clinical treatment, and follow-up management)

The article by Norcini, Hancock, Webster, Grosso & Shea (1988) also provides an example of how researchers have examined the psychometric properties of certification exams. At the time the article was written, many credentialing tests were norm referenced (i.e., the tests characterize a candidate's performance relative to that of other test takers), and the merits of criterion-referenced tests (i.e. tests that characterize a candidate's performance

¹² The term *factor analysis* refers to a series of techniques that are used to show whether a series of variables can be explained by a smaller number of constructs, known as *factors* (Norman & Streiner, 1997, p. 142). There are two main types of factor analysis. Exploratory factor analysis is used to look for a structure or pattern in a set of data, while confirmatory factor analysis is used to determine whether a specific, predefined model is compatible with the data (Everitt, 1996, p. 225).

¹³ An *eigenvalue* is a measure of the variance accounted for by each factor (Norman & Streiner, 1997, p. 145). As a matter of convention (a convention known as the *Kaiser-Guttman criterion*) a factor with an eigenvalue less than 1.0 is not retained for consideration (Norman & Streiner, 1997, p. 145).

relative to some domain of behaviour or knowledge) were widely debated. The authors conducted a study to shed more light on the use of criterion-referenced tests.

The data they examined came from the 1983 Certifying Examination in Cardiovascular Diseases, a criterion-referenced test of physician behaviour. The test is made up of three components: a multiple-choice test, a series of patient management problems (PMPs), and a series of performance items that required the test takers to read and interpret electrocardiographs (ECGs). The 1825 test takers were divided into three groups according to their educational background and experience with the exam: 1) first-time test takers who studied in high-ranking cardiology programs; 2) first-time test takers who studied in other cardiology programs; and 3) test takers who had taken and failed the exam at least once previously.

The ECG test required candidates to read and interpret 46 different items, and the cut-off score for the test was norm referenced (i.e., it was set at one standard deviation below the average score). For the purposes of the study, the authors devised a second ECG test, by selecting a subset of 30 items (ECGs that illustrated a simple condition, and that were typical for that condition), and by setting the cut-off score in a criterion-referenced manner (i.e., by using the *Angoff method*¹⁴).

¹⁴ The *Angoff method* (Angoff, 1971) is a procedure used to set the cut-off score for credentialing tests. In the method, a panel of subject-matter experts examines each item on the test at hand and estimates the lowest probability that a minimally competent examinee would answer each one correctly. The probabilities for all the items on the test are summed. The average of the panel experts' probabilities is calculated to determine the cut-off score for the test (Lee, 1999, p. 12).

To examine the merits of their criterion-referenced ECG test, the authors conducted three *generalizability analyses**¹⁵: 1) on the *standard setting** procedure; 2) on the scores from the test, and 3) on the two sources of information together. Each analysis comprised a generalizability study (G study), which allowed the researchers to estimate dependability in the multi-faceted test, and a decision study (D study), which allowed them to determine how to alter the test design in order to obtain more reliable results. The three G studies suggested that dependable decisions were obtained (the dependability indices were .898, .876, and .871 respectively). The results of the D studies suggested that if improved dependability were required, it could be achieved by increasing the number of items (30), but not by increasing the number of judges in the standard setting procedure (11). On the whole, the results seemed to indicate that dependable pass/fail decisions can be made from tests that are relatively short – good news for proponents of performance testing.

Norcini et al. also looked at the psychometric properties of the criterion-referenced ECG test in other ways. It had always been assumed that the three groups of candidates would obtain different results on the exam – first-timers from the high ranking programs would score highest, first-timers from other programs would score in the middle, and those that had written the test before would score lowest. To test this assumption, the authors examined the means of the three groups on four components of the exam (the multiple-choice test, the PMPs, the norm-referenced ECG test, and the criterion-referenced ECG test). The results showed the expected distribution of scores among the three groups.

¹⁵ A *generalizability analysis* (Cronbach, Gleser, Nanda & Rajaratnam, 1972) is normally made up of two components. The first is the *generalizability study* (G study), which allows researchers to measure the multiple sources of error inherent in the design of a test. The second is the *decision study* (D study), which allows

The authors conducted two final analyses. First, they calculated pairwise correlations between scores for the four components (i.e., a series of pairs of correlations was calculated until all components had been compared with one another). The correlation coefficients ranged from $r = .29$ to $r = .82$. Last, they examined the internal reliability of scores for each component. The multiple-choice test demonstrated the highest reliability ($\alpha = .87$), while the PMPs showed the lowest ($\alpha = .44$).

McGhan, Stimmel, Gilman & Segal (1982) report on the psychometric properties of a certification exam that was used as part of a pilot project that investigated pharmacists' ability to prescribe and administer drugs. The exam was made up of 204 multiple-choice items, and it tested candidates' knowledge of three content areas: clinical therapeutics, physical assessment, and the legislation and regulations surrounding the pilot project.

To ensure *content validity**, the test developers concentrated on constructing items that adequately represented the three content domains. Within clinical therapeutics, 12 subdomains were identified, and roughly 10 items were developed for each one (for a total of 127 questions). The items were modelled on sample review questions that appeared in a number of publications on clinical therapeutics. With regard to physical assessment, the developers noted that a course on the topic identified six subdomains. They constructed roughly ten items for each one (for a total of 62 questions). The items were modelled on those found in sample reviews in a number of textbooks. For the legal domain, developers

researchers to determine how large a sample size needs to be to achieve a given level of reliability.

constructed one item for each of the 15 sections in the legislation that allowed the pilot project to take place (for a total of 15 questions). After all the items were constructed, item analysis (which involved the calculation of *item difficulty** indices and *item discrimination** indices) was used to identify questions that were ambiguous and that discriminated poorly. These were either discarded or improved.

The cut-off score for the exam was determined by examining information from a number of sources. First, the exam was administered to a group of 30 pharmacists, and the questions for the first two domains were administered to a group of 14 physicians (the doctors could not reasonably be expected to know the legal material). *Independent samples t-tests**¹⁶ on clinical therapeutics, physical assessment, and total test scores determined that there was a significant difference between the two groups in physical assessment, with the physicians scoring higher. Measures were taken to ensure that future performances of pharmacists would attain the level suggested by the physicians' scores. Second, the mean score for the pharmacists was calculated, along with the standard deviation for the group. One standard deviation below the mean was equivalent to 60% of the total possible score. This number was adopted as the cut-off score.

The internal reliability of the exam as a whole was estimated, along with internal reliability for each of the domains. The overall estimate was high (KR 20 = .89), as were the domain estimates: clinical therapeutics (KR 20 = .836); physical assessment (KR 20 = .883); and legislation (KR 20 = .840).

¹⁶ An independent samples t-test is used to compare the means of two samples that are unrelated (Norman &

The final step taken by the authors was to investigate criterion validity. They collected a random sample of over 300 hypertension prescriptions, 169 of which were written by two pharmacists who had passed the certification exam, and 157 of which were written by three physicians. They then asked one physician and one pharmacist to evaluate the appropriateness of the prescriptions (the identities and professions of the prescribers were not revealed), using an evaluation form made up of five subscales. The internal reliability of the form was high ($\alpha = .844$), and interrater agreement (.65) was moderate. Independent samples t-tests were used to determine if the physicians' and pharmacists' scores on the five subscales were significantly different. Results showed differences for two of the subscales, and for the total appropriateness score, with the pharmacists scoring better. The authors suggest that pharmacists who pass the certification exam write prescriptions that are as appropriate as, and occasionally more so than, those written by doctors.

2.2.2 Research in Other Professions

Price & Wilkins (2001) begin their article by noting the lack of research on the psychometric properties of certification exams in their area of interest – scuba diving. They examine the National Association of Underwater Instructors (NAUI) entry-level scuba diving certification test. Taking their cues from the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999), the authors assert that the lack of research should be countered with three types of information: documentation on the

development of the exam, evidence of score reliability, and evidence of the validity of score interpretations.

Their discussion of development of the NAUI test is brief, but they do note a number of points. First, a group of subject matter experts was assembled, and these individuals put together the test specifications for the essential content areas. Second, a new group – made up of academics and representatives of the industry – were given the specifications and asked to develop a bank of 250 potential test items. Finally, a total of 85 items were eventually selected from the bank in order to evaluate the candidates on the six subscales of the exam (diving equipment, diving physics, diving medicine and physiology, decompression theory, diving environment, and diving skills and safety).

The authors then prepared their test scores for analysis. They divided their entire set of data (scores from 400 candidates) into two samples: the first was made up of scores from 200 candidates selected at random (the validation sample), and the second was formed from the scores of the remaining 200 test takers (the calibration sample).

To conduct a preliminary investigation of reliability, Price and Wilkins calculate a series of pairwise Pearson correlations between scores for the six subscales in the test. All but one of the correlations proved to be statistically significant ($p < 0.01$), however the magnitude of the coefficients was quite low (ranging from $r = .19$ to $r = .35$).

To examine the validity of the test constructs, the authors use a number of factor analytic methods. They begin with an exploratory factor analysis, conducted on the 85 item

scores for all 400 test takers, and the results of the analysis yield a three-factor solution (the authors do not specify the method of factor extraction). They then conduct a confirmatory factor analysis, testing a number of possible factor models (null, one-factor, two-factor, and three-factor models). Results indicated that the one- and two-factor models provided the best fit for the data, and that there was no significant difference between them. The authors indicate their preference for the two-factor model, since it corresponds well with the essential division of test items into two components – knowledge of diving theory, and application of that knowledge.

The authors use these results to further investigate score reliability. Applying a formula analogous to Cronbach's alpha, they calculate internal reliability coefficients for the theoretical and applied items, within both the validation and calibration samples. The results were encouraging (coefficients ranged from .90 to .94).

The final investigation the authors conduct addresses the issue of criterion validity. They compare results from the NAUI test with those obtained from an open-water diving performance test. Through the use of *logistic regression*¹⁷, they were able to demonstrate that the total test score obtained by an individual on the NAUI test is a significant predictor of success on the performance test.

¹⁷ Logistic regression is a statistical technique used to predict values of a dichotomous dependent variable from the knowledge of two or more independent variables (Norman & Streiner, 1997; George & Mallery, 2001).

Cobb, Shaw, Millard and Bomotti (1999) examine the validity of a battery of tests designed to license prospective teachers. The Program for Licensing Assessments for Colorado Educators (PLACE) evaluates test takers in terms of their performance on four component tests: Basic Skills, which examines candidates' reading, mathematics, and writing abilities; Liberal Arts and Science, which tests candidates' general knowledge of the physical sciences, history, social sciences, arts, humanities, and communication; Professional Knowledge, which assesses candidates' pedagogical skills at the elementary, middle, or secondary school level; and Content Areas, which measure candidates' proficiency in specific subject-matter areas.

The authors' primary objective was to investigate issues of validity related to the tests, and they did this by examining the tests' criterion validity, and also by looking at more novel considerations such as consequential validity (the effect that testing has on different groups of test takers) and utility (the usefulness of testing for the test takers and test users). To conduct their study, Cobb et al. used the scores from the various PLACE tests in combination with data provided by the institutions of higher learning the test takers attended, and feedback obtained from examinee focus groups. The final dataset they worked with contained scores and information from 1,930 examinees.

With regard to the issue of criterion validity, Cobb et al. performed a series of comparisons between the scores from the various tests in the PLACE battery and the institutional data they obtained. Their first efforts revolved around the Basic Skills tests. They correlated the scores from the three components of the test with scores from the American College Test (ACT), the Scholastic Aptitude Test (SAT), and the Graduate Record

Examination (GRE). They found that scores from the Basic Skills math component correlated relatively well with math scores from the other tests (ranging from $r = .66$ to $r = .68$). However, correlations between the Basic Skills reading test and the ACT reading and SAT and GRE verbal scores were noticeably lower (ranging from $r = .25$ to $r = .49$).

As a second means of evaluating the Basic Skills tests, the authors divided the institutional scores (ACT, SAT, GRE, and Grade Point Averages) into two groups, depending on whether they were linked to a candidate that had passed or failed the Basic Skills writing and reading tests. They then calculated *effect size*^{*18} differences for the institutional scores. In general, effect sizes based on the Basic Skills writing test proved to be larger than those based on the reading test. Also, the effect size for GPAs turned out to be particularly unreliable: for undergraduate GPAs, it was surprisingly low, and for graduate GPAs, it failed to reach significance.

The Liberal Arts and Science test scores were correlated with the relevant ACT and SAT scores, with acceptable results (ranging from $r = .50$ to $r = .65$). Scores from the three Professional Knowledge tests (for teaching in elementary, middle, and secondary schools) were correlated with ACT, SAT, and GRE scores, and also with candidates' grades in teaching and methods courses. Correlation with course grades was revealed to be somewhat low (ranging from $r = -.58$ to $r = .66$), while correlation with the other institutional data was

¹⁸ Effect size is a measure of the differences between two groups of observations. The most common means of estimating effect size is to subtract the mean of one group from that of the other, and to divide the difference by the standard deviation of either group. Effect size is expressed in terms of standard deviations, and not in terms of a specific unit of measurement, making it ideal for comparisons between different studies (Cohen, 1988).

marginally higher (ranging from $r = .15$ to $r = .99$). Correlations between the Content Area scores and student GPAs were generally low (ranging from $r = -.99$ to $r = .81$).

As a final step in the concurrent validation process, the authors divided the scores of students who had passed and failed the Content Area tests, and then correlated them with each candidate's mean GPA in his or her major. Among the examinees who passed, correlation was relatively high (.02 to .63), while among those who failed, correlation was somewhat lower (-.41 to .62). The authors conclude that these findings are positive, because the range of GPAs in their data set is somewhat limited.

Changing their tack somewhat, Cobb et al. then set out to investigate the PLACE battery's effect on different groups of test takers (consequential validity), and to evaluate the usefulness of the tests for test takers and test users (utility). The authors conducted five one-way analyses of variance and a series of *post hoc* tests to compare scores among three groups of test takers: White males, White females, and all others combined. The results showed that non-White candidates scored consistently lower than White candidates, suggesting in turn that the tests had the unintended effect of discriminating against minorities.

Interviews with focus groups also helped identify the tests' strengths and weaknesses. Interviewees consistently acknowledged the importance of a test of pedagogy (the Professional Knowledge test), but criticized its format (it was not performance based). They judged Content Area tests to be the most relevant to their prospective careers (other measures suggested that there were potential problems with the content validity of these tests), but they raised concerns about the Basic Skills tests. It appears that candidates spent a great deal of

time and effort preparing for these tests, yet test content largely duplicated that of other tests they were required to take.

Based on their findings, the authors draw a number of conclusions, both positive and negative. They indicate that they uncovered some evidence of criterion validity, mostly in particular components of other formalized tests. Yet other sources of information, such as grades and GPAs, generally gave a poor indication of an individual's performance on the PLACE tests. It was shown that the tests had unintended consequences on minority groups, and doubt was cast on the utility of the Basic Skills tests, which because of their content and cut-off scores failed to discriminate against all but the most grossly underprepared students.

Daniel & Siders (1994) present information on the psychometric properties of the Mississippi Teacher Assessment Instruments (MTAI), a certification exam used in education. The MTAI make use of three subscales to test would-be teachers on 14 distinct competencies: 1) the Teaching Plans and Materials (TPM) subscale assesses four competencies as they are demonstrated in a portfolio of teaching plans; 2) the Position Skills (PS) subscale assesses seven competencies through observation of the teacher in the classroom; and 3) the Interpersonal Skills (IS) subscale assesses three competencies through observation of the teacher in the classroom.

The authors note that earlier studies on the MTAI provided preliminary information on some of the test's psychometric properties. One study measured criterion validity by

using explained variance to correlate scores from the three subscales (TPM, PS, and IS). The results were not significant, and only a moderate level of correlation was found ($r^2 = .3880$). Another examined the issue of intra-rater reliability (evaluator ratings of the same teacher across two occasions of measurement), but found that the correlations – with values less than .20 – were unacceptably low (the correlation method was not specified). A third study attempted to examine content validity by measuring the correlations between each of the test items and the adjusted total score. The resulting values ranged between .33 and .71. One final study they cite used exploratory factor analysis to gather evidence of construct validity. However factor structure uncovered in the study (six factors) did not suitably match the expected factor structure (3 distinct subscales, 14 separate competencies).

In their study, the authors set out to do three things. First, they attempted to determine if conducting their own exploratory factor analysis would reveal a factor structure that could be interpreted in a satisfactory manner. Second, they attempted to determine if a confirmatory factor analysis using the expected factor model would provide a good fit for the existing data. Third, they attempted to use confirmatory factor analysis to identify other suitable factor models. In the study, they used a total of 760 complete ratings (four ratings for most of the 194 teachers in the data set).

In the exploratory factor analysis, the authors used a *principal components analysis**¹⁹ and *Cattell's scree test**²⁰ to extract four factors. The factors were then rotated to the *varimax criterion**²¹. This accounted for 41.1% of the variance in the data.

In conducting the confirmatory factor analysis, the authors assumed that there would be three factors (corresponding to each of the three subscales), and that they would be univocal and uncorrelated. In other words, the model had the following three expectations: 1) scores from the 12 items that test teaching plans and materials would display a factor structure coefficient of 1.00 for the TPM subscale and 0 for PS and IS subscales; 2) scores from the 20 items that test position skills would have factor structure coefficient of 1.00 for the PS subscale and 0 for the TPM and IS subscales; and 3) scores from the 10 items that test interpersonal skills would have a factor structure coefficient of 1.00 for the IS subscale and 0 for the TPM and PS subscales. However, the significance of the results ($p < .05$) and the low goodness-of-fit index²² (.777, less than the sought after value of .9) led the researchers to reject the theoretically established model. They were also unable to find an alternative model that had a high goodness-of-fit correlation and that was statistically significant.

¹⁹ Although there is much disagreement between statisticians and social scientists about the use of the term *principal components analysis*, it is being used here simply to designate one method of factor extraction in exploratory factor analysis. This definition of the term is found in other sources as well. See Pett, Lackey and Sullivan (2003, p. 89).

²⁰ Cattell's scree test (Cattell, 1966) is one of two common methods used to determine how many factors to keep in exploratory factor analysis. It involves plotting the eigenvalues on a graph, and through visual inspection identifying the point on the plot where the curve begins to level off. The location of the point indicates the number of factors to be retained. The other common method is the use of the Kaiser-Guttman criterion.

²¹ The varimax criterion is the most common method of orthogonal (angle-preserving) axis rotation used in factor analysis. Like other rotation methods, it provides a simpler structure and a more interpretable pattern of values (George & Mallery, 2001).

²² The goodness-of-fit index (GFI) compares observed data with values expected based on a statistical model.

In their conclusions, the authors note that the four factors identified in their exploratory factor analysis closely match those found in the factor analytic study cited at the beginning of their article. Moreover, one of the factors identified seems to conform to the TPM subscale, providing evidence that at least one of the subscales appears to have construct validity. They advocated further research into the construct validity of the test.

This short sample of the literature serves to illustrate a number of points. First, the tests examined in these studies are ultimately concerned with producers, and not products. The samples of work produced in the testing situation are exactly that – samples, which are used to gain a sense of the candidates' overall levels of competency in a pre-defined area. The ultimate goal in each test is to make a judgement about the test takers, to decide whether the level of competency demonstrated by each individual meets the minimum level required by the profession in which they wish to practice.

The second point raised in the credentialing literature is that the path from test taker to final decision is an indirect one. Test users are unable to peer into the candidate's head, isolate something called "a competency", and measure it the way they would a physical object. Instead, they recognize that the professional competencies they seek to evaluate are psychological constructs, and that they must be measured indirectly. Test users do this by constructing a chain of inductive arguments. This chain requires test users to define the competencies they are interested in carefully, to generate a hypothesis about the competencies based on the definition, to gather evidence that supports or refutes the

hypothesis, and to make a conclusion about the accuracy of the initial definition based on the evidence gathered.

We can see part of this process of indirect measurement in the credentialing studies reviewed in *Section 2.2*. The authors' discussions of subtests, domains of knowledge, and individual competencies are all designed to furnish the reader with a precise definition of the psychological constructs that each test purports to measure.

Yet test users need to exercise caution throughout their chain of arguments. The reasoning that supports their use of credentialing tests is based on a series of assumptions. In testing, it is assumed that 1) the content of the credentialing test adequately reflects the realities of the profession; 2) the competencies required for practice in the profession have been identified, and they are widely accepted within the profession; 3) a minimum level has been identified for each competency, and these levels are widely accepted within the profession; and 4) test takers' scores adequately reflect their levels in each competency, which in turn accurately describes their ability to perform in the profession.

There is obviously a certain danger in making these assumptions. If left unverified, each one could lead test users into difficulty. It is, after all, theoretically possible to construct a test where none of them is true. Test developers could very well set tasks for test takers that are unconnected to the activities of practicing professionals. Developers could likewise choose competencies at random, and they could set minimum levels arbitrarily. The scores generated by the exam might in fact bear no relationship to test takers' ability to practice in the profession.

Therefore, the third point that we must make about the credentialing literature is that it is very much concerned with demonstrating the accuracy of the four testing assumptions enumerated above. The authors of the studies reviewed above have attempted to demonstrate this accuracy by focusing on the psychometric properties of their particular test. They have shown through logical and empirical means that the measurements made in the test situation are stable (i.e., they are reliable), and that the decisions made about candidates on the basis of their test scores are appropriate (i.e., they are valid).

We should also point out that the empirical tools that have been used are relatively consistent from study to study. For example, Pearson's product-moment correlation and factor analysis were frequently used to determine if the mathematical relationships among sets of test scores matched those that were believed to exist among constructs. Cronbach's alpha was employed on a number of occasions to estimate the stability of the scores associated with a single construct, and item analysis was used on at least one occasion to ensure that all items on a test had similar levels of difficulty and a similar ability to discriminate between high and low scorers.

2.3 Establishing a Model for the Present Study

It is clear at this point that the two research perspectives we have examined in this chapter – TQA and professional credentialing – are in some ways similar. Both are used for

the purposes of assessment, that is, to collect, interpret, and synthesize data in order to make a decision (Gage & Berliner, 1991).

However, they differ in several important ways. TQA is primarily used to make a judgement about products, about individual translations or interpretations. In contrast, credentialing tests are used to make judgements about producers, that is, about individual practitioners being certified or licensed in a particular profession. TQA is largely concerned with the categorization of errors in the product, whereas credentialing examines internal qualities²³ in the producer. TQA attempts to make some statement about the product's quality, while credentialing is concerned with the quality of the test – and its ability to make accurate judgements about the “quality” of the producer, an ability which is most often demonstrated through acceptable levels of reliability and validity.

2.3.1 Choice of a Model

The differences between TQA and credentialing we identified above are important for methodological reasons. Our decision to review the two research perspectives was motivated by a need to find a suitable model for the present study, and the identifying features of each perspective will have an impact on our choice of model.

We have argued thus far that litigation (*Seltzer v. Foley*, 1980), professional opinion (Sawyer, 2000), and industry standards (AERA, APA & NCME, 1999) have all pointed to the need to investigate the psychometric properties of interpreter certification tests. The small sample of articles on credentialing reviewed in this chapter has provided us with several clear examples that illustrate how we may be expected to do this. It has drawn our attention to the importance of estimating the reliability of scores within a construct measure, and of comparing scores among construct measures as part of a validity argument. Not insignificantly, it has also identified specific statistical tests that may be used to achieve these goals.

However, this is only one piece of the puzzle. Citing Sawyer (2000) in *Chapter 1*, we noted that testing know-how was an important part of addressing current gaps in interpreting research. But we also noted the need for a source of information about interpreting that would allow us to identify the competencies interpreters require to perform their work.

It appears now that TQA is unable to assist us on either front. None of the TQA research we reviewed demonstrated to us how we might evaluate the internal qualities of test takers or how we might evaluate a test's ability to measure those qualities. And although some of the TQA literature did offer us a description of "good" translation and "good" interpreting, this description was always based on an observation of an individual translated

²³ Some might argue that a credentialing test evaluates professional performance, which can be measured directly, and is therefore not internal. While this is true, a credentialing body cannot certify or licence a performance – which is essentially a single instance of behaviour. They must instead look for evidence that the behaviour is the result of an internal quality (such as a professional competency). Once it has been demonstrated that the person possesses the quality, and can presumably use it to produce other instances of the same behaviour, the person is then licensed or certified.

text or an individual interpreted utterance. It was never crafted from the perspective of the translator or the interpreter. We therefore conclude that the TQA research is of little assistance to us in the present study. This is not meant as criticism of the TQA literature or of the people behind it. TQA research clearly has a place in both Translation Studies and Interpreting Studies. However, it is not reasonable to expect that TQA should form the theoretical framework of the present study, as some might suggest, because within TQA we will not find either of the two things most urgently required at this juncture – a practical means for evaluating decisions made about the qualities of test takers, or a comprehensive model of the interpreting process, presented from the perspective of the individual interpreter.

For this reason, we have made a conscious decision not to orient the study in the manner typical of the TQA literature. Instead, we have opted to construct the theoretical framework for the present study from two sources: the principles of measurement and evaluation laid out in the *Standards* (AERA, APA & NCME, 1999) and applied in the credentialing literature reviewed in this chapter; and the models of the interpreting process reviewed in *Chapter 3*.

2.3.2 Justification for the Choice of Model

Our decision to not conduct the present study from the perspective of TQA, and our choice to model it instead after examples of credentialing research are not immune to criticism. Some might claim that it is not appropriate to conduct such research under the

aegis of Interpreting Studies, and that it should be undertaken instead in another discipline such as educational or occupational psychology.

However, we would reject this claim, for two reasons. First, the present study was developed through the application of an important body of theory in Interpreting Studies, as *Chapter 3* will attest. Second, the psychometric approach we have adopted is consistent with at least some research in Interpreting Studies. There is a small number of articles, written by authors who are incontestably translation or interpreting scholars, that has distinct psychometric leanings. While they do not necessarily bear a strong resemblance to the credentialing literature reviewed earlier in this chapter, these articles do show some of the hallmarks of measurement and evaluation research. All the articles outline the development of a formal instrument designed to address testing needs in a professional context, and one of them discusses the concepts of reliability and validity in some detail. These are features that mark the research presented on the following pages and set it apart from the TQA literature reviewed above. The articles are presented separately below.

Two of the articles (Roberts, 1995; Roberts 1998) provide some much needed insight on the topic of the certification of *community interpreters**. Roberts describes two community interpreting tests that are used at the local level in Canada²⁴: CILISAT, a test

²⁴ Although there are national certification tests for conference interpreters and court interpreters in Canada, no such test currently exists for community interpreters.

used originally in the Ottawa region; and the Surrey Delta test, used originally in the area outside Vancouver²⁵.

The two tests resemble one another in the constructs they measure (language proficiency and interpreting accuracy) and in the types of test items they use (*sight translations** and *dialogue scenarios**). They were also prepared with the same set of purposes in mind – the certification of practicing interpreters, but also the screening of candidates entering and exiting interpreter training programs.

Where the two tests differ is in their scoring schemes. In the CILISAT test, separate scores are generated for language proficiency and interpreting accuracy. An interpreter's language proficiency is rated in a quasi-holistic manner, by using a *rating scale** to assess the use of vocabulary, technical terms, grammar, register, and pronunciation throughout the entire test. An interpreter's accuracy is rated analytically, by dividing the sight translations and dialogue scenarios into units of information, and using a rating scale to assess the interpreter's performance on each unit. The units are weighted according to their importance. In the Surrey Delta test, a joint score is generated for language proficiency and interpreting accuracy²⁶. Test items are scored analytically, by assessing such criteria as source-text comprehension (deducting points for mistranslations), accurate presentation of ideas

²⁵ Both tests are now used more widely. For example, beginning in 2000, Vancouver Community College took over the administration of the Surrey Delta test. With a few modifications (the addition of a written translation test, and an increase in the cut-off score from 65% to 70%) the test is now used to screen applicants into the college's community interpreting program, a 72-hour certificate required by many service providers that use interpreters in British Columbia (Carr, 2003).

²⁶ The second test also gathers additional information on language proficiency by using listening comprehension and memory exercises in both of the interpreter's languages.

(deducting points for additions, omissions, and distortions), handling of names and numbers, and use of grammar and syntax.

Roberts' articles are significant in that they convey a tremendous amount of detail about the development, administration and scoring of these two tests. In both cases, the reader is presented with a description of the logical, methodical approach that was used to build the tests.

A similar emphasis on process is also evident in Arjona (1985). The article outlines the development of the US Federal Court Interpreter Certification Exam, the same test that was discussed in *Chapter 1* with reference to the *Seltzer v. Foley* (1980) case.

The author explains that *test development** began with a thorough study of the existing conditions and practices within the interpreting profession as a whole. In the course of conducting this nation-wide job analysis, the test developers were able to identify a list of practices, tasks, needs, and skills required for court interpreting. It was discovered that court interpreters encounter a remarkable range of linguistic diversity in their work – both in terms of register and regional variation – and the developers felt that this placed a significant demand not only on their interpreting skills, but on their linguistic abilities as well. It was also revealed that the language spoken by the overwhelming majority of witnesses and defendants in the federal courts who needed interpretation was Spanish.

For these reasons, the developers decided to construct a pilot certification test from two components: a written subtest to screen for the requisite Spanish and English linguistic

abilities, and an oral subtest to evaluate actual interpreting ability between the two languages. To create the two subtests, the developers formed several task forces whose members included court interpreters, escort interpreters, representatives of interpreter associations, interpreter trainers, cultural anthropologists, linguists, and other researchers. The groups also comprised Latin Americans from a number of points of origin so that many varieties of Spanish were represented.

The written test measured several different constructs, including the knowledge of synonyms and antonyms, reading comprehension, sentence completion, and language usage. It was a norm-referenced test, set at a level of difficulty similar to the GRE. The test was piloted with the participation of 24 court interpreters, and it was scored objectively (i.e., without raters). The test scores were used to calculate difficulty indices for all the items, which in turn provided the developers with the necessary information to revise the written portion of the certification test.

The oral subtest examined candidates' proficiency in courtroom interpretation, and was therefore the more overtly "job-related" of the two components. It measured several specific constructs, including skill at sight translation, and the capacity to interpret opening statements, cross-examinations, and instructions to the jury. The subtest was criterion-referenced, and its items were constructed from transcripts of actual courtroom proceedings. To construct a scoring scheme, the developers actually interpreted all the materials themselves before the pilot test to create a list of cultural, semantic, and syntactic "pitfalls" (words or clusters of words that could potentially cause difficulty). After the subtest was administered, judges assigned an "objective" score to a candidate's performance by tallying

the grammatical errors, errors in vocabulary, and changes in meaning. In cases where the objective score put the candidate close to the cut-off mark, judges also took a “subjective score” into account. This involved rating the candidate’s delivery, pronunciation, and ability to adapt to the witness’ or defendant’s manner of speaking as “superior”, “acceptable”, or “unacceptable”.

These three articles (Roberts, 1998; Roberts, 1995; and Arjona, 1985) show evidence of an approach that might be described as “quasi-psychometric” Both authors have presented detailed explanations of their test instruments in a manner that is largely consistent with accepted test development models (see AERA, APA & NCME, 1999; and Newble et al., 1993). For example, there is ample evidence that steps were taken in the federal court interpreters test (Arjona, 1985) and the CILISAT test (Roberts, 1995) to ensure that test content was related to the realities of the profession. The use of job analysis, the recruitment of stakeholders in the professions, and the use of actual transcripts to develop test items all suggest that there is considerable evidence of validity related to the tests.

There is other evidence of a concern for measurement issues in Arjona’s work. The author notes that, as part of the piloting of the written subtest, developers conducted an item analysis – a procedure that was mentioned earlier in this chapter as part of our review of the credentialing literature (see *Section 2.2.1*). Arjona’s article also includes a short discussion of validity-related issues; she suggests that the written subtest was an important part of the complete test instrument, and that its use was supported by research on professional practice; she also admits that steps could be taken to increase the face validity of the subtest. Perhaps most importantly, Arjona mentions that while an empirical investigation of the test’s

psychometric properties was not included in the article, it was an important step in evaluating the test, and it was therefore planned for subsequent publication.

In short, these three articles adopt an approach that is different from that of the TQA literature, and that shows the beginnings of a regard for the issues that characterize the credentialing research. We therefore believe that a detailed examination of certification tests, even work that relies on empirical analysis of psychometric properties, is consistent with existing research in Interpreting Studies²⁷. The measurement-based perspective that we have adopted in the present study is the logical extension of research conducted by Roberts and Arjona, and we therefore feel that our choice of a psychometric approach to a problem in Interpreting Studies is justified.

2.4 Conclusion

In this chapter, we have reviewed two research perspectives that might initially seem important for a study of interpreter certification. We briefly examined a number of TQA models, both quantitative and qualitative, and we noted how TQA was characterized by a concern for translated texts or interpreted utterances, and by the categorization of errors within them. We also discussed a small sample of the credentialing literature, which

²⁷ The wider discipline of Translation Studies has also provided some indirect support for the measurement-based approach adopted in the present study. A number of articles on translation ability have made use of psychometric tests. Waddington (2001) uses factor analysis and correlation as part of a validation study of four different systems of evaluating the performance of students at a school of translation. Stansfield, Scott and Kenyon (1992) discuss the psychometric properties of the SEVTE, a test of job-related translation ability used by the FBI. They use KR-20 to examine internal reliability, and Pearson's product-moment correlation to look at inter-rater reliability, construct validity, and criterion validity. Campbell (1991) relies on correlation (the method of correlation is not mentioned) to identify three important components of translation ability exhibited

discussed certification and licensure tests in the health professions and in other areas, and which was concerned with evaluating test takers' internal qualities. The articles we reviewed were characterized by a concern for the psychometric properties of tests as a means of verifying that decisions made about test takers are accurate.

Also in this chapter, we argued that the credentialing literature constitutes a more appropriate model for the present study, because our principal aim is to investigate the psychometric properties of existing certification tests, and if need be, explore the ways in which they may be improved. We justified carrying out the present study under the *aegis* of Interpreting Studies by pointing to other research that shows the first vestiges of what might be described as a psychometric orientation. We therefore believe that the present study is appropriately conducted within Interpreting Studies, because it builds on existing "quasi-psychometric" work, and that it is novel, because it proposes to examine the psychometric properties of two interpreter certification tests.

With a model for the present study's methodology now in place, we turn our attention to the identification of the competencies that are necessary in professional interpreting. In the following chapter, we review a number of models of the interpreting process, and we attempt to use them to define a set of psychological constructs to form the basis of a new test instrument.

by candidates in a public translation exam.

3. The Identification of Interpreting Competencies

Up until this point, the term *competency* has been used in the present study in a rather vague manner, which presents us with an obvious difficulty. How can we identify an interpreting competency if we have not first defined what the term means? To solve this problem, we briefly examine the original research that defined the concept of competency, along with subsequent research that further elaborated the definition. Later in this chapter, we use the definition to identify the professional competencies suggested by a number of models of the interpreting process.

The term *competency* was first used with regard to testing by McClelland (1973). He noted at the time that a great deal of attention in measurement and evaluation was focused on one particular psychological construct – intelligence. It was widely used to predict performance in several areas, both educational and professional, yet McClelland was able to find evidence that intelligence actually bore only a weak relationship to professional or job-related success. The studies he cited appeared to show that people could obtain high scores on an intelligence test, but demonstrate a low level of ability when they actually found themselves in a professional or job setting. Hence, the average IQ test was not a very good predictor of on-the-job performance. This led McClelland to argue that instead of looking at intelligence, psychometricians interested in professional behaviour should think in terms of job-related competencies. McClelland further argued that identifying the competencies that were pivotal in a given profession or job required scholars to actually go out and observe

practitioners at work to determine what their job entails. “If you want to find out who will become a good policeman,” he reasoned, “go find out what a policeman does.”

This attitude is at the heart of McClelland’s three-step approach to identifying competencies. First, he advocates that researchers gather a criterion sample by identifying a number of clearly superior performers and a contrasting sample of average or poor performers for the job category in question. The level of performance is estimated by consulting stakeholders in the professional setting, such as bosses, peers, and clients. Second, he suggests that researchers hold *behavioural event interviews* with the subjects in the samples. In the interviews, subjects are asked to describe three of their peak successes and three of their major failures in short story fashion. Third, he recommends that researchers analyze the interviews to find the characteristics that differed between the two samples (i.e., those behaviours demonstrated by superior performers, but not by average or poor performers). This final step allows the researchers to identify the thoughts and behaviours causally related to successful outcomes in the job category (McClelland, 1993).

Building on McClelland’s work, Spencer and Spencer (1993) offer a more precise definition. They explain competency as an underlying characteristic in a person, which indicates that person’s way of behaving, thinking, or generalizing across situations, and which endures for a reasonably long period of time. They also describe the five basic types of competencies. *Motives* are those things that a person consistently thinks about or wants that cause him or her to take action (e.g., the drive to achieve). *Traits* are physical characteristics or consistent responses to situations (e.g., good eyesight, quick reaction time). *Self-concept* is made up of a person’s attitudes, values, or self-image (e.g. a person’s belief

that he or she can be effective in almost any situation is part of that person's concept of self). *Knowledge* constitutes the information that a person has in specific content areas (e.g., a surgeon's knowledge of nerves and muscles in the human body). Finally, *skill* is the ability to perform a certain physical or mental task (e.g., a dentist's skill at filling in a tooth without damaging a nerve).

Knowledge and skills tend to be visible, surface characteristics of people, while self-concept, traits, and motives are deeper, more hidden, and more central to an individual's personality. Knowledge and skills are acquired relatively easily, to the extent that many employers choose to foster them through training. Motives and traits are harder to develop, which explains why many employers select for them through hiring. Self-concept lies somewhere in between, in terms of its ease of development (Spencer & Spencer, 1993).

There is also a causal relationship among the various types of competencies. Motives, traits, and self-concept have been shown to predict skill and knowledge behaviour actions, which in turn predict the outcome of a person's performance in the work setting. What this means, for example, is that a professional may possess a certain skill, but without the requisite motive, they will not put it to use on the job. It is therefore important for organizations to pay attention to their employees' more central characteristics (Spencer & Spencer, 1993), even though the hidden nature of these types of competencies may make it more difficult to do so.

As part of their research on management competencies, Spencer and Spencer provide a "competency dictionary", that is, a meticulous description of the characteristics that

distinguish superior managers from mediocre ones. The contents of the description are not of particular interest to us in this study, but its structure is noteworthy. The authors organize the competencies they discuss into *clusters*, small groups of characteristics that are connected through their common relationship to performance in a particular profession or job. On a smaller level, each competency is broken down into several constituent elements, called *dimensions*, that describe in detail how a competency may be manifested.

The research on competencies reviewed here is obviously relevant for interpreter certification testing, but it seems unrealistic to follow its example to the letter. For instance, while McClelland's criterion-sampling approach could potentially offer a clear, logical, and defensible method for identifying interpreter competencies, it is unlikely that a professional association of interpreters would adopt it without reservation. This is because most interpreter associations have very few resources, either human or financial. They are largely staffed by volunteers, who have numerous other commitments in their professional lives. Membership bases are small, and budgets are limited (Cadieux, 2001; Warren, 2001). Implementing criterion-sampling would undoubtedly place a heavy demand on scarce resources, and before they would commit to such an undertaking, professional associations would need some evidence to demonstrate that it is worthwhile.

The results of the present study may provide that evidence. If our analysis of the new test suggests that the competencies upon which it is based are valid constructs, it could lay the groundwork for the kind of detailed study recommended by McClelland. However, since we have temporarily ruled out identifying interpreting competencies through empirical

means, we still need to determine which constructs we will attempt to measure in designing the new test.

Spencer and Spencer (1993) note that the first studies of management competencies were conducted on a conceptual basis, rather than an empirical one. This seems to be consistent with the position taken in the *Standards* (AERA, APA & NCME, 1999), which state that assumptions about the internal processes measured by a test should be supported with empirical or theoretical evidence (see *Section 1.1.2*). Consequently, we will consult a number of theoretical sources – models of the interpreting process – as a means of selecting the competencies that will form the foundation of the new test.

Our discussion of competencies will respect the definition and structure provided by Spencer and Spencer. We will concentrate on identifying skills and knowledge, rather than the other types of competencies. We believe this is justified for two reasons. First, because the present study is exploratory in nature (there is no prior research that estimates the psychometric properties of interpreter certification tests), an effort must be continually made to limit it to a reasonable size. An investigation of hidden competencies would obviously be more involved than an examination of surface ones, and we have therefore opted for the latter instead of the former. Second, the bulk of the existing research in Interpreting Studies appears to be more surface-competency oriented than hidden-competency oriented. Researchers regularly discuss aspects of interpreting that can be described in terms of the

skills required to carry them out – a fact that can be seen directly in the relevant literature¹, and indirectly in the models we review in this chapter – but they have only rarely mentioned characteristics that are more central to personality².

The remainder of this chapter is divided into four sections. The first three deal with a major model of interpreting or a major trend in model building: interpretive theory, information processing models, and discursive models. The final section in this chapter looks at how the components of the models we have reviewed may be translated into testing constructs, in a manner that is consistent with the preceding discussion of competency.

3.1 Interpretive Theory

The first model we will review is the earliest, and it dates from a point in time when conference interpreting was dismissed by many conference participants – and even linguists – as requiring merely a minor, mechanical skill dependent only on the knowledge of languages. Interpreting, they assumed, was nothing more than the substitution of the words

¹ To cite one example, Roberts (1992) argues that six competencies are necessary for successful interpreting: 1) language competency, the ability to manipulate with ease and accuracy the two languages involved in the interpreting process; 2) transfer competency, the ability to transfer the source-language message into the target language accurately, in accordance with stylistic conventions; 3) methodological competency, the ability to work in different modes of interpreting and to manage terminology; 4) subject matter competency, the knowledge to understand content of the message; 5) bi-cultural competency, deep knowledge and appreciation of the cultures underlying the working languages; and 6) technical competency, the ability to use equipment (e.g., headphones and microphones for conference interpreters) and other logistical factors (e.g., lighting and a stage or platform for sign language interpreters). All the competencies she lists are skill or knowledge based.

² Gerver (1976) provides examples of research on hidden competencies in interpreters, by citing two studies that examine interpreters' personality traits. The first, an unpublished study conducted by Howells, used an established psychological assessment instrument to determine that a small group of interpreter subjects scored highly on constructs such as assertiveness, independence, and self-sufficiency. The second, conducted by Gerver himself, used another established instrument to measure the connection between anxiety and performance under noise stress. It concluded that anxiety under moderately stressful listening conditions may

of one language for those of another, and therefore could be performed adequately without much training or preparation. This understandably prompted a strong reaction on the part of conference interpreters (Seleskovitch, 1968). Convinced that they could not rely on traditional academia to provide an accurate theory of interpreting, a number of interpreters set out to provide this for themselves.

The first attempt to do so has since become known as the *interpretive theory*. Its two main proponents, Seleskovitch and Lederer, outlined the theory in an impressive number of publications that spanned the period from the late 60s to the mid 80s. In the pages that follow, four important components of the theory are examined in detail: 1) the interpreting process; 2) the distinction between linguistic meaning and contextualized meaning; 3) units of meaning; and 4) the simultaneity of interpreting.

3.1.1 The Interpreting Process

Interpretive theory does not deny that interpreters do occasionally use a substitution-like linguistic process (“*transcodage*”) in their work. However, use of this process is much more limited than conference participants and linguists would imagine, being essentially restricted to what Seleskovitch refers to as *mots à signification unique** (1975, p. 11). With some lexical items – such as numbers, proper nouns, enumerations, and technical terms – only one meaning is possible. Interpreters consequently engage in the straightforward exchange of target-language items for source-language originals. “Fifty-two percent” becomes “cinquant-deux pour cent”; “London” becomes “Londres”; “first”, “second”, and

be an advantage, but when a greater degree of stress was involved, anxiety proved to be a liability.

“third” become “premièrement”, “deuxièmement” and “troisièmement”; and “carabiner” becomes “mousqueton”. Yet the rest of the time, interpretive theory argues, interpreters are engaged in a rather complex process that tells us more about communication than it does about language.

To explain this process, Seleskovitch has her readers consider the proceedings of a typical conference from the point of view of a participant who stands up to give a presentation. As he³ takes his place behind the podium, he addresses a well-defined group of attendees seated in the room in front of him. His speech is likely a reaction to what has just been said, and in delivering it he may have any number of goals: to explain a point, to defend an interest, to convince a fence-sitter, etc. His presentation has an end goal, and its own internal logic. When his allotted time is over, he rejoins the audience. Another of their number rises to become the new presenter, and gives a speech in reaction to what has gone before⁴ (Seleskovitch, 1968, p. 60). What this suggests is that a large part of understanding in this context is dependent on knowing who the participants are, how they assume their roles, and how the volley of messages back and forth affects meaning. Since interpreters are also conference participants – taking on the role of both receiver and producer – knowledge about the proceedings and the people in them is essential for interpreters as well.

³ To make the text easier to read, we have used masculine pronouns to refer to conference participants and feminine pronouns to refer to interpreters. This is a convention used by other authors (Wadensjö, 1998; Setton, 1999) who argue that it reflects the realities of professional interpreting.

⁴ Although at this point in her discussion Seleskovitch does not consider the case of participants who prepare written speeches in advance, it seems likely that her observation still holds true. In reading a prepared speech and keeping strictly to its contents, a participant may not produce a reaction to the points raised by the previous speaker on a given topic. However, at the time of writing, the participant does know his own position on the topic, he is presumably aware of the potential positions of other people, and he likely builds this knowledge and awareness into his speech. In this sense, he is “reacting to what has come before” even if that before is a previous conference, existing publication, etc.

Viewed in this light, it is clear that interpreting is more than the simple substitution of language forms. It is instead a complex process, which interpretive theory has described as comprising the following three steps.

Reception

The first step begins with the decoding of a series of source-language signifiers⁵. As the visual or acoustic signal reaches the interpreter, its basic components are identified, and meaning is attributed to them. The meaningful base components are in turn assembled into larger units, such as sentences, and additional meaning is derived from sources such as grammatical structure (Lederer, 1981).

Deverbalization

Interpretive theory posits that it is impossible for interpreters to remember a speech as a conglomerate of individual lexical meanings (Lederer, 1994). The second step in the process therefore requires interpreters to use their well-developed sense of critical analysis – honed through a constant effort to learn more about the world around them (Seleskovitch, 1968, p. 35) – to extract the contextualized meaning (“*sens*”) of the message. Interpreters must forget the syntactic structure and individual signifiers of the source language, and focus instead on its semantic content⁶ (Seleskovitch & Lederer, 1984), retaining a mental image of

⁵ Seleskovitch describes this first step as the “audition d’un signifiant linguistique, chargé de sens” (1968, p. 35).

⁶ The scope of this proposal does not allow for a discussion of interpretive theory’s assertion that form is distinct from meaning. Suffice to say that not all scholars in Translation Studies and Interpreting Studies are in agreement. See Berman’s (1999) distinction between “*mot*” and “*lettre*” and Meschonnic’s discussion (1973) of “*forme-sens*”.

the concepts and ideas behind the original words, and not the words themselves (Seleskovitch, 1968, p. 35).

Reformulation

The process ends with the production of target-language signifiers⁷ that convey not the words of the source-language speech, but its contextualized meaning. Emphasis here is on capturing the original message, but also on meeting the expectations of the target-language audience (Seleskovitch, 1968, p. 35). After all, the interpreter's production must still conform to the grammatical and stylistic norms of the "*langue d'arrivée*". To the layperson, walking this tightrope between form and content appears to be a daunting challenge, but for the interpreter, it is merely a question of expressing an idea that has been truly understood in a natural and spontaneous manner (Seleskovitch, 1968, p. 161).

Clearly, interpretive theory's explanation of the interpreting process rests in large part on the distinction between linguistic meaning and meaning in context. Without it, deverbalization and reformulation would be impossible, and the differentiation between "*trancodage*" and true interpreting would be meaningless. The distinction is therefore a central component in the theory, and it merits further examination.

3.1.2 Linguistic Meaning and Meaning in Context

Interpretive theory defines linguistic meaning as the conceptual contour of a word or grammatical structure, as defined by its semantic or lexical features (Seleskovitch & Lederer,

1984). The interpreter, or any other linguistically competent adult, arrives at linguistic meaning through an analysis that takes place out of context, purely at the level of language (Seleskovitch, 1975, p. 11). In contrast, meaning in context is linguistic meaning as it is understood within the act of communication (Seleskovitch, 1975, p. 12). It develops from the situation in which the word or structure was expressed, and from the surrounding “text” in which it appears (Seleskovitch & Lederer, 1984, p. 16).

The two types of meaning are associated with different types of memory. Linguistic meaning is linked to a component of short-term memory (“*mémoire verbale*”) designed to retain the sound structure⁸ of words. This link explains why the recollection of the actual words uttered typically lasts only for a short period of time – several seconds at best – although it is possible to increase this span of time by repeating a series of words over and over immediately after they are heard. However, if this repetition is impeded by the continuous flow of new input, as it is in interpreting, the retention time cannot be increased. This is what prevents an interpreter using the consecutive method from committing an entire speech to memory. Seleskovitch argues that the short duration of verbal memory is part of its “design”, in that syntactic patterns are only “intended” to be retained long enough for the hearer to analyze them (1975, p. 15). But using internalized linguistic ability, conceptual information, and knowledge of the situation at hand, the hearer is able to decipher and retain meaning in context. Because it is supported by a component⁹ of long-term memory

⁷ Seleskovitch explains this step as “la production d’un nouveau signifiant dans l’autre langue” (1968, p. 35).

⁸ In the early part of her career, Seleskovitch’s definition of interpreting was limited, taking only spoken language interpreting into consideration. At a later point in time, however, she adopted a more inclusive perspective. See Seleskovitch (1997).

⁹ Daró (1996) argues that long-term memory in particular should not be viewed as “a unitary function, but rather as a multifaceted system with several components” (p. 4).

(“*mémoire non-verbale*”), meaning in context is retained longer than the recollection of the actual words uttered. The resulting increase in retention gives the hearer the time to reach an understanding of the speaker’s intention (“*vouloir-dire*”) (Seleskovitch, 1975, p. 16).

The distinction between the two types of meaning also illustrates one of the credos of interpretive theory – “*les mots sont ambigus, mais la parole est univoque*”. For example, if an interpreter whose working languages include German and English were asked how to translate the German “*wenn*”, her answer would almost certainly be “it depends”. The single lexical item in the original language actually contains two separate meanings, which in English are expressed by two different words: *when* and *if*. However, if the same interpreter were told that a German mother had comforted her irate toddler in the Frankfurt airport by saying “*Wenn du groß bist, kannst du auch fliegen,*” she would likely produce her interpretation without any conscious thought about the choice between the two English words: “When you grow up, you’ll fly in a plane too.” In this example, it is clear that the use of a conditional is inappropriate (only under unusual circumstances would we speak to a child and begin a sentence with “*If you grow up...*”), and the interpreter chooses the more suitable meaning as a matter of reflex (Seleskovitch, 1975, p. 34).

However, the interpreter cannot always trust reflex. It can also lead her astray, causing her to produce a literal translation (“*traduction littérale*”) when a more thought-out response (“*traduction réfléchie*”) was called for. Seleskovitch and Lederer note that there are three instances when this is likely to happen: 1) when there is a formal resemblance between the source-language item and its potential translation (the English “realize” elicits “*réaliser*” in French, instead of the more likely “*se rendre compte*” or “*prendre conscience*”); 2) when

there is the popular perception of correspondence between words in two languages (“window” becomes “*fenêtre*”, even when “*vitrine*” might be more appropriate); or 3) when a foreign reality is imported and the language community has not yet adapted to it (“skyscrapers” were referred to as “*gratte-ciel*” until actual “*tours*” were constructed on French soil¹⁰). In each of these cases, the faulty translation is a result of a reliance on linguistic meaning, and a failure to analyze meaning in context (Seleskovitch, 1975, p. 53-56).

An explanation of interpreting in terms of linguistic meaning and meaning in context begs an interesting question. How much input is necessary before an interpreter can make the transition from the former to the latter? Observation of interpreters using the simultaneous method leads Seleskovitch and Lederer to note that the comprehension of meaning in context is not constant and continual, but rather that it occurs in short bursts (“*déclics*”). This in turn brings them to formulate an intriguing hypothesis – that simultaneous interpreting is largely a question of mentally collecting linguistic units into minimal groupings that allow the interpreter to understand the idea being expressed.

3.1.3 Units of Meaning

Interpretive theory refers to these groupings as units of meaning (“*unités de sens*”), and it defines them as segments that appear at irregular intervals in the mind of those who listen to a speech with the deliberate desire to understand it (Lederer, 1978, p. 330). The

¹⁰ Of course, a *calque* can be introduced to a language after the reality, but this possibility is not discussed by Seleskovitch and Lederer.

theory argues that units of meaning can be seen in the work of interpreters, particularly when they use the simultaneous method. Lederer (1981) observes that interpreters generally begin their turns in the booth by using “*transcodage*”, initially substituting target-language words for those they hear in the source language. However, as the speaker continues, interpreters experience irregular “eureka” moments of understanding, and they appear to switch to the three-step strategy of comprehension, deverbalization, and reformulation outlined above.

Lederer explains this change in strategy in the following manner. As the interpreter takes in the words of the speaker, she initially has little to work with but the linguistic meaning of those words. However, at some point she is able to associate what the speaker is saying with a “previous cognitive experience”, and she arrives at a clear understanding of the speaker’s intended meaning (Lederer, 1978, p. 330).

The cognitive experiences Lederer refers to may be of two types. The first is the interpreter’s prior knowledge (“*bagage cognitif*”). Understanding the topic at hand and analyzing the speaker’s position are in large part dependent on knowledge acquired from previous experience (Lederer, 1981, p. 194), whether that be formal study, independent reading, consultation with subject matter experts, or simply other interpreting work.

The second type of cognitive experience is an understanding of the situation (“*contexte cognitif*”) as it is progressively revealed through each conference participant’s presentation. Each of the interpreter’s eureka moments allow her to place a small fragment of understanding in her cognitive memory. These fragments accumulate over the course of each participant’s speech, throughout the conference proceedings. As her understanding grows,

the interpreter finds it easier to analyse subsequent linguistic forms and arrive at their meaning in context. In this manner, the interpreter is able to grasp a speaker's intention, even in the face of such daunting obstacles as heavy foreign accents and the poor expression of ideas. Although a particular utterance may be difficult to decipher, it bears a relationship to preceding utterances in the communicative event. If the interpreter has understood the previous utterances, the fragments accumulated in cognitive memory allow her to piece together the current speaker's intention, in spite of the difficulties she may face (Lederer, 1981, p. 191).

Prior knowledge and understanding of the situation also explain why the apparent size of units of meaning varies from individual to individual. If an interpreter knows the topic at hand well and understands the speaker's position on it, her units of meaning will likely be very small, and she will not need to wait until the end of the linguistic unit in question (a word, a phrase, a sentence, etc.) before producing her interpretation. The converse is also true – if an interpreter is relatively unfamiliar with the topic and has not yet worked out the speaker's position, her units of meaning will tend to be quite large (Lederer, 1994).

3.1.4 Simultaneity

Up until this point, our discussion of interpretive theory has made little mention of the feature of conference interpreting that has arguably garnered it the most attention from outside and inside Interpreting Studies – the carrying out of multiple tasks simultaneously.

This is not an oversight on our part, but rather a reflection of the importance that interpretive theory places on simultaneity in its overall account of the interpreting process.

Obviously, the theory cannot deny that several actions take place at the same time when interpreters are at work, for even the most haphazard observation of simultaneous interpreting will not fail to note that interpreters speak while they listen. The theory adds a level of precision to this rudimentary description by suggesting that three process co-occur: 1) reception of the speaker's source-language utterance; 2) conceptual analysis of the utterance's meaning and storage of that meaning in memory; and 3) the expression of a target-language utterance, with the aid of cognitive memory. Two other actions are also noteworthy, namely the awareness of the immediate surroundings (such that interpreters are capable of signalling in their interpretations when a new speaker has taken the floor), and the monitoring of their own output (such that they are able to correct a certain number of errors in their own production) (Lederer, 1981, p. 50).

However, the attention paid to the simultaneity of tasks in interpretive theory as a whole is minimal¹¹. In a discussion of interpreter pedagogy, Seleskovitch and Lederer (1984, p. 207) report that they do use exercises designed to train students to carry out simultaneous tasks. For example, they note that they have asked students to count backwards

out loud while listening to their instructor tell a story. Yet they also note – and devote more space to explaining – their use of exercises that take the emphasis off of simultaneity. For example, the authors indicate that they frequently introduce the simultaneous method to a group of students by bringing a native speaker of the students' *B language** or *C language** into the classroom and asking the speaker to recount a familiar story or fairy tale (such as Little Red Riding Hood) that exists in a multitude of languages. The students are required to simultaneously "interpret" the story into their *A language**. Students' familiarity with the tale in their native language makes its telling easy, despite the fact that they are speaking and listening simultaneously. This exercise illustrates an important tenet of interpretive theory – namely that using the simultaneous method is a straightforward matter, provided the interpreter is familiar with the topic (a fairy tale), the context in which it is presented (told by adults to children), and the reasons for presenting it (as a cautionary tale, to convey a set of morals, etc.).

The relatively minor importance of simultaneity to the theory can also be seen in other discussions. For example, it is often repeated in the interpretive theory literature that the consecutive method should be taught before the simultaneous method (Seleskovitch, 1975; Lederer, 1981; Seleskovitch & Lederer, 1989). This approach to teaching, the authors argue, allows student to first develop their analytical skills and improve their ability to

¹¹ Even Lederer's (1981) investigation of simultaneous interpreting places emphasis on aspects other than simultaneity. At one point, she discusses the work of some of the early proponents of the information processing approach (p. 240), who questioned whether interpreters make use of speaker's pauses to reduce the cognitive demands of carrying out several tasks simultaneously. Lederer dismisses this idea, citing instances in her corpus where the interpreter and conference participant speak and pause at the same time. She argues that understanding conference interpreting is not a matter of understanding simultaneity, but rather of figuring out how the "*déclic*" of understanding operates. She suggests that interpreters speak, not because the participant is silent, but rather because they have understood meaning in context. Similarly, interpreters pause, not because

understand meaning in context before then adding the simultaneity to the mix. Seleskovitch and Lederer repeatedly emphasize that when the students' analytical skills are in place, the simultaneous method generally poses few problems.

The elements of interpretive theory that we have examined in this short review – the degree of importance it attributes to simultaneity, its presentation of units of meaning, the distinction it makes between linguistic meaning and meaning in context, and its description of the interpreting process – give an overview of what is obviously a very significant body of work in Interpreting Studies. In the beginning, it directed scholarly attention away from a preoccupation with language learning and code substitution, and steered it instead towards understanding interpreting as the communication of a message in context. In the present day, it continues to offer scholars a comprehensive yet simple explanatory model, and to provide trainers with an appealing pedagogical tool (Moser-Mercer, 1994a).

Yet interpretive theory has not met with universal approval, nor has it been adopted by all as a research paradigm. Other scholars, influenced by the scientific method, have chosen to build theoretical models that make assumptions which are based on the findings of formal experiments, and that make predictions which can be tested empirically (Moser-Mercer, 1994a). In the pages that follow, we will review a number of these empirically oriented models and examine how they explain and describe the interpreting process.

the participant is speaking, but because they are still searching for the meaning in context (p. 242).

3.2 Information Processing

The second school of thought in interpreting research – information processing – is an extension of mainstream research in the fields of psychology and neuropsychology. Starting in the 1950s, a number of experimental psychologists began investigating the limits of the human ability to divide attention (i.e., to consciously perform more than one task at a time). Initially, these experiments were conducted on samples of the general population (see Broadbent, 1957; Treisman 1964 and 1969), but it did not take long before researchers became interested in interpreters and their use of the simultaneous method of interpreting (see Oléron & Nanpon, 1965; and Treisman, 1965).

As the information processing school of thought gathered momentum, and as it investigated the work of professional interpreters in ever greater detail, the objectives of the researchers involved began to change. Where once the primary goal had been to use interpreting – a decidedly extreme example of human language behaviour – to shed light on the workings and abilities of the human mind in general (Moser-Mercer, 1994a), the purpose of the new research was to generate a greater understanding of interpreting in its own right. The approach remained decidedly psychological in its orientation, with researchers borrowing theoretical models of general language processing from mainstream psychology. However, because they applied components of the models to interpreting, the researchers' work is clearly positioned well within Interpreting Studies. The studies in question have provided us with insight into a number of aspects of the overall interpreting process, including attention (Lambert, Daró, Fabbro, 1995; Daró, Lambert, Fabbro, 1996); memory

(Daró & Fabbro, 1994); and cerebral lateralization (Fabbro, Gran, Basso, & Bava, 1990; Fabbro, Gran & Gran, 1991; Clarici, Fabbro, Bava, & Daró, 1994; Corina & Vaid, 1994; Clifford, 1999).

This is not to suggest that proponents of information processing have not been involved in modelling the interpreting process in a more comprehensive manner. There are, in fact, a number of full-process models developed by researchers interested in this school of thought. In the pages that follow, we briefly review five such models: Gerver (1976), Moser (1978), Cokely (1992), Gile (1997), and Paradis (1994, 2000).

3.2.1 Gerver

Gerver's (1976) intent in publishing his model was not necessarily to provide a true-to-life account of interpreting, but rather to demonstrate how then state-of-the-art knowledge of human information processing could be used to create a formal representation of the interpreting process. He took information that had been tested in experiments with non-interpreter subjects and created a framework with which to study interpreting in an empirical manner (Moser-Mercer, 1997). Our review is concentrated on two primary features of the model, its memory components, and its account of decoding and encoding.

Memory Components

The model Gerver outlines contains two types of memory components: permanent structural features, such as short-term buffer storage, long-term storage of lexicons and

grammar, and short-term operational memory; and control processes that determine the distribution of attention. These components are discussed below.

Buffer storage, Gerver argues, is a prominent feature of both the input and output subprocesses. The ability of interpreters to translate while continually receiving source-language information suggests that there exists a type of storage that can accept input, bit by bit, as it arrives, and that can discard input, bit by bit, after it has been successfully processed. Similarly, the fact that interpreters monitor and correct their own target-language production suggests that a buffer is part of the output process. Output is also stored incrementally, and discarded incrementally as the interpreter proceeds through the act of interpreting. To determine whether the translation of a given segment is successful, the interpreter compares the contents of the input and output buffers. This running comparison continues throughout the interpreting process. Gerver places a great deal of emphasis on the comparison of buffer contents. He cites research on unilingual speech recognition and hypothesizes that analysis through synthesis is part of the normal comprehension process in non-interpreters. According to the hypothesis, when receiving verbal input, the average person is actually engaged in mentally generating a similar utterance, at any number of linguistic levels (phonological, morphological, syntactic, etc.). The results of this internal synthesis are then held up against the incoming message, hopefully resulting in successful decoding. If this hypothesis is valid, Gerver reasons, self-monitoring during interpretation may be more than an added afterthought, as other scholars have argued. Instead, it may actually be the linchpin of the interpreting process.

The decoding of the source-language input and the encoding of the target-language output (see below) are dependent on the long-term storage of a lexicon and grammar for each of the interpreter's working languages. Between the perception of the source language and the production of the interpretation, the interpreter must consult an internalized store of lexical items, and of grammatical rules. The specific items and rules are then activated, and held in some sort of short-term storage, known as "operational memory", for the duration of the time that they are required.

Gerver posits that the interpreter's attention, although a limited resource, can be shared among various tasks, depending on the demands of each. For example, under normal conditions (good sound quality, moderate rate of delivery, familiar topics) interpreters can easily divide their attention in an effective manner. However, under more challenging conditions, attention might be focused on a single task (e.g., decoding or encoding), leaving others (input monitoring, output monitoring) to suffer.

Decoding and Encoding

In his discussion of linguistic decoding, Gerver considers the possibility that prediction – the interpreter's ability to make probabilistic inferences about the future development of a sentence on the basis of what has just been said – plays a significant role in comprehension. He admits that interpreters will sometimes predict what the speaker is about to say, but he dismisses prediction as an adequate basis for describing and explaining the interpreting process. He assumes that prediction is based on the ability to recognize "sequences of words" from past experience, and argues that there are too many potential sequences for the interpreter to predict effectively and consistently. Also, he claims that

prediction is incompatible with self-monitoring. If interpreters are able to correct themselves, he argues, it must be because they can compare their probabilistic inference of the speaker's words with what the speaker has actually said. He reasons therefore that comprehension must involve a more complete analysis of input than prediction would allow (1976, p. 195-196).

Gerver suggests that during the initial stages of comprehension, interpreters access an internalized "code book" to decode the phonetic representation of each source-language segment. This subsequently allows them to understand the segment's "surface structure" (a network of linguistic or grammatical relationships indicated by the order of words in a sentence) and its "deep structure" (a network of logical relationships among concepts as established by the subject, predicate, and object) ¹².

The actual transfer from the source language to the target language in Gerver's model seems to be a product of the change from (source-language) surface structure to deep structure, and back to (target-language) surface structure again. He suggests that seasoned interpreters may occasionally be able to take short cuts, if experience allows them to generalize that a given surface structure in language X invariably leads to another given surface structure in language Y. This allows him to assert that interpreters need to know the

¹² To further clarify the difference between surface and deep structure, Gerver cites two examples. In the first, he notes that the sentences "Caesar crossed the Rubicon" and "The Rubicon was crossed by Caesar" differ in terms of their surface structure, but they share the same deep structure. The grammatical subject of the first sentence is "Caesar", while in the second it is "the Rubicon"; however, the agent – the logical, conceptual subject – in both cases is "Caesar". In the second example, the sentence "Flying planes can be dangerous" shows how one surface structure can represent two deep structures: "To fly planes can be dangerous" and "Planes that fly can be dangerous". Here the ambiguity is due to the fact that "flying" can function either as a gerund or as an adjective.

rules for generating sentences in both languages, but that they also need to know the extent to which the two rule systems correspond.

At the time it was introduced, Gerver's model made an important contribution to Interpreting Studies. It demonstrated how outside knowledge, particularly from the natural sciences, could contribute to an understanding of interpreting, and it provided a springboard from which to launch empirical research. It also inspired a number of other scholars to develop their own accounts of the interpreting process.

3.2.2 Moser

Shortly after Gerver presented his work, Moser published her own model of simultaneous interpreting. In doing so, her intent was not to supplant her colleague's ideas, but rather to complement them. She describes Interpreting Studies as being in a "pre-paradigm" stage¹³ and argues that the appearance of another model is a result of the variety of work being done in the new discipline. The two models, she argues, place emphasis on different aspects of the same process, and in so doing, they deepen scholars' understanding of the subject matter at hand (Moser, 1978).

¹³ Although Moser made this observation in the late 1970s, Interpreting Studies is still characterized today by the lack of a single, widely accepted paradigm. Moser-Mercer (1994a) recognizes two competing schools of thought in the subdiscipline, and Pöchhacker and Shlesinger (2002) suggest that Interpreting Studies "has developed into a remarkably heterogeneous series of loosely connected paradigms" (p. 4).

Moser's model is based largely on an information-processing account of speech comprehension developed by Massaro (1975). It attempts to map out, in a detailed manner, the progression of understanding from the initial awareness of an acoustic signal, to the eventual internal representation of the speaker's message. Moser selected this particular account as the foundation for her model of interpreting because of its focus on detail; interpreting, she maintains, is a complicated matter, and it requires a complex explanation.

Central to Moser's model is a hypothesis about the organization of conceptual information inside the mind of the interpreter. She argues that concepts are stored in long-term memory in a form that is at once separate from language, yet that at the same time allows linguistic structures from either of the interpreter's working languages to be mapped onto it. Long-term memory is composed of a network of "conceptual bases", each one containing language-independent semantic information about a particular concept. Tightly connected to each base are a source- and a target-language node that contain related sensory, phonetic and syntactic information. When an interpreter receives information of this type – let us say the phonetic contour of a source-language lexical item – it activates the source-language node, and, through their shared connection, the appropriate conceptual base. In this manner, the interpreter is able to retrieve the semantic information related to the phonetic contour she has just encountered. In addition, because the conceptual base is also tightly connected to a target-language node, the relevant information about a corresponding target-language form is activated, and the stage is set for translation in the process.

Moser suggests that connections exist not only between languages, but between concepts as well. Each conceptual base is arranged within a cluster of conceptual bases, and

each cluster is in turn connected to others like it. When one conceptual base is activated, as in the example above, it is able to activate the other bases in its cluster. The interpreter consequently gains access to a number of related concepts, and to their source- and target-language grammatical forms. The effect can potentially be fairly widespread, if activation extends from one conceptual cluster to others around it.

Understanding Moser's depiction of this conceptual network makes the rest of the model easier to follow. The interpreting process begins in a component called the auditory receptor system, which detects the presence or absence of an acoustic signal. If one is detected, it is fed into preperceptual auditory storage. The phonological rules of the source language are retrieved from long-term memory and are applied against the input. At the same time, the auditory input activates information stored in the appropriate source-language conceptual nodes. Because the node is tightly connected to the corresponding conceptual bases, the relevant concept contained within the bases is also activated.

Next comes a subprocess known as primary recognition. The input is synthesized into a percept (i.e., a syllable), which is stored in synthesized auditory memory (SAM). This is followed by secondary recognition, in which a sequence of synthesized syllables is recognized as a word or words. At times, secondary recognition may take place purely on the basis of phonological information. At others, the interpreter may recognize words with the help of semantic and syntactic cues. If this is the case, the relevant rules of syntax and semantics must be retrieved from long-term memory.

As input continues to arrive and go through primary and secondary recognition, a larger, more abstract linguistic unit, such as a string of words or a phrase, is eventually assembled and temporarily stored in generated abstract memory (GAM). GAM is typically able to store 7 ± 2 chunks of information, regardless of their size. However, if other parts of the process require more resources (e.g., if secondary recognition is difficult because of a speaker's foreign accent, or because of poor listening conditions), interpreters may not be able to store as much.

Moser hypothesizes that differences in information processing observed between experienced and beginner interpreters may be explained by the fact that the chunks experienced interpreters store in GAM tend to be made up of entire phrases, while the chunks stored by beginners tend to comprise individual words. In both cases, the interpreters are making full use of GAM's resources (7 ± 2 phrases, or 7 ± 2 words), but the experienced interpreters manage to use those resources more efficiently. As the chunks are stored in GAM, the interpreter attempts to connect the language forms they contain with conceptual bases in long-term memory.

By the time word string processing is complete, the interpreter has not only understood the semantic meaning of the language forms she has encountered, but she has likely also formed much of the target-language paraphrase that will make up her interpretation. This is because in analyzing the language forms of the source language, she has activated the corresponding conceptual bases, and through them, the related target-language nodes. However, if the relations between nodes and concepts are not numerous (because the

interpreter is unfamiliar with the topic, because she is inexperienced, etc.), this places a greater demand on the interpreter's mental resources.

Moser has also built a mechanism for prediction into her model. Recognizing that interpreters will engage in time-saving strategies whenever possible, she allows for the possibility that they will at times attempt to predict what the speaker will say. Doing so allows them to discard the current input, since they already "know" what is coming in. Auditory reception still occurs, but all the other processing stages are skipped. Prediction is therefore a function of how fast an interpreter can activate conceptual relations, and how many relations he or she can activate. The more an interpreter knows, the more she can predict.

The last stage of processing is auditory feedback, during which interpreters hear their own target-language production. Output becomes input, in the sense that the interpreter's production is picked up and analyzed by the same components of the model that handle the source-language input. Moser hypothesizes that processing the two messages poses little problem for the interpreter, provided the resources of the GAM are not exhausted. However, if GAM is overburdened, processing of the target-language message stops at the level of SAM.

Although Moser's model was ostensibly constructed in the same manner as Gerver's – existing research on speech recognition in non-interpreters was used as a basis for creating a model of simultaneous interpreting – it differs from her predecessor's work in two important ways. First, Moser's account of the workings of the mind of the interpreter is extremely

meticulous. It explains each step in the process in fine detail. Second, the model manages to incorporate a system for linguistic decoding together with a mechanism that explains prediction. In so doing, Moser has recognized that comprehension is not always the result of a purely linguistic process. This is an idea that is developed further in a number of other models.

3.2.3 Cokely

Cokely's (1992) theoretical description of simultaneous interpreting, like Gerver's and Moser's before it, is categorized here under the heading of information processing because it is primarily concerned with mapping the internal, psychological workings of the interpreting process¹⁴. However, it differs from the other models in the category in three ways. First, Cokely uses the model to describe and explain sign language interpreting rather than spoken language interpreting. Second, Cokely's goal of constructing an accurate internal model seems at times matched with a desire to shed light on interpreter pedagogy –

¹⁴ The mapping of the internal processes like understanding is something that interpretive theorists have repeatedly dismissed. They reject attempts to analyze understanding into smaller components because the components cannot be detected through introspection. Lederer (1994) clearly states this position:

Certains auteurs (...) postulent deux étapes dans la compréhension des textes. La première consisterait à comprendre la langue du texte, la seconde à en « inférer » le sens à l'aide de connaissances extra-linguistiques. L'hypothèse d'inférences logiques, allant de la compréhension des significations à la déduction du sens, postule une opération en deux temps dont la réalité psychique ne peut pas être démontrée. Elle est sans doute inspirée des problèmes que l'ordinateur a eu à résoudre, mais n'est pas étayée par l'observation du comportement humain. La saisie du sens n'est pas le produit d'étapes successives mais d'une seule démarche d'esprit. On ne comprend pas un texte d'abord au niveau de la langue, puis à celui du discours, mais d'emblée au niveau du discours. Ce phénomène psychologique marque en fait toutes nos perceptions : en rentrant chez moi je ne vois pas d'abord *une* maison puis *ma* maison; en rencontrant *un* ami je ne vois pas tout d'abord *un* homme puis M. X. Le concept général (*maison*, *homme*) n'est pas perçu séparément et préalablement à la reconnaissance de telle maison ou de tel homme. Seul le sens particulier atteint suffisamment le conscient pour marquer le souvenir (p. 25-26).

each of the components in the model identify a point where interpreting can break down, and they point to a way in which interpreter trainers might intervene to avoid such problems. Third, Cokely is interested in providing a “sociolinguistically sensitive” model, one that describes how interpreters mediate not only between languages, but also between individuals and communities. These three differences will be further illustrated in the outline of the model that follows.

The chapter that contains Cokely’s model is entitled “Probable Miscue Causes”, and it describes seven components of the interpreting process, each of which can potentially lead an interpreter into difficulty. During *message reception*, a portion of auditory or visual SL input first reaches the interpreter. It then undergoes *preliminary processing*, where the phonological rules of the source language are applied, and meaningless signals are discarded. Small elements of the source-language input (such as lexical items) are identified, perhaps through the application of syntactic and semantic context. Next is *short-term message retention*, where the small elements are stored until further analysis (at the phrase, sentence and discursive levels) can process them into larger “chunks” of information. When analysis of the portion of input is complete, *semantic intent* is realized, and the interpreter arrives at some form of comprehension. However, the interpreter must still determine a *semantic equivalent*, since there are often differences between the SL and TL cultures that must be compensated for if meaning is to be conveyed successfully. During *syntactic message formulation*, the interpreter chooses the form of the TL message according to linguistic and cultural norms. In the final component, *message production*, the TL message is at last articulated, either in speech or in sign.

Cokely's most significant contribution comes perhaps in the form of his notion of semantic equivalence, which is overtly sociolinguistic. The search for equivalence, he argues, must be carried out with the awareness that different groups of people use language forms in different ways. To effectively bridge the gap between hearing and Deaf people, sign language interpreters must negotiate these differences well. To illustrate his point, he gives the example of a sign language interpreter who says the word "Deaf", and he suggests that most other hearing people will unfortunately understand "deaf". In other words, the interpreter's use of the lexical item is patterned after the Deaf community's use of the corresponding sign, which designates a proud linguistic and cultural minority. However, most hearing people will likely interpret the same lexical item as the designation of a group of disadvantaged and disabled individuals they have learned to pity (Cokely, 2001). It is the interpreter's role, Cokely argues, to prevent such misunderstandings by making lexical choices that convey the missing sociolinguistic information.

In advocating this position, Cokely is in fact focusing attention on a feature of interpreting that is not always explored fully or explained clearly in information processing – the communication of a linguistic message by individuals in a given situation. Like Moser before him, he seems aware that understanding meaning requires more than the analysis of the linguistic components in an utterance. Larger situational factors – the different social positions of the participants in the interpreted event is but one – must also figure prominently in an accurate depiction of simultaneous interpreting. In taking this stance, Cokely lays the foundation for the discursive approaches to interpreting that will follow.

3.2.4 Gile

So far, the information processing models reviewed have attempted to map out the mental processes of interpreting by examining current information on human language processing, and incorporating it into a coherent account of simultaneous interpreting. Gile's model (1997) has a slightly different origin, in that the author begins by observing interpreter behaviour, and then constructs an account of interpreting based on those observations.

In his previous research, Gile consistently noted significant numbers of errors in the performance of expert interpreters. This was true even of interpreters who had excellent professional reputations, and even when they worked in conditions that presented no obvious difficulty (e.g., noise, fast delivery in source language, technical complexity, complex syntax, etc.). Other observations Gile made were also interesting. When more than one interpreter worked on the same speech, the errors made by each did not occur at the same points during the speech. Also, when a single interpreter interpreted the same speech twice in succession, she often made errors the second time in segments that were handled correctly the first time around. This led Gile to hypothesize that the errors he observed were unrelated to the objective features of the speech in question, but rather that they were due to some built-in difficulty in the act of interpreting itself. In other words, the interpreters appeared to be facing a kind of cognitive management problem.

To model conference interpreting in terms of cognitive management, Gile describes it as a process that is made up of four distinct efforts.

1. The Listening and Analysis Effort (L)

Listening and analysis comprises all parts of comprehension, such as analysis of sound waves, identification of words, and final decisions about the meaning of a sentence.

2. The Production Effort (P)

The second effort includes all phases of production, such as initial mental representation, speech planning, and implementation of the speech plan.

3. The Memory Effort (M)

This is made up of all the demands placed upon short-term memory, by such factors as the lag between sound wave analysis and final comprehension, the lag between speech planning and speech implementation, and the tactical delays used by the interpreter (e.g., waiting for more context to compensate for unclear logic).

4. The Coordination Effort (C)

This is the effort required by the interpreter to coordinate cognitive capacity among listening and analysis, production, and memory.

In order for simultaneous interpreting to be successful, the capacity available for each effort must exceed the capacity required. The total capacity available must also exceed the total capacity required. However, since an interpreter's total available capacity is finite, it may at times be insufficient to perform all interpreting tasks correctly. As a result, errors occur.

In the course of outlining his model, Gile also discusses some of the factors that place an increased demand on capacity requirements. Chief among these is information density. Speech segments that contain a great deal of information require the interpreter to understand and produce more in a shorter amount of time. Information density is typically increased by the following:

1. Fast delivery of the speech;
2. Enumerations (which tend to be devoid of logical connectors);
3. External factors (e.g., poor sound); and
4. Prior written preparation (prepared speeches are delivered with fewer false starts and hesitations than spontaneous language production, and they have intonation patterns that are less “helpful”) (Gile, 1997, p. 205).

Information density may be the most frequent source of problems in an interpretation, and it therefore merits close attention in any account of the interpreting process.

The novelty of Gile’s model is twofold. First, it developed out of the direct observation of interpreter error, and it is one of the few to focus attention on performance limitations as a phenomenon in its own right. Second, the effort model is used not only to provide an account of the simultaneous method, but also of the consecutive method, sight translation (during which the interpreter reads a written source-language text and produces an oral target-language version), and “sight interpreting” (which involves the source-language producer reading aloud from a written text – albeit with occasional spontaneous

additions – while the interpreter, who has a copy of the text, produces an oral target-language version).

Gile's writings also offer evidence of the importance of factors that lie outside traditional linguistic analysis. While his description of the listening and analysis effort places a certain emphasis on linguistic segments (understanding is equivalent to the recognition of words, meaning is contained within the sentence, etc.), Gile also recognizes the role of suprasegmental features. For instance, he notes that prepared speeches are notoriously difficult to interpret, and he suggests that this is because they lack "helpful" intonation patterns. Other researchers have shown how intonation is used to shape oral discourse into structures larger than the sentence and to provide listeners with cues to help them understand meaning¹⁵. Although Gile does not explore these ideas in his discussion of the model, he does seem – intuitively at least – to recognize their importance.

3.2.5 Paradis

Like Gile, Paradis (1994, 2000) bases his model of simultaneous interpreting on observational data. However, the observations in question come not from the interpreting booth, but rather from the hospital examination room. Paradis turns to clinical literature – for the most part on aphasia – and uses a number of hypotheses that have been put forward to

¹⁵ Couper-Kuhlen (1997) has demonstrated how speakers use intonation to let listeners know whether they are voicing their own perspective or reporting the speech of others. She argues that speakers typically establish a vocal deixis in conversation, that is, a set of prosodic features – relative loudness, pitch, tempo and voice quality – that are understood to be their own conversational norms. A noticeable departure from these reference

explain bilingual language processing. The result is a loose framework that describes the interpreting process, and that is made up of five component hypotheses.

The first stages of interpreting, Paradis argues, can be explained by a hypothesis that distinguishes between *linguistic decoding* and *mnemonic encoding*. As the acoustic¹⁶ signal of the source message reaches the interpreter, the sounds of the utterance are kept in echoic memory for only a second or so. During this time, the signal is parsed into words, which are stored in short-term memory until subsequent levels (phrase, sentence, utterance, etc.) are also decoded. When decoding is complete, the acoustic form is discarded, and only the meaning is retained in short-term memory. Over time, the meaning may be transferred to long-term memory. This first hypothesis explains why, after a conference has ended, interpreters may be unable to retrieve an utterance verbatim, even though they can recall its meaning. Content, not form, is preserved. In this regard, Paradis' account bears a resemblance to interpretive theory.

When considering the interpreter's general ability to function in a bilingual environment, Paradis makes reference to the *subset hypothesis*. It argues that the interpreter's two languages are subserved in the brain by two subsystems that are part of a larger cognitive system. This makes the subsystems both distinct from one another, and yet interrelated at the same time. Each subsystem forms its own network of connections that can independently be activated or inhibited. (Paradis notes the cases of bilingual aphasics who

values will be heard as a shift in deixis; interlocutors will assume the speaker is animating another person through reported speech.

¹⁶ Paradis only considers spoken language interpreting.

suffer impairment in only one of their languages.) However, as language systems, the two subsystems more closely resemble one another than they do other cognitive systems in the brain. Because language is organized in this way, bilinguals can choose to use their languages together (as in code switching), or to use them one at a time.

The *threshold activation hypothesis* is useful in explaining how bilinguals avoid using their languages inappropriately. Paradis suggests that linguistic segments are stored in the brain behind individualized barriers – or “thresholds” – that when fully raised impede the use of the segments. Each time a segment is activated, its threshold is lowered. When activation is finished, the threshold begins to rise to its normal position. If a segment is activated frequently or recently, there is a greater chance that its threshold will be lowered, and therefore easier to access. A similar process affects entire language systems. When a normal bilingual (i.e., one without brain injury) speaks, the activation threshold of the selected language is lowered, while that of the non-selected language remains raised. Thus, the bilingual is able to use one language with minimal interference from the other. Paradis argues that threshold activation is precisely what makes the simultaneous method of interpreting so difficult – it requires the thresholds of both the source language (to decode the incoming message) and the target language (to encode the outgoing message) to be lowered at the same time. Untrained bilinguals generally avoid doing this, as it places great demands on the cognitive system.

A great deal of experimental research on bilingual language processing has examined the issue of cerebral lateralization; many authors have claimed that the first or dominant language (L1) and non-dominant language (L2) of bilinguals are represented more

asymmetrically (i.e., that there is greater right hemisphere involvement) than the L1 of unilinguals. The question of lateralization has also been extensively discussed with regard to interpreting (Fabbro, Gran, Basso & Bava, 1990; Fabbro, Gran & Gran, 1991; Corina & Vaid, 1994; Lambert, Daró & Fabbro, 1995; Clifford, 1999). Paradis also wades in on the discussion, by making an important distinction between *linguistic* and *pragmatic competence*. Clinical research has repeatedly demonstrated that implicit linguistic competence – which Paradis defines as the “decoding and encoding of phonological, morphological, syntactic and lexical properties” (1994, p.328) – is subserved by areas of the left hemisphere, and that this is true for unilinguals and bilinguals alike. Similarly, it is well established that pragmatic competence – which Paradis defines as the “interpretation of what is implicitly meant, though not explicitly said, such as inferences from general knowledge and situational context, affective prosody, and other paralinguistic features...” (1994, p. 328) – is subserved by areas of the right hemisphere. Moreover, the greater amount of right hemisphere activation in bilinguals occurs when they are using their L2. Paradis explains the lateralization conundrum by suggesting that when unilinguals or bilinguals process communication in their L1, they make use of a grammar that is “hard wired” into the brain. They do not need to make extensive use of pragmatic competence to understand the message. In contrast, when bilinguals process communication in their L2, they have more difficulty using purely linguistic resources to reconstitute the message, and must rely more heavily on pragmatic competence. The supposed differences noted by other researchers, Paradis argues, are the results we would expect to see from languages acquired with different levels of completeness.

The final hypothesis Paradis explores is one that makes a distinction between *implicit language competence* (ILC) and *meta-linguistic knowledge* (MLK). ILC is acquired incidentally, without conscious knowledge of its contents, and it is used automatically. Bilinguals develop ILC through interaction with speakers of the language in situational contexts. MLK is learned consciously, is available for conscious recall, and is applied in a controlled manner. Bilinguals, for the most part, learn MLK in school.

With the framework in place, Paradis then proceeds to model the interpreting process. The interpreter breaks the incoming signal down into “chunks” (syntactic phrases and/or semantic units) that are held temporarily in echoic memory. The chunks are decoded linguistically to arrive at their meaning, at which point they are encoded in the target language. The target language chunk is produced orally, and picked up by the ear of the interpreter. The target language chunk is held temporarily in echoic memory, monitored for correctness, and decoded to derive its meaning. The meaning of the target language chunk is compared with that of the original source language chunk. Processing takes place in an ongoing loop; while the first chunk is produced in the target language, a second chunk is being encoded in the target language, a third chunk is being decoded from the source language, and so on.

Paradis’ work makes several contributions to the study of interpreting. First, it promotes an exchange of ideas between Interpreting Studies and an outside discipline not previously considered by information processing – neuroscience. Because his model is derived from clinical observations of brain-injured patients, it is unlike the others reviewed thus far, and it offers interpreting scholars a fresh perspective on a familiar topic. Second, the

information Paradis reviews offers new insight into the differences between what he terms linguistic and pragmatic competence. The studies he cites have concluded that linguistic tasks (the analysis of phonemes, morphemes and other linguistic units) and pragmatic tasks (the analysis of suprasegmental features that allows us to understand meaning in context) are processed in different hemispheres of the brain. This in turn suggests that there is a neuroanatomical justification for distinguishing between linguistic and discursive abilities.

The information processing models that we have reviewed in this section were responsible for introducing several important ideas to Interpreting Studies. They have attempted to peer inside the “black box” that interpretive theory avoids¹⁷ (Setton, 1999) and map out the inner workings of the mind of the interpreter. They have offered us detailed descriptions of source-language comprehension, of target-language expression, and of the resources needed to carry out these tasks simultaneously. Proponents of information processing, far from being content with introspective, intuitive accounts of the interpreting process (Moser-Mercer, 1994a), have demonstrated how Interpreting Studies might borrow a page from the sciences and conduct research empirically and with the rigour common in experimental psychology and neuroscience.

¹⁷ Setton (1999) argues that interpretive theory has neglected to offer an account of “intermediate representation” (cognitive processes beyond language recognition). This account has been attempted by the proponents of information processing, although the results have not been entirely satisfactory. As he explains, “The [interpretive theory] model deliberately leaves the intermediate stage of cognitive processing unspecified, in contrast with the [information processing] theories which see a conceptual-semantic network integrated with a cross-linguistic lexicon, through which the interpreter must find her way” (p. 43).

However, in constructing and testing models through strictly controlled experimentation, the proponents of information processing often designed studies that investigated discrete parts of the overall process, and that consequently bore little resemblance to interpreting as it is actually practiced (Seleskovitch, 1975). The information processing literature often examined the interpreter's ability to carry out linguistic tasks¹⁸, and in the process, it often appeared to lose sight of the interpreter's role as a communicator in a very special context (Setton, 1999). It is hardly surprising then that, as we have seen, the processing of discursive features of communication has only been incorporated into the information processing models in a limited way.

This limitation is not typical of all the theoretical models available in the literature. In fact, a recent trend in the research suggests that a new school of thought may be forming, one which pays explicit attention to the study of discourse and its relevance for professional interpreting. To illustrate this growing trend, we have reviewed a number of discursive models of the interpreting process in the following section.

¹⁸ Several studies are illustrative of this tendency. In the first part of their experiment, Fabbro, Gran, Basso and Bava (1990) asked interpreter and non-interpreter subjects to say the names of the days of the week out loud. In the second part, interpreter subjects were instructed to interpret either "word for word" or in a "semantic" manner. Corina and Vaid (1994) asked their subjects to interpret a list of individual English words and ASL signs. Daró, Lambert and Fabbro (1996) had subjects in their experiment interpret "texts" made up of a series of sentence groups, where each group contained four sentences that shared a common topic. None of the researchers asked their subjects to interpret the kinds of utterances they regularly encounter (e.g., conference speeches, doctor-patient dialogues, etc.).

3.3 Discursive Models

Before we begin our review of the final set of models, we believe it is important to clarify what we mean by *discourse* and to explain how it relates to other perspectives on language. Discursive research is generally understood to be a response to the perceived shortcomings of structuralist (Saussurian) and theoretical (Chomskyan) linguistics. The former, discourse analysts would assert, views language as a set of relationships between language forms (“*signifiants*”) and their corresponding concepts (“*signifiés*”). The latter argues that linguists should study the tacit knowledge of the abstract rules of language (“competence”) (Schiffrin, 1994). Discourse analysts choose instead to understand language as a system integrated with the speaker’s knowledge of the world and society, and they assert that the system should be described in linguistic, cognitive, and social terms, and in terms of the conditions under which speakers use it (de Beaugrande, 1996).

In her seminal work on the topic, Schiffrin (1994) goes to great lengths to define discourse, and to describe the nature of discursive research. She does so by highlighting issues of both form and function. The formalist perspective, she explains, views discourse as “language above the sentence or clause”, and those who adopt it tend to be interested in investigating large structures of language, such as paragraph or text organization, question-answer pairs, and intonational contours. The functionalist perspective views discourse as “language use in context”, and its proponents focus on the effects language users and the contexts of language use have on language itself (Schiffrin, 1994).

The author also points out that this dichotomous definition of discourse, while providing a concise introduction to the topic, fails to describe the vast range of research that may be considered discursive, and to provide clear examples of what that research might entail. To resolve these shortcomings, she devotes considerable attention to outlining six approaches that have characterized discursive research, and to identifying the researchers who have been instrumental in forming them. Below, we have listed the six approaches to discourse as they are categorized by Schiffrin and provided a brief description of each. (Expanded descriptions of the approaches, and references to seminal works for each, are provided in *Appendix I*.)

1. *Speech Act Theory**

Speech act theorists study the underlying conditions for production and interpretation of acts through words.

2. *Interactional Sociolinguistics**

Scholars engaged in this approach to discourse argue that people from different cultures may share grammatical knowledge of a language, but may differ in the background knowledge they use to make inferences about meaning, such that they understand different things about the same utterance.

3. *The Ethnography of Communication**

Proponents of this approach examine the tacit social, psychological, cultural, and linguistic knowledge governing appropriate use of language.

4. *Pragmatics**

Pragmatists assert that people work with minimal assumptions about one another, and that those assumptions form the basis from which they draw inferences about one another's intended meanings¹⁹.

5. *Conversation Analysis**

Conversation analysts seek to discover the methods by which members of a society produce a sense of social order through conversation.

6. *Variation Analysis**

Variationists posit that language variation is patterned both socially and linguistically, and that such patterns can be discovered through systematic investigation of a speech community.

For the purposes of the present study, we have defined discourse in the following way: it is a system that combines linguistic knowledge with a practical understanding of the real world and of the ways in which people interact in it. This system is used by producers to negotiate communication with receivers in specific settings. To study discourse, we must therefore expand our focus beyond what is traditionally considered linguistic, and this expansion may take place in one of two ways: 1) formally, by examining the language units

¹⁹ Schiffrin's discussion of *pragmatics* is limited to the work of H.P. Grice and of those inspired by him. Her definition of the term is therefore less inclusive than that of other scholars, who define *pragmatics* in much the same way we have defined *discourse* in the present study.

beyond the level of syntax, units which we believe are socially and culturally defined; and 2) functionally, by examining all language as it is used by producers to communicate with receivers in particular situations, for particular purposes. This second expansion will often question “meaning” as the sole property of language in isolation and require us to consider other potential sources.

We believe that the six approaches outlined by Schiffrin, together with the *Bakhtinian approach**, provide clear examples of discursive research. We have therefore chosen models of the interpreting process inspired by these approaches for inclusion in this section. There are three such models: Wadensjö (1998), Roy (2000), and Setton (1999).

3.3.1 Wadensjö

Wadensjö’s (1998) study of spoken language interpreting in the courtroom and in the community was one of the first to adopt a discursive approach, and it stands in contrast to some of the more traditional views in Interpreting Studies. At the time she was writing, most work in the discipline was concerned with explaining how interpreters analyzed message content in the source language and conveyed it in the target language. The result was a view of interpreting as a text-like, unidirectional process of transfer, from producer to interpreter to receiver. Wadensjö suggests that this traditional view may be appropriate for conference interpreting, where opportunities to participate in the communication event are regulated and limited, but that it is insufficient to account for the complexities of interpreting in the courtroom and in the community.

The author then sets out to explore alternative approaches, and she finds a promising one in the work of Mikhail Bakhtin. Like the work of other discourse-oriented thinkers, Bakhtin's writings on communication are in many ways a response to contemporary linguistic thought. He deplores the fact that linguists regard language from the restrictive standpoint of a single, isolated speaker and focus their attention exclusively on units of language such as words, phrases, and sentences. The role of language in communication is considered secondary, and if mentioned at all, it is typically explained in some graphic or schematic representation that shows the transmission of a message between a speaker and a listener. Bakhtin points out several problems with this view. First, to talk of a speaker presupposes not only the existence of a language system that the speaker is using (a point emphasized by traditional linguists), but also the existence of communication between participants in a sphere of human activity (a point neglected by traditional linguists). Second, the typical representation of communication fails to note that the listener is not perpetually condemned to a passive role – she eventually responds to the speaker – and that the speaker himself is a respondent to what has been said before (Bakhtin, 1986). The study of language has most often been conducted as though the speaker were delivering a monologue, but a more accurate approach would see the speaker for what he truly is – a participant in an ongoing dialogue.

As part of his “dialogical” approach, Bakhtin argues that scholarly attention should focus not on the units of language so favoured by linguists, but rather on the *utterance*. He defines the utterance as a unit of speech communication, the boundaries of which are determined by a change in speaker. All utterances, regardless of their length, have a clearly

definable beginning and end: their beginning is preceded by the utterance of the speaker who came before, and their end is likewise followed by the utterance of the speaker who comes after. As a result, “[a]ny utterance is a link in a very complexly organized chain of other utterances” (Bakhtin, 1986, p. 69). It necessarily enters into a relationship with preceding utterances, because it builds on them, because it polemicizes with them, or because it simply presumes that they are known to the listener.

Each new utterance is not only connected backwards in time, to the utterances that preceded it, but forwards in time as well. It is created with a particular listener in mind, and it is constructed in a way that takes into account the reactions that the listener may have. One of the markers of the utterance, therefore, is its *addressivity* – its quality of being directed at someone (Bakhtin, 1986).

The traditional focus on language units, Bakhtin asserts, also obscures our understanding of meaning. When we consider a sentence isolated from communication – “The sun has risen,” for example – we can say that it is comprehensible, because we understand its language meaning. However, we only grasp the sentence’s true sense when it is used in context, in what Bakhtin refers to as the “whole of the utterance” (1986, p. 82). Only when we see the sentence’s role in the utterance do we understand the speaker’s motivation for producing it. Consider the same sentence in two different utterances:

1. “The sun has risen. It’s time to get up.”
2. “The sun has risen. But it’s still very early. Let’s get some more sleep.”

Each utterance has a different meaning and will elicit a very different response from the listener, and yet each one contains the same initial sentence. Clearly, to understand an utterance, we need to know the neutral dictionary meanings of the words of a language, because they provide us all with a common ground. But we also need to know how the words are used in live speech communication to fully comprehend. Bakhtin explains this necessity by stating that words exist for the speaker in three aspects: 1) as units of a language in their neutral dictionary form; 2) as the words of another, belonging to another, and filled with echoes of another's utterance; and 3) as the speaker's words chosen in a particular situation, with a particular purpose in mind, and imbued with the speaker's expression²⁰.

For Wadensjö, Bakhtin's ideas provide an efficient tool for understanding how prior research in Interpreting Studies, based largely on conference interpreting, has explained the interpreting process, and how new research, based on court and community interpreting, might explain the process differently. Using Bakhtin's terminology, she contrasts the traditional "monological" approach (which, as she claims, views "talk as text") with a new "dialogical" alternative (which sees "talk as activity").

She points out a number of important differences between the two, particularly when it comes to explaining 1) what language use is; 2) what people are doing by talking; 3) what an utterance is; and 4) where meaning comes from. When we observe language use from a

²⁰ There are some parallels between Bakhtin's ideas and interpretive theory. Bakhtin's talk of "neutral, dictionary" forms roughly corresponds to Seleskovitch and Lederer's explanation of "*signification*", and Bakhtin's discussion of references to prior and future utterances is reminiscent of Seleskovitch's claim that a conference participant's speech is both a response and an invitation to respond. However, Bakhtin's assertion that the conversation is the archetype of speech communication, and his emphasis on the negotiation of meaning distinguish him from interpretive theory, which sees "*sens*" as monolithic, produced by the speaker's

monological perspective, we tend to see it as a producer's creation of different types of texts; adopting a dialogical outlook, we are likely to instead focus on the interplay between people, which occurs in the midst of other human activities. Through the monological lens, we see producers using the tools of language (lexicon, syntax, prosody, etc.) to create meaning. From a dialogical viewpoint, we see that the participants in a conversation are trying to achieve their own goals within a particular situation (an immigration officer uses an interview to assess an applicant's truthfulness; the applicant uses the interview to get the officer to approve a request). Monologically, an utterance is a language unit made up of smaller units of meaning (phonemes, morphemes, lexemes, etc.). Dialogically, an utterance is part of a situated interaction, and each one makes sense to the people involved. Meanings, from the monological perspective, are the properties of lexical items, and sequences of lexical items, chosen through the intention of the speaker. From the dialogical perspective, meanings are continually established and re-established between people in an encounter.

Face-to-face interpreting, Wadensjö argues, is best explained through a combination of the two views. The data she has collected seem to indicate that interpreters regularly encounter two types of problems in their work. On one hand, they must bridge linguistic gaps between the two languages in use; on the other, they must bridge the social gaps between two or more language users in the encounter. As a result, interpreters' utterances realize two different functions – translation and coordination – and the author has created a taxonomy that describes each in detail.

intent, decipherable in its entirety by the listener.

Taxonomy of Translation Functions

1. Close Renditions

The propositional content in the interpretation is equally found in original utterance, and the style is approximately the same.

2. Expanded Renditions

The interpretation includes more explicitly expressed information than the original utterance.

3. Reduced Renditions

The interpretation includes less explicitly expressed information than the original utterance.

4. Substituted Renditions

The interpretation contains some reductions and some expansions.

5. Summarized Renditions

The interpretation corresponds to two or more original utterances.

6. Multi-Part Renditions

The interpretation corresponds to one original utterance, but the interpretation is split into parts by other, interjected originals that are not translated.

7. Non-Renditions

An utterance of this type is the interpreter's initiative; it does not correspond to an original utterance.

8. Zero Renditions

The original utterance is left untranslated

Taxonomy of Coordination Functions

1. Implicitly Coordinated

Interpreters coordinate the conversation simply by taking their regular turn.

Their turns are normally designed to prepare an addressee to receive an utterance, or to elicit an utterance from an addressee.

2. Explicitly Coordinated

These utterances are coordinating non-renditions (i.e., they are initiated by the interpreter) that can be divided into two categories.

a. Text Oriented

This category includes requests for clarification, requests for time to translate, requests to stop talking, and comments on translations.

b. Interaction Oriented

This category includes requests to observe turn taking, invitations to start or continue talking, and requests for solicited but not yet provided information.

In outlining the two functions, Wadensjö has underlined the importance of traditional research, and at the same time drawn attention to an aspect of interpreting that was previously neglected. The author argues that by expanding their focus beyond the text to the wider situation, interpreters may be in a position to better serve their clients. She cites the example of a paediatrician working with a young boy through an interpreter. The boy had been injured by a bomb blast, and the doctor was trying both to determine if he had any recollection of the event and to establish a rapport with the boy. However, the translation of the doctor's questions sparked a conversation among the interpreter, the boy, and his mother. The conversation was eventually summarized for the doctor, and details about the bomb were relayed. Unfortunately, the doctor had no way of telling who the source of the information was, and an opportunity to make a connection with the boy was short-circuited. Interpretive theory and information processing might have seen this instance of interpreting as successful, since the interpreter efficiently relayed message content between the two languages. Yet in discursive terms, the encounter was unsuccessful, because a participant's main goal was repeatedly thwarted.

Wadensjö's contribution to Interpreting Studies is noteworthy. While research on conference interpreting is slowly gaining momentum, there is still a paucity of information on court and community interpreting. The author's recording and observation of interpreting in the courtroom and in the community has helped to address serious gaps in our understanding. In addition, through the application of a Bahktinian approach to her work, she had underlined the importance of discourse to the study of interpreting. Her taxonomies are a first step in building a discursive model of the interpreting process, and she has sparked interest in issues beyond the word, beyond the sentence, and into the larger situation at hand.

3.3.2 Roy

As we have already seen, discourse is defined in formalist terms as “language above the sentence”. However, this definition begs an important question: what exactly is language above the sentence? Finding an answer, with regard to written language at least, is relatively straightforward. We can all point to a number of possible structures, such as paragraphs and entire texts, that would be worthy of study in discourse analysis. Yet with the spoken word, matters are more complicated. This is for two reasons: 1) it is difficult to identify a spoken-language unit beyond the level of syntax; and 2) even the sentence itself is suspect as a stable unit of analysis. Spontaneous utterances are rarely shaped into the forms dictated by formal grammars. They are instead defined by features such as intonation patterns, completeness of thought, relation to other utterances (e.g., question and answer pairs), and communication function (e.g. telling a story).

It is not surprising then that Roy’s (2000) discourse analysis of sign language interpreting avoids a discussion of the sentence altogether and focuses instead on another unit – turns in a dialogue. In her study, the author records and observes a number of one-on-one interpreted encounters, with the goal of analyzing how the participants use language to create, distribute, and maintain turns. Her focus is on the participants as members of different cultures, on the cross-cultural communication between them, and on the interpreter’s contribution to the management of the interaction.

Roy sees the interactive aspect of interpreting as key. Traditional categorizations of interpreting – made on the basis of such criteria as institutional setting (educational, medical,

legal, etc.) language group (ASL, Spanish, etc.), mode of interpreting (simultaneous or consecutive), number or type of participants (conference, community, liaison, platform, etc.)²¹, or social occasion (wedding, graduation, etc.) – have all ignored interactivity. This leads the author to argue that perhaps the most important distinction to observe in interpreting is that between a single-producer discourse event and a conversational discourse event.

In a single-producer discourse event, there is minimal turn taking, and a more intense focus on message content. The topic at hand is chosen by the producer, who adds features that guide receivers' interpretations and that create interest in the content. The producer wields a great deal of control over the event, deciding when to start and when to stop, and when (or even if) to answer receivers' questions. Receivers are thus unable to take an active conversational role. As might be expected, the study of interpreting in this kind of event centres on the transfer of message content.

In a conversational discourse event, transmitting information is only one of the participants' goals. They are also trying to achieve more interactional objectives (e.g., to obtain a loan by convincing authorities of trustworthiness, to protect a lending institution's interests by assessing risk, etc.). The encounter typically involves a small number of participants, and changes in topic are frequent. The event is also characterized by turn-taking, and it is usually shaped into utterance pairs (questions and answers). Participants are not on equal standing, and the roles that they are cast in are very influential. Commands,

²¹ The terminology used here is Roy's.

apologies and compliments underline much of what is said. The study of interpreting in this kind of event stresses the interpreter's active management of communication. It is this form of interpreting that is the subject of Roy's study.

To guide her analysis, Roy constructs a theoretical framework. She briefly outlines three currents of research in discourse analysis – *interactional sociolinguistics*, *conversation analysis*, and *ethnography of communication* – and explains why each is helpful in understanding the interpreting process.

Interactional sociolinguistics, as portrayed in the work of John Gumperz and Erving Goffman, focuses on the interplay between language, society, and culture. It argues that producers and receivers with similar cultural backgrounds share a number of *contextualization cues*, aspects of surface form attached to message content that function as a way to understand what was said. Contextualization cues help receivers to determine which comments are jokes and which are indications of anger, which information is foreground and which is background, and what the relationships are between comments. In observing communication, interactional sociolinguists must ask themselves a number of questions. What was the utterance? How was it understood by the people who heard it? How did the participants arrive at their interpretation? What grammatical knowledge do they have? What socio-cultural knowledge do listeners rely on to understand the message? For Roy, this approach is important because it suggests that receivers in an interpreted encounter are continually assessing contextualization cues – which because of cultural differences may or may not be familiar to them – and attempting to construct appropriate responses. Part of the

interpreter's responsibility is to be aware of cues that are not shared between participants and to steer the communication away from misunderstanding.

Conversation analysis, an application of Harold Garfinkel's ethnomethodology to the study of conversation, argues that all conversations are rule governed, and that the mechanisms underlying producer-receiver coordination can be studied empirically. One of the goals of conversation analysis is to identify the aspects of conversation that are context free, so that we can better understand the underlying structure of turn-taking. Analysts attempt to make generalizations about "typical conversational activities". For example, some have stated that turn-taking in a conversation is organized in such a way as to oblige producers to demonstrate an understanding of the previous turn in their own contribution (e.g., by answering a question that was asked, by mentioning a point that was raised, etc.). For Roy, conversation analysis is a useful tool, in that it suggests that it is possible to identify the features that are common to all interpreted encounters.

The ethnography of communication, based largely on the writings of Dell Hymes, posits that culture is a general "world view", a set of assumptions and beliefs that orient how people think, feel, and act. It also suggests that people communicate this world view through language. A key concept within this approach is the notion of *communicative competence*. To communicate effectively, it is not enough for producers to have explicit knowledge of grammatical rules – otherwise known as *linguistic competence*. They must also have knowledge governing the appropriate and meaningful use of language; they must also have *communicative competence*. It is this knowledge that allows producers to engage easily in conversation, to tell interesting stories, to relay pertinent information, and to present effective

arguments. Yet communicative competence is not universal. Behaviour that is considered evidence of communicative competence in one culture is not necessarily considered evidence of the same in another. It is up to the interpreter to understand what constitutes communicative competence in both cultures represented, in order to navigate successfully between them.

Cross-cultural communication in an interpreted encounter is further complicated by the fact that the participants are not on equal standing. Typically, the meeting is conducted according to the rules of only one of the communities represented, which means that one of the participants is cast in the role of member of the majority culture, while the other is cast in the role of member of the minority culture. This fact influences the way the encounter plays out in three ways. First, it is normally the case that discourse events in the majority culture are only slightly familiar to the minority members. Second, even if minority members are somewhat familiar with discourse events in the majority culture, there is always the possibility – particularly with ongoing encounters – that some unfamiliar aspect or expectation of the majority culture will come to light. Third, the minority member may be familiar with a “typical” discourse event in the majority culture, but may not know how to contend with variation in that event. In each of these situations, the minority member may be unable to decide how to behave “appropriately”. Responsibility then falls to the interpreter to mediate the event.

What all of this suggests – and this is a point already raised by Wadensjö – is that interpreters do more than transfer the language content of messages. They are also responsible for negotiating the way messages are understood by participants in the encounter.

Roy's contribution to the discipline lies in the fact that she has demonstrated that discursive issues are central to an understanding of sign language interpreting, and that she has laid the groundwork for a thorough model of the interpreting process as it is carried out in a different mode.

3.3.3 Setton

Setton (1999) describes his account of interpreting as a cognitive-pragmatic analysis, and this epithet indicates its hybrid nature. In many ways, Setton's work is a successful marriage of the three schools of thought reviewed thus far in the present study. Like the interpretive theorists, Setton underlines the importance of observing interpreting the way it is actually performed, and he makes a clear distinction between language and contextual variables as sources of meaning. Like the proponents of information processing, he consults state-of-the-art information in the cognitive sciences as a foundation for his work, and he ensures that the components of his model are both operationalized and verifiable. Like the discursive model builders, he relies on one of the approaches we earlier identified as discourse-based – in his case, Gricean pragmatics – and his model clearly describes how meaning can be derived from sources other than the linguistic units of the source language.

Before reviewing the model itself, we felt it important to describe the frameworks, pragmatic and cognitive, that make up its foundation. Consequently, we have briefly outlined Setton's use of theoretical research in three areas: relevance theory, mental models, and frame semantics.

1. Relevance Theory

Using Grice's work (1975) on the cooperative establishment of meaning as a starting point (see the entry under *pragmatics* in the Glossary), relevance theory (for examples, see Blakemore, 1987; Sperber & Wilson, 1995) argues that communication should be explained in cognitive terms. Specifically, human understanding should be seen as an attempt to obtain the greatest cognitive effect with the least processing effort. To do this, individuals must focus their attention on the most relevant information. Certain communicative features in language, such as focus, prosody, and optional word order, direct receivers to the relevant context so they can process new information and infer meaning

2. Mental Models

The work of Johnson-Laird (1983) provides a hypothesis about how meaning is organized in working memory. It suggests that in thinking, humans create temporary, internal representations of the world around them. This creation can take place in a number of stages. It may begin, for example, with the decoding of linguistic information (e.g., phonemes) and the construction of a propositional representation. There may be a number of intermediate representations, but the final construct is a full-scale mental model – which requires the recognition of correspondence between internal representations and external events.

3. Frame Semantics

Fillmore (1985) argues that understanding is not a linguistic process, but rather a cognitive one. He takes issue with the belief that meaning is language internal, and suggests instead that to understand the significance of words in language requires the construction of conceptual structures he calls *frames*. For Fillmore, frames “provide the background and motivation for [the] existence [of words] in the language and for their use in discourse.” (Fillmore, 2001, ¶ 1) The processing of lexical units in an utterance evokes their corresponding frames, leading in turn to comprehension.

What these three theoretical frameworks share is a rejection of the notion that meaning can be found in language. Instead, they assert that meaning is located in the receiver’s head, and that it is constructed from the interplay of grammar, lexicon, *and* inferences from context.

With these frameworks as a starting point, Setton constructs his model of the interpreting process. Like the information processing models that precede it, Setton’s work is cognitive; it is primarily concerned with mapping internal processes. However, taking a cue from interpretive theory, he also stresses the importance of analyzing producer’s intent in context. Like the other discursive models, Setton’s makes contextual factors explicit by building them into a certain number of the model’s components. The model can roughly be divided into two segments. The first accounts for the comprehension of SL input, while the second maps out the production of TL output.

At the beginning of the comprehension process, the interpreter receives different types of perceptual input, including information on the immediate physical environment, on the prosodic contours of the SL utterance, and on its phonetic content. The phonetic content undergoes preliminary decoding – in the form of word recognition – before it is sent to a component called the *assembler*, which itself contains an important subcomponent known as the *parser*.

The parser is capable of recognizing individual words as belonging to larger strings (propositions), and of deriving meaning from them. However, the meaning in question is only that which follows from linguistic knowledge. When the syntax of a proposition “proposes alternative analyses”²², the parser is incapable of judging among them. That task falls to the assembler, which is assisted in the endeavour by another component – the *mental model*.

Located in working memory, the mental model uses general world knowledge and specific situational knowledge to carry out primary pragmatic processing. It creates a representation of the “who”, “what”, and “whom” in a given situation, by providing a set of probable referents along with information on the current discourse topic. This representation is fed to the assembler, which uses it to build an *elaborated proposition* (a linguistic string informed by contextual information), and perhaps even to predict possible referents in the following proposition.

The elaborated proposition is at last presented to the *executive*, the final major component in the comprehension process. The executive has a number of responsibilities, chief among which is secondary pragmatic processing. It uses the rules of Relevance Theory to make assumptions about the producer's intent, and then judges the match between that intent and the elaborated proposition.

As it solicits information from memory stores, one of these components – assembler, mental model, or executive – may also evoke a *frame*. By tapping into a single word or concept, it likely accesses a cluster of concepts linked in memory by association. As a result, the system is provided with even more contextual information with which to decode the incoming message.

At the beginning of the production process, the executive is once more called into play. Another of its responsibilities is addressee orientation – it uses its judgement to determine whether to include or exclude, highlight or tone down SL textual material in the TL production. It thus begins the planning of the communicative intention and speech act of the eventual output.

Actual formulation of the TL production is carried out by two components. The first, the *formulator*, is responsible for microplanning the TL units. It selects word order and focus, based not only on the SL input, but also on the TL output as it has been realized thus far. The second component, the now-familiar parser, carries out the encoding of TL

²² Setton does not provide an example of such an instance, but the sentence cited by Gerver ("Flying planes can

production by retrieving items from the lexicon. It then assembles the items into phrases and subphrases, with appropriate regard to syntax and semantics.

Finally, the model also allows for the possibility of a more direct translation route. There is a path in the model that leads from word recognition and SL lexicon, to TL lexicon and TL parsing that would explain what Setton refers to as “mixed-level processing”. In other words, like the interpretive theorists, he allows for a one-to-one substitution of forms between SL and TL that does not call upon a deeper level of processing.

The accuracy of the cognitive-pragmatic model, Setton argues, can be seen in the ability of conference interpreters to go beyond the information contained within the producer’s original utterance and to supplement their understanding with inferences based on real-world knowledge. In so doing, interpreters are able to compensate for contrasts between the source-language and target-language structure. The experimental study in Setton’s book examines interpreting between English and German, and between English and Chinese. In both cases, the differences in structure between the language pairs would serve as an obstacle to interpreters if they were not in some way processing discursive features. For example, working from German to English would require interpreters to wait for the verb at the end of a clause (e.g., in a dependent clause or with a complex verbal construction) before being able to process meaning and produce the English translation. This is clearly not what actually happens; interpreters work effectively despite such differences in source-language and target-language syntax. (On this same point, see also Seleskovitch & Lederer, 1984, p. 122.)

be dangerous”) illustrates Setton’s point adequately.

The most convincing argument Setton offers in support of his cognitive-pragmatic model has to do with improved interpreting ability. The information processing school has long suggested that the key to training interpreters lies in helping them use cognitive resources effectively. Juggling a variety of language tasks simultaneously, they contend, is a complex skill. If interpreters are to improve their ability, they must learn to better meet the demands of simultaneous processing. Setton's research explains improved performance differently. He argues that there is a limit to how well interpreters can allocate the resources needed to process simultaneous tasks. Instead, most improvements in interpreting ability are better explained by an increased ability to manage discursive features. With training and experience, interpreters learn to use relevance cues, mental models, and the information from semantic frames to interpret better.

Setton's model makes an important contribution to Interpreting Studies in three ways. First, unlike the interpretive theorists, he peers into the "black box" in the mind of the interpreter and attempts to model intermediate representation in a formal manner. Second, unlike the proponents of information processing, he offers a cognitive account of interpreting that explains the interpreter's incorporation of clues in the physical environment and of real-world knowledge. Third, unlike the other proponents of a discursive approach, he provides a full-fledged discursive model, which explains and describes spoken-language interpreting in the conference setting. Together, these three contributions make Setton's model one of the most comprehensive in Interpreting Studies (Pöchhacker & Shlesinger, 2002).

3.4 Conclusion

Throughout our reading of the various models, it appeared to us that a number of components were mentioned repeatedly, and, as a result, a relatively cohesive view of interpreting began to emerge from the review as a whole. We were able to identify three common components, which, while not being a part of every model we examined, were nevertheless representative of most models overall. The three components are enumerated below.

First, all the models included the analysis of linguistic segments in one manner or another. The interpretive theorists touch on the analysis of language at two points in the material we reviewed. They describe *reception* as the decoding of source-language signifiers, which are assembled into grammatical structures. They also explain *transcodage* as an operation conducted on language forms, without any sort of critical analysis. For their part, the proponents of information processing place a great deal of emphasis on the analysis of linguistic units: recall Gerver's description of surface and deep structures, and Moser's detailed account of acoustic signals, phonological rules, perception of syllables, word recognition, and syntactic strings. Among the discursive model builders, Wadensjö argues that interpreting is in part a monologic exercise carried out on language forms, and that interpreters must bridge linguistic gaps between the participants in an interpreted encounter. Setton too acknowledges the importance of the linguistic aspects of interpreting when he outlines his account of the reception of phonetic content and its subsequent transformation into a grammatical proposition. Together, these various assertions indicate that part of an

interpreter's task requires him or her to demonstrate a skill that is very much based in language.

Second, a number of the models have made mention of the interpreter's use of information about her immediate surroundings to understand meaning. Paradis' account of interpreting is based on compelling evidence that what he terms "pragmatic competence" plays an important role in second language processing. The research he cites indicates not only that what we refer to as discursive ability is distinct from linguistic ability, but that the two are subserved by different hemispheres in the brain. He also argues that "pragmatic" ability is pivotal in communication, particularly for bilinguals using their non-dominant language. Although he does not further subdivide discursive ability, his work lends general support to other authors who do. One such subdivision is discussed by interpretive theorists, who argue that the "*contexte cognitif*" – the interpreter's understanding of the conference proceedings and the different players in them – is an important key to unlocking meaning in context. This idea is further developed by Setton, who argues that part of the interpreting process involves complementing the grammatical proposition with "who", "what", and "whom" information about the interpreting environment. In these three models, there is evidence to suggest that perceiving the relationship between language forms, their users, and their context of use is a skill that is central in interpreting.

Third, many of the models we reviewed claim that using general knowledge to understand the producer's communicative intention is also part of the interpreting process. For instance, interpretive theorists argue that if an interpreter knows a topic well and consequently understands the producer's position, it will make a great deal of difference in

the resulting interpretation. The amount of source-language input required to provoke the interpreter's "*déclat*" of understanding will be appreciably smaller, and the interpretation will be more successful. For his part, Cokely also underscores the importance of general knowledge in his discussion of "sociolinguistic sensitivity". He suggests that the interpreter's knowledge of Deaf identity allows her to understand that in referring to himself as "Deaf", a client is intending to express pride in belonging to a cultural group, and not to evoke feelings of pity. She therefore formulates her target-language production in a manner that helps her accurately communicate this intention to hearing interlocutors. A similar argument is made by Wadensjö, who, in adopting a Bakhtinian approach, emphasizes the producer's attitude towards an utterance as a source of meaning. Knowledge of the "neutral, dictionary form" of language units must be complemented with knowledge of the way words have been used by others in the past, in order to understand utterances in communication. Participants in an interpreted dialogue all have a goal that they want to achieve through talk; if an interpreter fails to recognize and convey these goals, she will also fail to interpret adequately. Finally, Setton maintains that interpretation of inference is an essential component of interpreting. In many cases, the ambiguity of the syntax in a source-language proposition will make it impossible to determine the meaning of the proposition through linguistic means alone. The interpreter must necessarily fall back on her own set of internalized rules about communicative behaviour to interpret the producer's intention and decipher the proposition's meaning. Even when no syntactical ambiguity exists, Setton points out that an interpreter will regularly clarify aspects of the source-language message by adding information based on her own knowledge of the topic or of the conference proceedings²³.

²³ Setton (1999) discusses several instances in his corpus where the interpreter's target-language utterances

These three components are present across a number of models, but in constructing the new test instrument, we have elected to work with one in particular – Setton’s cognitive-pragmatic model – for a number of reasons. First, although there was some level of agreement about the components across many models, there was also significant level of disagreement. Working with one model allowed us to enter into the specifics of each component and to avoid any theoretical contradictions. Second, Setton’s model was the only one that offered a rigorous account of all three components as part of a coherent account of cognitive processing beyond language recognition. Only some of the information processing models even hint at “intermediate representation”, and interpretive theory does not give a satisfying explanation of how discursive information is incorporated into the interpreter’s understanding in a particular instance²⁴.

Third, the cognitive-pragmatic model offered advantages that the others did not. It explains and describes the analysis of meaning from sources other than linguistic segments, it offers a coherent, state-of-the-art account of the cognitive inner workings of such an analysis²⁵, and it deals with spoken-language interpreting in the conference setting.

contained pertinent information not included in the source-language original. He notes that a Chinese delegate’s vague reference to “1989” was made more explicit by the interpreter’s mention of “the Tiananmen Square incident”. Likewise, a German speaker’s discussion of the “Federal High Court of Justice” was clarified for the English audience when the interpreter explained the court’s location (Karlsruhe) (p. 178).

²⁴ Setton (1999) states that the information processing models, notably Moser’s, explain intermediate representation in a manner that remains “close to linguistic form” (p. 49). In reference to interpretive theory, he asserts that “the concept of *sens* is underspecified, and unassailable in that it tacitly conflates utterance-meaning and hearer-meaning (i.e. interpreter-meaning), implying full determinacy and perfect symmetry between brain states: there is no account of how relevant contexts are chosen” (p. 48-49)

²⁵ Setton (1999) notes how many models of the interpreting process “have not been recognisably in step with development in linguistics and cognitive psychology” (p. 59). He accuses the interpretive theorists of not having kept up to date with advances in experimental psycholinguistics or (Gricean) pragmatics, and the

Using Setton's model as a guide, we were then able to sketch the structure of the competencies we wished to measure with the new test. Following the example outlined in Spencer and Spencer's (1993) "competency dictionary", we structured our constructs in a three-tiered hierarchy. At the top is the *cluster**, the group of constructs connected through their relevance to a particular job or profession. In the middle are the competencies – and in interpreting it makes most sense to examine skills and knowledge rather than the other types of competencies²⁶ – that are central to successful performance in the profession. At the bottom are the *dimensions**, the smaller components that make up each individual competency.

It was decided that the new test developed in the present study would examine a cluster of competencies associated with spoken-language conference interpreting. The specific competencies chosen were linguistic parsing, the assignment of reference, and the interpretation of inference. All three are skill- and/or knowledge-based competencies that have been consistently identified by researchers as important to conference interpreting. Each competency may be made up of several dimensions. For example, the linguistic parsing competency may involve the recognition of lexical units, or the analysis of syntactic structures. The exact nature of the three competencies, along with the manner in which they were measured, is discussed in more detail in *Section 4.2*.

proponents of information processing of neglecting information about speech processing through inference and the construction of frames (p. 59).

²⁶ We have already given our rationale for not targeting the "hidden competencies" (motives, traits, and self-concept) – they would unnecessarily broaden the scope of this exploratory study and take us out of the realm of most existing work on interpreting.

We recognize that the terminology we have chosen to describe our construct hierarchy puts us at odds with much of the existing testing literature. Instead of a single “cluster” comprising various “competencies”, it would be more usual to analyze one “competency” in its various dimensions²⁷. To be more consistent with the psychometric literature, we might have described interpreting as a competency, and the aspects we have targeted in our tests as dimensions. However, dimensions are normally considered to be part of a single construct, and that the scores associated with them are usually thought of as being related. To describe the skills measured by the new test in this way would have been inaccurate. As we have previously mentioned, there is ample evidence that linguistic parsing is unrelated to discursive skills such as the assignment of reference and the interpretation of inference. Paradis argues convincingly that linguistic tasks and “pragmatic” tasks are processed in different hemispheres of the brain. Also, academia has traditionally insisted in making a very large distinction between the study of “language” and the study of “discourse”²⁸. If we were to blur the line between the linguistic and the discursive, we would be ignoring both their neuroanatomical substrate and their epistemological history. In addition, Setton’s model situates the assignment of reference within the *assembler*, and the interpretation of inference within the *executive*. These are two separate components of his

²⁷ To cite but two examples, Pett, Lackey and Sullivan (2003) describe factor analysis as the “identification of linear combinations or clusters of related observed variables that represent specific underlying dimensions of a construct” (p. 2), and Cortina (1993) proposes that factor analytic techniques and measures of internal correlation be used to shed light on the *dimensionality** of test data.

²⁸ Couper-Kuhlen (2001) notes that it was traditional to think of “language” and of other features “outside of language”. This can be seen in the fact that many British universities set up departments of “Linguistics and Phonetics”; the study of phonetics subsumed intonation and other discursive phenomena that Couper-Kuhlen discusses in detail. She argues that this traditional cleavage has its roots in the competence-performance dichotomy introduced by generative linguistics.

cognitive-pragmatic model. It therefore seems inaccurate to describe the two discursive skills as dimensions of a single construct.

Describing linguistic parsing, the assignment of reference, and the interpretation of inference in terms of clusters, competencies, and dimensions allows us to be consistent with the theoretical literature from which they were taken. In addition, the terminology we have chosen is consistent with the concept of competency as it was put forward by its original proponent (McClelland, 1973, 1993) and his collaborators (Spencer & Spencer, 1993). We have therefore elected to describe our constructs in this way, despite the fact that it appears to put us at odds with some of the measurement and evaluation literature.

In summary, we have examined in this chapter how a competency-based approach may be used as a means of clearly defining the constructs that will form the basis of the new test. We have reviewed a number of important models of the interpreting process, and we have identified a number of constructs that are central to successful performance in conference interpreting. We have described those constructs in a manner that is consistent with a competency-based approach, in other words, in terms of the relevant cluster, competencies, and dimensions. In the pages that follow, we will take this understanding of interpreter competencies and outline the methods we will use to investigate the issue of content validity with regard to interpreter certification.

4. Validation Methodology

In order to test the general and specific hypotheses formulated in *Chapter 1*, we designed and conducted a study made up of three components: 1) the analysis of an existing test; 2) the construction and analysis of a new test; and 3) the comparison of the two analyses. Each of the three components is described in detail below.

4.1 The Existing Test

The existing test is a national certification exam for conference interpreters administered by a professional association ("Association 1"). In principle, the administration of the test is on an annual basis. In practice, the test is offered only when a substantial number of candidates have registered to take it.

4.1.1 Test Description

The certification exam is a performance test, which is made up of three items (from here on referred to as items A, B, and C). Each item is a ten-minute, videotaped excerpt of a speech given by a single speaker on a topic that deals with public affairs, and that is relevant in a national context. The speech is non technical in nature. Candidates are asked to interpret the speech using the simultaneous method, and their performance is recorded on audiotape.

In each administration, the test is offered in a variety of language pairs and directions. The exact pairs and directions depend on the requirements of the candidates who register to

take the exam. For instance, if candidates with French as their A language and English as their B language register for a particular administration of the exam, a version will be developed and administered that evaluates interpretation from English into French. In contrast, if no candidates with English as an A language and French as a B language register, a version evaluating French into English will not be developed for that particular administration. (It is expected that candidates will normally interpret from their B language into their A language.)

The raters chosen to score the exams are themselves certified interpreters, and they are selected to evaluate candidates who work with the same language pairs, and in the same direction. Because the interpreting community is small, and because there is a danger that raters may recognize the voices of candidates and give a biased evaluation, raters are asked to score the performances of candidates who live in a jurisdiction other than their own.

Candidates' performances on each item are given three scores. These are based on their perceived achievement on three rating scales, which are represented in their entirety in *Appendix II – Existing Test Documentation*. The competencies on which the scales are based are discussed briefly below.

1. Intelligibility

If candidates demonstrate acceptable intelligibility, it means that their interpretations sound like “ordinary speech”. Their production is clear, grammatically correct, and stylistically acceptable. They use words and expressions appropriately, the “idea” is understandable, and there are no uninterpreted words or phrases.

2. Informativeness

If candidates show appropriate levels of informativeness, their TL speeches contain an amount of “information” that is similar to or greater than that found in the SL speech. Their interpretations are not missing any “meaning”, either at the “word level” or in terms of “sentence structure”. There may be additions in the interpretation, but these are included to make it “more understandable”.

3. Style

If candidates have good interpreting style, they are able to handle both linguistic concerns and delivery concerns appropriately. Linguistically, they use an appropriate level of language, the relevant vocabulary, acceptable grammar, and interesting imagery (metaphors, similes, etc.). With regard to delivery, they have good diction, speak at an appropriate speed, avoid hesitations and excessive self-corrections, and have a pleasing voice quality.

The dataset for the existing test contains scores for three administrations of the exam. In the first administration, each candidate received 27 separate scores. This is because each candidate was evaluated by three raters, each of whom scored all three test items (A, B, and C) using all three rating scales (intelligibility, informativeness, and style) (3 raters x 3 items x 3 scales = 27 scores). In the second and third administrations, each candidate received 18 separate scores. This is because each candidate was evaluated by two raters, each of whom scored all three test items (A, B, and C) using all three rating scales (intelligibility, informativeness, and style) (2 raters x 3 items x 3 rating scales = 18 scores)¹.

4.1.2 Assumptions about the Test Data

Two assumptions were made about the data in the existing test. First, we assumed that the data were multidimensional. The items in the existing test are intended to measure intelligibility, informativeness, and style, which are three largely unrelated abilities (the distinct nature of the competencies is discussed further in *Section 5.1.1*). There was therefore an expectation that scores associated with the competencies would be largely unrelated as well. In other words, in performing our analysis, we were looking to “provide evidence on whether scores on a construct warrant the interpretation that it is separate from other constructs” (Shavelson, Hubner & Stanton, 1976, p. 417).

We recognize that analyses of other tests are often carried out with different assumptions. In the measurement and evaluation literature, we found indications that many

¹ For further clarification of the scoring scheme, see the actual rating sheet in *Appendix II*. In the first administration, three sheets were filled out per test taker, while only two sheets per test taker were filled out in

tests are assumed to be unidimensional (i.e., that only one ability is measured by the items on the test). The expectation of unidimensionality appears to be common in Item Response Theory, a particular school of thought in measurement and evaluation (Hambleton, Swaminathan & Rogers, 1991, p. 9)², and we saw at least one study in our own literature review that also held this expectation. McGhan, Stimmel, Gilman & Segal (1982), whose work we reviewed in *Section 2.2.1*, examine a single test (for pharmacists) that evaluates three domains (clinical therapeutics, physical assessment, and knowledge of legislation and regulations). At one point in their analysis, the authors estimate the internal reliability of the test items associated with each of the three domains separately, but they also estimate the internal reliability of all the test items together. The authors' expectation was that ability in each of the three domains would be related to some larger, underlying factor. In this case, information about the test constructs described them as related, and evidence of a single factor in the data was interpreted as supporting construct validity.

Yet there was also ample evidence in the literature that not all tests are unidimensional. Cohen, Swerdlik and Phillips (1996) discuss this possibility, noting that tests may either measure a single trait (what they refer to as a "homogeneous test") or that tests may be composed of items that measure more than one trait (a "heterogeneous test" in their terminology) (p. 153). Also, in *sections 2.2.1* and *2.2.2*, we reviewed two studies that looked at multidimensional tests. Violato, Salami and Muiznieks (2002) examined a single test (for massage therapists) that purported to measure three constructs (professional knowledge,

the second and third administrations.

clinical judgement, and clinical skills). The exploratory factor analysis they conducted showed that there were three separate factors in the test data, and it confirmed beliefs about the test constructs. Daniel & Siders (1994) also investigated a single test (for teachers) that made use of three separate subscales (teaching plans and materials, position skills, and interpersonal skills). In their investigation, the authors used confirmatory factor analysis with the expectation that a model consisting of three univocal and uncorrelated factors would best fit the data. While their results did not confirm the three-factor model, the authors did report evidence that their data were multidimensional. An earlier study had already shown that there was only a moderate level of correlation between items for the three subscales, and that a six-factor structure had been found in an exploratory factor analysis. In both of these cases, information about the constructs described them as separate and unrelated, and evidence of separate and unrelated factors in the data was interpreted as supporting construct validity. It appears then that our expectations of multidimensionality in the present study, while not universal, are nevertheless common. In our analysis of the existing test, outlined in *Section 4.1.4*, we examined the test scores for evidence that they measured more than one ability.

The second assumption made about the existing test data was that they are measurements made on an interval scale. Like many who use quantitative methods, Leavitt (1991) asserts that measurements can be made in four different ways. If measurements are made on a nominal scale, they are slotted into categories that have no quantitative value, but that act simply as labels. For example, a variable that indicates the sex of participants in an

² The authors suggest that unidimensional models are prevalent in Item Response Theory because multidimensional IRT models are more complex, and because they have not been well developed (Hambleton, Swaminathan & Rogers, 1991, p. 9).

experiment (i.e., male or female) is considered nominal-level data. If measurements are made on an ordinal scale, they are slotted into categories that have some rank order. For example, a variable that indicates a runner's place in a road race (i.e., 10th place, 5th place, 1st place) is considered ordinal-level data. If measurements are made on an interval scale, they are ordered, and the intervals between the measurements represent equal quantities of what is being measured. For example, a variable that indicates participant's scores on a test of introversion (i.e., where a score of 15 is midway between a score of 10 and a score of 20) is considered interval-level data. Finally, if measurements are made on a ratio scale, they have interval-level properties, plus a zero point. For example, a variable that measures outside temperature in degrees Celsius (i.e., where there can be a value of zero) is considered ratio-level data (Leavitt, 1991, p. 80).

As the rating scales outlined in *Appendix II* clearly show, scores from the existing test indicate rank-order (e.g., a score of 8 on the intelligibility scale is clearly lower than a score of 9 on the same scale), and they indicate degree of difference (e.g., a score of 5 on the informativeness scale is defined as midway between a 4 and a 6). We also have reason to believe that the concept of interval-level data was clear in the minds of the raters. Not only did they use scores to rank-order test taker performances, but several raters awarded half points; there were obviously instances where the raters felt that a particular test taker's performance was midway between points on the rating scale. Because it seems evident that

the existing test data are on an interval scale, our analysis of the test, outlined in *Section*

4.1.4, makes use of a number of statistical tests that are *parametric*³ in nature.

4.1.3 Sample

The existing exam was administered to a total of 26 candidates. Twelve candidates took the exam in the first administration, while the numbers for the second and third administrations were seven and seven. The version of the exam that candidates received depended on the year of administration, as well as on the candidates' language pair/direction. Similarly, the rater group assigned to evaluate a candidate was also dependent on the administration, and on the language pair/direction. The precise breakdown of candidates, administrations, language pair/direction, and rater groups is outlined in *Table 1* below.

Table 1: Candidate Information

Candidate	Administration	Language Pair/Direction	Rater Group
1	Year 1	English to French	1
2	Year 1	French to English	2
3	Year 1	French to English	2
4	Year 1	English to French	3
5	Year 1	English to French	1
6	Year 1	French to English	2
7	Year 1	English to French	1
8	Year 1	English to French	1
9	Year 1	English to French	3
10	Year 1	English to French	3
11	Year 1	French to English	2
12	Year 1	French to English	2
13	Year 2	French to English	4
14	Year 2	French to English	4
15	Year 2	German to English	5
16	Year 2	German to English	5
17	Year 2	English to Cantonese	6
18	Year 2	English to Mandarin	7

³ When measurements are made on an interval or ratio scale, it is assumed that the data follow a bell curve, and that the statistics used will focus on the centre and width of the curve. These statistics are referred to as parametric (Norman & Streiner, 1997, p. 5)

19	Year 2	English to Russian	8
20	Year 3	English to French	9
21	Year 3	English to Romanian	10
22	Year 3	English to Spanish	11
23	Year 3	Spanish to English	12
24	Year 3	French to English	13
25	Year 3	English to Portuguese	14
26	Year 3	English to Spanish	11

The *Candidate* column lists our identification numbers for the 26 candidates. The *Administration* and *Language Pair/Direction* columns indicate the year each candidate took the exam and the language pair and direction for which they were tested. The *Rater Group* column shows which trios or pairs of raters evaluated the test takers. For example, Candidate 1 was assessed by the same three raters⁴ as Candidate 5; likewise, the same two raters⁵ evaluated candidates 22 and 26. In contrast, candidates 1 and 4 were evaluated by different trios of raters, and different pairs of raters assessed candidates 14 and 24 (even though they were tested on the same language direction).

The dataset for the existing test was highly heterogeneous. It contained 13 different versions of the test (two in the first year of administration, five in the second year, and six in the third year). To further complicate matters, not all candidates who took the same version of the test were scored by the same group of raters. For example, candidates 1 and 4 both received the same version of the test (they took the test in the same year, and were tested for the same language pair/direction), but they were evaluated by different groups of raters.

⁴ There were three raters because candidates 1 and 5 took the test in Year 1.

⁵ There were two raters because candidates 22 and 26 took the test in Year 3.

For the purposes of this study, the features of the dataset outlined above make it less than ideal. In order to eliminate sources of measurement error, our analysis should have first been conducted within uniform groups of test takers – those individuals who received the same test version, and who were evaluated by the same raters. If we had subsequently discovered that scores across uniform groups were relatively equivalent, we might then have considered an analysis with all candidates pooled together. Unfortunately, however, the uniform groups of candidates were extremely small. There was one group of five individuals, one group of four, one group of three, and two groups of two. The remaining “groups” were made up of single individuals.

Obviously, this situation presented us with a serious methodological difficulty. It would have been unproductive to use statistical measures to analyze the scores of uniform groups because of their size; with sample sizes of five or less, the statistical power of our tests would have been greatly reduced, and the chance of error greatly increased (Leavitt, 1991, p. 206). Yet without such an analysis, there was no way to determine if inter-group differences were a significant source of error. For example, if we pooled all the candidate scores together, performed our statistical tests, and discovered noteworthy findings, we would have no way of knowing if the findings were accurate, or if they were unduly influenced by differences among test versions, among groups of raters, or both.

We faced two additional methodological difficulties. The first of these was the design of the existing test. Scores for the three test constructs are given for each one of the three test items, which are scored by several raters. In other words, construct scores were nested within item scores, which were nested within rater scores. This meant, for example, that if we

attempted to group together all the scores for one construct and examine their similarity, we would be uncertain whether we were actually measuring other facets of the text design, such as items or raters. We were therefore unable to evaluate the test in the precise, logical manner typically used in studies of other certification tests, where the design is different.

The second factor was sample size. At 26, the total number of candidates in the sample was much smaller than typically seen elsewhere. Subdivided to control for the design of the test, sample numbers would have been reduced to impractical levels.

And yet, we have every indication that the sample size in the existing test is reflective of the population as a whole. Information provided by professional associations and government institutions illustrates this point effectively. For example, the AIIC directory lists 112 certified interpreters for all of Canada (AIIC, 2003), the federal government's Translation Bureau has an in-house complement of roughly 30 interpreters (Card, 2003), ATIO indicates that there are 50 certified interpreters in Ontario (ATIO, 2003), and OTTIAQ indicates that there are roughly 30 interpreters who work with either English or French in Quebec (OTTIAQ, 2003). Based on this information, we believe that the question of sample size, which was problematic in the present study, is likely to complicate any future study on interpreter certification as well.

When faced with this reality, it is not acceptable for researchers to throw up their hands and declare that interpreter certification tests defy analysis. Without some form of evaluation, these tests pose a risk to candidates, to the public, and to the professional associations that are responsible for testing. If competent candidates are not certified, they

could potentially be unfairly deprived of success in their chosen livelihood, or of their livelihood altogether. If incompetent candidates are certified, consumers stand to receive inferior quality service, which in some settings (e.g., medical, legal) may have dire consequences. Both of these outcomes threaten the reputation of the profession, and through it, that of the professional association. Also, dissatisfied candidates and dissatisfied consumers may take further action against professional associations, such as initiating litigation proceedings (see *Chapter 1*).

Evaluating interpreter certification tests is clearly important. We therefore elected to pursue the present study, accounting for the two sources of methodological difficulty as best we could. It was not within our power to alter the design of the existing test, but we were able to introduce measures to attempt to compensate for some of its less desirable aspects (see our use of the multitrait-multimethod matrix, detailed below). Also, we constructed the new test (see *Section 4.2*) based on our understanding of the existing test's drawbacks. With regard to sample size, we decided to conduct our analyses on candidate scores pooled together in one group. In doing so, we recognize that we have failed to control for measurement error resulting from differences in test version and specific rater groups. However, we feel that some steps were taken to minimize these sources of error in the test's initial development. For example, Association 1 does have an established set of guidelines for selecting the material that is used to make up the different test versions. Also, the association selects raters who are certified interpreters with experience in the profession, and it provides them with written instructions for rating candidates' performances.

4.1.4 Procedure

To test the specific hypotheses related to the existing test, a number of steps were carried out. Several of these involved using statistical tests that are parametric in nature, which we feel is appropriate because the existing test scores are measurements made on an interval scale (see *Section 4.1.2*). The steps we took are enumerated below, along with a justification for each one.

1. Information was gathered on the development of the existing test, and the theoretical origins of the exam constructs were evaluated as logical evidence of their validity. It seemed warranted to begin the analysis this way, since the *Standards* (AERA, APA & NCME, 1999) outlined in *Section 1.1.2* indicate that examining the theoretical rationale behind the use of test constructs is an important step in the validation process.
2. A multitrait-multimethod matrix was constructed to determine if there was justification for analyzing items and constructs separately. Pearson's product-moment correlation was then used to examine the similarity of scores for the three items (A, B, and C), followed by a Fisher's z test to determine if there were any significant differences between the correlations. Use of the multitrait-multimethod matrix seemed particularly appropriate given the structure of the existing test (constructs nested within items) and the intent behind the original development of the matrix – it was introduced by Campbell and Fiske (1959) as a means of further understanding construct

validity by breaking it into convergent validity (the degree to which constructs purported to be related to one another appear in reality to be related) and discriminant validity (the degree to which constructs purported to be unrelated to one another appear in reality to be unrelated). As for Pearson's correlation, it was repeatedly used in the credentialing literature we reviewed in *Section 2.2* to compare scores from different test components (see Violato, Salami & Muiznieks, 2002; Price & Wilkins, 2001; Cobb, Shaw, Millard & Bomotti, 1999; and Norcini, Hancock, Webster, Grosso & Shea, 1988). The comparison using Fisher's z (Guilford & Fruchter, 1973) was a safeguard designed to ensure that it was justifiable to make similar conclusions about all the correlation coefficients we obtained, based on their magnitude (i.e., to provide an indication that there were no statistically significant differences among the three coefficients).

3. Pearson's product-moment coefficients were calculated to assess the similarity of the scores for the three constructs (intelligibility, informativeness, and style) and a comparison using Fisher's z was carried out to reveal any significant differences between the correlations. The justification for the use of these two tests is outlined above.
4. Principal components analysis was used to determine if the observed score component structure resembled the structure predicted by the rationale behind the three constructs (which was outlined in the first step of the procedure), and this was followed by a calculation of Cronbach's alpha for the same data.

We paired the two techniques together based on Cortina's (1993) claim that both are needed to confirm the unidimensionality of a test. We also found support for using both techniques in the literature we reviewed: principal components analysis, or a similar factor analytic technique, was employed repeatedly (see Violato, Salami & Muiznieks, 2002; Price & Wilkins, 2001; and Daniel & Siders, 1994), as was Cronbach's alpha (see Violato, Salami & Muiznieks, 2002; Norcini, Hancock, Webster, Grosso & Shea, 1988; and McGhan, Stimmel, Gilman & Segal, 1982).

To interpret the principal components analysis, both the *Kaiser-Guttman criterion** and Cattell's scree analysis (Norman & Streiner, 1997, p. 145) were used. Statistical significance was set at the 95% level, which is a common convention in the social sciences.

To help us with the interpretation of the magnitude of the correlations we obtained in the analyses, we turned to the work of Guilford. The author argues against imposing any hard and fast rule, and suggests instead that interpreting a coefficient of correlation is largely a matter of convention. The choice of convention depends on the correlation's use. For instance, he argues that researchers may have one set of expectations about a correlation used to indicate the relationship between scores on an aptitude test and scores for some other measure of vocational success (predictive validity), and another set of expectations about the relationship between scores on two forms of the same test (*alternate forms reliability**). In the first case, they may be quite happy with a coefficient with a magnitude of .70, because this is on the high end of the range of values conventionally found with validity coefficients.

In the second case, a magnitude of .70 may be cause for concern, since reliability coefficients are usually much higher (Guilford, 1965).

Somewhat paradoxically, Guilford took a different stance at an earlier point in his career and published a scale of values for interpreting correlations that has since become known as “Guilford’s rule of thumb” (Guilford, 1950). We elected to use the scale of values for several reasons. First, with five clear and distinct categories, it is easy to use. Second, it is frequently cited in the literature (Williams, 1992; Borg & Gall, 1983). Third, its author is widely recognized for his contributions to the field of psychological measurement.

Guilford’s scale is outlined below.

Correlation value less than .20	Slight correlation, only a negligible relationship
Correlation value between .20 and .40	Low correlation, definite but small relationship
Correlation value between .40 and .70	Moderate correlation, substantial relationship
Correlation value between .70 and .90	High correlation, marked relationship
Correlation value between .90 and 1.00	Very high correlation, very dependable relationship

Although we have applied this “rule of thumb” to our interpretation of the results of the present study, we have done so cautiously. We have not drawn conclusions related to the categorization of the correlations without additional evidence.

4.2 The New Test

The new test was designed to determine if a set of competencies based on a discursive approach to interpreting would show greater evidence of validity than the competencies measured by the existing test. The goal was not to develop a viable, alternative certification test, nor to suggest that the testing methods used in this study should be used in professional certification. Rather, the goal was to evaluate whether further research on the use of discursive competencies in certification testing is warranted.

To begin, we examined Setton's (1998) cognitive-pragmatic model of the interpreting process. As it was noted in *Chapter 3*, the model has three important components that are used to explain the comprehension of input and the production of output during simultaneous interpreting.

1. The Parser

During comprehension, this component derives the meaning of words and propositions from linguistic knowledge. During production, it assembles TL lexical items into grammatical subphrases and phrases.

2. The Assembler

During comprehension, this component integrates a set of probable referents into the linguistic proposition. During production, it is presumably activated to some degree, since it houses the parser.

3. The Executive

During comprehension, this component makes assumptions about the producer's intent, and judges the match between that intent and the elaborated proposition. During production, it makes decisions about including, excluding, highlighting, and attenuating parts of the message based on information about the receiver.

Clearly, the model suggests that comprehension and production are processed in a very similar manner, and that the components of the model involved in both are – for the most part – the same. On both the input and output sides of the model, processing is carried out in a series of steps that progress from one to the other. Input is first analyzed at the level of language, then wider conclusions are drawn about the grammatical proposition and its connection to the real world, and finally the utterance's place in the larger context (which includes both previous utterances and the general situation) is decided by applying a set of internalized rules to interpret its relevance. The production of output is essentially the reverse. Decisions are first made about the discursive structure of the upcoming utterance (so as to ensure that it is relevant to the situation and to previous utterances), then word order and focus structure are chosen, and finally the utterance is executed at the level of language. The parser and the executive play a vital role in both the input and output processes, and the assembler is also active to a large extent.

Because of the similarity between comprehension and production, it did not seem essential to examine both processes with the new test, particularly when the exploratory

nature of the present study is taken into account. As mentioned above, our goal was not to produce a viable certification test, but rather to determine if there was enough evidence to validate a set of discursive constructs and thereby justify further discourse-based research on interpreter certification testing. We decided that this evidence could easily be obtained by evaluating only one side of the model, and our decision seemed pragmatic in light of our limited resources and our need to continually monitor the scope of the study, in order to prevent it from getting out of hand.

When faced with a choice between evaluating comprehension or production, we chose the former, for several reasons. First, testing comprehension is a commonplace occurrence, and most people are familiar with the format of a comprehension test. It was much less obvious to us how we might go about testing production, and any effort we made in that regard would require us to devote time and energy explaining the test format to the study's participants. Second, because we had the input speech ahead of time, we were able to create the comprehension-related multiple-choice items for the new test instrument before we met with the participants. If the new test had been production-based, we would have needed the participants' output to develop multiple-choice test items⁶. This, in turn, would have meant structuring the administration of the test in two sessions – one to record the subjects' production, and another to test them on it somehow. Given that the number of possible participants was low to begin with (see below), we felt that the attrition of participants between the two sessions would be an unacceptable risk.

⁶ It was important for the test format to be multiple-choice. Recruiting human raters for another test format would have taken additional time and energy, and it would have potentially reduced our pool of participants (i.e., an interpreter who agreed to serve as a rater could not also be a participant).

4.2.1 Test Description

The new test is made up of two components: 1) a ten-minute speech, the content of which is non technical in nature and relevant to the national context; and 2) a 45-item multiple-choice questionnaire. The items were designed to test interpreters' understanding of the SL communication by targeting the competencies associated with the three components of Setton's model outlined in *Section 3.3.3*: linguistic parsing, the assignment of reference, and the interpretation of inference. Fifteen items were developed for each of the three competencies, and each item contained a question stem with four possible answer options.

The items were designed to be scored in a weighted manner. All answer options are technically correct, but they vary incrementally in their "correctness". The items were developed in this manner to provide interval-level data, which allowed us to examine the results using parametric tests. This facilitated our study in two ways. First, it enabled us to structure the analyses for the existing and new tests in the same manner, and in some cases to use the same statistical tests (e.g. Pearson's product-moment correlation). Second, it allowed us to make a direct comparison between the results of the two tests (i.e., using Fisher's z to determine significant differences between correlations).

Weighted scoring is discussed repeatedly in the measurement and evaluation literature, predominantly as a means of improving the psychometric properties of test instruments. However, the improvements reported by most researchers are modest. Essex (1976) found that weighted scoring increased internal consistency reliability and decreased variance, but not significantly. Cross, Ross & Geller (1980) noted that weighted scoring

resulted in test specific effects on internal consistency reliability and on predictive validity (i.e., they could not make a generalization because the results varied according to the tests they examined), but they also point out that the items used in their study were not written specifically for weighted scoring. Harasym, Norris & Lorscheider (1980) reported that under certain conditions weighted scoring increased mean score, but reduced variance. In recent years, it would seem that interest in this scoring method has waned⁷, largely because it is labour intensive (Essex, 1976) and offers only minimal gains in return for the added effort (Cross, Ross & Geller, 1980).

In spite of these findings, weighted scoring was retained as a scoring method in this study, for several reasons. First, while none of the research on the topic pointed to a net increase in measurement quality, neither did it indicate a net decrease. Second, in the literature, the authors' objective in using weighted scoring was to improve reliability and validity; in this study the objective was to obtain interval-level data so that the same parametric tests could be performed with both the existing and newly developed tests. Third, it seemed reasonable to believe that the problems associated with weighted scoring of items that were designed to be scored dichotomously would not likely affect this study if the items on the newly developed test were designed specifically with weighted scoring in mind. The development of the items and the scoring method are discussed below.

⁷ While none of the authors recommended using weighted scoring in a context where decision-making was the ultimate goal, they were all quick to point out its usefulness as a pedagogical tool. Most test takers surveyed by the authors preferred weighted scoring over other scoring methods, and it appeared to offer valuable information about test takers who were performing below an optimal level.

Item Type A

Items of this first type were designed to target linguistic parsing (as conducted by the parser). The items are intended to test participants' understanding of the linguistic elements of the speech they have interpreted. To ensure that the items were evaluating a single competency, it was decided that all items of this type should test lexicon, and not syntax.

To develop the test items, a series of lexical units was selected from the chosen speech, and for each one an explanation was written that was specific to the context of the speech, and that contained four key elements. This then made writing the answer options straightforward. The answer option that received the highest number of points was the one that contained all four elements, the option that received the next highest number of points contained three elements, and so on. An example of this item type is given below.

In Canada, "Parliament" is

- a) located in the city of Ottawa;
- b) synonymous with the House of Commons in Ottawa;
- c) the two federal legislative chambers located in the city of Ottawa; or
- d) the House of Commons and the Senate on Parliament Hill in Ottawa.

Option	Worth	Justification
a)	1 point	Contains one element of the full explanation (general location)
b)	2 points	Contains two elements of the full explanation (general location + House)
c)	3 points	Contains three elements of the full explanation (general location + House + Senate)
d)	4 points	Contains all four elements of the full explanation (general location + House + Senate + specific location)

Item Type B

Items of this second type were designed to target the assignment of non-linguistic reference (as carried out by the assembler). The items are intended to test participants' ability to correctly attribute specific references to linguistic elements of the speech they have interpreted.

To develop these items, portions of the speech were selected that contained examples of one of three discursive features. They are listed and explained below.

1. Synecdoche

In using this feature, the producer refers to a part to designate a whole, or to a whole to designate a part. For instance, he may mention the capital of "Ottawa", when he is actually referring to the entire country of Canada, or he

may talk about the “the government” when he actually means the Prime Minister in particular.

2. Deixis

Deixis refers to those parts of speech that do not refer to a constant and objective notion. They must instead be defined by a relationship between producer and receiver in a particular time and in a particular place. For example, words such as “table” and “tree” can be explained out of context, but there is no way of knowing *a priori* what the word “I” means until it has been used in a particular “instance of discourse”. Other examples of deixis include lexical items such as “this”, “that”, “here”, and “now”⁸.

3. Pronouns Lacking an Immediate Anaphoric Reference

The producer may use a pronoun without ever mentioning the noun that it replaces. Alternatively, the noun in question may have been mentioned at an earlier point in time, but the span between noun and pronoun exceeds the capacity of the receivers’ working memory. The receivers must use their understanding of the immediate context to provide the reference. For example, the producer may suggest that “it is such an archaic institution”, leaving the receiver to understand that he is referring to the Canadian parliament.

⁸ This conception of deixis is taken from Benveniste’s (1966) discussion of “*La nature des pronoms*”. It is also supported by Fillmore’s (1997) lecture “Deixis I” and by Cruse’s (2000) accounts of “person deixis”, “spatial deixis”, and “temporal deixis”.

Answer options were then obtained by either increasing or decreasing the level of specificity, as dictated by context. In the following example, the producer mentions our country's "leadership". At the most general level, he is referring to public figures. More specifically, he is talking about the government, and about a particular political party. However, a careful analysis of his speech reveals that at the most specific level he is referring to the Prime Minister and to members of his cabinet. This example was used to create the item below.

By mentioning "the stumbling and bumbling of some of our *leadership*" Manning means

- a) slip-ups made by public figures;
- b) mistakes made by the government;
- c) errors committed by the leader of a political party; or
- d) gaffes made by the Prime Minister and his cabinet.

Option	Worth	Justification
a)	1 point	Minimum specificity (leaders generally are responsible)
b)	2 points	Moderate specificity (the government is responsible)
c)	3 points	High specificity (the leader of a party is responsible)
d)	4 points	Maximum specificity (the PM and cabinet are responsible)

Item Type C

Items of this third and final type were designed to target the analysis of communicative intentions (as performed by the executive). In the midst of communication, producers frequently imply a meaning that the receiver must then infer. The third item type was intended to test participants' ability to make this kind of inference.

To develop these items, portions of the speech were selected where the producer appeared to be communicating his point indirectly. Answer options were written that represented varying degrees of ability in interpreting inference. For example, the option with the lowest weight describes the linguistic function of the portion of the speech in question, while the option with the highest weight describes the portion in terms of its likely intended discursive effect on the receivers. An example of this item type is given below.

When Manning says "...on the democratic front. *Just to illustrate the democracy deficit* – as it's now referred to in Ottawa – my last vote in the House of Commons was in December of last year. As you know as Christmas comes close...", he is

- a) peppering his speech with political terminology;
- b) keeping his audience up to date on political jargon;
- c) interrupting his main line of reasoning and introducing a tangent; or
- d) allowing listeners to see that the tangent supports his main point.

Option	Worth	Justification
a)	1 point	Deals with the function of the producer's use of language units (producer introduces political jargon)
b)	2 points	Deals with receiver's understanding of the use of language units (receivers are meant to learn the jargon)
c)	3 points	Deals with producer's communicative intent (producer signals that he will interrupt his argument and introduce an aside)
d)	4 points	Deals with receiver's interpretation of communicative intent (receivers understand that the aside supports the argument structure)

Before the test instrument was administered to the study's participants, the order of the answer options in each item was randomized, as was the order of items in the instrument. (For a copy of the final instrument, see *Appendix III – New Test Documentation*.)

4.2.2 Assumptions about Test Data

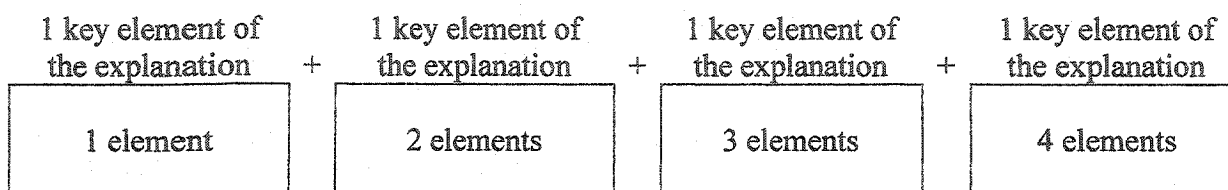
In the new test, we made two assumptions about the test data. First, we assumed that the three test competencies are separate constructs. It seems plausible that an interpreter may excel at linguistic parsing and yet show lower levels in the two discursive abilities, the assignment of reference, and the interpretation of inference. To support this assumption, we note that Setton (1999) sees improvement in interpreting competence as the result of an increased ability to manage the discursive elements of communication (see *Section 3.3.3*). What Setton is suggesting is that the observable skill difference between a beginner interpreter and her more experienced colleague is not likely the result of differences in their respective abilities to parse language units. Instead, the skill differential is probably due to the experienced interpreter's more developed referencing and inferencing abilities. Our belief

in the separate nature of the competencies is also supported by Paradis (1994, 2000), who points out that linguistic abilities are neuroanatomically distinct from the “pragmatic” abilities, such as the referencing of general and situational knowledge and the use of inference (see *Section 3.2.5*). In light of this information, our analysis of the new test (outlined in *Section 4.2.4*) looked for evidence of multidimensionality in the new test data. As it was explained in *Section 4.1.2*, there are multiple references in the literature to tests whose items measure more than one ability, and to analyses that show more than one construct in test data. We feel therefore that our belief in the competencies’ independence and our expectation about multidimensional data are justified.

The second assumption that we made about the new test data was that they are measured on an interval scale. Below, we explain how this is true for each of the three item types, keeping in mind the points raised by Leavitt (1991) about rank-order and degree (p. 80).

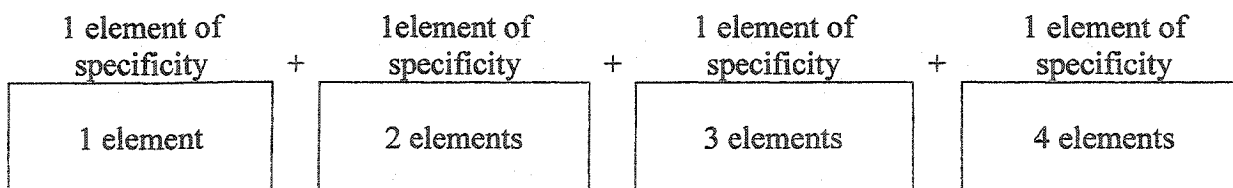
Item type A

Scores for answer options of Type A items constitute interval-level data because they have been quantified. They therefore express both rank-order (the answer option with four key elements of an explanation ranks higher than the answer option with three key elements) and degree (the answer option with two key elements is clearly situated midway between the answer option with one key element and three key elements).



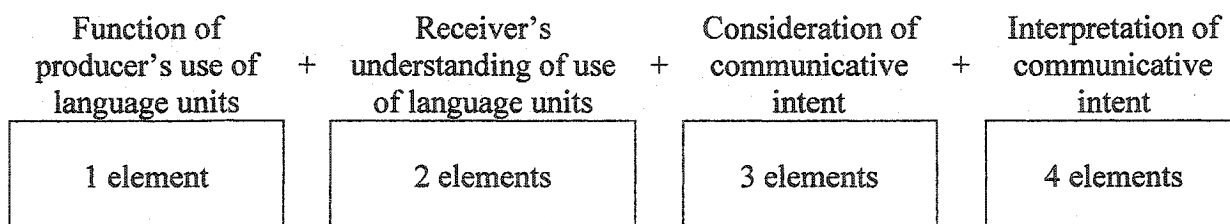
Item type B

The scores for answer options of Type B items have also been quantified, and therefore also constitute interval-level data. In the case of a reference that is quite specific (for example, the sample item given in *Section 4.2.1*, in which the producer uses the term “leadership” to refer to the Prime Minister specifically), the answer option with minimum specificity is ranked lower than the answer option with moderate specificity. Also, the moderately specific answer option contains two “elements of specificity” (in the same sample item, the moderately specific option encompasses the concepts of “leadership” and “government”), which clearly situate it midway between the minimally specific option with one element (the concept of “leadership” alone) and the highly specific option with three elements (the concepts of “leadership”, “government”, and “political party”).



Item type C

Scores for answer options of Type C items are similar to those of the other two types. There is a clear order to the answer options, and each one represents a different level of competence in the interpretation of inference. An answer option that explains a passage of the speech in terms of the producer's use of language, for example, ranks lower than one that explains the passage in terms of the communicative intent behind it. Type C items also contain a system of elements not unlike those of the other two items types. Each answer option adds an element not included in the previous one. The various answer options differ in terms of degree; the option that explains the producer's communicative intent contains three elements, and is therefore situated midway between the option with two elements, and that with four elements.



Because it seems evident that the new test data are on an interval scale, our analysis of the test, outlined below in *Section 4.2.4*, makes use of a number of statistical tests that are parametric in nature.

4.2.3 Sample

The new test was administered to 15 trainee or practising conference interpreters: four students, two interns, and nine experienced professionals. Four of the interpreters indicated that English was their dominant language, while the other 11 all indicated French. Three interpreters were men, and the others were women. All the interpreters in the sample shared the same working languages, and all were asked to work in the same direction (English to French), even though this meant that some participants interpreted into their B language.

In taking this measure, we did run the risk of introducing a potential source of measurement error, because some participants received the text in their dominant language, while others did not. However, it should be stressed that the Francophones all listed English as a B language (an active language into which they occasionally interpret) and not a C language (a passive language into which they never interpret). We believe this indicates that they have a high degree of proficiency in English. This belief was further supported by informal conversations with the participants before and after the experiment – most of the Francophones demonstrated near-native fluency in English, and some were indistinguishable from native speakers.

Having all the interpreters work in the same direction also provided an added benefit. It allowed us to work with a single version of the test and to eliminate version differences as a potential source of measurement error.

The size of this sample – although it is three times as large as the largest uniform group of candidates in the existing test – may still be considered small for statistical purposes. However, it is reflective of the size of the overall statistical population.

4.2.4 Procedure

To test the specific hypotheses related to the new test, a number of steps were carried out. Several of these involved using statistical tests that are parametric in nature, which we feel is appropriate because the new test scores are measurements made on an interval scale (see *Section 4.2.2*). The steps we took are enumerated below, along with a justification for each one.

1. The theoretical origins of the exam constructs were examined as logical evidence of validity. This was also the first step in the analysis of the existing exam, and it is consistent with the *Standards* (AERA, APA & NCME, 1999) outlined in *Section 1.1.2*.
2. Item difficulty indices, item discrimination indices, and *item distractor analyses** were calculated for each of the 45 items in the new test, in order to select a sample of viable items. To calculate the difficulty index for an item, we used the approach described by Thompson & Levitov (1985). The number of participants who received the maximum number of points on that item was divided by the total number of participants (15). To calculate the discrimination index for an item, we followed the method advocated by Ebel

& Frisbie (1986). The participants were first divided into three equal groups (5 high scorers, 5 mediocre scorers, and 5 low scorers), based on their total scores on the test. Then, the number of participants in the low scoring group who received the maximum number of points for the item was subtracted from the number of participants in the high scoring group who received the maximum number of points for the item. Finally, the difference was divided by the number of participants in a group (5). To conduct the distractor analysis, we followed the method outlined by Carey (2001). We calculated the percentage of participants in the high and low scoring groups that chose each answer option. (See *Appendix III* for details.)

3. Pearson's product-moment coefficients were calculated to assess the degree of the relationship among scores for the three constructs in the sample, and a comparison of related correlations, based on Fisher's z , was carried out to assess the similarity of coefficients that appeared to be somewhat different in terms of their magnitude. As we noted in *Section 4.1.4*, Pearson's correlation was used in a similar manner in the credentialing literature, and the Fisher's z comparison allowed us to determine whether the coefficients were significantly different.
4. Cronbach's alpha coefficients were calculated for items within each of the three constructs in the sample in order to provide an estimation of internal reliability. This test was used repeatedly in the credentialing literature we reviewed to examine the consistency of scores within test components (see

Violato, Salami & Muiznieks, 2002; Norcini, Hancock, Webster, Grosso & Shea, 1988; and McGhan, Stimmel, Gilman & Segal, 1982). Ideally, we would have liked to pair the calculation of alpha with a principal components analysis, as recommended by Cortina (1993). However, with 15 items and only 15 test takers, we were too far from the recommendation of 10 to 15 subjects per item made by Pett, Lackey and Sullivan (2003). (To meet the recommendation, we would have needed between 150 and 225 test takers.)

In a discussion of item discrimination, Ebel & Frisbie (1986) suggest that the following scale be used to determine if items on a classroom test discriminate effectively between high and low scorers.

Index value below 0.2	Poor items (to be rejected or revised)
Index value from 0.2 to 0.29	Marginal items (to be revised)
Index value from 0.3 to 0.39	Reasonably good items (to be considered for revision)
Index value of 0.4 and up	Very good items

In the new text, we have evaluated the acceptability of the items based on this scale of discrimination.

The difficulty of an item was considered acceptable if its difficulty index was between 0.37 and 0.87. This range represents 25% above or below 0.62, the so-called ideal for a four-option multiple-choice item (Thompson & Levitov, 1985). To interpret the correlations obtained from the analysis, the scale of values presented in *section 4.1.4* above was used. Significance was set at the 95% level.

4.3 Comparison of the Two Analyses

To test the general hypothesis, the final component of the study compared the analyses of the existing test and the new test. This comparison comprised three steps:

1. A Fisher's z-test for unrelated correlations (Guilford & Fruchter, 1973) was used to compare the correlations of the construct scores from the existing test and the new test, in order to determine if there were significant differences between them;
2. The specific and general hypotheses were reviewed, in order to determine if they were confirmed or refuted by the results of the study; and
3. All the findings were discussed and interpreted in the context of the original statement of the problem.

4.4 Conclusion

This chapter has outlined the methodology used to conduct the present study. With regard to the first component of the study, we have described the general make-up of the existing test, the data obtained from its administration, and the procedure that was used to test specific hypotheses about the existing test. With regard to the second component, we have described the construction of the new test, the data obtained from its administration,

and the procedure used to test specific hypotheses about the new test. With regard to the third component, we have listed the steps taken to compare the analyses of the two tests. The results of the first two components of the study are described in *Chapter 5*, while *Chapter 6* contains the results of the third component.

5. Validation Results

This chapter presents the results obtained from two different analyses. First, we examine the qualitative and quantitative evidence that was gathered as a means of validating the constructs in the existing test. Second, we study similar types of evidence compiled in an effort to validate the constructs in the newly developed test.

5.1 The Existing Test

The analysis of the existing test consisted of four parts.

1. Examination of the Constructs' Theoretical Origins

Because a validity argument should determine if there are good theoretical reasons why the constructs used in the test should predict the performance of candidates in a particular field (AERA, APA & NCME, 1999), the theoretical origins of the test constructs were examined.

2. Estimation of Construct-Irrelevant Variance

Because of the structure of the scoring scheme, it was necessary to estimate the degree to which factors other than the test constructs affected score variance.

Since test constructs were nested within test items, a multitrait-multimethod matrix was used to initially distinguish item variance from construct variance.

Pearson's product moment correlation was then used to examine the degree of

similarity among the scores for the test items (A, B, and C) and Fisher's z to identify any significant differences between pairs of correlations.

3. Estimation of Construct Variance

To provide some measure of construct validity, Pearson's product-moment correlation was employed to examine the degree of similarity among the scores for the three constructs (intelligibility, informativeness, and style). Fisher's z was also used to identify any significant differences between pairs of correlations.

4. Secondary Estimation of Construct Variance

Principal components analysis was used, along with Cronbach's α , to determine if relationships among the data suggested the presence of the test constructs.

5.1.1 Examination of the Constructs' Theoretical Origins

The existing test evaluates candidates' performance according to three competencies which are purportedly unrelated constructs – intelligibility, informativeness, and style, which were briefly explained in *Section 4.1.1*. The origins of the three constructs are explored below in a discussion of the work of three authors.

Anderson (1979)

The rating scales used in the existing test by Association 1 were first developed for use by another professional association of translators and interpreters (hereafter known as

“Association 2”). Sometime around 1979 or 1980, Association 2 asked Linda Anderson to develop a certification exam for them (Anderson, 2001), and she proposed a test instrument based on a scoring scheme she had used in her M.A. thesis (Anderson, 1979). The scoring scheme was borrowed from Carroll (1966) and was made up of two rating scales designed for machine translation: the first rated the intelligibility of a translation on a 10-point descriptive scale; the second rated the informativeness of a translation on a 10-point descriptive scale. Intelligibility and informativeness are described by Carroll as independent constructs (see below). When the scales were applied to conference interpreting, some members of Association 2 felt that exam candidates should also be rated on a third competency that had up until that point not been addressed in the test’s scoring scheme – style. Despite her opposition to this request, Anderson accommodated it by creating a third 10-point descriptive scale. She removed some features from the intelligibility scale as written by Carroll and used them to create rating scale for style. The intent behind the addition of style was to end up with a test that measured three separate constructs, but Anderson admits that there may be a high degree of “cross-leaking” between the intelligibility and style scales (Anderson, 2001).

Carroll (1966)

The material that makes up the three rating scales was first published in Carroll’s (1966) discussion of the evaluation of machine translation. In the article, the author attempts to determine how to measure, and subsequently proceeds to measure, two features that were considered highly desirable in automated translation at the time. The first of these was intelligibility, which Carroll describes as the degree to which a sentence is as understandable in the TL as it was originally in the SL. The second was fidelity, which was measured by

reading the target text, reading the source text, and then estimating the informativeness of latter as compared to the former. If a translation has high fidelity, reading the original should provide little information to supplement or controvert the prior reading of the target text. Carroll indicates that intelligibility and informativeness are “conceptually separable variables”. In other words, it is possible for a translation to be perfectly intelligible, but at the same time be only minimally informative. The opposite is also possible – a translation can be largely unintelligible, yet highly informative (1966, p. 57). Carroll notes that he chose these two evaluation criteria in discussion with the Automatic Language Processing Advisory Committee of the U.S. National Academy of Sciences, but offers little additional information.

Nida (1964)

Some have speculated that Carroll’s article was indirectly influenced by the work of Nida (1964) on Bible translation, and this suggestion appears to be plausible on the surface. In developing his well-known distinction between formal and dynamic equivalence, Nida discusses the two opposing goals that guide translators in their work. First, he notes that translators may simply want to pass along information, either of a cognitive or emotional nature. As the ultimate example of an informative translation, Nida cites the gloss, which is often accompanied by footnotes and appendices, and which is intended to teach readers about the source language¹. Second, he points out that translators may want to produce a particular type of behaviour in their target-language readers, and under such circumstances

¹ As some have pointed out, however, a gloss may not always be the archetype of informativeness. If it relays a great deal of information about the linguistic structure of the source language, it may do so at the expense of the readability of the source text’s content.

they are likely to try and make the text as intelligible as possible. For example, to effectively transmit Christian behaviour to other cultures, Bible translators may need to make adaptations. The metaphor “as white as snow” may have little meaning in a tropical culture, and may need to be replaced with a local equivalent (e.g. “as white as egret’s feathers”). However, it should be noted that informativeness and intelligibility in Nida’s view are competing purposes; translators may decide to either make a text informative or make it intelligible. In this sense, his use of the two criteria differs from that of Carroll.

The information gathered from the three authors’ work does not appear to support the constructs’ validity for use in interpreter certification. In a best-case scenario, the rating scales used in the existing test can be traced back as far as Nida’s work. Although this would connect them to the general discipline of Translation Studies, it also suggests that the scales would be more appropriately used in certain types of written translation. The professional field to which Nida himself had the closest connection was that of Bible translation. Through his own experience, and through his observation of the work of others, he presumably gained a significant understanding of the competencies that Bible translators must develop, but the same experience and observation would leave him ill-prepared to understand the world of the conference interpreter. There is little evidence in his writings to suggest that his conceptualization of informativeness and intelligibility bear a connection to the profession of conference interpreting.

In a worst-case scenario, the scales can only be traced back to Carroll’s work. Carroll is a cognitive psychologist who is interested in applied linguistics (Plucker, 2003), and who conducted a certain amount of research in the field of machine translation. The job-related

areas that he has likely had the most first-hand knowledge of and the greatest opportunity to study are information technology and the language professions, and this is evidenced in his writing. His view of the translation process is a linguistic one – it is limited to the substitution of target language units (predominantly lexical and semantic) for source language ones. This view fails to see translation as the communication of a message in context, and it has been vigorously criticized by those in the field of conference interpreting². Therefore, we must conclude that there is little evidence linking Carroll's conceptualization of informativeness and intelligibility to conference interpreting.

Some may argue that a connection between interpreting and at least one of the scales was established in Anderson's work, since the third rating scale – style – was included at the behest of several members of Association 2, who arguably had important first-hand knowledge of the relevant professional field. However, there is no record of how many members made the request, nor is there any indication that all were conference interpreters themselves (Anderson, 2001). Perhaps most importantly, there is no evidence to demonstrate how widely their opinion was shared in the field of interpreting as a whole. It should also be remembered that the material for the style rating scale came ultimately from the intelligibility rating scale as first published by Carroll.

² The view of translation as an exercise in word substitution between the source and target languages is something that Seleskovitch and Lederer, for example, object to in nearly all of their writings. One single citation (among many possible ones) illustrates this point effectively, "Les elements transcodables sont myriades mais le transcodage qu'ils appellent est toujours ponctuel; pour être clairement réexprimé, le texte dont ils font partie doit toujours être interprété grace à des connaissances autres que linguistiques" (1984, p. 8).

At this juncture, our examination of the existing test turns from the logical to the empirical. The following sections report the findings of statistical analyses of the scores obtained from the test.

5.1.2 Estimation of Construct-Irrelevant Variance

Because scores for the three constructs (intelligibility, informativeness, and style) are nested within the items (A, B, and C, each one a ten-minute extract of a recorded speech), it is not possible to begin the empirical analysis of the existing test with a simple comparison of scores for each of the constructs. If any discrepancies were noted in the comparison of the constructs' scores, it would not be immediately apparent whether they were the result of differences among the constructs, or of differences among the items. What is needed at this juncture is a tool that makes it possible to interpret relationships between sets of scores that are nested within one another, as is the case here with construct and item scores.

Consequently, we have used a multitrait-multimethod matrix (Campbell & Fiske, 1959) as a preliminary means of approaching the data. The matrix allows us to examine a series of *traits* (in this case, the three constructs, intelligibility, informativeness, and style) that are measured by a series of *methods* (the three test items, A, B, and C) by organizing the correlations for all the possible combinations of traits within methods in a meaningful way.

The matrix allows us to look for evidence of two types of construct validity: 1) convergent validity, which is supported if constructs that theoretically should be related are interrelated in reality; and 2) discriminant validity, which is supported if constructs that

theoretically should not be related are, in fact, not related in reality. Information gathered on the existing test suggests ideally that scores for the three test constructs should be largely unrelated (although there may be some degree of relationship between scores for intelligibility and scores for style), and that scores for the three items should be related (because the items are purportedly similar in difficulty).

Table 2: Multitrait-Multimethod Matrix of Correlations for Constructs within Item Scores

		Item A			Item B			Item C		
		Int	Inf	Sty	Int	Inf	Sty	Int	Inf	Sty
Item A	Int	1.000								
	Inf	.720	1.000							
	Sty	.734	.716	1.000						
Item B	Int	.836	.634	.655	1.000					
	Inf	.712	.787	.731	.750	1.000				
	Sty	.747	.699	.761	.636	.675	1.000			
Item C	Int	.795	.595	.643	.869	.697	.653	1.000		
	Inf	.704	.778	.710	.725	.886	.668	.765	1.000	
	Sty	.717	.564	.739	.706	.652	.742	.725	.694	1.000

To interpret the matrix, featured in *Table 2*, we must examine its four components.

1. The Reliability Diagonal

The one sloping line of perfect correlations in the matrix, represented in each case by a value of 1.000, shows correlations that compare scores for the same constructs within the same items.

2. The Heterotrait-Monomethod Triangles

Made up of solid lines, these three triangles contain the correlations that compare scores for different constructs within the same items.

3. The Heterotrait-Heteromethod Triangles

Made up of dotted lines, these six triangles contain the correlations that compare scores for different constructs within different items.

4. The Validity Diagonals

The three sloping lines of three correlations sandwiched between the dotted triangles show correlations that compare scores for the same constructs within different items.

Our interpretation of the matrix hinges on our answers to four questions.

1. Are the correlations in the validity diagonals significantly different from zero, and are they of a sufficient magnitude (see *section 4.1.4*) to encourage further examination of validity?

2. Is each value in the validity diagonals higher than the values in its column and row in the adjacent heterotrait-heteromethod triangles?
3. Is each value in the validity diagonals higher than all the values in the heterotrait-monomethod triangles?
4. Is the same pattern of trait relationships seen within all the triangles?

The answer to the first question is yes. In fact, all the correlations in the validity diagonals are of a sufficient magnitude to be considered evidence of moderate or high correlation. This suggests that scores given for an individual construct are relatively similar, regardless of which item we examine. In other words, the intelligibility scores are similar for items A, B, and C; the informativeness scores are similar for items A, B, and C; and the styles scores are similar for items A, B, and C. This finding is consistent with one of our expectations, in that it suggests there are no obvious differences in difficulty across the three items.

The answer to the second question is also yes. For instance, the validity diagonal intelligibility comparison between items A and B (.836) is higher than the values in its adjacent heterotrait-heteromethod row (.634 and .655) and in its adjacent heterotrait-heteromethod column (.712 and .747). However, the differences between the values are not as great as those seen in other studies (Campbell & Fiske, 1959). This suggests that relationships in the data are stronger where we expect to see them (the same construct scores

for different items) than where we do not expect to see them (different construct scores for different items). However, the difference in strength of relationship is not as stark as we might have hoped. This is the first sign in the data that there is a general similarity across all scores.

The answer to the third question is no. Several correlations in the heterotrait-monomethod triangles are higher than values in the validity diagonals. (The values of .742 for style in the comparison of items B and C, and of .739 for style in items A and C are lower than the values of .765 for the item C comparison of intelligibility and informativeness and of .750 for the item B comparison of intelligibility and informativeness.) This suggests that relationships in the data are not always stronger where we expect to see them (the same construct scores for different items) than where we do not expect to see them (different construct scores for the same items). This is the second sign in the data that there is a general similarity across all scores.

Finally, the answer to the fourth question is also no. The pattern appears to be random. This is not consistent with our expectation – if we were to find similarity between scores for different constructs, we expected that it would be between intelligibility and style (this is what Anderson predicted). There are nine triangles altogether, and the highest correlation occurs between intelligibility and style in two triangles, between informativeness and style in three triangles, and between intelligibility and informativeness in four triangles.

Within this particular matrix, some of Campbell and Fiske's interpretation criteria have been met, while others have not. There are high correlations for the same construct

across all the items, but the correlations that correspond to other score comparisons are nearly as high. This suggests there is a high degree of similarity in the exam scores, across both items (which is consistent with our expectations about the data) and constructs (which is not). To demonstrate this more clearly, we have analyzed the two sources of variance separately. Correlations between item scores have been examined immediately below, while correlations between construct scores are analyzed in the *section 5.1.3*.

Ideally, the three items on the exam (A, B, and C, each one a ten-minute excerpt of a conference speech) should be of a similar level of difficulty, and test takers' scores should be consistent across items. We used Pearson's product-moment correlation to verify if this was the case. In preparation for the analysis, all the data from the existing test were arranged in three columns, according to the item (A, B, or C) for which the scores were generated. The tables below present the descriptive statistics for exam items, and the correlations between them.

Table 3: Descriptive Statistics for the Existing Test Items

Item	Number of Observations	Mean	Standard Deviation
A	198	6.523	1.684
B	198	6.184	1.717
C	192	6.194	1.691

A quick survey of the means for the three items already reveals that they are quite similar, and that the strongest resemblance is between items B and C.

Table 4: Correlations among the Existing Test Items

	Item A	Item B	Item C
Item A			
Pearson Correlation	1.000	.790	.774
Significance (2-tailed)	.	.000	.000
N	198	198	192
Item B			
Pearson Correlation	.790	1.000	.818
Significance (2-tailed)	.000	.	.000
N	198	198	192
Item C			
Pearson Correlation	.774	.818	1.000
Significance (2-tailed)	.000	.000	.
N	192	192	192

In *Table 4*, the correlations among all the pairs of exam items (A and B, A and C, and B and C) are all statistically significant (i.e., with a significance of .000 they are obviously below the .05 level), and all three (.790, .774, and .818) are evidence of high correlation. (Refer to *section 4.1.4* for interpretation of correlation magnitude.)

While the three correlations were all of a similar magnitude, we nevertheless investigated the possibility that there were statistically significant differences between them using a test for correlated coefficients (Guilford & Fruchter, 1973).

Table 5: Comparison of Correlated Coefficients for the Existing Test Items

Correlations Compared	Coefficients	Z	α	Accept Null Hypothesis?
AB and AC	.790 and .774	.643	1.96	Yes
BC and AC	.818 and .774	1.755	1.96	Yes
BC and AB	.818 and .790	1.112	1.96	Yes

Because the value for z was lower than the α for a two-tailed significance level of .05, we accepted the null hypothesis and concluded that there were no significant differences between the coefficients.

The bulk of the evidence examined in this section – the multitrait-multimethod matrix, the Pearson’s product-moment correlation coefficients, and the results of the Fisher’s test for correlations – suggests that the three test items (A, B, and C) are highly similar. This leads us to believe that the scores for the three items are quite uniform, and that the variation among candidates is likely not due to differences in item difficulty. It therefore appears reasonable to conclude that construct-irrelevant variance is within an acceptable range.

5.1.3 Estimation of Construct Variance

A series of correlation coefficients was also calculated for the three item constructs. Carroll’s (1966) writing on the subject would lead us to expect a low level of correlation between intelligibility and informativeness, because they are considered separate constructs. Anderson’s (2001) account of the development of the existing test’s scoring scheme leaves

us with an expectation that is less clear. Ideally, style should be a construct separate from intelligibility and informativeness (the members of Association 2 felt style was not sufficiently accounted for by the rating scales for intelligibility and informativeness), but because of the way the style scale was developed, we may expect to see a higher level of correlation between intelligibility and style (the latter scale was constructed by borrowing material from the former). The descriptive statistics for the three variables are presented in the table below.

Table 6: Descriptive Statistics for the Existing Test Constructs

Construct	Number of Observations	Mean	Standard Deviation
Intelligibility	196	6.584	1.610
Informativeness	196	6.389	1.639
Style	196	5.931	1.793

A simple examination of the means already reveals some interesting findings. The scores assigned for the three constructs appear to be similar, with the greatest resemblance between intelligibility and informativeness (with means of 6.584 and 6.389 respectively). This is surprising, because it runs contrary to our expectations about the data.

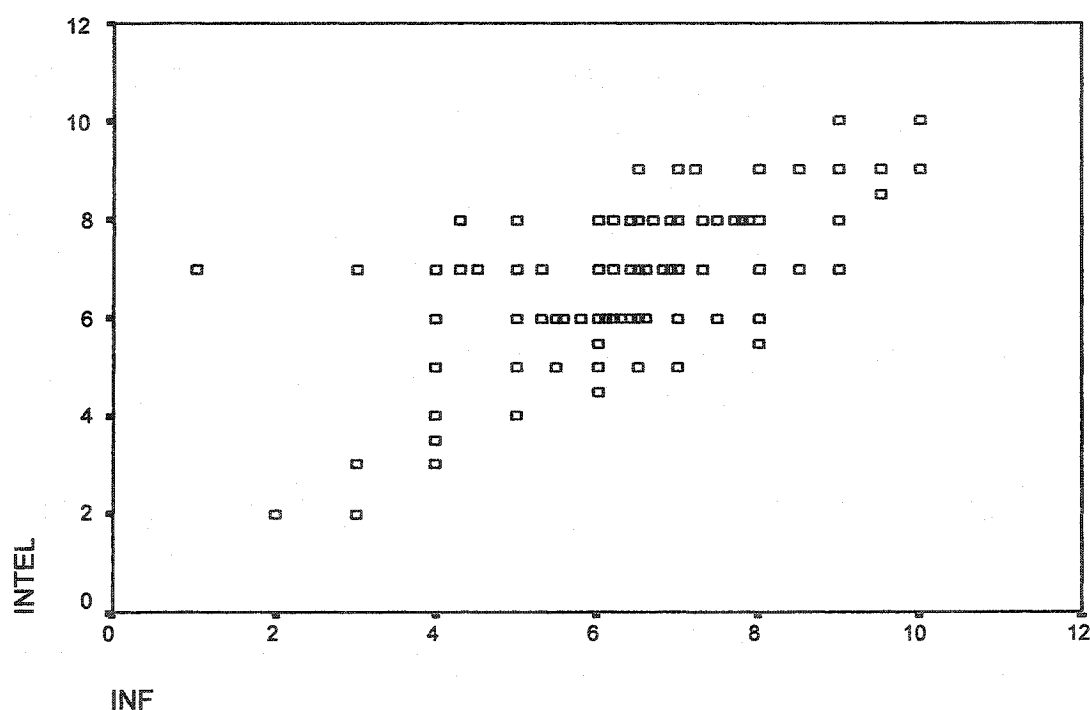
Table 7: Correlations among the Existing Test Constructs

	Intelligibility	Informativeness	Style
Intelligibility			
Pearson Correlation	1.000	.746	.690
Significance (2-tailed)	.	.000	.000
N	196	196	196
Informativeness			
Pearson Correlation	.746	1.000	.691
Significance (2-tailed)	.000	.	.000
N	196	196	196
Style			
Pearson Correlation	.690	.691	1.000
Significance (2-tailed)	.000	.000	.
N	196	196	196

All three correlation coefficients are statistically significant (i.e., with a significance of .000 they are obviously below the .05 level), and all three (.746, .690, and .691) are evidence of high correlation. (Refer to *Section 4.1.4* for interpretation of correlation magnitude.) The greatest correlation in Table 5 (.746) is between intelligibility and informativeness, and it suggests that there is tremendous overlap between the two constructs. This finding is consistent with our initial examination of the means above, yet both are unexpected. Information on the origins of the constructs (see *Section 5.1.1*) would lead us to assume the greatest overlap would be between intelligibility and style. The correlations between intelligibility and style (.690), and between informativeness and style (.691) also suggest a high degree of relatedness, although not as high as between the first two construct scores.

Another way to understand the relationship between the intelligibility and informativeness scores is to plot it on a graph. This was done with a scatter-plot (see below), where intelligibility is represented on the y-axis, and where informativeness is represented on the x-axis. The high degree of correlation is apparent on the graph – it appears that in only one instance was a high intelligibility score paired with a low informativeness score.

Figure 1: Scatter-Plot of Scores for Intelligibility and Informativeness



The magnitude of the correlation between intelligibility and informativeness appeared to suggest that the relationship between the two constructs was in some way more significant than the others outlined in *Table 7* (i.e., the relationships between intelligibility and style, and between informativeness and style). To investigate this possibility, we once again

conducted a test for significant differences between correlated coefficients (Guilford & Fruchter, 1973).

Table 8: Comparison of Correlated Coefficients for the Existing Test Constructs

Correlations Compared	Coefficients	Z	α	Accept Null Hypothesis?
Int/Inf and Int/Sty	.746 and .690	1.595	1.96	Yes
Int/Inf and Inf/Sty	.746 and .691	1.566	1.96	Yes
Int/Sty and Inf/Sty	.691 and .690	.029	1.96	Yes

Because the z values calculated for each pair of correlations was below 1.96 (the two-tailed α value for a .05 significance level), we accepted the null hypothesis and concluded that there were no significant differences between the coefficients. This in turn suggests that the nature of the relationship between intelligibility and informativeness is similar to that between intelligibility and style, and to that between informativeness and style.

In other words, the weight of the evidence examined thus far leads us to believe that there is a general similarity among all the scores, regardless of what construct they purport to measure. There are some variations in the strength of that similarity, but the variations are not significant. This is not consistent with our expectations about the test data. The test is intended to measure three competencies that are believed to be distinct constructs. The data therefore should have been multidimensional. Instead, all the evidence examined up to this point indicates that the data are unidimensional.

5.1.4 Secondary Estimation of Construct Variance

The final step in the analysis of the existing test data was to make a second estimation of the test scores' unidimensionality, that is, of the degree to which they appear to be explained by a single underlying factor. To determine unidimensionality, Cortina (1993) recommends first using a principal components analysis. If the results of the analysis indicate the presence of a single common factor in a set of items, he suggests confirming the findings by calculating alpha for the same set of items.

Principal components analysis is often used by researchers to examine the variables in their data and to group related variables together. The resulting list of grouped variables or *principal components* helps to provide a new interpretation of the data. Consider, for example, a test that uses a large number of items to measure a smaller number of constructs. Ideally, a principal components analysis of the variables (in this case, the scores for the large number of test items) should reveal a set of components that is the same, or similar to, the number of constructs that the test purports to measure.

Cronbach's alpha is used by researchers to estimate the internal consistency of scores. It provides a measure of the average inter-item correlation, and it is sensitive to the number of items on a test. Ideally, alpha should only be calculated for items that measure the same thing.

Although it might initially seem unorthodox to use this statistical procedure with a small number of test takers (26)³, the datasheet was organized in a manner that allowed us to take advantage of the test's complex design. We considered that there were nine different variables:

1. the intelligibility score for item A;
2. the informativeness score for item A;
3. the style score for item A;
4. the intelligibility score for item B;
5. the informativeness score for item B;
6. the style score for item B;
7. the intelligibility score for item C;
8. the informativeness score for item C; and
9. the style score for item C.

We also arranged the data in terms of observations (an interaction between candidates and raters), which allowed us to increase the N from 26 to 66. It would have been exceedingly problematic to organize such that each candidate's scores were displayed in a single row in our table of data, because more than one rater scored each candidate's performance, and because the number of raters was not consistent.

³ Pett, Lackey and Sullivan (2003) recommend having at least 10 to 15 subjects per test item before using factor analytic techniques. However, it is unclear to what extent this recommendation is followed in published research. Violato, Salami and Muiznieks (2002) were able to meet the 10-subject threshold, but not the 15-subject one; they performed an analysis of nine variables with data from 112 test takers. Price and Wilkins (2001) did not meet either threshold; their analysis was conducted on 85 items with data from 400 test takers.

The charts presented below illustrate what our table of data might have looked like. With all a candidate's scores in one row, we would have been forced to artificially expand the number of variables from nine to 27 in some cases (e.g., Candidate 2), and from nine to 18 in others (e.g., Candidate 13). This would have made it appear that we were working with a larger number of items than was actually the case, and it would have obscured the relationship between different raters' scores for the same item. For example, Raters 4, 5, and 6 all gave an intelligibility score for Item A, but the organization of the table makes it appear that the scores are attributed to separate variables.

	Rater 4									Rater 5									Rater 6								
	Item A			Item B			Item C			Item A			Item B			Item C			Item A			Item B			Item C		
C2	I	I	S	I	I	S	I	I	S	I	I	S	I	I	S	I	I	S	I	I	S	I	I	S	I	I	S
	n	n	t	n	n	t	n	n	t	n	n	t	n	n	t	n	n	t	n	n	t	n	n	t	n	n	t
	t	f	y	t	f	y	t	f	y	t	f	y	t	f	y	t	f	y	t	f	y	t	f	y	t	f	y

	Rater 13									Rater 14								
	Item A			Item B			Item C			Item A			Item B			Item C		
C13	I	I	S	I	I	S	I	I	S	I	I	S	I	I	S	I	I	S
	n	n	t	n	n	t	n	n	t	n	n	t	n	n	t	n	n	t
	t	f	y	t	f	y	t	f	y	t	f	y	t	f	y	t	f	y

Arranging candidate data in a single row would also have forced us to work with a smaller data set. We would either have had to choose to include only those candidates who, like Candidate 2, had been scored by three raters (a total of 12 candidates), or we would have had to elect to discount a third rater's scores for the first 12 candidates (so that the dataset included only two sets of scores for each candidate). This second option would also have presented us with the difficulty of deciding which rater's scores to eliminate.

Daniel and Siders (1994) also failed to meet the recommended numbers; they used data from 194 test takers to

The table below lists the principal components extracted by the analysis, along with their corresponding eigenvalues. The eigenvalues provide a measure of the variance associated with a particular principal component. In keeping with the Kaiser-Guttman criterion, only those components with an eigenvalue greater than one were considered in our interpretation of the results.

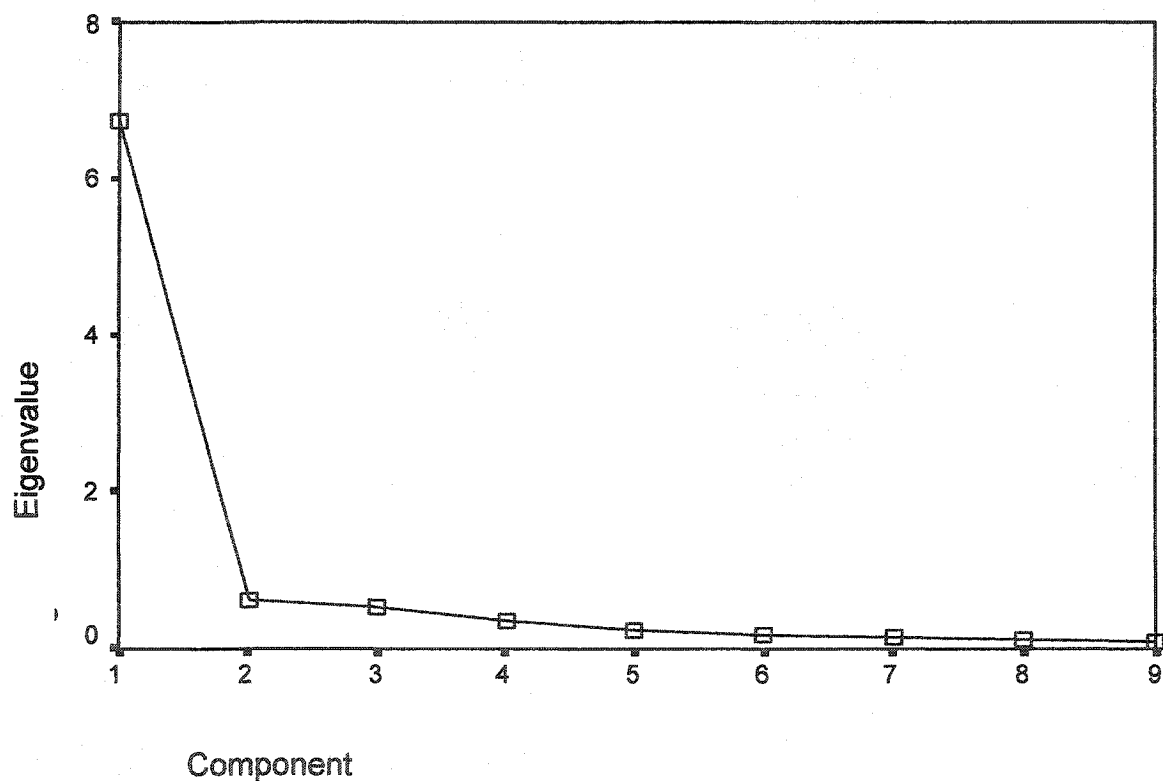
Table 9: Principal Components Analysis of Existing Test Constructs

Component	Eigenvalue	Initial Eigenvalues	
		% of Variance	Cumulative %
1	6.749	74.992	74.992
2	.615	6.836	81.828
3	.534	5.937	87.765
4	.340	3.781	91.546
5	.240	2.668	94.214
6	.180	1.997	96.211
7	.141	1.568	97.779
8	.125	1.384	99.163
9	7.534E-02	.837	100.000

The analysis extracted a total of nine components from the data, but only component number one has the required eigenvalue magnitude ($6.749 > 1$), and there appears to be a very stark contrast between it and all of the remaining components. This suggests that there is only one principal component in the data.

Another way to view these results is to graph them on a scree plot. The point at which there is an obvious change in the slope of the line indicates the number of components that should be retained in the interpretation of the results. The plot appears in the figure below.

Figure 2: Scree Plot of Eigenvalues for Factor Analysis of Constructs



The graph in *Figure 2* supports our interpretation of the results in *Table 9*. Both appear to indicate that the scores for the three constructs (intelligibility, informativeness, and style) actually contain only one principal component. This finding is consistent with the results of our correlation (see *Section 5.1.3*), and it reinforces the suggestion that there is a very high degree of overlap among the three exam constructs.

On the basis of these results, we also calculated Cronbach's alpha for the nine variables. Again, it might initially seem unusual to do so, since the variables are purported to be measures of different constructs. However, all of the analyses we have conducted thus far indicate the opposite and suggest that the variables are all measuring the same thing. We therefore went ahead with the calculation of this statistic, and the results are presented below.

Table 10: Coefficient Alpha for Existing Test Constructs

Construct	Number of Cases	Number of Items	Coefficient Alpha
All test variables	64	9	.9571

The magnitude of the coefficient suggests that it is evidence of very high correlation. It also indicates that there is a high degree of consistency within the nine variables, and it appears to be further proof of their unidimensionality. However, we make this observation cautiously. We recognize that we have used both the principal components analysis and Cronbach's alpha under less-than-ideal circumstances, and that the findings are therefore somewhat suspect.

Yet even if we cannot have total confidence in the results of the last two tests, they do present a picture that is consistent with everything else that we have seen up to this point. The data collected provide a clear image of the constructs used in the existing test, and of the pertinence of their use in an interpreter certification test. We have seen little evidence to connect the use of the three existing test competencies to the profession of conference interpreting, and we have seen little evidence to confirm that the competencies are separate

constructs, as the existing test purports. With our analysis of the existing test complete, we next turn our attention to the new test and to the constructs that support it.

5.2 The New Test

The analysis of the new test consisted of four parts.

1. Examination of the Constructs' Theoretical Origins

We examined the theoretical origins of the new test constructs in order to determine if there were good theoretical reasons why they should predict interpreter performance.

2. Estimation of Construct-Irrelevant Variance

We calculated discrimination and difficulty indices, and we conducted a distractor analysis, all as a means of distinguishing between item variance and construct variance.

3. Estimation of Construct Variance

We used Cronbach's alpha to determine the degree of intra-construct score similarity.

4. Secondary Estimation of Construct Variance

We employed Pearson's product-moment correlation to determine the degree of similarity among scores associated with different constructs.

5.2.1 Examination of the Constructs' Theoretical Origins

The new test is based on three constructs – linguistic parsing, the assignment of reference, and the interpretation of inference. The constructs are derived from Setton's (1999) model of the interpreting process, but their presence can also be seen in a number of other previous models that influenced Setton's work. We have discussed the origins of the constructs by tracing their development back through the various models. We have also attempted to establish the models' connection to the interpreting profession by noting the professional credentials of their authors, and the participation of other interpreters in efforts to test them.

Setton (1999)

Setton is a practicing conference interpreter who teaches at the *École de traduction et d'interprétation* at the *Université de Genève* (Pöchhacker & Shlesinger, 2002). His cognitive-pragmatic model was the outcome of his analysis of a corpus of spoken-language interpreted discourse communicated in a conference setting. The model has three key components that have been discussed in more detail above (see *Section 4.2*): 1) the parser, which processes linguistic input to form a basic proposition; 2) the assembler, which houses the parser, assigns non-linguistic reference, and constructs an elaborated proposition for

further processing; and 3) the executive, which interprets inference and judges the match between the inference and the elaborated proposition as a final stage in the comprehension of input.

The novelty of Setton's model does not lie in the outright creation of its components, for the origin of each one can be traced to earlier work. Instead, Setton's contribution resides in his ability to successfully blend ideas from a number of competing schools of thought in Interpreting Studies. He incorporates certain aspects of interpretive theory into an information-processing approach, in the process constructing a model that may also be described as discursive (Pöchhacker & Shlesinger, 2002), because it can be used to describe and explain elements of communication that are not strictly rooted in linguistic structure. The contributions that these other schools of thought have made to Setton's model are discussed below.

Moser (1978)

A trained conference interpreter and active AIIC member, Moser-Mercer has led a distinguished career as an interpreter trainer at both the Monterey Institute of International Studies, and the *École de traduction et d'interprétation* of the *Université de Genève* (Pöchhacker & Shlesinger, 2002). A long-time proponent of the information-processing approach, she constructed one of the first psychological models of the interpreting process. She used her model to develop a test for interpreting aptitude, and the results of her test were shown to predict student performance in an actual interpreter training program (Moser-Mercer, 1984).

Two of the constructs in the new test – the linguistic parsing, and the interpretation of inference – have predecessors in Moser’s work. Her model describes linguistic processing in great detail, by meticulously outlining the analysis of an incoming message. Primary phonological recognition, syllabic reconstitution, assembly of detected syllables into recognized words, assembly of word strings, and the recognition of meaningful phrases are all spelled out clearly in the model.

The interpretation of inference also finds a place in Moser’s work, if only in a tentative way. The author notes that interpreters’ understanding depends not only on what they hear or see, but also on the relevant knowledge that they already possess. She explains that this “prior information”, when combined with “received information” allows interpreters to predict upcoming input. However, this idea is incorporated into the model in a limited fashion; actual examples of how prediction might work are all linguistic in nature (e.g., prediction as a function of differences in word order between English and German).

Gile (1997)

Moser-Mercer is not the only proponent of the information-processing approach mentioned by Setton. He also examines Gile’s (1997) effort model. Gile is an experienced conference interpreter and interpreter trainer who teaches at the *Institut supérieur d’interprétation et de traduction* of the *Institut catholique de Paris*. His model is based on extensive observation and qualitative analysis of interpreters in the field.

Gile’s model, like the others discussed already, suggests that linguistic processing is

an important part of interpreting. The first of the four “efforts” that comprise the model, listening and analysis, describes how interpreters arrive at an understanding of the source language input. All of the tasks listed under this effort are linguistic in nature: the analysis of sound waves, the identification of words, and the determination of the meaning of a sentence. In this sense, Gile’s model reinforces the notion that interpreting is a linguistic exercise.

Seleskovitch & Lederer (multiple publications)

The information-processing approach was not the only school of thought to influence Setton’s work, since elements of interpretive theory can also be clearly seen in his cognitive-pragmatic model. The originators of interpretive theory, Seleskovitch and Lederer, were both practicing conference interpreters, and they both trained interpreters for many years at the *École supérieure d’interprètes et de traducteurs* of the *Université de Paris III* (Pöchhacker & Shlesinger, 2002, p. 120, 130). Their model of the interpreting process was based on extensive observation of conference interpreting, and it was tested against actual corpora of conference interpretations (Seleskovitch, 1975; Lederer, 1981).

There are elements of interpretive theory that correspond to all three constructs used in the new test. First, interpretive theory suggests that interpreting ability depends, in part, on linguistic ability. Lederer (1978) argues that “...since [interpreting] implies a comparison of speech acts in different languages it also obliges us to tread on linguistic ground” (p. 323). This is a point that both she and Seleskovitch further underscore in a later discussion of interpreter pedagogy. They note that a majority of students in the ESIT’s programs arrive at the school with a high degree of competency in their working languages, but that they still

need to continue their language training. Particularly when dealing with the technical subjects imposed by the conference setting, students regularly demonstrate difficulty with phoneme recognition, rare syntax, or unusual verb inflections (Seleskovitch & Lederer, 1984: 217). This clearly suggests that part of the interpreting process is linguistic in nature.

Second, interpretive theory suggests that an interpreter's linguistic analysis must be combined with situational information. In the early years of conference interpreting, conference organizers gave little regard to the needs of the interpreters in the conference hall. Organizers often placed the interpreting booth in a location from which the interpreters had only an obstructed view of the proceedings, and at times they placed the interpreters in another room so that they had no view of the proceedings at all. Seleskovitch (1968) criticizes these practices, explaining that an unobstructed sightline of the conference floor is an absolute necessity. Interpreters need to know who the conference participants are, to understand the roles they play in the larger discussion, and to follow the back and forth of their statements and responses in the proceedings. When the conference participants make reference to people, places, and things that are either physically present in the conference hall or that have been previously mentioned in the proceedings, interpreters must have a means of reconstituting these references. Interpreters are continually called upon to correctly identify "this esteemed colleague of ours", "the block of seats to my right", "the place from which I'm speaking", or even "the point raised by the previous presenter". And since the linguistic input alone is not enough to allow the interpreters to do this, they must make use of situational information. This in turn suggests that the assignment of reference is a competency distinct from linguistic ability.

Finally, interpretive theory also suggests that an interpreter's ultimate comprehension of meaning depends on an ability to understand the intentions behind a conference participant's words. In the same discussion noted above, Seleskovitch (1968) remarks that all the participants in the conference proceedings have an objective or end goal in mind. The final – and interpretive theorists would argue the most important – part of an interpreter's task is to understand that end goal and use it to decipher meaning. In other words, interpreters must use some other source to infer information that will fill in the gaps left by pure linguistic code and immediate deixis.

This short discussion allows us to draw two conclusions about the three constructs used in the new test. First, the constructs are rooted in a successive series of models that date as far back as some of the earliest scholarly writings on conference interpreting (i.e., interpretive theory, starting with Seleskovitch, 1968). Second, the models from which the constructs are derived are connected to the interpreting profession, both through their authors' personal experience as interpreters, and through their authors' quantitative or qualitative analysis of actual interpreting.

With these conclusions drawn, we now turn to our quantitative examination of the data gathered from the new test. To the extent that it was possible, we endeavoured to model this analysis after the one used to examine the existing test. For this reason, we will begin with an attempt to distinguish item variance from construct variance, before using a number of measures to establish the validity of the constructs.

5.2.2 Estimation of Construct-Irrelevant Variance

Although items on the new test were nested within constructs (instead of the other way around, as was the case with the existing test), there was still a risk that inconsistencies across items could have clouded an examination of construct scores. We therefore tried to determine to what degree the overall score variance could be explained by item variance, rather than construct variance. In other words, we began this part of the analysis with two expectations that mirrored those guiding our analysis of the existing test. First, we expect the test data will be multidimensional, because the three new test competencies are believed to be three separate constructs. Setton (1999) suggests that linguistic parsing is functionally separate from discursive abilities (see the discussion improvement of skill in *Section 3.3.3*), and Paradis (1994, 2000) adds that the two sets of abilities are neuroanatomically separate (see *Section 3.2.5*). Setton also indicates that the assignment of reference is functionally separate from the interpretation of inference; the two skills are governed by different components of his model.

Second, we expect that variance in the data will be caused by candidates' different levels of ability in the three competencies, and not by inconsistencies in the test items.

To examine this second expectation, we conducted an item analysis of the new test, which comprised three different procedures: the calculation of item difficulty indices, the calculation of item discrimination indices, and the analysis of item distractors. To facilitate

the item analysis, participants were ranked in terms of their final score and arranged into three score groups: high, moderate, and low. The raw scores, difficulty indices, discrimination indices, and distractor analyses for all the items on the new test are presented in *Appendix III – New Test Documentation*. (The size of the tables prevented them from being presented in this chapter.)

To calculate the difficulty indices, we used the method advocated by Thompson & Levitov (1985). We counted the number of participants who had chosen the answer option worth four points (the highest scored option) and divided that number by the total number of participants (15). It was determined that 22 items had a difficulty index that ranged 25% above or below .62. This suggested that nearly half the items piloted had a difficulty level within the ideal range (Thompson & Levitov, 1985).

To calculate the discrimination indices, we used the method described by Eble & Frisbie (1986). We subtracted the number of participants in the low scoring group who chose the answer option worth four points from the number of participants in the high scoring group who chose the same option. The difference was divided by the number of participants in each group (five). It was determined that ten items out of the total pool of 45 had discrimination indices greater than or equal to .4, which in turn suggested that these items discriminated effectively between high and low scorers on the test. This finding appears to be consistent with those of other professional certification tests. Ebel and Frisbie (1986) argue that certification exams and other competency tests tend to have lower variability than that of classroom tests, which often results in discrimination indices of zero or near zero.

To analyze the distractors, we used the method explained in Carey (2001). We calculated the percentage of participants in the high scoring group and in the low scoring group who chose each answer option. If the distractors were working effectively, the high scoring group should have outperformed the low scoring group (by choosing the “correct” answer), and all the answer options should have been chosen by participants in either group. It was discovered that the high scoring group performed best on 20 items, that the low scoring group performed best on 10 items, and that the two groups were equal on 15 items. We also learned that there were 13 items where all four options were chosen, that there were 15 items where only three options were chosen, that there were eight items where only two options were chosen, and that there were nine items where only one option was chosen.

In the rest of our analysis of the new test, we opted to work with a sample instead of with the entire pool of 45 items. Our relative inexperience at weighted multiple-choice item writing and the lack of opportunity to satisfactorily pilot test the instrument (because the statistical population was so small, conducting a pilot test would likely have reduced the number of participants in the actual administration) produced a level of inconsistency in the total pool of items. By working with a sample, we were better able to control for the inconsistencies.

To select the sample, we chose five items of each type (Item Type A, Item Type B, and Item Type C). Our first priority was finding items that had difficulty indices in the ideal range stated above. Our second priority was identifying items that had discrimination indices above .4. Our third and lowest priority was selecting items where the distractors were working effectively. We arranged our priorities in this manner because we believed that

inconsistent item difficulty would have the strongest effect on our attempts to measure construct validity in subsequent tests, and because there was no discussion in the relevant literature of the importance of discrimination indices and distractor analyses as they relate to weighted scoring.

When the items were chosen, three item-type subtotals were calculated for each candidate by adding the five sample scores within each item type (i.e., for Candidate One, the five Item A sample scores were added together, the five Item B sample scores were added together, the five Item C sample scores were added together, etc.).

The raw scores for the items in the sample, along with the corresponding difficulty and discrimination indices are presented in *Table 11*.

Table 11: Raw Scores for Items in the Sample with Indices and Subtotals

Item	High Scorers					Moderate Scorers					Low Scorers					Discrim	Difficulty
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15		
8a	4	2	4	2	4	2	4	4	4	2	4	4	2	2	3	0.2	53
18a	4	4	4	3	4	3	1	4	4	4	4	4	2	4	4	0	73
27a	4	1	4	4	1	4	4	1	4	4	4	4	4	4	3	-0.2	73
34a	4	3	4	4	4	3	4	4	4	4	4	4	3	3	4	0.2	73
44a	4	4	1	3	4	4	3	4	4	4	4	4	4	4	2	-0.2	73
Type A Subtotal	20	14	17	16	17	16	16	17	20	18	20	20	15	17	16		
12b	2	4	3	4	1	4	4	4	2	4	4	1	4	4	1	-0.2	60
17b	4	4	4	4	4	4	3	4	4	1	2	4	2	1	4	0.6	67
20b	4	4	4	4	4	4	3	3	4	3	4	4	4	4	4	0	80
28b	4	4	3	4	4	4	4	3	1	4	4	4	4	2	1	0.2	67
32b	4	4	4	4	2	4	4	4	4	4	4	4	4	3	2	0.2	80
Type B Subtotal	18	20	18	20	15	20	18	18	15	16	18	17	18	14	12		
4c	1	4	4	4	4	4	4	4	4	2	4	2	4	2	3	0.4	67
15c	4	4	4	2	4	2	4	4	2	4	2	4	4	4	2	0.2	67
19c	4	4	4	4	4	4	4	4	4	4	2	4	1	4	3	0.6	80
22c	4	4	3	4	4	4	2	4	4	4	4	4	4	4	4	-0.2	87
41c	3	3	3	3	4	4	4	3	4	4	1	3	3	4	3	0	40
Type C Subtotal	16	19	18	17	20	18	18	19	18	18	13	17	16	18	15		

The information in *Table 11* reflects our priorities in selecting the items for our sample. As the values in the *Difficulty* column show, the item difficulty indices are all in the preferred range. Unfortunately, as the values in the *Discrim* column attest, four of the items in the sample have negative item discrimination indices.

Normally, a negative discrimination value should be viewed as a “red flag” to test developers. It indicates that the item is not challenging enough – low scorers are able to answer it correctly, but high scorers are wary of an easy answer and are fooled into making an alternative interpretation of a distractor (Cohen, Swerdlik & Phillips, 1996, p. 236-238). However, we were not overly concerned by the negative discrimination values in this dataset, for two reasons.

First, strictly speaking, the answer options worth 1, 2, and 3 points are not “distractors”. They are not incorrect answers written in an attempt to “distract” test takers from selecting a solitary correct response. Instead, they are all valid answers that vary only in their degree of completeness. Our multiple-choice items were not scored dichotomously (i.e., as “right” or “wrong”); we instead used weighted scoring. Furthermore, because we were unable to locate any references in the literature to item analysis in tests with weighted scoring, it was not readily apparent to us to what degree we should expect the analysis of the new test to conform to assumptions about item analysis in dichotomously scored tests.

Second, it seems to us that the negative discrimination indices may simply have been a manifestation of the fact that the test is measuring three separate constructs. Cohen,

Swerdlik and Phillips (1996) point out that it is sometimes difficult to interpret and compare two scores from the same heterogeneous test. One test taker may score high on construct A and low on construct B. Another test taker may score low on construct A and high on construct B. Yet the two test takers may obtain the same overall test score (p. 153). The differences in abilities between the two test takers are not readily apparent in an examination of their overall scores.

In *Table 11*, the test takers are ranked from highest to lowest according to their overall score on the test, which is made up of scores from all three item types, which purportedly measure different constructs. It is therefore plausible that a candidate could score well on items that measure the first two constructs, score poorly on items that measure the third construct, and still be classed as a high scorer. It is also plausible that a candidate could score poorly on items that measure the first two constructs, score well on items that measure the third construct, and be classed as a low scorer. In this case, the discrimination index for an item measuring the third construct could very well turn out to be negative. Such a result would not necessarily indicate that there is a problem with the item, only that there is a problem in our interpretation of the overall score.

To test for this possibility, we reorganized the data in *Table 11*. For each of the three item types, we ranked the candidates from highest to lowest, based on their subtotal scores (yet to avoid confusion, and to allow for comparison with the analysis above, we did not change the candidate identification numbers). We then calculated a new set of item discrimination indices, based on the new rankings of candidates within each item type. The new item discrimination indices are listed *Table 12*, in the column labelled *Subtotal Discrim.*

The previous set of item discrimination indices, calculated from rank based on overall score, are listed in the column labelled *Total Discrim* in order to provide a basis for comparison.

Table 12: Sample Items Comparing Discrimination based on Totals and Subtotals

Item	High Scorers					Moderate Scorers					Low Scorers					Total Discrim	Subtotal Discrim
	C1	C9	C11	C12	C10	C3	C8	C5	C14	C6	C4	C7	C15	C13	C2		
8a	4	4	4	4	2	4	4	4	2	2	2	4	3	2	2	0.2	0.6
18a	4	4	4	4	4	4	4	4	4	3	3	1	4	2	4	0	0.6
27a	4	4	4	4	4	4	1	1	4	4	4	4	3	4	1	-0.2	0.4
34a	4	4	4	4	4	4	4	4	3	3	4	4	4	3	3	0.2	0.4
44a	4	4	4	4	4	1	4	4	4	4	3	3	2	4	4	-0.2	0.6
Type A Subtotal	20	20	20	20	18	17	17	17	17	16	16	16	16	15	14		
Item	High Scorers					Moderate Scorers					Low Scorers					Total Discrim	Subtotal Discrim
	C6	C2	C4	C1	C3	C7	C8	C13	C11	C12	C10	C9	C5	C14	C15		
12b	4	4	4	2	3	4	4	4	4	1	4	2	1	4	1	-0.2	0.2
17b	4	4	4	4	4	3	4	2	2	4	1	4	4	1	4	0.6	0.4
20b	4	4	4	4	4	3	3	4	4	4	3	4	4	4	4	0	0.2
28b	4	4	4	4	3	4	3	4	4	4	4	1	4	2	1	0.2	0.4
32b	4	4	4	4	4	4	4	4	4	4	4	4	2	3	2	0.2	0.6
Type B Subtotal	20	20	20	18	18	18	18	18	18	17	16	15	15	14	12		
Item	High Scorers					Moderate Scorers					Low Scorers					Total Discrim	Subtotal Discrim
	C5	C2	C8	C6	C10	C7	C9	C3	C14	C4	C12	C13	C1	C15	C11		
4c	4	4	4	4	2	4	4	4	2	4	2	4	1	3	4	0.4	0.4
15c	4	4	4	2	4	4	2	4	4	2	4	4	4	2	2	0.2	0.2
19c	4	4	4	4	4	4	4	4	4	4	4	1	4	3	2	0.6	0.6
22c	4	4	4	4	4	2	4	3	4	4	4	4	4	4	4	-0.2	0
41c	4	3	3	4	4	4	4	3	4	3	3	3	3	3	1	0	0.6
Type C Subtotal	20	19	19	18	18	18	18	18	18	17	17	16	16	15	13		

It is interesting to note that discrimination appeared to improve for many of the items in the subsample. Whereas previously, only three items appeared to be discriminating effectively between high and low scorers (i.e., they had an index value greater than or equal to 0.4), with the rank adjusted for item type, 11 of the 15 items appeared to be effective. Also, all four of

the items that were previously “red flags”, because of their negative item discrimination indices, appeared to “improve” substantially. One item (22c) showed a zero value (not an optimal finding, but better than the previous result), while three (27a, 44a, and 12b) “became” positive. Two of the three (27a and 44a) showed values in the effective range. These findings appeared to confirm our suspicions that the items in the sample were discriminating between high and low scorers, but that this fact was being obscured by the method used to rank the test takers.

To complete our examination of the new test items, we conducted a distractor analysis for the items in the sample. Our findings are presented in *Table 13*.

Table 13: Distractor Analysis for Items in the Sample based on Total Score

Item	High Scoring Participants				Low Scoring Participants			
	Option A	Option B	Option C	Option D	Option A	Option B	Option C	Option D
8a	0.40	0.00	*0.60	0.00	0.40	0.20	*0.40	0.00
18a	0.00	0.00	0.20	*0.80	0.00	0.20	0.00	*0.80
27a	0.00	*0.60	0.40	0.00	0.20	*0.80	0.00	0.00
34a	0.00	0.00	0.20	*0.80	0.00	0.00	0.40	*0.60
44a	*0.60	0.00	0.20	0.20	*0.80	0.20	0.00	0.00
12b	0.20	*0.40	0.20	0.20	0.00	*0.60	0.00	0.40
17b	0.00	0.00	*1.00	0.00	0.00	0.40	*0.40	0.20
20b	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00	0.00
28b	0.00	*0.80	0.00	0.20	0.20	*0.60	0.20	0.00
32b	0.00	*0.80	0.00	0.20	0.20	*0.20	0.40	0.20
4c	0.00	0.00	0.20	*0.80	0.40	0.20	0.00	*0.40
15c	0.00	0.20	0.00	*0.80	0.00	0.40	0.00	*0.60
19c	0.00	0.00	0.00	*1.00	0.20	0.20	0.20	*0.40
22c	0.00	*0.80	0.20	0.00	0.00	*1.00	0.00	0.00
41c	0.80	0.00	0.00	*0.20	0.60	0.00	0.20	*0.20

*Indicates the answer option that was awarded the most points.

If the distractors in a multiple-choice item are working effectively, they will not succeed in fooling the high scorers, but they will appear more attractive to the low scorers. In an ideal distractor analysis, the high scorers' choices are centred mainly on the correct answer option, while the low scorers' choices are more evenly distributed across the answer options. If a distractor is not selected at all, it likely indicates that it is in need of revision.

According to these criteria, the distractors in the new test are not functioning well. In *Table 13* there are only five items (44a, 28b, 32b, 4c, and 19c) where candidates have chosen all the answer options. Also, there are eight items where the high scorers' choices are either as widely distributed as those of the low scorers, or more widely distributed than those of the low scorers (18a, 27a, 34a, 44a, 12b, 20b, 15c, and 22c). However, the distractor analysis in *Table 13* was conducted with the test takers ranked according to their overall scores on the new test. To see if this had an effect on the distractors, we conducted a second analysis based on a ranking according to subtotal scores for each item type. The results of this analysis are presented in *Table 14*.

Table 14: Distractor Analysis for Items in the Sample based on Subtotals

Item	High Scoring Participants				Low Scoring Participants			
	Option A	Option B	Option C	Option D	Option A	Option B	Option C	Option D
8a	0.20	0.00	*0.80	0.00	0.60	0.20	*0.20	0.00
18a	0.00	0.00	0.00	*1.00	0.20	0.20	0.20	*0.40
27a	0.00	*1.00	0.00	0.00	0.20	*0.60	0.20	0.00
34a	0.00	0.00	0.00	*1.00	0.00	0.00	0.40	*0.60
44a	*1.00	0.00	0.00	0.00	*0.40	0.20	0.40	0.00
12b	0.20	*0.60	0.20	0.00	0.20	*0.40	0.00	0.40
17b	0.00	0.00	*1.00	0.00	0.00	0.00	*0.60	0.40
20b	*1.00	0.00	0.00	0.00	*0.80	0.20	0.00	0.00
28b	0.00	*0.80	0.00	0.20	0.40	*0.40	0.20	0.00
32b	0.00	*1.00	0.00	0.00	0.20	*0.40	0.00	0.40
4c	0.20	0.00	0.00	*0.80	0.20	0.20	0.20	*0.40
15c	0.00	0.20	0.00	*0.80	0.00	0.40	0.00	*0.60
19c	0.00	0.00	0.00	*1.00	0.20	0.20	0.20	*0.40
22c	0.00	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00
41c	0.20	0.00	0.00	*0.60	0.80	0.00	0.20	*0.00

Once again, the subtotal-based approach has yielded some interesting results. The number of items where all the distractors were chosen did not increase (it stayed at five). However, in *Table 14* there are only four items where the high scorers' choices are either as widely distributed as those of the low scorers, or more widely distributed than those of the low scorers (12b, 15c, 22c, and 41c).

Although this is an improvement over the results in *Table 13*, it is still not an ideal finding. We believe that it is nevertheless acceptable. As we indicated above, the answer options in the new test are not typical distractors, and the literature does not provide an account of the sort of results that should be expected from item analysis in a test with weighted scoring. In the absence of expert opinion on this matter, it seems reasonable to conclude that the analysis of items in the new test would not reveal findings that are entirely consistent with those of dichotomously scored multiple-choice tests.

In light of the findings presented in this section, we feel that our construction of a sample from the original larger pool of items, has allowed us to accomplish three things. First, we have ensured that the items in our subsequent analysis were all of a similar difficulty. Second, we have demonstrated that the items discriminated between high and low scorers in an effective manner. Third, we have controlled, to the best of our ability, for items with poorly written distractors. We are therefore able to state that our expectation of minimal inconsistencies in the test items has largely been met through the construction of the sample. It therefore also seems reasonable to state that we have taken steps to reduce the likelihood that item variance will affect the results of the statistical tests that follow.

5.2.3 Estimation of Construct Variance

Information about the new test competencies suggests that they are separate constructs, and we therefore expect that their scores will be multidimensional. To examine this expectation, the next step in the analysis of the new test was the estimation of the degree of similarity among the scores for the three competencies. A high degree of score similarity would suggest that the test does not measure three separate constructs. This in turn could indicate that the theoretical model is invalid. In contrast, a low degree of score similarity would suggest that the test does measure three separate constructs, and could it confirm the model.

The degree of score similarity was estimated by calculating Pearson's product-moment coefficients among the three item-type subtotals (each one the sum of the scores for the five items retained in the sample) for each participant. Descriptive statistics for the analysis are presented first, followed by the correlation coefficients.

Table 15: Descriptive Statistics for the New Test Constructs

Construct	Number of Participants	Mean	Standard Deviation
Item Type A	15	17.27	1.944
Item Type B	15	17.13	2.326
Item Type C	15	17.35	1.759

At first glance, an examination of the means might lead us to conclude that there is high degree of similarity among scores for the three constructs. However, because the mean is a measure of central tendency, it does not provide an indication of the range of scores in each case, and it may be the case that the three sets of scores differ considerably in terms of their high and low values. Indeed, the standard deviations for the three constructs seem to indicate that there is a greater range of values in the new test data than was seen in the existing test (compare with *Table 5* and *Table 8*), suggesting that the constructs are not as similar as they might first appear.

Table 16: Correlations among the New Test Constructs in the Sample

	Item Type A	Item Type B	Item Type C
Item Type A			
Pearson Correlation	1.000	-.230	-.341
Significance (2-tailed)	.	.411	.214
N	15	15	15
Item Type B			
Pearson Correlation	-.230	1.000	.076
Significance (2-tailed)	.411	.	.789
N	15	15	15
Item Type C			
Pearson Correlation	-.341	.076	1.000
Significance (2-tailed)	.214	.789	.
N	15	15	15

The results of the correlation seem to support our interpretation of the standard deviations. None of the three coefficients are statistically significant (i.e., .411, .214, and .789 are all greater than .05). Two of them (-.230 and -.341) are negative, suggesting an inverse relationship between the constructs in question (as scores for one tend to increase, scores for the other tend to decrease). In any case, all of the coefficients are of a low magnitude, offering evidence of low or slight correlation. (Refer to *section 4.1.4* for interpretation of correlation magnitude.) This suggests that whatever relationship exists among the constructs, it is small to negligible. These findings are a first indication of multidimensionality in the new test scores, and they are therefore consistent with our expectation.

As we did in the analysis of the existing test, we investigated the possibility that the nature of the relationships between pairs of constructs was not the same, using a test for correlated coefficients (Guilford & Fruchter, 1973).

Table 17: Comparison of Correlated Coefficients for the Existing Test Constructs

Correlations Compared	Coefficients	Z	α	Accept Null Hypothesis?
BC and AB	.076 and -.230	.665	1.96	Yes
BC and AC	.076 and -.341	.979	1.96	Yes
AB and AC	-.230 and -.341	.308	1.96	Yes

Once again, the test revealed no significant differences between the correlation coefficients.

This is further evidence that leads us to believe that our interpretations of the negative correlations -.230 and -.341 should be the same as our interpretation of the correlation .076.

It therefore seem reasonable to state that all the correlations are consistently low, and that they indicate a low level of similarity among scores for the three new test competencies.

These findings appear to support our beliefs about the multidimensional nature of the test scores.

5.2.4 Secondary Estimation of Construct Variance

In order to confirm our findings, we elected to calculate a second estimation of construct variance. Ideally, we would have liked to follow Cortina's (1993) suggestion of combining a principal components analysis with the calculation of Cronbach's alpha to determine the unidimensionality of the scores. Unfortunately, we did not have enough data to justify using a factor analytic technique. In *Section 5.1.4*, we were willing to cautiously flout recommendations about the number of cases required (at 66, we would have had to increase the number of cases by less than 50% to reach the recommended minimum of 90, or 10 cases per nine items). However, in this section, it seemed completely unjustifiable to repeat this

course of action (at 15, we would have had to increase the number of cases by 1000% to reach the recommended minimum of 150, or 10 cases per 15 items).

We therefore contented ourselves with using the second of the techniques recommended by Cortina, Cronbach's alpha. We calculated the internal consistency of the five item scores for each item type, confident that this was a more orthodox application of the statistical test, because each item type attempts to measure a single construct. The results of our calculations are presented below.

Table 18: Coefficient Alpha for New Test Constructs

Construct	Number of Participants	Number of Items	Coefficient Alpha
Item Type A	15	5	.9958
Item Type B	15	5	.9961
Item Type C	15	5	.9845

In *Table 18*, the three coefficients all indicate that there is a very high degree of correlation, which in turn suggests a great deal of consistency within each item type. (Refer to *Section 4.1.4* for interpretation of correlation magnitude.)

Cortina (1993) cautions against reaching a conclusion about dimensionality based on the results of Cronbach's alpha alone⁴. However, the results do appear to be consistent with all the other evidence reviewed thus far. The low or slight degree of score similarity between item types (seen in *Section 5.2.3*), coupled with the apparently very high degree of score

⁴ Others are less cautious in this regard. See, for example, Academic Technology Services (2003).

similarity within item types (seen here), leads us to believe that there are in fact three separate constructs in the data.

In any case, it is certainly reasonable to conclude that the results in *Table 18* provide solid evidence of internal reliability. In itself this is an important finding, since reliability is a precondition for validity. It cannot be argued that the interpretation of a measure is appropriate unless it is first demonstrated that the measure is consistent.

5.3 Conclusion

This chapter has outlined the results obtained from our validation of both the existing and new test instruments. The validation used qualitative evidence (in the form of information about the origins of the test constructs) and quantitative evidence (in the form of statistical analyses of test scores) in an attempt to determine whether each set of constructs is used appropriately for the purposes of interpreter certification. In the next chapter, the results obtained from the existing and new tests will be compared, and they will be interpreted in a broader context.

6. Interpretation

In this chapter, we will attempt to gain some sense of whether our effort to improve the construct validity of interpreter certification tests by working with discursive constructs has been successful. To make that determination, we will take two steps. First, we will make a quantitative comparison between the correlations of constructs from the existing and new tests. Second, we will revisit the hypotheses introduced in *Section 1.2* and determine if they have been confirmed or refuted.

6.1 Comparison of Test Results

Our principal aim in the present study has been to learn whether the current state of testing in interpreting, represented by the existing test and the competencies it borrowed from machine translation, can be ameliorated by a new method, represented by the new test and its discursive competencies. To make this determination, we examined the theory related to the competencies the two tests purport to measure, and we analyzed actual measurements gathered from administrations of the two tests. Both the existing test and the new test measure three competencies (intelligibility, informativeness, and style in the case of the former, and linguistic parsing, assignment of reference, and interpretation of inference in the case of the latter) that are described in the relevant theoretical literature as separate and largely unrelated constructs. If this is in fact the case, scores from the two tests should be multidimensional. In other words, we should be able to detect the presence of factors in the scores of each test that correspond to the constructs each test purports to measure. Scores from both tests were therefore examined to determine if there was evidence to support the

theoretical claims. This evidence, if revealed, would suggest that interpretations of the tests' scores are valid.

One of the principal forms of evidence we considered was correlation. The coefficients we calculated in *Chapter 5* provided an indication of the similarity among the competencies within the existing test, on the one hand, and within the new test, on the other. Strictly speaking, since none of the coefficients was equal to zero, we discovered some level of construct similarity within each of the two tests. The scores for intelligibility, informativeness, and style revealed some level of relatedness with one another, as did the scores for linguistic parsing, assignment of reference, and interpretation of inference. Yet up until this point, the two tests have been considered separately, and since we are ultimately interested in comparing the two tests, we must ask ourselves a logical follow-up question: how similar are the similarities?

We already have part of this answer. When we presented the correlations, we gave some interpretation of their magnitude, by stating that construct correlations on the existing test were "high", while construct correlations on the new test were "slight" and "low". This application of "Guilford's rule of thumb" (see *Section 4.1.4*) appears to suggest that the nature of construct relationships on the two tests can be categorized as being different. In other words, we have some evidence to suggest that the existing test scores are unidimensional (a finding that does not support claims about the test constructs) and that the new test scores are multidimensional (a finding that does support claims about the test constructs).

However, remembering Guilford's later admonition (1965) concerning the use of a hard and fast rule about the numerical value of coefficients, we reserved judgement. The categorization of magnitude is somewhat arbitrary, since it was based on a relatively decontextualized reading of the value of the coefficients. We therefore attempted to determine whether a second method of comparing the two sets of coefficients would also suggest that they were categorically different. This would furnish us with a more complete answer to our question about similarity between the two tests.

To do this, we used a variation of the Fisher's z test for unrelated coefficients (Guilford & Fruchter, 1973). We selected this variation because the coefficients in the two tests come from different datasets, and they are therefore unrelated. To carry out this step in our analysis, we compared each one of the correlations between the existing test constructs (intelligibility, informativeness, and style) that we presented in *Table 7* with each one of the correlations between the new test constructs (linguistic parsing, measured by Item Type A; assignment of reference, measured by Item Type B; and interpretation of inference, measured by Item Type C) that we presented in *Table 16*. The comparisons are listed below in *Table 19*.

Table 19: Comparison of Unrelated Coefficients for Constructs on Both Tests

Correlations Compared			Coefficients	Z	α	Accept Null Hypothesis?
Existing Test		New Test				
Int/Inf	and	Item Type A/Item Type B	.746 and -.230	4.020	1.96	No
Int/Inf	and	Item Type A/Item Type C	.746 and -.341	4.424	1.96	No
Int/Inf	and	Item Type B/Item Type C	.746 and .076	2.978	1.96	No
Int/Sty	and	Item Type A/Item Type B	.690 and -.230	3.637	1.96	No
Int/Sty	and	Item Type A/Item Type C	.690 and -.341	4.040	1.96	No
Int/Sty	and	Item Type B/Item Type C	.690 and .076	2.594	1.96	No
Inf/Sty	and	Item Type A/Item Type B	.691 and -.230	3.637	1.96	No
Inf/Sty	and	Item Type A/Item Type C	.691 and -.341	4.040	1.96	No
Inf/Sty	and	Item Type B/Item Type C	.691 and .076	2.594	1.96	No

Table 19 shows that in each of the nine cases above the value for z is larger than the corresponding α . We have therefore rejected the null hypothesis (which states that the result was obtained by chance) in all nine cases, which in turn leads us to conclude that the differences between the unrelated coefficients are significant in every comparison. This result occurs consistently throughout the table, and it stands out in contrast with the comparisons carried out on constructs within the existing test, and within the new test, where the null hypothesis was consistently accepted (see *Table 8* and *Table 17*).

In light of the comparison, it seems reasonable to state that the relationship between the constructs in the existing test is not similar to the relationship between the constructs in the new test. In our initial description of the correlation coefficients, we used Guilford's "rule of thumb" to classify coefficients from the existing test as "high" and coefficients from the new test as "slight" or "low". Now we have additional evidence that the two sets of

coefficients are categorically different. Our results suggest that there is a degree of relatedness among construct scores in the first dataset that is unlike that found in the other.

This new information supports the findings obtained in *Chapter 5*, which suggested that the scores or the constructs in the existing test are relatively dependent, while the scores for the constructs in the new test are relatively independent. The comparison conducted in this section has provided further evidence to indicate that the scores from the existing test are unidimensional, which in turn makes it appear that the three competencies measured by the test are not separate constructs. We also have further evidence to indicate that the scores from the new test are multidimensional, which in turn makes it appear that the three competencies measured by the test are separate constructs.

6.2 Hypothesis Testing

At this juncture, we are still far from being able to reach a general conclusion about our effort to improve the construct validity of an interpreter certification test. To reach that conclusion we now revisit the hypotheses we formulated in *Section 1.2* and discuss each one in light of our results.

Specific Hypotheses regarding the Existing Test

An examination of the origins of the constructs of the existing test will reveal a weak link to the interpreting profession.

This hypothesis was confirmed. We were unable to find an obvious link between the existing test constructs (intelligibility, informativeness, and style) and the interpreting profession. Association 1 adopted the use of the three constructs when it began administering the test developed for Association 2 by Anderson, which was based on her unpublished master's thesis (Anderson, 1979). Anderson borrowed two of the constructs – intelligibility and informativeness – and their corresponding rating scales from Carroll (1966), who notes that they were developed in consultation with members of the US National Academy of Sciences for an evaluation of machine translation. Any suggestion that the constructs can be traced further back in time appears highly speculative, both because Carroll gives no indication that there is an ultimate source, and because the most likely source (Nida, 1964) conceptualizes intelligibility and informativeness in a manner that is incompatible with Carroll's work.

An examination of construct-irrelevant variance will suggest that factors other than the constructs may explain patterns in the test scores.

This hypothesis was refuted. We used three statistical measures to examine construct-irrelevant variance (multitrait-multimethod matrix, Pearson's product-moment correlation,

and Fisher's z-test), and all three suggested that there was a significant degree of similarity among the three test items (A, B, and C).

The multitrait-multimethod matrix indicated that there was widespread similarity both in areas where we expected to find it (the same competencies measured by different items) and in areas where we did not (different competencies measured by the same items, different competencies measured by the same items)¹. The Pearson's correlation indicated that there was a high degree of correlation among scores for the three items, and the Fisher's z-test indicated that there were no significant differences among the three correlations. Together, this evidence leads us to believe that factors such as differences in item difficulty had a minimal effect on the candidates' scores.

An examination of construct variance will suggest that the constructs do not adequately explain patterns in the test scores.

This hypothesis was confirmed. Correlations calculated between scores for the three constructs (intelligibility, informativeness, and style) suggested that there was a high degree of similarity among all of them, and that the most similarity appeared to occur between intelligibility and informativeness. This result did not correspond to the pattern predicted by the theoretical information on the constructs.

¹ In other studies that have used the multitrait-multimethod matrix (see Campbell & Fiske, 1959), a very different pattern of relationships has appeared. In a soundly constructed test, there is both discriminant validity

Carroll's (1966) discussion of the intelligibility and informativeness scales indicates that he considered the competencies to be "conceptually separable variables". Anderson's (2001) account of the creation of the style scale suggests that it too should – in principle – be viewed as a separate construct. This is because the members of Association 2 felt there needed to be a third scale to measure a competency not accounted for in the first two. (In practice, however, the means by which Anderson constructed the style scale would lead us to expect some degree of relationship between style scores and intelligibility scores).

By all accounts, the existing test should therefore have been "heterogeneous" (Cohen, Swerdlik & Phillips, 1996) since its items were intended to measure more than one ability. We should therefore have seen results analogous to those reported by Violato, Salami and Muiznieks (2002) and by Daniel and Siders (1994) (see the reviews in *Section 2.2.1* and *Section 2.2.2*). Yet instead of indicating multidimensionality, as the tests used in those studies did, our first estimation of construct variance suggested unidimensionality, a finding which undermines claims made about the test.

Secondary estimation of construct variance will also suggest that the constructs do not adequately explain patterns in the test scores.

(i.e., constructs that theoretically should not be related to one another are, in fact, not related in reality) and convergent validity (i.e., constructs that theoretically should be related to one another are related in reality).

This hypothesis appears to have been confirmed. The principal components analysis suggested that there is only one component in the data, and Cronbach's alpha indicated that there was a high degree of interrelation among all the variable scores. Together, these findings are evidence that the data are unidimensional. In other words, it would seem that raters are in actuality assigning a single score to the candidates they evaluate, and not three separate scores based on their evaluation of three different aspects of the candidates' performances. Our second estimation of construct variance supports our first, and it indicates that the test scores do not confirm the contention that the test measures three separate constructs.

However, we make this claim somewhat cautiously. It is not recommended to use factor analytic techniques with such small numbers of participants (Pett, Lackey & Sullivan, 2003)² nor is alpha typically calculated for scores that do not purport to be of the same measure. We must therefore admit that the accuracy of our results may have been affected.

Specific Hypotheses regarding the New Test

An examination of the origins of the constructs of the new test will reveal a strong link to the interpreting profession.

² In our defence, we point out again that it appears that Pett, Lackey and Sullivan's recommendation is not always followed in published research. Violato, Salami and Muiznieks (2002) performed an analysis of nine variables with data from 112 test takers. Price and Wilkins (2001) conducted an analysis on 85 items with data from 400 test takers. Daniel and Siders (1994) also failed to meet the recommended numbers; they used data from 194 test takers to examine 43 items.

This hypothesis was confirmed. The new test constructs (linguistic parsing, assignment of reference, and interpretation of inference) appear to be linked to the interpreting profession in two ways. First, the constructs were developed by people with personal experience as interpreters. The constructs were based on a model presented by Setton (1999), who thoroughly reviewed the work of other researchers, including Moser (1978), Gile (1997), and Seleskovitch and Lederer (various publications). All of the authors have first-hand experience as conference interpreters.

Second, the constructs were validated in research situations where interpreters served as participants. Setton, Seleskovitch, and Lederer each assembled a corpus of interpreted speeches, and each used his or her corpus to test a model of the interpreting process. Gile used his qualitative, long-term observations of working interpreters to construct his account of interpreting, while Moser's model was tested in a quantitative study of interpreting aptitude. All the authors invited interpreters to participate in their research, and in so doing, they gave themselves a second opportunity to ensure that their work accurately reflected the realities of the interpreting profession.

An examination of construct-irrelevant variance will suggest that other factors do not explain patterns in the test scores.

This hypothesis was refuted. Our relative inexperience with writing multiple-choice items, particularly for weighted scoring, and the lack of satisfactory pilot testing produced

test items that were somewhat irregular. *Item difficulty* indices for the entire pool of items were not as stable as might have been desired, *item discrimination* indices were somewhat low, and *item distractors* (i.e., the multiple-choice items that were not worth full marks) in a number of cases were not working effectively.

However, we believe that the irregularity of the test items may not be as serious an issue as it first appears, for a number of reasons. First, we were able to control for problems with item difficulty by selecting a sample of 15 items, five of each item type, from the original pool. All the items selected have item difficulty indices in the desirable range.

Second, we have reason to believe that the low discrimination indices originally discovered were the result, not of problems with the test items, but of our initial interpretation of the overall score. Cohen, Swerdlik and Phillips (1996) suggest that test users need to be cautious when interpreting and comparing the overall scores obtained from a multidimensional test. Overall scores obscure candidate performances on individual constructs – two candidates could obtain the same overall score, but attain different levels of achievement on different constructs. It appears that this was the case in the new test. When test taker rankings were reorganized according to performance on each item type, and when item discrimination indices were recalculated, the results were substantially different from those of our first calculation. We were able to eliminate all the negative indices, and 11 of the 15 items had indices in the effective range.

Third, we have evidence to suggest that our initial distractor analysis was also affected by the method used to rank test takers. When the candidates were ordered according

to item type, the results of a subsequent distractor analysis were more satisfactory. In the end, not all distractors appeared to be working effectively, but we do not believe that this is a cause for concern. The new test was not dichotomously scored, the way traditional multiple-choice tests are. We opted instead to use weighted scoring, and we developed a series of answer options that were all correct, but varied in the number of points that they were worth. Consequently, none of the answer options were – strictly speaking – “distractors”. They were not designed to “distract” test takers from the correct answer, but rather to provide different amounts of accurate information. The measurement and evaluation literature does not discuss the use of distractor analysis with weighted scoring. In the absence of careful study in the literature, it seem unreasonable to us to assume that the answer options in the new test should exhibit the same properties as distractors in a standard multiple-choice test.

Still, we recognize that the results of our item analysis have an effect on the degree of certainty we may have about our findings overall. We therefore draw our subsequent conclusions with a great deal of caution.

An examination of construct variance will suggest that the constructs adequately explain patterns in the test scores.

This hypothesis was confirmed. Using Pearson’s product-moment correlation, we determined that the degree of similarity among scores for the three item types was low or slight. This finding suggested that the item type scores were relatively independent of one

another, and it was consistent with theoretical information about the three new test competencies.

Paradis (1994, 2000) indicates that there is a neuroanatomical basis for distinguishing between linguistic competence (which subsumes the ability we describe as linguistic parsing) and “pragmatic” competence (which Paradis indicates includes referencing outside knowledge and interpreting inferences). Setton (1999) further distinguishes the assignment of reference and the interpretation of inferences by locating them in two separate components of his model (the assembler and the executive, respectively). Information about the three competencies therefore leads us to believe that they are separate and unrelated constructs.

We further expected that the test scores would be multidimensional, i.e., that they would show evidence of three factors that corresponded largely to the three competencies. This expectation was met by our first estimation of construct variance, which indicated a very low degree of relationship between scores of the three item types.

Secondary estimation of construct variance will also suggest that the constructs adequately explain patterns in the test scores.

This hypothesis appears to have been confirmed. We attempted to repeat the fourth step in the analysis of the existing test and use principal components analysis and Cronbach's alpha to examine dimensionality in the new test data. However, the limited number of test

takers ruled out the possibility of a principal components analysis, and we had to content ourselves with only Cronbach's alpha.

It was revealed that there was a very high degree of consistency within each item type in the sample. We feel that this is a particularly robust finding, because the number of items within each item type was low (five items), and because Cronbach's alpha is sensitive to the number of items (the higher the number of items, the higher the value for alpha is likely to be) (ATS, 2003). Cortina (1993) states that internal consistency is a necessary but not sufficient condition for unidimensionality. We are therefore not able to state conclusively that each item type is measuring a single construct.

However, we do feel that we have collected as much evidence as possible, given the small sample size of the new test dataset. We have demonstrated that there is a low degree of similarity between scores for the three different item types, and we have demonstrated that there is a high degree of similarity among the scores within each item type. While this information is not as definitive as we may have liked, it does appear to be consistent with our expectations about the multidimensionality of the new test scores. It therefore also appears to be consistent with our beliefs about the new test competencies.

General Hypothesis

There will be greater evidence of validity of score interpretations in an interpreter certification test when the test constructs are based on a discursive model of the interpretation process than when they are based on research in machine translation.

Before we draw our conclusions about the present study's general research hypothesis, we believe it is useful to briefly consider some of the information presented in the *Standards* (AERA, APA & NCME, 1999) about validation. The Standards go to considerable length to clarify the sort of evidence that should normally be offered to support the validity of test score interpretations, but three points are of particular relevance here.

First, the *Standards* indicate that "When validation rests in part on the appropriateness of test content³, the procedures followed in generating and specifying test content should be described and justified in reference to the construct the test is intended to measure..." (p. 18). In other words, when questions have been raised about parts of the test instrument – including its scoring scheme – test users must demonstrate how they were developed and obtained, and they must show that they are justified and appropriate. If users are not able to do this, the validity of the score interpretations remains in question.

³ The Standards define test content as "the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring" (p. 11). We therefore believe that it is justified to describe the existing test's rating scales as "test content".

Second, the *Standards* discuss the situation where a test purports to measure more than one construct, noting that “When a test provides more than one score, the distinctiveness of the separate scores should be demonstrated, and the interrelationships of those scores should be shown to be consistent with the constructs being assessed” (p. 20). If test users fail to demonstrate distinctiveness, or if they are unable to use information about the constructs to successfully describe interrelationships, users must conclude that score interpretations are invalid.

Third, the *Standards* also discuss the situation where a test is designed to measure a single construct: “It might be claimed that a test is essentially unidimensional. Such a claim could be supported by a multivariate statistical analysis, such as a factor analysis, showing that the score variability attributable to one major dimension was much greater than the score variability attributable to any other identified dimension” (p. 20). Presumably, if test users are claiming the contrary – that a test is multidimensional – and actual analysis reveals the presence of only one major factor in the data, this is evidence that test score interpretations are not valid.

With these points in mind, we return to our discussion of the general hypothesis. Recall that when we analyzed the existing test, our examination of the qualitative and quantitative evidence led us to state that 1) theoretical information about the test constructs (intelligibility, informativeness, and style) did not appear to justify their use in a test for conference interpreters; 2) separate construct scores did not appear to be distinct; 3) the interrelationships of the scores were not consistent with statements about the related

constructs; and 4) score variability appeared to be largely attributable to a single component in the data.

When we analyzed the new test, our examination of the qualitative and quantitative evidence revealed that 1) theoretical information on the new test constructs (linguistic parsing, assignment of reference, and interpretation of inference) did appear to justify their use in a test for conference interpreters; 2) separate construct scores did appear to be distinct; 3) the interrelationships of the scores were consistent with statements about the related constructs; and 4) score variability appeared to be attributable to three components in the data.

What this suggests is that it is not justifiable to use the existing test constructs in interpreting, and that scores from the existing test do not give accurate information about the constructs. It also indicates that it does appear to be justifiable to use the new test constructs in interpreting, and that scores from the new test do give accurate information about the constructs. To put this another way, our examination of the new test constructs, which are discursive in nature, has yielded greater evidence of validity than a similar examination of the existing test constructs, which are based on research into machine translation. Since this is the result predicted by the present study's general research hypothesis, we therefore conclude that the hypothesis has been confirmed.

To explain the conclusion in more concrete terms, let us consider the outcome of the two tests' use. Scores derived from the existing test do not appear to measure what they purport to measure, and it seems doubtful that interpretations based on them will be

appropriate. In other words, if test users look at a given test taker's score on the test, and if they use it to make a judgement about the test taker's ability to meet the minimum standards of the interpreting profession, they will be basing their judgement on criteria that at best have a tenuous link with the profession, and on scores that reveal little about the criteria. In contrast, scores derived from the new test do appear to measure what they purport to measure, and it seems likely that interpretations based on them will be appropriate. If someone were to use the scores from the new test to make a certification decision, there is evidence that their decision would be based on criteria strongly linked to the interpreting profession, and that the scores provide accurate information about the criteria.

6.3 Conclusion

The present study has uncovered evidence to suggest that discursive constructs, along with conscious attention to psychometric principles, can improve the state of affairs in interpreter certification testing. It was demonstrated that a better understanding of measurement and evaluation can be successfully combined with knowledge from Interpreting Studies to produce test instruments that are sound and defensible. In the final chapter of this study, we discuss these findings in a broader context, and we suggest how they might serve as a springboard to future research.

7. Conclusion

In the present study, we have seen how legal precedent, academic opinion, and professional standards all underscored the importance of estimating the psychometric properties of interpreter certification tests. We have argued that conducting this kind of research required us to part company with the proponents of TQA, and to turn instead to studies of other credentialing tests in order to find suitable examples to follow. This new approach, we have suggested, is not inconsistent with work that has already been carried out in Interpreting Studies.

Existing work in Interpreting Studies may also have a role to play in the new psychometric research, by helping us identify the specific competencies on which certification testing should focus. Although these competencies could theoretically be derived from any of the existing theoretical models of the interpreting process, we have argued that it would be unwise, at our current stage of understanding, to ignore the analysis of discursive factors as part of interpreting, given that the discursive models of the interpreting process are some of the most comprehensive currently in existence.

The present study also contained a qualitative and quantitative analysis of a particular certification test, which showed that the test failed to demonstrate acceptable levels of validity. In so doing, the study has provided some support for the contention that current testing practices in the interpreting profession are not sound (Sawyer, 2000), and for the belief that there is a practical need for continued research. Our qualitative and quantitative

analysis of a new discourse-based test showed that it demonstrated acceptable levels of validity. The study therefore appeared to confirm the explanatory power of Setton's cognitive-pragmatic model of the interpreting process in particular, and the benefits that can be derived from a discursive approach to interpreter testing in general.

The present study was only intended as a preliminary investigation into the merits of combining a discursive understanding of interpreting with psychometric principles. We believe that the study has been successful, and that it has established the merits of our approach. Furthermore, we feel that there is a need for more work in this area. But what would future research in this area look like? What form would it take? What topics would it investigate? As part of our overall conclusion, we take a moment to consider these questions.

7.1 Future Research

In the present study, we were working with limited time and limited resources, which had a direct effect on the design of the study. We were obliged to formulate a concise general hypothesis that was restrictive in its scope, to adopt a theoretical model that was very discrete and tightly defined, and to develop an instrument that was simple and efficient. Our goal was not necessarily to develop an instrument that could be put to direct use in the interpreting profession.

While the present study was able to offer evidence of the worth of a discursive approach to interpreter testing, there were also a number of important questions that we were forced to ignore. A number of these are discussed briefly below.

The Production Side of Setton's Model

Although we feel we were justified, for theoretical as well as practical reasons, in restricting our focus to the comprehension side of Setton's model, a thorough approach to model testing would require researchers to examine production as well. It stands to reason that the findings of this type of research should reflect those that were found for the comprehension side in the present study. However, this hypothesis should be determined empirically and not simply be left to speculation.

More Job-Related Instruments

In developing the new test, we chose to work with a multiple-choice format. Doing so allowed us to avoid the difficulties associated with human raters, and it enabled us to use simple and well-established techniques (e.g., item analysis) to evaluate the instrument. Yet we do not advocate using a test comprising *selected-response items** on its own to evaluate interpreting competencies. Although such a test can be designed and demonstrated to have valid score interpretations, it may still suffer from other difficulties. In the *Seltzer v. Foley* (1980) case, evidence based on test content clearly supported the use of the written component of the US Federal Court Interpreter Certification Exam, yet candidates reacted to its lack of face validity (Arjona, 1985). This suggests that a test format that is more overtly related to on-the-job activity, such as a performance-based test, is likely the wiser choice for

interpreter certification. Future research should investigate the development of performance tests, and the methods that can be used to evaluate their psychometric properties.

Other Sources of Validity Evidence

In the present study, we restricted our scope by focusing exclusively on the issue of construct validity, and we concentrated on examining evidence based on internal structure. Yet the *Standards* (AERA, APA & NCME, 1999) identify a number of other types of evidence that may be used to construct a validity argument. These include, but are not limited to, the following:

1. Evidence Based on Test Content

This type of evidence should demonstrate that the content of a test is related to the profession in question and gives an accurate representation of it¹.

2. Evidence Based on Relations to Other Variables

This type of evidence should demonstrate that performance on the test is an accurate predictor of another criterion measure. Examples of an appropriate criterion may include an employer's review of the candidate's performance in the profession, or a second validated test that measures the same constructs.

¹ The present study did examine some parts of test content, most notably the existing test's rating scales. It could therefore be said that we did take evidence based on test content into account. However, our consideration of test content was a far cry from the more global studies of validity based on test content depicted by the *Standards*. Such a study would likely examine the entire development of a test instrument – the initial job analysis, the construction of a test blueprint, the development and review of test items, the standard setting procedure, etc. – to show how test content supports the test users' validity argument. (In our examination, test content seemed to indicate a lack of validity.) For more information on test development and validity, see the *Standards* (AERA, APA & NCME, 1999) and Newble et al. (1993)

3. Evidence Based on the Consequences of Testing

This type of evidence should demonstrate that the test does not discriminate inappropriately against groups of test takers. For example, it is appropriate for test scores to discriminate among candidates according to a variable such as competency, but not according to a variable such as race or gender.

Future research may be designed to evaluate the validity of score interpretations by examining sources of evidence such as these.

Research of this type clearly presents a number of advantages, both social and scientific. It can potentially benefit the interpreting profession by reducing the likelihood of *false positives** and the harm they can do to the public and the reputation of the profession, and of *false negatives** and the harm they can do to candidates who are denied the opportunity to earn a livelihood in their desired profession. It can potentially benefit Interpreting Studies by enriching the subdiscipline with theory and methods from other fields of study, and by focusing interpreting scholars' attention on an interesting topic that deserves further study. It may even eventually contribute to the wider area of measurement and evaluation by demonstrating how to assess the competencies that are necessary, not only for interpreting, but for human communication in general.

Such research would continue what we feel is a fascinating and important avenue of inquiry. If the present study is any evidence, delving simultaneously into discourse and measurement furnishes us with a unique opportunity to consider how it is that we use

language and languages to establish communication, and how it is that we determine if our communication has been successful.

References

- Academic Technology Services. (2003). *What does Cronbach's alpha mean?* Retrieved July 12, 2003, from the UCLA Web site: <http://www.ats.ucla.edu/stat/spss/faq/alpha.html>.
- American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C: APA Inc.
- Anderson, L. (1979). *Simultaneous interpretation: Contextual and translation aspects*. Unpublished master's thesis, Concordia University, Montreal.
- Anderson, L. (March 29, 2001). Personal communication.
- Angoff, W. (1971). Scales, norms, and equivalent scores. In R. Thorndike (Ed.), *Educational measurement*, (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.
- Arjona, E. (1985). The Court Interpreters Certification Test Design. In L. Elias-Olivares, E.A. Leone, R. Cisneros & J.R. Gutierrez (Eds.), *Spanish language use and public life in the United States*, (pp. 181-200). Mouton: Berlin.
- Association internationale des interprètes de conférence. (2003). *AIIC Online*. Retrieved March 19, 2003, from the Web site: <http://www.aiic.net/>.
- Association of Translators and Interpreters of Ontario. (2002). *ATIO On-line*. Retrieved March 19, 2003, from the Web site: <http://www.atio.on.ca>.
- Association of Visual Language Interpreters of Canada. (2003). *Welcome to AVLIC*. Retrieved March 19, 2003, from the Web site: <http://www.avlic.ca/>.
- Austin, J.L. (1962). *How to do things with words*. London: Oxford UP.
- Bakhtin, M. M. (1986). The problem of speech genres. In C. Emerson & M. Holquist (Eds.), *Speech genres and other late essays* (pp. 60-102). Austin: University of Texas Press.
- Benveniste, E. (1966). *Problèmes de linguistique générale*, Tome I. Paris: Gallimard.
- Berger, M.J. & Simon, M. (1995). *Programmation et évaluation en milieu scolaire : notes de cours, PED 3713 95/96*. Unpublished manuscript, Université d'Ottawa.
- Berman, A. (1999). *La traduction et la lettre ou l'auberge du lointain*. Paris: Seuil.
- Blakemore, D. (1987). *Semantic constraints on relevance*. Oxford: Basil Blackwell.

- Borg, W.R. & Gall, M.D. (1983). *Educational research: An introduction* (4th ed.). New York: Longman.
- Broadbent, D.E. (1957). Listening and speaking simultaneously. *Journal of Experimental Psychology*, 43, 267-273.
- Brunning, R.H., Schraw, G.J. & Ronning, R.R. (1999). *Cognitive psychology and instruction* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.
- Cadieux, F. (February 16, 2001). Personal communication.
- Campbell, S.J. (1991). Towards a model of translation competence. *Meta*, 36(2/3), 329-343.
- Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.
- Card, C. (February 17, 2003). Personal communication.
- Carey, L.M. (2001). *Measuring and evaluating school learning*, 3rd Ed. Boston: Allyn & Bacon.
- Carr, S. (September 4, 2003). Personal communication.
- Carroll, J.B. (1966). An experiment in evaluating the quality of translations. *Mechanical Translation*, 9, 55-66.
- Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Clarici, A., Fabbro, F., Bava, A. & Daró, V. (1994). Effects of speaking speed on cerebral lateralization of speech assessed by a dual-task interference paradigm. *Perceptual & Motor Skills*, 78(3), 947-953.
- Clifford, A. (1999). *La Préférence hémisphérique dans deux modes d'interprétation simultanée*. Unpublished master's thesis, Université Laval, Québec City.
- Cobb, R.B., Shaw, R., Millard, M. & Bomotti, S. (1999). An examination of Colorado's teacher licensure testing. *The Journal of Educational Research*, 92(3), 161-175.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum.
- Cohen, R.J., Swerdlik, M.E. & Phillips, S.M. (1996). *Psychological testing and assessment*, (3rd ed.). Mountain View, CA: Mayfield.

- Cokely, D. (2001). *The cognitive process model of interpretation*. A videoconference given at the meeting of the Toronto Association of Visual Language Interpreters, February 21, 2001, Toronto, Canada.
- Cokely, D. (1992). *Interpretation : A sociolinguistic model*. Burtonsville, MD: Linstok Press.
- Corina, D., & Vaid, J. (1994). Lateralization for shadowing words versus signs: A study of ASL-English interpreters. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation*, (pp. 237-248). Amsterdam: Benjamins.
- Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98-194.
- Couper-Kuhlen, E. (2001). Intonation and discourse: Current views from within. In D. Schiffrin, D. Tannen & H. Hamilton (Eds.) *The handbook of discourse*, (pp. 13-34). Oxford, UK: Blackwell.
- Couper-Kuhlen, E. (1997). Coherent voicing: On prosody in conversations reported speech. In W. Bublitz, U. Lenk & E. Ventola (Eds.), *Coherence in Spoken and Written Discourse*, (pp. 11-33). Amsterdam and Philadelphia: John Benjamins.
- Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley & Sons, Inc.
- Cross, L.H., Ross, F.K & Geller, E.S. (1980). Using choice-weighted scoring of multiple-choice tests for determination of grades in college courses. *Journal of Experimental Education*, 48(4), 296-301.
- Cruse, A. (2000). *Meaning in language: An introduction to semantics and pragmatics*. Oxford: Oxford UP.
- Daniel, L.G. & Siders, J.A. (1994). Validation of teacher assessment instruments: A confirmatory factor analytic approach. *Journal of Personnel Evaluation in Education*, 8(1), 29-40.
- Daró, V. (1996). Memory in conference interpreting. Unpublished manuscript.
- Daró, V. & Fabbro, F. (1994). Verbal memory during simultaneous interpretation: Effects of phonological loop. *Applied Linguistics*, 15, 365-381.
- Daró, V., Lambert, S. & Fabbro, F. (1996). Conscious monitoring of attention during simultaneous interpretation. *Interpreting*, 1(1), 101-124.

- de Beaugrande, R. (1996). The story of discourse analysis. In T. van Dijk (Ed.), *Introduction to discourse analysis* (pp. 35-62). London: Sage.
- Delisle, J. (1980). *L'Analyse du discours comme méthode de traduction*. Ottawa: Éditions de l'Université d'Ottawa.
- Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Elkind, D. (1991). Developmentally appropriate practice: A case study of educational inertia. In S.L. Kagan & K.J. Rehg (Eds.), *Ninetieth yearbook of the National Society for the Study of Education: Part 1. The care and education of America's young children: Obstacles and opportunities* (pp.1-16). Chicago, IL: The University of Chicago Press
- Essex, D.L. (1976). A comparison of two item-scoring procedures and student reactions to them. *Journal of Medical Education*, 51, 565-572.
- Everitt, B.S. (1996). *Making sense of statistics in psychology*. Oxford: Oxford UP.
- Everitt, B.S. & Wykes, T. (1999). *A dictionary of statistics for psychologists*. London: Arnold.
- Fabbro, F., Gran, B. & Gran, L. (1991). Hemispheric specialization for semantic and syntactic components of language in simultaneous interpreters. *Brain & Language*, 41, 1-42.
- Fabbro, F., Gran, L., Basso, G. & Bava, A. (1990). Cerebral lateralization in simultaneous interpretation. *Brain & Language*, 39, 69-89.
- Fillmore, C. (1985). Frames and the semantics of understanding. *Quaderni de semantica*, 6(2), 222-254.
- Fillmore, C. (1997). *Lectures on deixis*. Stanford, CA : CSLI Publications.
- Fillmore, C. (2001) *The frame semantic basis*. Retrieved September 19, 2003, from University of California at Berkeley, International Computer Science Institute Web site: <http://www.icsi.berkeley.edu/~framenet/>
- Fine, S.A. (1986). Job analysis. In R.A. Berk (Ed.) *Performance assessment: Methods & applications* (pp. 53-81). Baltimore, MA : Johns Hopkins UP.
- Gage, N.L., & Berliner, D.C. (1991). *Educational psychology* (5th ed.). Boston: Houghton, Mifflin.

- Gallagher, J.D. (1998). *Classroom assessment for teachers*. Upper Saddle River, NJ : Prentice-Hall.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.
- Gentile, A., Ozolins, U. & Vasilakakos, M. (1996). *Liaison interpreting: A handbook*. Melbourne: Melbourne UP.
- George, D. & Mallery, P. (2001). *SPSS for Windows step by step: A simple guide and reference* (3rd ed.). Boston: Allyn and Beacon.
- Gerver, D. (1976). Empirical studies of simultaneous interpretation: A review and a model. In R.W. Brislin (Ed.), *Translation: Applications and research* (pp. 165-207). New York: Gardner.
- Gerver, D. & Sinaiko, W.H. (1978). *Language interpretation and communication. Proceedings of the NATO Symposium on Language Interpretation and Communication, Venice, 1977*. New York: Plenum Press.
- Gile, D. (1994). Opening up in Interpretation Studies. In M. Snell-Hornby, F. Pöchhacker & K. Kaindl (Eds.), *Translation Studies: An interdisciplinary* (pp. 149-158). Amsterdam: John Benjamins.
- Gile, D. (1997). Conference Interpreting as a Cognitive Management Problem. In J.E. Danks, G.M. Shreve, S.B. Fountain & M.K. McBeath (Eds.), *Cognitive Processes in Translation and Interpreting* (pp. 196-214). London: Sage.
- Gile, D. (2001). L'évaluation de la qualité de l'interprétation en cours de formation. *Meta*, 46(2), 379-393.
- Goffman, E. (1974). *Frame analysis: an essay on the organization of experience*. Cambridge, Mass: Harvard UP.
- Gouadec, D. (1989) Comprendre, évaluer, prévenir : Pratique, enseignement et recherche face à l'erreur et à la faute en traduction. *TTR*, 2 (2), 35-54.
- Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics: Volume 3, speech acts* (pp. 41-58). New York: Academic Press.
- Guilford, J.P. (1950). *Fundamental statistics in psychology and education* (2nd ed.). New York: McGraw-Hill.
- Guilford, J.P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw-Hill.

- Guilford, J.P. & Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill.
- Gumperz, J.J. (1982). *Discourse strategies*. New York: Cambridge University Press.
- Hambleton, R.K. (1999). Criterion-referenced testing principles, technical advances, and evaluation guidelines. In T.B. Gurtin & C.R. Reynolds (Eds.) *Handbook of school psychology* (3rd ed.). New York: Wiley.
- Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.
- Harasym, P.H., Norris, D.A. & Lorscheider, F.L. (1980). Evaluating student multiple-choice responses. *Evaluation & the Health Professions*, 3(1), 63-84.
- House, J. (1977). *A model for translation quality assessment*. Tübingen: Gunter Narr Verlag.
- House, J. (1997). *Translation quality assessment: A model revisited*. Tübingen: Gunter Narr Verlag.
- House, J. (2001a). Translation quality assessment: Linguistic description versus social evaluation. *Meta*, 46(2), 243-257.
- House, J. (2001b). Pragmatics and translation. In F. Nistal and J.M. Bravo Gozalo (Eds.) *Pathways of translation studies* (pp. 57-77). Valladolid, Spain: Centro Buendía.
- Huitt, W. (2002). Educational psychology interactive: Assessment, measurement & evaluation. Retrieved September 19, 2003, from the Valdosta State University Web site: <http://www.valdosta.edu/~whuitt/psy310/310msevl.html>.
- Hymes, D. (1964). Introduction: Toward ethnographies of communication. *American Anthropologist*, 66(6), 12-25.
- Hymes, D. (1972). On communicative competence. In J.B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp.). Harmondsworth, UK: Penguin Books.
- Johnson-Laird, P. (1983). *Mental models*. Cambridge: Cambridge UP.
- Jones, R. (1998). *Conference interpreting explained*. Manchester, UK: St. Jerome.
- Kopczynski, A. (1994). Quality in conference interpreting: Some pragmatic problems. In M. Snell Hornby, F. Pöchhacker & K. Kaindl (Eds.), *Translation studies: An interdisciplinary* (pp. 189-198). Amsterdam: John Benjamins.
- Kussmaul, P. (1995). *Training the translator*. Amsterdam/Philadelphia: John Benjamins.

- Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
- Labov, W. (1972). *Language in the inner city*. Philadelphia: University of Pennsylvania.
- Lambert, S. (1989). Simultaneous interpreters: One ear may be better than two. *The Interpreters' Newsletter*, 2, 11-16.
- Lambert, S. (1991). Aptitude testing for simultaneous interpretation at the University of Ottawa. *Meta*, 36(4), 586-594.
- Lambert, S., Daró, V. & Fabbro, F. (1995). Focalized attention on input vs. output during simultaneous interpretation: Possibly a waste of effort! *Meta*, 40, 39-46.
- Larose, R. (1998). Méthodologie de l'évaluation des traductions. *Meta*, 43(2), 163-186.
- Lauscher, S. (2000). Translation quality assessment: Where can theory and practice meet? *The Translator*, 6(2), 149-168.
- Leavitt, F. (1991). *Research methods for behavioral scientists*. Dubuque, IA: Wm. C. Brown.
- Lederer, M. (1978). Simultaneous interpretation – units of meaning and other features. In D. Gerver & H.W. Sinaiko (Eds.), *Language interpretation and communication* (pp. 323-332). New York: Plenum Press.
- Lederer, M. (1981). *La traduction simultanée : expérience et théorie*. Paris: Lettres modernes Minard.
- Lederer, M. (1994). *La traduction aujourd'hui : Le modèle interprétatif*. Paris: Hachette.
- Lee, A.C. (1999). The establishment of pass/fail standards for high stakes tests. Unpublished manuscript.
- Lee-Sing, A.C. (1999). *Performance standard determination in the health professions: A comparison of judgmental-based methods and cluster analytic techniques*. Unpublished doctoral thesis, University of Calgary, Calgary.
- Martínez Melis, N. & Hurtado Albir, A. (2001). Assessment in translation studies: Research needs. *Meta*, 46(2), 272-287.
- Massaro, D.W. (1975). *Understanding language*. New York: Academic Press.
- McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist*, 28 (1), 1-14.
- McClelland, D. C. (1993). Introduction. In L.M. Spencer & S.M. Spencer, *Competence at work: Models for superior performance* (pp. 3-8). New York: John Wiley and Sons.

- McGhan, W.F., Stimmel, G.L., Gilman, T.M. & Segal, J.L. (1982). Pharmacists as drug prescribers: Validation of certification exams and evaluation instruments. *Evaluation & the Health Professions*, 5(2), 115-129.
- Meschonnic, H. (1973). *Pour la poétique II*. Paris: Gallimard.
- Metzger, M. (1999). *Sign language interpreting: Deconstructing the myth of neutrality*. Washington, DC: Gallaudet University Press.
- Mikkelsen, H. (2000). *Introduction to court interpreting*. Manchester: St. Jerome.
- Morissette, D. (1984). *La mesure et l'évaluation en enseignement*. Sainte-Foy : Les Presses de l'Université Laval.
- Morson, G. S. & Emerson, C. (1990). *Mikhail Bakhtin: Creation of a prosaics*. Stanford: Stanford UP.
- Moser, B. (1978). Simultaneous interpretation: A hypothetical model and its practical application. In D. Gerver & H. W. Sinaiko (Eds.), *Language interpretation and communication* (pp. 353-368). New York: Plenum.
- Moser-Mercer, B. (1984). Testing interpreting aptitude. In W. Wilss & G. Thome (Eds.), *Translation theory and its implementation in the teaching of translating and interpreting* (pp. 318-325). Tübingen: Narr.
- Moser-Mercer, B. (1994a). Paradigms gained or the art of productive disagreement. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation*, (pp. 17-22). Amsterdam: Benjamins.
- Moser-Mercer, B. (1994b). Aptitude testing for conference interpreting: Why, when and how. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation*, (pp. 58-68). Amsterdam: Benjamins.
- Moser-Mercer, B. (1997). Beyond curiosity: Can interpreting research meet the challenge? In J.H. Danks, G.M. Shreve, S.B. Fountain & M.K. McBeath (Eds.), *Cognitive processes in translation and interpreting* (pp. 176-195). London: Sage.
- Mossop, B. (June 3, 2003). Personal communication.
- Newble, D., Dauphinee, D., Dawson-Saunders, B., MacDonald, M., Mulholland, H., Page, G., Swanson, D., Thomson, A. & Van der Vleuten, C. (1993). Guidelines for the development of effective and efficient procedures for the assessment of clinical competence. In D. Newble, B. Jolly & R Wakeford (Eds.), *The Certification and Recertification of Doctors: Issues in the assessment of clinical competence* (pp. 69-91). Cambridge: Cambridge UP.

- Nida, E. (1964). *Toward a science of translation*. Leiden: Brill.
- Norcini, J. (1994). Research on standards for professional licensure and certification examinations. *Evaluation & the Health Professions*, 17(2), 160-177.
- Norcini, J.J., Hancock, E.W., Webster, G.D., Grosso, L.J. & Shea, J.D. (1988). A criterion-referenced examination of physician competence. *Evaluation & the Health Professions*, 11(1), 98-112.
- Nord, C. (1996). El error en la traducción: categorías y evaluación. In A. Hurtado Albir (Ed.) *La enseñanza de la traducción* (pp. 91-108). Castello: Publicacions de la Universitat Jaume I.
- Norman, G.R. & Streiner, D.L. (1997). *PDQ statistics* (2nd ed.). St. Louis, MI: Mosby.
- Oléron, P. & Nanpon, H. (1965). Recherches sur la traduction simultanée. *Journal de psychologie normal et pathologique*, 62, 73-94.
- Ordre des traducteurs, terminologues et interprètes agréés du Québec. (2003). *Bienvenue à l'OTTIAQ*. Retrieved March 19, 2003 from the Web site : <http://www.otiaq.org>
- Paradis, M. (1994). Toward a neurolinguistic theory of simultaneous translation. *International Journal of Psycholinguistics*, 10 (3), 319-335.
- Paradis, M. (2000). Prerequisites to a study of neurolinguistic processes involved in simultaneous interpreting: A synopsis. In B. Englund Dimitrova & K. Hyltenstam (Eds.), *Language processing and simultaneous interpreting* (pp. 17-24). Amsterdam: John Benjamins.
- Pearson Education Development Group (2002). *Authentic assessment overview*. Retrieved September 19, 2003, from Teachervision.com Web site: <http://www.teachervision.com/lesson-plans/lesson-4911.html>.
- Pett, M.A., Lackey, N. & Sullivan, J.J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage.
- Peytard, J. (1995). *Mikhail Bakhtine: Dialogisme et analyse du discours*. Paris: Bertrand-Lacoste
- Plucker, J.P. (2003). John B. Carroll. Retrieved May 16, 2003, from the Indiana University Web site: <http://www.indiana.edu/~intell/carroll.html>.
- Pöschhacker, F. & Shlesinger, M. (2002). *The interpreting studies reader*. London: Routledge.

- Price, L.R. & Wilkins, C. (2001). Confirmatory factor analysis of a test used for certification in sport scuba diving. *Perceptual and motor skills*, 92(3), 869-880.
- Pyrzczak, F. & Bruce, R. R. (2003). *Writing empirical research reports: A basic guide for students of the social and behavioural sciences*. Los Angeles: Pyrczak Publishing.
- Quigley, S., Brasel, B. & Montanelli, D. (1973). Interpreters for Deaf people : Selection, evaluation & classification. Final Report HEW SRS 14-P-55400/5 (as cited in Strong & Rudser, 1985).
- Registry of Interpreters for the Deaf. (2001). *RID Homepage*. Retrieved September 19, 2003, from Registry of Interpreters for the Deaf Web site: <http://www.rid.org>.
- Reiss, K. & Vermeer, H.J. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: M. Niemeyer.
- Roberts, R.P. (1987). Spoken language interpreting vs. sign language interpreting. In K. Kummer (Ed.), *Across the gap: Proceedings of the 28th annual conference of the American Translators Association* (pp. 293-306). Medford, NJ: Learned Information Inc.
- Roberts, R.P. (1992). Student competencies in interpreting: Defining, teaching and evaluating. In E.A. Winston (Ed.), *Student competencies: Defining, teaching, and evaluating: Proceedings of the ninth national convention of the Conference of Interpreter Trainers* (pp. 1-17). Washington: Conference of Interpreter Trainers.
- Roberts, R.P. (1995). An assessment tool for community interpreting. In P.W. Krawutschke (Ed.), *Connections: Proceedings of the 36th annual conference of the American Translators Association* (pp. 135-145). Medford, NJ: Information Today Inc.
- Roberts, R.P. (1998). Interpreter assessment tools for different settings. In R.P. Roberts, S.E. Carr, D. Abraham & A. Dufour (Eds.), *The Critical Link 2: Interpreters in the Community* (pp. 103-120). Amsterdam: John Benjamins.
- Roy, C.B. (2000). *Interpreting as a discourse process*. New York: Oxford UP.
- Rudestam, K.E. & Newton, R.R. (2001). *Surviving your dissertation: A comprehensive guide to content and process* (2nd ed.). Thousand Oaks, CA: Sage.
- Sawyer, D.B. (2000). Towards meaningful, appropriate, and useful assessment: How the false dichotomy between theory and practice undermines interpreter education. *ATA Chronicle*, 29(2), 32-40.
- Schegloff, E. (1987) Some sources of misunderstanding in talk-in-interaction. *Linguistics*, 25, 201-218.
- Schiffrin, D. (1994). *Approaches to discourse*. Oxford UK: Blackwell.

- Schjoldager, A. (1995). Assessment of simultaneous interpreting. In C. Dollerup & V. Appel (Eds.), *Teaching translation and interpreting 3: New horizons* (pp. 187-195). Amsterdam: John Benjamins.
- Schweda-Nicholson, N. (1986). Language planning and policy development for court interpretation services in the United States. *Language Problems & Language Planning*, 10(2), 140-157.
- Searle, J.R. (1969). *Speech acts: an essay in the philosophy of language* London, Cambridge UP.
- Seleskovitch, D. (1968). *L'interprète dans les conférences internationales: problèmes de langue et de communication*. Paris: Minard Lettres Modernes.
- Seleskovitch, D. (1975). Langage, langues et mémoire : étude de la prise de notes en interprétation consécutive. Paris: Minard Lettres Modernes.
- Seleskovitch, D. & Lederer, M. (1984). *Interpréter pour traduire*. Paris: Didier Érudition.
- Seleskovitch, D. & Lederer, M. (1989). *Pédagogie raisonnée de l'interprétation*. Paris: Didier Érudition.
- Seleskovitch, D. (1997). Interview de Mme Arlette Morel, présidente de la Fédération nationale des sourds de France. *Meta*, 42(3), 560-563.
- Seltzer v. Foley*, 502 F. Supp. 600 (SDNY 1980)
- Setton, R. (1999). *Simultaneous interpretation: A cognitive-pragmatic analysis*. Amsterdam: John Benjamins.
- Shavelson, R.J., Hubner, J.J. & Stanton, G.C. (1976). Self-Concept: Validation of construct interpretations. *Review of Educational Research*, 46(3), 407-441.
- Smith, I.L. and Hambleton, R.K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practices*, 9(4), 7-10.
- Speers, E. (1997). *The comparative analysis of the certification processes of five professional associations*. Unpublished doctoral thesis, University of Toronto.
- Spencer, L.M. Jr. & Spencer S.M. (1993). *Competence at work: Models for superior performance*. New York: John Wiley and Sons.
- Sperber, D. & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell.

- Stansfield, C.W., Scott, M.L. & Kenyon, D.M. (1992). The measurement of translation ability. *The Modern Language Journal*, 76(iv), 455-467.
- Strong, M. and Rudser, S.F. (1985) An assessment instrument for sign language interpreters. *Sign Language Studies*, 49, 343-361.
- Tamblyn, R.M. (1994). Is the public being protected? Prevention of sub-optimal medical practice through training programs and credentialing examinations. *Evaluation & the Health Professions*, 17, 198-221.
- Thompson, B. & Levitov, J.E. (1985). Using microcomputers to score and evaluate items. *Collegiate Microcomputer*, 3(2), 163-168.
- Thorndike, R., & Hagen, E. (1986). *Measurement and evaluation in psychology and education* (4th ed.). New York: Wiley.
- Tombari, M.L. & Borich, G.D. (1999). *Authentic assessment in the classroom: Applications and practice*. Upper Saddle River, NJ: Merrill.
- Treisman, A. M. (1964). Selective attention in man. *British Medical Bulletin*, 20(1), 12-16.
- Treisman, A. M. (1965). The effects of redundancy and familiarity on translating and repeating back a foreign and a native language. *British Journal of Psychology*, 56(4), 369-379.
- Treisman, A.M. (1969). Strategies and Models of Selective Attention. *Psychological Review*, 76(3), 282-299.
- Viezzi, M. (1993). Consideration on interpretation quality. In C. Picken (Ed.), *The vital link. Proceedings of the XIIIth World Congress of FIT, Vol. 1* (pp. 389-397). London, Institute of Translation & Interpreting.
- Violato, C., McDougall, D. & Marini, A. (1992). *Educational Measurement and Evaluation*. Dubuque, Iowa: Kendall-Hunt.
- Violato, C. Salami, L. & Muiznieks, S. (2002). Certification examinations for massage therapists: A psychometric analysis. *Journal of Manipulative and Physiological Therapeutics*, 25(2), 111-115.
- Waddington, C. (1999). *Estudio comparativo de diferentes métodos de evaluación de traducción general (Inglés-Español)*. Madrid: Publicaciones de la Universidad Pontificia Comillas.
- Waddington, C. (2001). Different methods of evaluating student translations: The question of validity. *Meta*, 46(2), 311-325.
- Wadensjö, C. (1998). *Interpreting as interaction*. London: Longman.

- Warren, E. (February 12, 2001). Personal communication.
- Wiggins, G. (1990). *The case for authentic assessment*. Retrieved March 11, 2002, from Practical Assessment, Research & Evaluation, 2(2) Web site: <http://ericae.net/pare/getvn.asp?v=2&n=2>.
- Williams, F. (1992). *Reasoning with statistics*, 4th Ed. Fort Worth: Harcourt Brace.
- Williams, M. (1989). The assessment of professional translation quality: Creating credibility out of chaos. *TTR*, 2(2), 13-33.
- Williams, M. (2001a). The application of argumentation theory to translation quality assessment. *Meta*, 46(2), 326-344.
- Williams, M. (2001b). *An argumentation-centred approach to translation quality assessment*. Unpublished doctoral thesis, University of Ottawa, Ottawa.

Appendix I – Glossary

Section A

A Language

In conference interpreting, the label *A language* designates an interpreter's first and dominant language. It is expected that an interpreter will most often interpret from another, non-dominant language (*B language* or *C language*) into her A language. For example, a native English-speaking interpreter who also lists French as a working language would likely work from French (her *B* or *C language*) into English (her A language). See also *B language* and *C language*.

Alternate-Forms Reliability

Reliability can be measured by producing two separate instruments that are based on the same test framework and are rated with the same scoring scheme. The instruments are administered to the same group of test takers, and the results are then compared mathematically. See also *reliability*, *internal reliability*, *rater reliability*, *split-half reliability*, and *test-retest reliability*.

Angoff Method

The *Angoff method* (Angoff, 1971) is a procedure used to set the cut-off score for credentialing tests. In the method, a panel of subject-matter experts examines each item on the test at hand and estimates the lowest probability that a minimally competent examinee would answer each one correctly. The probabilities for all the items on the test are then summed. Finally, the average of the panel experts' probabilities is calculated to determine the cut-off score for the test (Lee, 1999, p. 12).

ASL

ASL (American Sign Language) is the naturally occurring sign language developed and used by Deaf people in the United States and most of Canada. It is linguistically related to some sign languages, such as LSF (*Langue des signes française*) or LSQ (*Langue des signes québécoise*), and unrelated to others, such as BSL (British Sign Language). See also *sign language*, and *Signed English*.

Assessment

Assessment is defined as the process of collecting, interpreting, and synthesizing data in order to make decisions (Gage & Berliner, 1991). See also *evaluation* and *measurement*.

Authentic Assessment

Authentic assessment aims to evaluate test takers' abilities in real-world contexts. Examples of authentic assessment formats include performance assessment and portfolio assessment (Pearson Education Development Group, 2002). See also *performance assessment*, *portfolio assessment*, and *traditional assessment*.

Section B

B Language

In conference interpreting, the label *B language* designates a second language that the interpreter knows well enough to interpret out of, and occasionally into. For example, a native English speaker who has French as a B language would most often interpret from French to English, and occasionally the other way around. See also *A language* and *C language*.

Bakhtinian Approach to Discourse

Like the work of proponents of other approaches to discourse, Bakhtin's work can be seen as an attempt to shift the focus in linguistics away from language as a system and towards an analysis of the communication between actors in a socially defined context (Peytard, 1995). Bakhtin objected to Saussure's distinction between *langue* (the system) and *parole* (the individual speech act) because it failed to adequately account for the *utterance*. Far from being a simple instantiation of a language system, he argued, the utterance is instead best defined as a unit that is extralinguistic in nature (Morson & Emerson, 1990).

To speak of an utterance is to presuppose the existence of an author, and an addressee. The utterance is produced by a particular individual, in a certain time and place, and is destined for a given audience. The utterance is also dialogical. It is connected backwards in time to utterances that came before, and it preserves a trace of them. It is also connected forwards in time, to a future utterance that will be formulated as a response (Bakhtin, 1986).

Section C

C Language

In conference interpreting, the label *C language* designates a third language that the interpreter knows well enough to interpret out of, but not into. For example, a native English speaker who has Portuguese as a *C language* would interpret from Portuguese to English, but not the other way around. See also *A language* and *B language*.

Cattell's Scree Test	<i>Cattell's scree test</i> (Cattell, 1966) is one of two common methods used to determine how many factors to keep in exploratory factor analysis. It involves plotting the eigenvalues on a graph, and through visual inspection identifying the point on the plot where the curve begins to level off. The location of the point indicates the number of factors to be retained. See also <i>factor analysis</i> , <i>eigenvalue</i> , and the <i>Kaiser-Guttman criterion</i> .
Certification	<i>Certification</i> is a type of credentialing used in professions where there is no perceived risk to a consumer's life. In such professions, governments are not directly involved in setting standards. Instead, professional associations may choose to set and ensure the achievement of standards through certification (Speers, 1997).
Checklist	A <i>checklist</i> is a series of statements, related to the test framework, that describes test takers' performances. Raters determine whether each item in the checklist is either present or absent from the performance in question. This type of scoring scheme has the advantage of being easy to use and easy to read, but it is susceptible to rater bias, and it does not distinguish an adequate performance from an excellent one (Gallagher, 1998). See also <i>rating scale</i> and <i>rubric</i> .
Cluster	Within the construct hierarchy used in the present study (Spencer & Spencer, 1993), a <i>cluster</i> is located at the top level. It groups together several related <i>competencies</i> , which are in turn made up of a number of connected <i>dimensions</i> .
Community Interpreting	<i>Community interpreting</i> takes place in a variety of community settings – such as classrooms, doctors' offices, and social service interviews – when a client and service provider are not comfortable expressing themselves in the same language. Discussions are often of a general nature, although some technical issues may arise. <i>Community interpreters</i> usually use the consecutive method to carry out their work.
Competency	A <i>competency</i> is an underlying characteristic in a person, which indicates that person's way of behaving, thinking, or generalizing across situations, and which endures for a reasonably long period of time (Spencer & Spencer, 1993). Competencies are most often studied in professional settings, as a means of identifying individuals who have the characteristics required to succeed in a given job or profession.

Conference Interpreting	<i>Conference interpreting</i> takes place in conferences, congresses, or assemblies when not all delegates or members are comfortable expressing themselves in the same language. Discussions are often technical, touching on political, economic, or scientific topics. Most often, <i>conference interpreters</i> use the simultaneous method, and equipment such as microphones, headphones, and insulating booths to carry out their work.
Construct	A <i>construct</i> is the concept or characteristic that a test is designed to measure. It is a theoretical variable inferred from multiple types of evidence (AERA, APA & NCME, 1999).
Construct Validity	Traditionally, one of the ways in which a validity argument was constructed was to show that test data accurately reflected information about the constructs the test purported to measure. For example, if the internal test structure, observations of response processes, or interrelations of the test scores with other variables were consistent with beliefs about the constructs, test users might argue that their score interpretations possessed construct validity. The latest version of the <i>Standards</i> (AERA, APA & NCME, 1999) suggests that validity is a unitary concept. There may be many different sources of evidence that support a validity argument, but only one type of validity. As a result, the <i>Standards</i> recommend that all test scores be viewed as measures of some construct. Despite this recommendation, much research still refers to construct validity. See also <i>validity</i>, <i>content validity</i>, and <i>criterion validity</i>.
Constructed-Response Item	A <i>constructed-response item</i> is one that requires test takers to formulate their own response to a test item. Examples include short answer questions, essay questions, and performance items.

Content Validity

Traditionally, one of the ways in which a validity argument was constructed was to show that the content of a test accurately depicted the universe of situations the test was intended to represent. For example, if the items of a credentialing test adequately sampled the range of activities in a given profession, or if the weighting of the items reflected their importance within the profession, test users might argue that their score interpretations possessed *content validity*.

The latest version of the *Standards* (AERA, APA & NCME, 1999) suggests that validity is a unitary concept. There may be many different sources of evidence that support a validity argument, but only one type of validity. As a result, the current version of the *Standards* (AERA, APA & NCME, 1999) now recommends the term *evidence based on test content*.

Despite this recommendation, much research still refers to content validity. See also *validity*, *construct validity*, and *criterion validity*.

Conversation Analysis

This approach to discourse was inspired by Garfinkel (1967), a sociologist who advocated studying mundane occurrences with the same attention normally devoted to extraordinary events. Garfinkel hoped to better understand the ways in which everyday life is organized.

Court Interpreting

Court interpreting takes place in the courtroom or at quasi-judicial hearings when a defendant or witness is not comfortable in the language of the proceedings. Discussions can be of a general nature, or they may address legal issues. *Court interpreters* generally use whispered simultaneous at the defence table, and consecutive at the witness stand.

Credentialing	<i>Credentialing</i> is a formal testing procedure that determines membership in a professional group (Lee-Sing, 1999). It is used to protect the public by assuring them that the professionals whose practice directly affects the public are at least minimally competent (Tamblyn, 1994). See also <i>licensure</i> , <i>certification</i> , and <i>accreditation</i> .
Credentialing Body	A <i>credentialing body</i> is the group that is responsible for granting credentials to professional practitioners. In the case of <i>licensure</i> , the credentialing body is often part of a government, or regulated by legislation. In the case of <i>certification</i> , the credentialing body is often a professional association. Note that while the credentialing body is ultimately responsible for the credentialing process, it may not necessarily develop the test used in the process. This work is carried out by the <i>test developers</i> .
Criterion Validity	Traditionally, one of the ways of constructing a validity argument was to show that the interpretations made from test scores matched those of an accepted, outside measure or <i>criterion</i>. For example, if a contemporary or future indication of performance yielded the same results as the test in question, test users might argue that their score interpretations possessed <i>criterion validity</i>. The concept was sometimes subdivided into <i>concurrent validity</i> (when the criterion was a contemporary measure) and <i>predictive validity</i> (when the criterion was a future measure) The latest version of the <i>Standards</i> (AERA, APA & NCME, 1999) suggests that validity is a unitary concept. There may be many different sources of evidence that support a validity argument, but only one type of validity. As a result, the current version of the <i>Standards</i> (AERA, APA & NCME, 1999) now recommends the term <i>evidence based on relationships to other variables</i>. Despite this recommendation, much research still refers to criterion validity. See also <i>validity</i>, <i>construct validity</i>, and <i>content validity</i>.
Cronbach's Alpha	<i>Cronbach's alpha</i> measures the internal consistency of a test (Everitt & Wykes, 1999). It shows the degree to which scores for a series of test items are similar. Values for alpha (represented by the symbol α) typically range between 0 and +1, although negative values are also possible (George & Mallery, 2001). Cronbach's alpha is often used to provide some indication of unidimensionality in the test data (Cortina, 1993). The higher the value of α (the closer it is to +1), the more likely it is that variation in the data is due to a single underlying factor.

Section D

Dialogue Scenario

In interpreting, a *dialogue scenario* is a situation in which an interpreter is asked to participate in a conversation between two (or more) parties that do not share a common language. Typically, one party speaks or signs and then pauses, allowing the interpreter to use the consecutive method. The second party responds and makes a similar pause for interpretation. Turn taking throughout the dialogue follows this pattern.

Dimension

Within the construct hierarchy used in the present study (Spencer & Spencer, 1993), a *dimension* is located at the bottom level. Several related dimensions make up a *competency*, while related competencies are grouped together to form a *cluster*.

Dimensionality

Describing the *dimensionality* of a set of test scores entails demonstrating, through statistical evidence, the number of variables present in the data. If analysis reveals the presence of a single variable, the data are said to be *unidimensional*. If analysis reveals the presence of several variables, the data are said to be *multidimensional*.

Discourse

In the present study, *discourse* is defined broadly as language above the clause or sentence (e.g., paragraphs, texts, question-answer pairs, intonational contours, etc.) or as language use in context (Schiffrin, 1994).

To give the reader a clearer sense of what the term means, we referred to the six types of discursive research identified by Schiffrin (*speech acts theory, interactional sociolinguistics, the ethnography of communication, pragmatics, conversation analysis, and variationist linguistics*), or to other research that holds similar aims (e.g., the *Bakhtinian approach to discourse*).

Section E

Effect Size

Effect size is a measure of the differences between two groups of observations. The most common means of estimating effect size is to subtract the mean of one group from that of the other, and to divide the difference by the standard deviation of either group. Effect size is therefore expressed in terms of standard deviations, and not in terms of a specific unit of measurement, which makes it ideal for comparisons between different studies (Cohen, 1988).

Eigenvalue	In factor analysis, an <i>eigenvalue</i> is a measure of the variance accounted for by each factor (Norman & Streiner, 1997, p. 145). As a matter of convention (a convention known as the <i>Kaiser-Guttman criterion</i>) a factor with an eigenvalue less than 1.0 is not retained for consideration (Norman & Streiner, 1997, p. 145). See also <i>factor analysis</i> .
Ethnography of Communication	This approach to discourse is based largely on the work of Hymes (1972). Its proponents view language as a system whose rules and norms are constrained by culture, and they argue therefore that communication through language does not operate in the same way in all cultures. Taking their cues from anthropology, they assert that understanding a speech community requires observation of its members in a wide range of life activities. This allows researchers to learn how to think, act, and believe, all according to the norms of the speech community (Schiffrin, 1994). Proponents of this approach take issue with Chomsky's notion of linguistic competence, arguing that language users need more than grammatical competence to be able to communicate effectively; they also need to know how language is used by the members of their speech community. This knowledge – or communicative competence – allows them to use language in the concrete situations of everyday life, for example, to engage in a conversation, to shop in a store, or to interview for a job (Schiffrin, 1994).
Evaluation	<i>Evaluation</i> is the process of making judgements about the value or quality of data by comparing it to a standard (Huitt, 2002). See also <i>assessment</i> and <i>measurement</i> .
Section F	
False Negative	A <i>false negative</i> is an error in which an individual test taker is assessed or predicted not to meet the criteria for inclusion in a particular group (such as the group of test takers who pass), but in truth does meet these criteria (AERA, APA & NCME, 1999).
False Positive	A <i>false positive</i> is an error in which an individual test taker is assessed or predicted to meet the criteria for inclusion in a particular group (such as the group of test takers who pass), but in truth does not meet these criteria (AERA, APA & NCME, 1999).

Factor Analysis

The term *factor analysis* refers to a series of techniques that are used to show whether a series of variables can be explained by a smaller number of constructs, known as *factors* (Norman & Streiner, 1997, p. 142). There are two main types of factor analysis. Exploratory factor analysis is used to look for a structure or pattern in a set of data, while confirmatory factor analysis is used to determine whether a specific, predefined model is compatible with the data (Everitt, 1996, p. 225). See also *eigenvalue* and *principal components analysis*.

Field Test

A *field test* is a test administered to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting. A *field test* is generally more extensive than a pilot test. See *pilot test* (AERA, APA & NCME, 1999).

Fingerspelling

Like most sign languages, ASL can represent each letter of the written alphabet with a particular handshape. An English word (or a word from another spoken language) represented in this way is said to be *fingerspelled*. Fingerspelled words can be used on singular occasions (such as when a signer is trying to communicate with another person who does not know ASL), or they can be borrowed into ASL and lexicalized to varying degrees.

Section G**Generalizability Analysis**

A *generalizability analysis* (Cronbach, Gleser, Nanda & Rajaratnam, 1972) is normally made up of two components. The first is the *generalizability study* (G study), which allows researchers to measure the multiple sources of error inherent in the design of a test. The second is the *decision study* (D study), which allows researchers to determine how large a sample size needs to be to achieve a given level of reliability.

Section I

Interactional Sociolinguistics

This approach to discourse is largely based on the work of two scholars. Gumperz (1982) focuses attention on the ways in which differences in contextualization affect the communication process. When two people from different cultures attempt to communicate, they apply their background knowledge to make inferences about the meaning of an utterance (i.e., they create their own contextualizations). As a result, although they may share similar grammatical knowledge of a language, they will come to understand very different things about the utterance.

Goffman (1974) suggests that people use frames – constructed and organized definitions of situations – in order to make sense out of the events they participate in. Goffman's work can be understood as an attempt to isolate some of the more basic frames that people use in society, including those used in talk. He is also interested in analyzing the vulnerabilities inherent in frames that provoke misunderstanding.

Internal Reliability

One means of determining measurement consistency is to analyze the statistical relationships among scores from different test items or from different parts of a test. See also *reliability*, *alternate-forms reliability*, *rater reliability*, *split-half reliability*, and *test-retest reliability*.

Interpreting

Interpreting may be defined as the act of communicating a spoken or signed message between two groups of people who do not share, or who do not feel comfortable using, the same language. At times, the act of interpreting may also involve the oral or signed communication of a written message in another language.

Roberts (1987) suggests a number of criteria that may be used to characterize interpreting. Three of them are explained briefly below.

Mode of Communication

There are three main modes of interpersonal linguistic communication: spoken, written, and sign. Interpreting may take place within single mode (e.g., spoken to spoken) or between modes (e.g., sign to spoken, or written to spoken).

Method of Interpreting

There are two main methods of interpreting: consecutive and simultaneous. In the consecutive method, the interpreter gives her interpretation after the producer has finished his communication. In the simultaneous method, the interpreter gives her interpretation at virtually the same time as the communication is given by the producer.

Interpreting Setting

There are three main settings in which interpreting is used: the community, the courtroom, and the conference hall. These are explained elsewhere in the glossary. See *community interpreting*, *conference interpreting*, and *court interpreting*.

Item Difficulty

To calculate an index of *item difficulty*, test developers divide the number of test takers who answered an item correctly by the total number of test takers. Note that a high item difficulty actually suggests that the item was easy for the test takers to answer.

Items tend to improve test reliability when the percentage of students who correctly answer the item is halfway between the percentage expected to correctly answer if pure guessing governed responses and the percentage (100%) who would correctly answer if everyone knew the answer. For example, on a four-alternative multiple-choice item, if everyone responded purely on the basis of guessing, theoretically 25% of the students would correctly answer the item on the basis of chance alone. The value halfway between this percentage and 100% would be 62.5% ($25\% + (100\% - 25\%)/2$) (Thompson & Levitov, 1985).

Item Discrimination

To calculate an index of *item discrimination*, test developers divide the test takers into three equal groups based on their total score: low scorers, middle scorers, and high scorers. They then subtract the number of low scorers who answered a dichotomous item correctly from the number of high scorers who answered the item correctly. The difference is subsequently divided by the number of test takers in each group (Ebel & Frisbie, 1986)

Based on their experience with classroom tests, Ebel and Frisbie suggest that discrimination indices be interpreted in the following manner:

0.40 and up – very good items

0.30 to 0.39 – reasonably good but possibly subject to improvement

0.2 to 0.29 – marginal items, usually needing and being subject to improvement

below 0.19 – poor items, to be rejected or improved by revision

The authors also note, however, that many good items used in content-referenced measures (e.g., minimum competency tests, and professional certification tests) may have discrimination indices of zero or near zero. The explanation for this phenomenon relates to the fact that score distributions from content-referenced measures tend to be quite negatively skewed and low in variability. The upper and lower criterion groups tend to be very similar in terms of total test score (Ebel & Frisbie, 1986)

Item Distractor

In a dichotomously scored multiple-choice item, the correct answer is supplied, along with a number of incorrect ones, which are meant to distract the test takers. The incorrect answers are therefore referred to as *distractors*.

Item Distractor Analysis

To analyze item distractors, test developers divide the test takers into three equal groups based on their total score: low scorers, middle scorers, and high scorers. They then calculate the percentage of participants in the high and low scoring groups that chose each answer option (the correct answer and all distractors) (Carey, 2001).

Developers use the analysis to ensure that the distractors are working effectively. Ideally, the high scorers should outperform the low scorers by selecting the correct answer more often, and all the answer options should have been chosen by test takers in either group. If this is not the case, it may indicate that the test item needs to be revised.

Section K

Kaiser-Guttman Criterion

Application of the *Kaiser-Guttman criterion* is one of two common methods used to determine how many factors to keep in exploratory factor analysis. The criterion suggests that only those factors with eigenvalues greater than or equal to 1 should be retained. See also *factor analysis*, *eigenvalue*, and *Cattell's scree test*.

Section L

Licensure

This is a type of credentialing that required when unqualified practice may pose a risk to a consumer's life, health, or safety. Licensed professions are often regulated by public statutes and have governing bodies that are solely responsible for granting licences (Speer, 1997).

Section M

Measurement

Measurement is the process of quantifying assessment data. It involves three steps: 1) identifying and defining the quality or attribute that is to be measured; 2) determining a set of operations by which the attribute may be made manifest and perceivable; and 3) establishing a set of procedures or definitions for translating observations into quantitative statements of degree or amount (Thorndike & Hagen, 1986). See also *assessment* and *evaluation*.

Mot à signification unique

A *mot à signification unique* is an element of the source-language discourse that interpreters can "transcode" directly into the target language. Numbers, proper names, and technical terms can all be processed this way because they are all "universal knowledge" (Seleskovitch, 1975).

Section P

Pearson's Product-Moment Correlation

Pearson's product-moment correlation measures the degree to which two continuous variables are related (George & Mallery, 2001, p. 359). Values for Pearson's (designated by a lower-case r) typically range between -1 and +1. A negative value indicates that as one variable increases, the other tends to decrease; the closer the value is to -1, the stronger is that tendency. A value of 0 indicates that there is no relationship between the variables. A positive value indicates that as one variable increases, the other also tends to increase. The closer the value is to +1, the stronger is that tendency (George & Mallery, 2001, p. 112-113).

Performance Assessment	<i>Performance assessments</i> are product- and behaviour-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied (AERA, APA & NCME, 1999). See also <i>authentic assessment</i> , and <i>portfolio assessment</i> .
Pilot Test	A pilot test is a test administered to a sample of test takers to try out some aspects of the test or test items, such as instructions, time limits, item response formats, or item response options. See <i>field test</i> (AERA, APA & NCME, 1999).
Portfolio Assessment	A <i>portfolio assessment</i> makes a decision about a candidate based on a structured collection of work products, which the candidate selects and assembles him- or herself. See also <i>authentic assessment</i> , and <i>performance assessment</i> .
Pragmatics	This approach to discourse is based largely on Grice's (1975) work. He suggests that a strictly logical analysis of utterances, based on the stable senses of words (their "semantic meanings"), would make them appear to be a succession of disconnected remarks. Yet they are not perceived this way. Instead, receivers of an utterance are generally able to interpret the communicative intention (the "speaker meaning") behind the utterance (Schiffrin, 1994). This led Grice to formulate his "principle of cooperation", which posits the existence of a number of underlying assumptions that govern all communicative exchanges. Receivers assume that an utterance will respect the maxims of quantity (it will contain an appropriate amount of information), quality (it will be truthful), relation (it will be relevant), and manner (it will be clear, brief, and orderly). Because all parties share these implicit beliefs, producers are able to lead receivers to an interpretation of their communicative intent.
Principal Components Analysis	Although there is considerable disagreement between statisticians and social scientists over the meaning of the term <i>principal components analysis</i> , in the present study, it is used to indicate a method of factor extraction in exploratory factor analysis (Pett, Lackey & Sullivan, 2003, p. 89). See also <i>factor analysis</i> .

Psychometrics

Some authors in the field of educational psychology make a distinction between two philosophies of education and testing. The first, which they label “psychometric”, is the product of associationist psychology (Brunning, Schraw & Ronning (1999) and posits a belief in the existence of innate mental qualities such as intelligence (Elkind, 1991). It also favours the use of traditional assessment methods. The second, sometimes labelled “edumetrics” or the “developmental approach”, is the product of cognitive psychology (Brunning, Schraw & Ronning, 1999), and posits a belief in the development of mental competencies (Elkind, 1991). It advocates the use of authentic assessment. See also *authentic assessment* and *traditional assessment*.

Elsewhere, the term is used in a much more basic way. As its components “psycho” and “metrics” might suggest, it refers merely to the measurement of mental abilities. In this sense, anything consistent with, or anyone sympathetic to, the principles of measurement and evaluation may be described as *psychometric*. It is in this broader sense that the term is used in the present study.

Section R

Rater Reliability

Reliability can be established by evaluating the consistency of raters’ judgements. When candidates are scored by more than one rater, these scores may be compared (inter-rater reliability). When an individual rater scores many candidates, these scores may be compared (intra-rater reliability). See also *reliability*, *alternate-forms reliability*, *internal reliability*, *split-half reliability*, and *test-retest reliability*.

Rating Scale

A *rating scale* is a type of scoring scheme, which may have different levels of complexity. At its simplest, the scale may be numerical (e.g. it may ask raters to evaluate a performance on a scale from 1 to 10) or nominal (e.g. it may ask raters to evaluate a performance as poor, acceptable, or good). More complex rating scales may take the form of a rubric – where different criteria (taken from the objectives) are listed in a table and are cross-matched with descriptions of those criteria for different levels of achievement (Gallagher, 1998). See also *checklist* and *rubric*.

Reliability	<i>Reliability</i> is defined as the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure, and hence are inferred to be dependable and repeatable for an individual test taker. It also represents the degree to which scores are free of errors of measurement for a given group (AERA, APA & NCME, 1999). The concept of <i>reliability</i> has traditionally been broken down into several subcategories. See also <i>alternate-forms reliability</i> , <i>internal reliability</i> , <i>rater reliability</i> , <i>split-half reliability</i> , and <i>test-retest reliability</i> .
Rubric	A <i>rubric</i> is a set of criteria, rules, principles, and illustrations used in scoring responses to individual items or clusters of items. The term is usually used in reference to assessment tasks that do not provide enumerated responses from which test takers make a choice. Rubrics vary in the degree of judgement entailed, in the number of distinct score levels defined, and in other ways. It is common for a rubric to contain examples of performances at each of the score levels to help clarify the criteria (AERA, APA & NCME, 1999). See also <i>checklist</i> and <i>rating scale</i> .
Section S	
Selected-Response Item	A <i>selected-response item</i> is one that enumerates a series of responses from which test takers make a choice. Examples include multiple-choice items, true-false items, matching items, etc.
Sight Translation	To perform a <i>sight translation</i> , an interpreter reads a text written in the source language and provides an oral or signed version of it in the target language, often with little or no preparation time.
Sign Language	A <i>sign language</i> is a naturally occurring language that develops within a community of Deaf people. Like spoken languages, sign languages can be used to communicate a limitless number of concepts and ideas. They can also be analyzed in terms of their linguistic structure – phonology, morphology, lexicon, syntax, etc. – the same way any spoken language can. Examples of sign languages include <i>ASL</i> and <i>LSQ</i> .

- Signed English** This is the name given to a number of artificial linguistic systems that use manual signs to represent English grammar and syntax. Traditionally, hearing teachers used these systems to teach English to Deaf students. They still exert considerable influence in the Deaf community today (see *Pidgin Signed English* and *Signing Exact English*). There are also signed versions of other spoken languages (e.g., Signed French).
- Speech Act Theory** Based on the work of Austin (1962) and Searle (1969), this approach to discourse is largely concerned with the underlying conditions for production and interpretation of acts through words. Austin's contribution to the current was the identification of the three features of the speech act: locution (the utterance itself), illocution (the expression of the producer's attitude towards the speech act), and perlocution (the effect the speech act has on the receiver).
- Searle's contribution to the approach took the form of his *principle of expressibility*. He argued that speakers are always able, in principle at least, to say what they mean through the use of language. This means that every illocutionary act speakers attempt is literally expressible, and that an analysis of the logical forms of sentences will reveal the presence of *illocutionary force markers* (such as verbal mood, sentence type, punctuation signs, etc.). In formulating the principle, Searle effectively provided a means for connecting speech act theory, which had up until that point largely been a philosophical concern, with the study of language (Schiffrin, 1994).
- Split-Half Reliability** This is one way of determining internal reliability. If a test instrument has a large number of items, scores from the items can be divided into two groups (for instance, scores for even-numbered items versus scores from odd-numbered items) and then compared statistically. See also *reliability*, *alternate-forms reliability*, *internal reliability*, *rater reliability*, and *test-retest reliability*.
- Standard Setting** To determine which scores will fail and which scores will pass, test developers must establish a cut-off score, or *standard*. They cannot choose this point arbitrarily; instead, they must set a standard that reduces the likelihood of error in the categorization of test takers – the risk that competent test takers will fail (a false negative) or that incompetent test takers will pass (a false positive). Categorization error can never be completely eliminated, but there does need to be some reasonable process in place to arrive at the standard that is used (Hambleton, 1999).

Section T

Test

A test is an evaluative device or procedure in which a sample of an examinee's behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardized process (AERA, APA & NCME, 1999). In common usage, a test is simply a series of individual test items; in a technical sense, however, the term is used more comprehensively. A test comprises not only the test items, but also the test framework, the scoring scheme, pilot testing results, a performance standard, a cut-off score, and evidence of psychometric properties.

Test-Retest Reliability

Reliability can be measured by assessing candidates on two separate occasions – they are not allowed the opportunity to learn or practice between the administrations – and by comparing the results mathematically.

Test Developer

Test developers are responsible for building the *test*, and for documenting its technical quality for an intended purpose (AERA, APA & NCME, 1999). Developers may be the same party as the *credentialing body*, or they may be contracted by the credentialing body to build an instrument.

Test Development

Test development refers to the complete set of processes required to plan, construct, evaluate, and modify a test. Developers must ensure that the content, format, item properties, administration, scoring and technical quality are appropriate for their intended purpose (AERA, APA & NCME, 1999).

Traditional Assessment

Traditional assessment uses tests constructed from simple and efficient indirect items, and it relies on the items' demonstrated validity – rather than an overt simulation of real-world tasks – to make inferences about test takers' abilities (Wiggins, 1990). Typically, this has meant constructing test instruments made up primarily of large numbers of *selected-response items*. See also *authentic assessment*.

Translation Quality Assessment (TQA)

This is the name given to a current of research in Translation Studies. It attempts to define the features of "good" translation (House, 2001), to categorize any number of major or minor errors (Williams, 2001), and to collect evidence to support various theories of translation (Waddington, 2001). The TQA approach has at times been extended to consider interpreting.

Section V

Validity

Many introductory texts (Linn & Gronlund, 1995; Jackson, 1996; and Tombari & Borich, 1999) suggest that *validity* is an indication that a test measures what it is intended to measure. A more technical definition would state that *validity* is the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test (AERA, APA & NCME, 1999).

The current version of the *Standards* takes the position that validity is a unitary concept. Many sources of evidence may be considered in making a validity argument, but there is only one type of validity (AERA, APA & NCME, 1999). However, many researchers still refer to the traditional types of validity. See also *construct validity*, *content validity*, and *criterion validity*.

Variation Analysis

Variationists believe that heterogeneity in language is influenced by both linguistic and social forces, and that these forces can be identified through systematic study of a language community. This approach has largely been shaped by the work of Labov (for a classic example, see Labov, 1966).

Variationists have traditionally been interested in understanding the social and linguistic constraints on “semantically equivalent variants”. While this has typically meant focusing attention on traditional linguistic concerns (morphology, syntax, lexicon, etc.), the approach has also been extended to the study of entire texts (Schiffrin, 1994). This extension can be seen, for example, in Labov (1972). The author examines narratives as a discourse unit, demonstrating that they have a linear structure, and that different sections of a narrative have different functions.

Varimax Criterion

The *varimax criterion* is the most common method of orthogonal (angle-preserving) axis rotation used in factor analysis. Like other rotation methods, it provides a simpler structure and a more interpretable pattern of values (George & Mallery, 2001). See also *factor analysis*.

Appendix II – Existing Test Documentation

The documents contained in the Appendix are verbatim copies of those given to raters.

Rater Instructions

Marche à suivre pour les Membres du jury d'examen

- A. Lire les critères d'évaluation – les 3 échelles ci-jointes – après avoir lu les instructions ci-après et apprécier qu'il y a un ORDRE À SUIVRE RIGOREUSEMENT.
- B. Au moyen d'un appareil d'écoute en stéréophonie ou de deux lecteurs mono (à la rigueur un seul), écouter la (ou les) cassette(s) soumise(s) pendant 5 à 10 minutes au maximum, à titre de mise en train.
- C. Écouter attentivement, dans l'intention de juger (crayon en main) l'interprétation seulement de l'extrait indiqué de 5 minutes environ.
- D. Noter selon l'échelle d'INTELLIGIBILITÉ.
- E. Écouter maintenant avec le même soin l'original seulement du même extrait.
- F. Écouter à la fois l'original et l'interprétation par segments de 30 secondes du même extrait pour vérifier la correspondance informative; noter au moyen de l'échelle d'analyse du CONTENU INFORMATIF.
- G. Réécouter l'interprétation du même extrait et noter selon l'échelle de FORME.
- H. Reprendre le même processus, de B à F pour tout autre extrait de cassette soumis à l'examen.
- I. Livrer les résultats à l'état brut en utilisant les formules et les enveloppes fournies (une note par extrait et par échelle; 9 notes au total). Ils seront calculés proportionnellement,
 - L'échelle d'intelligibilité comptant pour 30%
 - L'échelle d'informativité comptant pour 50%
 - L'échelle de forme comptant pour 20%.
 La note finale sera la moyenne de celles des 3 membres du jury.
- N.B. Si plus d'un des 3 membres du jury donne une note éliminatoire dans une échelle quelle qu'elle soit, le candidat a échoué, même si sa moyenne générale est au-dessus de 60%.

Intelligibility Rating Scale

10. Perfectly clear and intelligible. Sounds like ordinary speech; is neat, clear and trenchant.
9. Perfectly or almost clear and intelligible, but contains instances of minor grammatical or stylistic imprecision and/or unusual word usage that could, nevertheless, be easily clarified.
8. Generally clear and intelligible, but clarity in style, word choice and/or syntactical arrangement is somewhat poorer than in category 9.
7. The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by a certain vagueness of style, poor word choice, inappropriate expressions, uninterpreted words and phrases, and incorrect or equivocal grammatical arrangements.
6. Same as above except that the listener is obliged to post-edit to make the message understandable.

** _____

5. The general idea is intelligible only by dint of considerable attention, but thereafter one is fairly confident to have understood. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present but constitute mainly "noise" through which the main idea is still perceptible.
4. Masquerades as intelligible speech, but actually it is more unintelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement and/or alternative expressions are generally bizarre, and there may be critical words left uninterpreted.
3. Generally unintelligible; it tends to sound like nonsense, but with a considerable amount of concentration and reflection, one can at least hypothesize the idea intended.
2. Almost hopelessly unintelligible, even with painstaking attention and effort. Nevertheless it does not seem completely nonsensical.
1. Hopelessly unintelligible. It appears that no amount of attention and effort would ever reveal the thought of the message.

** A candidate will be automatically failed if given a score below this mark by more than one judge.

Informativeness Rating Scale

Having listened SEPARATELY to the original, you conclude that:

10. The original contains, if anything, less information than the interpretation. The interpreter may have added certain meanings, apparently to make the passage more understandable.
9. The original is not informative at all; no new meaning is added nor is the reader's confidence in his understanding increased or enhanced.
8. The original adds no new meaning either at the word level or the grammatical level, but the listener is somewhat more confident to have apprehended the meaning as intended.
7. By correcting one or two possibly critical meanings, chiefly at the word level, the original gives a slightly different "twist" to the meaning conveyed by the interpretation. It adds no new information about sentence structure, however.
6. In contrast to 7, the original adds a certain amount of information about the sentence structure and syntactical relationships. It may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words.

**

5. Between 6 and 4.
4. The original is clearly informative and adds considerable information about sentence structure and individual words, putting the listener "on the right track" as to the meaning intended.
3. Between 4 and 2.
2. The original is very informative, contributing a great deal to the clarification of the meaning intended. By correcting sentence structure, words, and phrases, it makes a great change in the listener's impression of the meaning intended, although not to the point of changing or reversing the meaning completely.
1. The original is extremely informative. It makes "all the difference in the world" in comprehending the meaning intended. (A rating of 1 should always be assigned when the original completely changes or reverses the meaning conveyed by the interpretation.)

** A candidate will be automatically failed if given a score below this mark by more than one judge.

Style Rating Scale

- 10/9. Outstanding: overall impression is of excellent style with correct language level throughout. Gives evidence of varied and aptly chosen vocabulary, with semantic/syntactic arrangements of words and phrases perfectly attuned to the language used. Imagery may be presumed perhaps even more forceful than that of the original; metaphors, similes, sayings, etc., all expertly rendered. Delivery is assured and hesitations, self-corrections, no "er's" or blanks, etc. Excellent diction, right speed and voice qualities. "Easy listening", with interpreter in complete and obvious control of the interpretation.
- 8/7. Generally very good and fully satisfactory: language level is correct, but style less generally outstanding, though flashes of brilliance are possible, and vocabulary perhaps somewhat more limited, with occasional minor lapses in words and phrases chosen. Delivery quite fluent and confident, with good diction, but contains occasional minor hesitations, delays, false starts, repetitions, etc.; may be slightly uneven or monotonous. Voice itself may be less easy to attend to.
- 6/5. Good, satisfactory performance; style is workmanlike and language level generally appropriate. Vocabulary unexceptional but apt, with occasional semantic or syntactic flaws, some unusual words or expressions. Imagery is present, but listener may have the occasional suspicion of the original's being slightly muted. Diction is sufficiently clear, but delivery less fluent and assured than in 8/7, with some hesitations, false starts, self-corrections, "er's", occasional dangling word, etc., not however disturbing to the general effect; may be a little too fast or too slow and voice qualities may be poorer than in 8/7.

**

- 4/3. Minimally satisfactory; general impression is of style being adequate but only fair and occasionally lame. Language level may be fluctuating or inappropriate, and vocabulary relatively weak. The force of the original can be suspected to be distinctly dampened. Some grammatical and semantic errors, bizarre words, expressions, or sentences are in evidence, and metaphors and sayings may go awry. Delivery is adequate but shows some strain, with diction less clear; may be uneven, staccato or dragging, but not enough to be very obtrusive; occasional blanks, and words or phrases may sometimes be left dangling. Voice qualities may leave something to be desired. Needs to improve.
- 2/1. Not acceptable: style is distinctly inept, with language level jarringly inappropriate, possibly. Vocabulary is poor and repetitive and imagery weak, leaving listener with a definite impression that original has been seriously marred. Semantic anomalies and syntactic/grammatical errors in abundance; bizarre words and expressions dot the interpretation. Metaphors and maxims, if attempted, routinely misfire or are incomprehensible. Delivery is decidedly strained, may be spasmodic, rapid-fire or soporific, replete with false starts, blanks, "er's", malapropisms, dangling syllables, helter-skelter phrasing and aborted sentences much in evidence. Diction may be poor, voice tone and timbre may be grating, continually changing, etc., with overall effect being abrasive.

** A candidate will be automatically failed if given a score below this mark by more than one judge.

Rating Sheet

Evaluation Sheet

Candidate _____

Excerpt	A	B	C
Intelligibility			
Informativeness			
Style			

Jury Member

Comments:

Appendix III – New Test Documentation

Transcript of the Ten-Minute Speech

Applying Market-Based Principles to Public Policy:
A speech given by Preston Manning to the Fraser Institute
Toronto, November 26, 2002

I resigned my seat in Parliament at the beginning of this year and I've been spending time at the university campuses, in Calgary and the University of Toronto. We've just had a great experience here in getting to know a large number of people. U of T is a huge intellectual resource, and it's been marvellous for us not just only meeting the professors, but meeting some very, very bright students.

You might find it interesting that after one of my lectures to one of the students – err – this political science class – I always try to invite feedback from the students. Someone came up to me and said that, “You know, whenever you talk to us about federal politics here, you should spend the first five minutes telling us why we should care.” This was a political science student. It's a bit of an indication of how tuned off, err, tuned out of national politics our young people are. There was a bit of a hard edge to his comment too, that “We don't find it attractive, this subject. You've got to work at it to make us” ...err ... So I have been working at it, and hopefully, I might be successful.

As Michael mentioned, I've just finished this book and I've been on the customary book tour, called *Think Big*. It basically is – it's published by McClelland & Stewart, the publisher likes me to get that in. It's basically the story of a small group of people who, 15 years ago, wanted to change the national agenda. You may or may not agree with the way we wanted to change it, but the basic idea is that we disagreed with certain things on the national agenda. We wanted a change. We wanted to make deficit, err, budget balancing and deficit reduction, a big issue, which it wasn't. It wasn't high enough on the priorities of the federal parties at that time. We wanted to do that to get to the point of tax reduction and debt reduction. We didn't want to pursue deficit reduction as an end in itself.

We wanted to make certain constitutional changes. At least to get some institutional and attitudinal changes in Ottawa that gave more attention to the concerns of all of the regions – more regional balance in Canadian federalism. That was one of the big motivations behind our pushing for Senate reform. We wanted to democratize the federal institutions, make them more accountable, get freer voting in the Parliament, less party discipline, better balance between the Prime Minister's office and the legislature.

And we ...uh... used the tools to try and change the agenda that we used, are the tools that are available to everybody, and that are available today. Freedom of speech, freedom of association, freedom to try to persuade your friends and neighbours to vote this way instead of that way, or to crusade for this idea instead of that idea. And we had, I think – you'll have to be the judge, I'm not objective on this subject – but I think we had some

success. We were able to make at least the budget balancing enough of an issue that particularly the Liberals who had formed the government in '93 had to address it more seriously than they would have otherwise. We had some success on the constitutional front in at least stiffening the federal government's approach to separatism. We argued that you had to be crystal clear about what the consequences of secession were, but offer a better, more constructive federalism as an alternative. Some of that became part of the government's position.

The front where we had the least success, and is therefore still a very important item for the national agenda, is on the democratic front. That was the area where, despite numerous attempts to get changes in the procedures and approaches in the House of Commons, we had the least success. So there's still a lot of work to be done, that's what I'm saying, on the democratic front.

Just to illustrate this democracy deficit – as it's now referred to in Ottawa – my last vote in the House of Commons was in December of last year. As you know, as Christmas comes close, and adjournment comes close, Parliament gets a little bit like a school. Everybody's anxious to get out. There's a huge pile of votes that pile up, and you vote half the night to try to get through them. But on this one evening, the bells rang, it was snowing outside, and the members faithfully trooped down to the Chambers. You know in our Chamber, we still have a lot of standing votes. It takes half an hour to take a total standing vote in the House, like, we have the most archaic... – but that's another subject, and we'll get into that another day.

But anyway, on this evening, there was a supply bill that was one of the first items on the agenda. Most members don't even know – the supply is voting money to the government for its expenditures – most members don't even know how much those are for anymore. It's not – unless you ask, you don't know how much it's for. I tend to make it a point of knowing how much it's for, and this particular one was for \$6 billion. It attracted no interest at all from anyone in the House. It's just \$6 billion, the government is for it, the opposition is against it, the vote goes on, and it is carried.

But following that bill there was a motion, a private member's motion, requesting that the federal government disclose why it reduced the strychnine level in gopher poison from five percent to three percent. Some prairie member that had a mad-on with respect to gophers wanted this information, managed to get it on the order paper as a motion. He had obviously done his homework, because there's members all on the opposition side were starting to vote for this motion. And then, lo and behold, the Government starts to vote, and there's Ontario members are voting for this gopher motion. There's a quickening of interest in the House. There's a rallying together: French speakers, English speakers, federalists, separatists are coming together on this ground-breaking issue. And, lo and behold, the gopher motion passed, which is unusual, for an opposition member's motion to pass. And members are throwing papers in the air, and jumping over their desks to slap on the back the member that had proposed this ground-breaking motion. I went out of the House, it was my last time in the House, saying like, "What is wrong with this place, when a gopher motion actually – because members had to make up there minds on how to vote on it, you see, it's a private member's bill – actually stimulated more interest and attention than a \$6 billion

supply bill. So there's still a lot of work to do – is what I'm telling you – on the democracy front.

Now I... uh.. what I thought I might touch on briefly today – and we're going to have a bit of a question period as well, so this may stimulate your thinking. The national agenda today – I'm sure there's people here who would like to change it. Who would have different priorities. Who would like see things higher on the list than they are. Maybe see some other things off the list. I would like to touch on just three areas of... uh... where I hear a lot people saying they want to change the national agenda, and then just reflect a little bit on how our experience at trying to change the national agenda might help.

One of the areas is national security, defence, and foreign policy. Those of you who have heard a speech that Mike Harris has given – I've heard him once in Toronto and once in Calgary, I think he's giving it in several different places in the country – touches on this fact that it is self-evident that Canada's influence in the area of defence and international relations has declined. The starving of our military on the funding side for years and for years has got, of course, to the point where we can't even make, let alone keep, international commitments. The stumbling and bumbling by some of our leadership on the international stage, I think, and others have said it, has become an embarrassment. Everything from the hesitant reaction on Canada's part in the hours right after September 11th, to the recent "moronic" comments by the Prime Minister's secretary – Canada's influence on... ability to influence, particularly the United States, on foreign policy or defence measures has declined from the day when Lester Pearson was the one that presented the solution to a Middle East crisis and was listened to by statesmen around the world. So I think one of the things that has got to be done on the national agenda, we have a need to restore our commitment and capability to defence, and our competence on the international stage.

On the area of health care reform, it is self-evident, now to everybody, that health care is in trouble. This study that the Fraser Institute published in *Fraser Forum* a month or so ago that shows Canada is the number one spender on a *per capita* basis on health care among the 23 OACD countries, but not number one on anything else. Not number one on outcomes, not number one in any one of the health care categories. There's something wrong with that when you're spending more money on a *per capita* basis than anyone else, but you're not number one when it comes to levels of care in dozens – 15 – different categories. Now I believe it is unlikely that Mr. Romanow will recommend the reforms that are necessary. We'll have to wait and see, we'll see this later this week. But it seems to be that there's going to be a rude... uh... need there for crusading for health care reform, to raise reforms that will actually improve the system, on the national agenda. Two of those that the Fraser Institute has, and others, have pushed for years – one is to get more flexibility in the system, to open up the *Canada Health Act* and let the provinces at least experiment with different approaches. If they make mistakes, let their people discipline them. They'll throw them out of office if they make a mess of medicare. But it's far better to have them disciplined by their electors than by the distant Health Care Minister. If there's more money required in the health care system, you're only going to get it from one of two sources: you're either going to get it out of the tax payers' pockets; or you're going to allow more private capital, private resources into the system. Those are two reforms that at least some of

the provinces should be able to experiment with, and I think a lot of people would like to push higher on the national agenda. Higher than they will be even after the Romanow report.

And then thirdly, a subject that Fraser has had a lot to say about is the whole Kyoto accord. I agree with the – I was in Parliament when the Kyoto was... the first commitment was made to Kyoto, and when the first “take note” debate was held in the Canadian Parliament in 1997, on Kyoto. It was the worst debate that I was in, in over nine years. Towards the end of it, I proposed that the pages in the House take the place of the members, and we would probably raise the level of the debate. Mainly because there was no effort made by the federal government at all to explain the science behind it. I mean, this is a science-based issue whether you like it or not. There was no effort made at all, in the beginning, to explain the science to the House. Some attempt at the committee level, but even that was... And no attempt to explain what the cost of Kyoto might be on the economic side. It has taken almost five years for the liberal members of Ontario to figure out that it's the people that burn the hydrocarbons that have got a lot of the problem with greenhouse gases, that Ontario is the biggest hydrocarbon-consuming province, that therefore it's Ontario that will pay both the biggest portion of the environmental cost and the economic cost in any agreement like Kyoto. That is finally starting to sink in, maybe in the last three months, but for five years it didn't. I think I agree – and I think that many people in this room would agree – that there is a need for a complete rethinking of our approach to how we handle the products of burning hydrocarbons. Nobody is denying that aren't any problems there. But how we handle them, and how we implement a program that is both environmentally responsible and doesn't destroy your economy at the same time is something that somebody is going to have to raise higher on the national agenda than it is.

So suppose one wanted then to change the... Canada's approach to defence, and our status internationally, suppose one wanted to change our approach to national health care, suppose one wanted to come up with a better approach to dealing with the products of burning hydrocarbons than what is proposed in Kyoto, how might one go about that?

Complete Test Instrument with Option Weighting

INTERPRETING COMPREHENSION TEST

This test is designed to measure your understanding of Preston Manning's speech, which you have just interpreted. It is made up of 45 multiple-choice questions. **However, the questions are of a variety that you have likely not seen before.** All the options to each question are correct, but they vary in terms of their "correctness".

To complete the test, circle the answer that is most complete, that is most appropriate for the context of the speech, or that best explains the meaning of a passage. Take as much time at you need to complete all 45 questions.

1. In stating that "*we* have the most archaic system", Manning is making reference to
- himself and some other, unidentified people;
 - the citizens of Canada;
 - himself and his audience; or
 - the members of the Canadian Alliance.

A = 1	B = 4	C = 2	D = 3
Manning + some	Canadians	Manning + many	CAP* members

*The Canadian Alliance Party

2. If you mount a "crusade", it means that you
- go on an important campaign;
 - mount a passionate campaign for an idea or cause;
 - campaign vehemently for a political cause; or
 - launch a campaign with religious fervor.

A = 1	B = 3	C = 4	D = 2
Campaign	Idea	Political nature	Fervor

3. In saying "what I thought I might touch on briefly today – and *we're* going to have a bit of a question period as well...", Manning is referring to
- everyone present at the speech;
 - himself and members of the Canadian Alliance Party;
 - himself alone; or
 - himself and his supporters.

A = 4	B = 3	C = 1	D = 2
all	Manning + CAP	Manning	Manning + some

4. When Manning says "...on the democratic front. *Just to illustrate the democracy deficit* – as it's now referred to in Ottawa – my last vote in the House of Commons was in December of last year. As you know as Christmas comes close...", he is
- keeping his audience up to date on political jargon;
 - interrupting his main line of reasoning and introducing a tangent;
 - peppering his speech with political terminology; or
 - allowing listeners to see that the tangent supports his main point.

A = 2	B = 3	C = 1	D = 4
Keep listeners current (understanding)	Introduce tangent (intent)	Use political jargon (use)	Illustrate main point (interpretation)

5. The national "agenda" is
- a set of matters discussed on the national political stage;
 - a set of issues repeatedly debated in federal politics;
 - a set of questions that is important for discussion; or
 - a set of topics that garners a great deal of political attention.

A = 3	B = 4	C = 1	D = 2
National scope	Recurrence	Issues	Political nature

6. In announcing "...then just reflect a little bit on how *our* experience at trying to change the agenda might help" Manning is referring to
- himself and his supporters;
 - himself and people with similar goals;
 - the Canadian Alliance Party; or
 - himself alone.

A = 2	B = 3	C = 4	D = 1
Manning + some	Manning + many	CAP	Manning

7. When Manning says "and we ...uh... used *the tools* to try and change the agenda that we used, are *the tools* that are available to everybody" he is
- repeating himself for stylistic effect;
 - drawing the audience's attention to a particular word in his speech;
 - letting people know it is important to listen to an upcoming list; or
 - signaling that he is about to make an important enumeration.

A = 2	B = 1	C = 4	D = 3
Draw attention to a word (understanding)	Repeat a word (use)	Enumerate to help audience achieve similar goals (interpretation)	Introduce enumeration (intent)

8. When you compare spending on a "*per capita* basis", it means that you
- consider the amount spent for each person living in a jurisdiction;
 - are describing per-person spending in proportional terms;
 - consider per person spending as a basis for comparing two jurisdictions; or
 - set aside a sum of money for each individual under consideration.

A = 2	B = 3	C = 4	D = 1
Jurisdiction	Proportion	Comparison	Division of money

9. When Manning mentions that “the *opposition* was starting to vote” he means

- a) the representatives of the Canadian Alliance party;
- b) members who were not part of the government;
- c) the people who were against a particular position; or
- d) those individuals who were not receptive to a certain ideal.

A = 3	B = 4	C = 2	D = 1
Official opposition	All opposition parties	Specific position	General definition

10. When he says that “If there’s more money required in the health care system, you’re only going to get it from one of two sources: you’re either going to get it out of the taxpayers’ pockets; or you’re going to allow more private capital, private resources into the system,” Manning is

- a) steering the audience to the conclusion that privatization is better;
- b) hinting at his preferred means for resolving the health care crisis;
- c) indicating that there are two possible solutions to the problem; or
- d) inviting the audience to decide which solution is better.

A = 4	B = 3	C = 1	D = 2
Direct to private health care (interpretation)	Suggest his choice (intent)	Indicate two possibilities (use)	Invite audience to choose (understanding)

11. In saying, “...let the provinces at least experiment with different approaches. If *they* make mistakes, let their people discipline *them*. They’ll throw *them* out of office if *they* make a mess of medicare...” Manning is referring to

- a) the provincial governments;
- b) authorities in the provinces;
- c) elected provincial officials; or
- d) residents of the provinces.

A = 3	B = 2	C = 4	D = 1
Provincial government	Provincial officials	Elected officials	Provincial population

12. In discussing “the day when Lester Pearson was the one who presented the solution to a *Middle East crisis*”, Manning is referring to

- a) problems taking place in the Holy Land;
- b) the conflict between Israelis and Palestinians;
- c) the creation of the State of Israel; or
- d) the region on the eastern shore of the Mediterranean.

A = 2	B = 4	C = 3	D = 1
Specific region	Israel + Palestine	Israel	General region

13. When he says, "I was there when the first commitment was made to Kyoto,"

Manning is talking about

- a) the accord signed in Kyoto;
- b) the countries represented at a summit in Kyoto;
- c) a city in Japan; or
- d) the people present at a summit in Kyoto.

A = 4	B = 3	C = 1	D = 2
Accord at summit	Countries at summit	City	Summit

14. In making the following statement, "We had, I think – you'll have to be the judge,

I'm not objective on this subject – but I think we had some success," Manning is

- a) pointing the audience towards a conclusion about goals he stated earlier;
- b) inviting the listeners to make an evaluation of the Canadian Alliance party;
- c) stating upfront his own bias towards his political party; or
- d) suggesting his own conclusions about the party's effectiveness.

A = 4	B = 2	C = 1	D = 3
Help audience conclude (interpretation)	Invite audience evaluation (understanding)	State bias (use)	Introduce own evaluation (intent)

15. When he explains "...a gopher motion... actually stimulated more interest and attention than a \$6 billion supply bill. So there's still a lot of work to do – is what I'm telling you – on the democracy front ..." Manning is

- a) interrupting his own statement;
- b) trying to clarify the point that he is making for his listeners;
- c) indicating that he had a reason for telling the gopher anecdote; or
- d) directing his listeners to his reason for telling the gopher anecdote.

A = 1	B = 2	C = 3	D = 4
Interrupt own statement (use)	Help listeners understand (understanding)	Indicate reason for anecdote (intent)	Direct listeners to reason for anecdote (interpretation)

16. In remarking that "*You might find it interesting* that after one of my lectures to one of the students – err – this political science class ..." Manning is in fact

- a) signaling that he will drift away from his point to tell an anecdote;
- b) trying to capture his audience's interest;
- c) suggesting that it is worthwhile for the audience to pay attention; or
- d) beginning to tell an entertaining story.

A = 3	B = 2	C = 4	D = 1
Signal drift away from point (intent)	Capture audience interest (understanding)	Get audience to pay attention (interpretation)	Start a story (use)

17. By mentioning “the stumbling and bumbling of some of our *leadership*” Manning means

- a) errors committed by the leader of a political party;
- b) mistakes made by the government;
- c) gaffes made by the Prime Minister and his cabinet; or
- d) slip-ups made by public figures.

A = 3	B = 2	C = 4	D = 1
A party	The government	The PM and cabinet	Leaders generally

18. In Canada, “Parliament” is

- a) located in the city of Ottawa;
- b) synonymous with the House of Commons in Ottawa;
- c) the two federal legislative chambers located in the city of Ottawa; or
- d) the House of Commons and the Senate on Parliament Hill in Ottawa.

A = 1	B = 2	C = 3	D = 4
Location	House or Senate	House and Senate	The hill

19. When Manning says “I’m sure there’s people here today who would like to change the agenda... then just reflect a little bit on how our experience at trying to change the agenda might help,” he is

- a) implying that the Canadian Alliance’s story is illustrative;
- b) telling his listeners that he knows something about changing the national agenda;
- c) remarking on the number of people who are unhappy with the national agenda; or
- d) aiding people to conclude that his story might be instructive for their own efforts.

A = 3	B = 2	C = 1	D = 4
Imply CAP story is illustrative (intent)	Show he is knowledgeable (understanding)	Indicate people are unhappy (use)	Demonstrate people can be effective (interpretation)

20. When he suggests that “some of our input became part of the *government’s* position,” Manning is referring to

- a) the present liberal administration;
- b) the party currently in power;
- c) a body that governs a state or community; or
- d) the elected officials on Parliament Hill.

A = 4	B = 3	C = 1	D = 2
Government, elected officials, current party + liberals	Government, elected officials + current party	Government body	Government + elected officials

21. To “resign your seat” means that you
- give up a government responsibility;
 - leave the government;
 - step down from a position in political office; or
 - make a decision to quit.

A = 3	B = 2	C = 4	D = 1
Withdrawal from government and responsibility	Withdrawal from government	Withdrawal from government, responsibility, and position	Withdrawal

22. In telling his audience “There was a motion, a *private member’s motion*, requesting that the federal government disclose *why* it had reduced the *strychnine* level in *gopher* poison from *five* percent to *three* percent. Some prairie member that had a mad-on with respect to gophers wanted this information,” Manning is
- making use of staccato intonation;
 - underlining the precise and meticulous nature of the member’s request;
 - relaying some information about the character of one Member of Parliament; or
 - emphasizing some words more than others.

A = 1	B = 4	C = 3	D = 2
Use intonation (use)	Help listener understand precise information (interpretation)	Illustrate someone’s character (intent)	Emphasize part of the message (understanding)

23. In talking about “...the hesitant reaction on *Canada’s* part in the hours right after September 11th...” Manning is referring to
- the members of the House of Commons;
 - the government;
 - the liberal administration; or
 - the country.

A = 3	B = 2	C = 4	D = 1
Members of Parliament	The government	The government members	The country

24. In telling his audience “I always try to invite feedback from the students. Someone came up to me and said that ‘You know, whenever you talk about federal politics here, you should spend the first five minutes telling us why we should care.’ *This was a political science student*,” Manning is
- helping listeners understand where the student was coming from;
 - suggesting that listeners should find this worrisome;
 - placing importance on the student’s field of study; or
 - providing some basic information about the student who was speaking.

A = 2	B = 4	C = 3	D = 1
Understand student’s position (understanding)	Use information to draw conclusion (interpretation)	Focus on important information (intent)	Provide information (use)

25. When Manning argues that "...we have a need to restore *our* commitment and capability to defence, and our competence on the international stage," he means

- a) all members of the Canadian Alliance party;
- b) everyone concerned with politics in this country;
- c) everyone present at his speech; or
- d) Canada as a whole.

A = 2	B = 3	C = 1	D = 4
CAP members	All those interested in politics	His audience	The whole country

26. A statement like "*I think* I agree – and *I think* many people in this room would agree – that there is a need for a complete rethinking of our approach to how we handle the products of burning hydrocarbons," shows that Manning is

- a) allowing the audience to understand his position on global warming;
- b) drawing attention to his belief;
- c) helping the audience see the relative importance of his belief; or
- d) providing a basic outline of his thoughts on the topic of global warming.

A = 2	B = 3	C = 4	D = 1
Inform audience of his position (understanding)	Draw attention to his belief (intent)	Allow audience to determine relative importance (interpretation)	Outline his thoughts (use)

27. Achieving "debt reduction" means that you

- a) use budget surpluses to pay back money you owe;
- b) use budget surpluses to repay government debt;
- c) resolve to pay back money outstanding; or
- d) pay the money you owe with any surplus you have.

A = 3	B = 4	C = 1	D = 2
Budget	Government debt	Debt payment	Surplus

28. When he states "but that's another subject, and *we'll* get into that another day," Manning means

- a) everyone present during the speech;
- b) Manning himself;
- c) members of the Canadian Alliance; or
- d) Manning and his supporters.

A = 1	B = 4	C = 2	D = 3
His audience	Manning alone	CAP members	Manning + some

29. To describe a system as "archaic" means that it

- a) has been in existence for a long time;
- b) is old, outdated, and in need of reform;
- c) is both ancient and limited in its use; or
- d) is not particularly up to date.

A = 2	B = 4	C = 3	D = 1
Old	In need of reform	Of limited use	Not current

30. In the following quote, "...you've got to work at it to make us'... err... So I have been working at it, and hopefully I might be successful," Manning is
- making a transition from reported speech back to his own words;
 - allowing audience to understand the different "voices" in his speech;
 - allowing his listeners to conclude he will make his speech relevant for them;
- or
- suggesting a similarity between the university audience and the one at hand.

A = 1	B = 2	C = 4	D = 3
Transition from reported speech (use)	Allow audience to know who's speaking (understanding)	Infer he will make speech relevant (interpretation)	Point out similarity of the two (intent)

31. If you are not pursuing a policy as "an end in itself", it means that you
- are using the policy to attain another goal;
 - have an ultimate goal in mind;
 - have very complex motivations; or
 - are not merely interested in your immediate goal.

A = 4	B = 3	C = 1	D = 2
Policy	Ultimate goal	Complexity	Immediate goal

32. In making the following argument, "But it's far better to have them disciplined by the electors than by *the distant health care minister*," Manning is referring to
- the elected official in charge of a health care system;
 - the federal minister of health;
 - a government official; or
 - the head of a health care system in a jurisdiction.

A = 3	B = 4	C = 1	D = 2
Elected head	Federal minister	Government official	Head of ministry

33. When Manning says, "...a lot of work left to do on the democracy front... Now I... uh.. *what I thought I might touch on briefly today...*" he is
- allowing the audience follow the structure of his talk;
 - giving cues about the structure of his speech;
 - providing an introduction for his remarks; or
 - signaling his listeners that he is beginning his remarks.

A = 4	B = 3	C = 1	D = 2
Make audience understand structure (interpretation)	Convey speech structure (intent)	Introduce remarks (use)	Signal beginning of remarks (understanding)

34. Pursuing a policy of “deficit reduction” means that you

- a) attempt to reduce your expenditures;
- b) decide to reduce expenditures in comparison to income;
- c) try to achieve a balanced budget; or
- d) reduce spending relative to income in a government budget.

A = 1	B = 2	C = 3	D = 4
Spending alone	Spending and income	Budget	Government budget

35. In asking “What is wrong with *this* place?” Manning is referring to

- a) the House of Commons;
- b) Parliament Hill;
- c) Ottawa; or
- d) the federal government.

A = 4	B = 3	C = 1	D = 2
The House of Commons	The whole legislature	The city	The whole government

36. When Manning states, “I believe it is unlikely that *Mr. Romanow* will recommend the reforms that are necessary. We’ll have to wait and see,” he is referring to

- a) the head of a Royal Commission on health care;
- b) a well-known figure in national politics;
- c) the former premier of Saskatchewan; or
- d) a long-time member of the New Democratic Party.

A = 4	B = 1	C = 2	D = 3
Present role	Celebrity	Former Premier	NDP member

37. “Hydrocarbons” are

- a) chemical substances that are often in the news;
- b) the waste products that are responsible for global warming;
- c) fossil-fuel pollutants Canada has committed to reducing; or
- d) chemicals that are the byproducts of fossil fuel consumption.

A = 1	B = 3	C = 4	D = 2
Chemical substance	Effect	Reduction	Source

38. The statement “It’s *basically* the story of a small group of people who, 15 years ago, wanted to change the national agenda”

- a) involves the audience by telling a story;
- b) introduces a narrative element to the speech;
- c) shows how the speaker is introducing focus; or
- d) signals the audience to pay attention to what follows.

A = 2	B = 1	C = 3	D = 4
Involve the audience (understanding)	Tell a narrative (use)	Introduce focus (intent)	Signal audience to focus (interpretation)

39. In Canada, the term “secession” refers to
- the ending of a political alliance;
 - the formal termination of an agreement;
 - a formal withdrawal from a federal union; or
 - Quebec’s departure from Canada.

A = 2	B = 1	C = 3	D = 4
Political alliance	General agreement	National federation	Canadian federation

40. A “page” is
- a young person who acts as a servant;
 - a teenager delivering messages to important people;
 - a young messenger to members of Parliament; or
 - a youthful attendant to someone of high rank.

A = 1	B = 3	C = 4	D = 2
Youth	Messenger	Messenger in Parliament	Servant

41. “One of the areas is national security, defence, and foreign policy... On the area of health care reform, it is self evident now to everyone... And then *thirdly*, a subject that Fraser has had a lot to say about is the whole Kyoto accord,” is Manning’s way of

- making a series of ordered points;
- putting emphasis on certain phrases to make things interesting for his listeners;
- using intonation to stress certain parts of his argument; or
- assisting his audience to better follow the structure of his speech.

A = 3	B = 2	C = 1	D = 4
Show structure in speech (intent)	Emphasize for audience’s sake (understanding)	Emphasize some aspects of message (use)	Lead audience to see structure in speech (interpretation)

42. Manning tells his audience, “what I thought I might touch on *briefly* today, and we’re going to have a *bit* of a question period as well... I’d like to touch on *just* three areas,” in order to

- indicate that it is worth it for people to listen;
- ensure that people understand the point he is making;
- repeat a point so that the people present will understand it; or
- repeatedly suggest that his talk will be short in duration.

A = 4	B = 2	C = 1	D = 3
Indicate importance of point (interpretation)	Ensure audience understands point (understanding)	Repeat a point (use)	Communicate purpose behind repetition (intent)

43. In Canada, if you “democratize” an institution like Parliament, you

- a) return some power over the institution to the people;
- b) alter the way power over the institution is distributed;
- c) make the institution more accountable to the people; or
- d) ensure that the institution is accountable to the people.

A = 2	B = 1	C = 4	D = 3
Return power	Alter power distribution	Make more accountable	Make accountable

44. If young people are “tuned out” with regard to federal politics, it means that they are

- a) feeling uninvolved, and therefore apathetic and disinterested;
- b) unconcerned by what politicians have to say, and therefore not listening ;
- c) feeling like they do not matter, and therefore not paying attention; or
- d) not paying attention to what politicians are saying to them.

A = 4	B = 2	C = 3	D = 1
Lack of involvement	Lack of concern	Affect	Inattention

45. Obtaining “regional balance” involves

- a) equipping government with the means to address regional concerns;
- b) finding a mechanism for resolving regional issues; or
- c) recognizing that there is regional diversity in a political jurisdiction; or
- d) understanding the issues important to people in different political regions.

A = 4	B = 3	C = 1	D = 2
Government	Mechanisms	Regionalism	Issues

New Test Raw Scores with Indices

Item	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	Discrim	Difficulty
1b	4	4	4	4	4	4	4	4	4	4	4	1	4	4	4	0.2	93
2a	2	3	4	2	3	3	3	3	4	3	4	4	3	3	3	-0.4	27
3b	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	0	100
4c	1	4	4	4	4	4	4	4	4	2	4	2	4	2	3	0.4	67
5a	2	3	3	1	3	3	3	3	1	4	2	2	2	3	3	0	7
6b	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	0	93
7c	3	3	3	3	1	1	1	2	3	2	4	1	1	1	2	-0.2	7
8a	4	2	4	2	4	2	4	4	4	2	4	4	2	2	3	0.2	53
9b	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	0	100
10c	4	1	3	3	1	3	4	3	4	1	3	3	1	1	1	0.2	20
11b	3	3	3	3	3	3	4	3	3	3	3	4	3	3	3	-0.2	13
12b	2	4	3	4	1	4	4	4	2	4	4	1	4	4	1	-0.2	60
13b	4	4	4	4	4	4	4	4	4	3	4	4	4	4	4	0	93
14c	3	3	1	2	2	4	3	2	1	3	3	3	3	1	2	0	7
15c	4	4	4	2	4	2	4	4	2	4	2	4	4	4	2	0.2	67
16c	2	4	4	1	3	2	2	2	1	3	4	1	1	4	3	0	27
17b	4	4	4	4	4	4	3	4	4	1	2	4	2	1	4	0.6	67
18a	4	4	4	3	4	3	1	4	4	4	4	4	2	4	4	0	73
19c	4	4	4	4	4	4	4	4	4	4	2	4	1	4	3	0.6	80
20b	4	4	4	4	4	4	3	3	4	3	4	4	4	4	4	0	80
21a	4	4	4	4	4	4	4	4	4	1	4	4	4	3	4	0.2	87
22c	4	4	3	4	4	4	2	4	4	4	4	4	4	4	4	-0.2	87
23b	4	4	4	2	4	2	2	2	2	4	2	4	2	1	2	0.6	40
24c	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	0	100
25b	4	4	3	4	4	4	4	4	4	3	4	4	4	4	4	-0.2	87
26c	4	4	2	2	2	1	4	3	4	2	2	1	2	3	1	0.4	27
27a	4	1	4	4	1	4	4	1	4	4	4	4	4	4	3	-0.2	73
28b	4	4	3	4	4	4	4	3	1	4	4	4	4	2	1	0.2	67
29a	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	0	100
30c	3	4	2	4	1	3	2	1	4	4	1	1	1	1	1	0.2	27
31a	4	2	3	2	4	4	3	4	1	2	4	2	4	3	4	-0.2	47
32b	4	4	4	4	2	4	4	4	4	4	4	4	4	3	2	0.2	80
33c	2	1	1	4	4	3	1	1	2	1	1	4	2	3	1	0.2	20
34a	4	3	4	4	4	3	4	4	4	4	4	4	3	3	4	0.2	73
35b	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	-0.2	93
36b	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	0	100
37a	4	4	2	4	2	2	4	4	2	4	2	2	3	3	4	0.4	47
38c	3	4	4	3	2	2	1	1	3	1	1	2	1	1	1	0.4	13
39a	4	4	4	4	4	4	4	4	4	4	3	4	4	1	3	0.6	80
40a	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	0	100
41c	3	3	3	3	4	4	4	3	4	4	1	3	3	4	3	0	40
42c	3	3	3	2	3	2	3	3	3	3	1	3	3	3	3	0	0
43a	4	4	3	4	4	3	1	2	2	2	4	2	2	4	1	0.4	40
44a	4	4	1	3	4	4	3	4	4	4	4	4	4	4	2	-0.2	73
45a	4	2	3	4	4	2	2	2	2	4	1	2	4	1	1	0.4	33
Total	159	157	152	151	150	148	147	147	147	145	144	144	139	136	130		

Distractor Analysis for New Test Raw Scores

Item	High Scoring Participants				Low Scoring Participants			
	Option A	Option B	Option C	Option D	Option A	Option B	Option C	Option D
1b	0.00	*1.00	0.00	0.00	0.20	*0.80	0.00	0.00
2a	0.00	0.40	*0.20	0.40	0.00	0.60	*0.40	0.00
3b	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00	0.00
4c	0.00	0.00	0.20	*0.80	0.40	0.20	0.00	*0.40
5a	0.60	*0.00	0.20	0.20	0.40	*0.00	0.00	0.60
6b	0.00	0.00	*1.00	0.00	0.00	0.00	*1.00	0.00
7c	0.00	0.20	*0.00	0.80	0.20	0.60	*0.20	0.00
8a	0.40	0.00	*0.60	0.00	0.40	0.20	*0.40	0.00
9b	0.00	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00
10c	*0.20	0.40	0.40	0.00	*0.00	0.40	0.60	0.00
11b	1.00	0.00	*0.00	0.00	0.80	0.00	*0.20	0.00
12b	0.20	*0.40	0.20	0.20	0.00	*0.60	0.00	0.40
13b	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00	0.00
14c	*0.00	0.40	0.20	0.40	*0.00	0.20	0.20	0.60
15c	0.00	0.20	0.00	*0.80	0.00	0.40	0.00	*0.60
16c	0.20	0.20	*0.40	0.20	0.20	0.00	*0.40	0.40
17b	0.00	0.00	*1.00	0.00	0.00	0.40	*0.40	0.20
18a	0.00	0.00	0.20	*0.80	0.00	0.20	0.00	*0.80
19c	0.00	0.00	0.00	*1.00	0.20	0.20	0.20	*0.40
20b	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00	0.00
21a	0.00	0.00	*1.00	0.00	0.20	0.00	*0.80	0.00
22c	0.00	*0.80	0.20	0.00	0.00	*1.00	0.00	0.00
23b	0.00	0.20	*0.80	0.00	0.00	0.60	*0.20	0.20
24c	0.00	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00
25b	0.00	0.20	0.00	*0.80	0.00	0.00	0.00	*1.00
26c	0.60	0.00	*0.40	0.00	0.40	0.20	*0.00	0.40
27a	0.00	*0.60	0.40	0.00	0.20	*0.80	0.00	0.00
28b	0.00	*0.80	0.00	0.20	0.20	*0.60	0.20	0.00
29a	0.00	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00
30c	0.20	0.20	*0.40	0.20	1.00	0.00	*0.00	0.00
31a	*0.40	0.20	0.00	0.40	*0.60	0.20	0.00	0.20
32b	0.00	*0.80	0.00	0.20	0.20	*0.20	0.40	0.20
33c	*0.40	0.00	0.40	0.20	*0.20	0.20	0.40	0.20
34a	0.00	0.00	0.20	*0.80	0.00	0.00	0.40	*0.60
35b	*0.80	0.20	0.00	0.00	*1.00	0.00	0.00	0.00
36b	*1.00	0.00	0.00	0.00	*1.00	0.00	0.00	0.00
37a	0.00	0.00	*0.60	0.40	0.00	0.40	*0.20	0.40
38c	0.20	0.00	0.40	*0.40	0.20	0.80	0.00	*0.00
39a	0.00	0.00	0.00	*1.00	0.00	0.20	0.40	*0.40
40a	0.00	0.00	*1.00	0.00	0.00	0.00	*1.00	0.00
41c	0.80	0.00	0.00	*0.20	0.60	0.00	0.20	*0.20
42c	*0.00	0.20	0.00	0.80	*0.00	0.00	0.20	0.80
43a	0.00	0.00	*0.80	0.20	0.40	0.20	*0.40	0.00
44a	*0.60	0.00	0.20	0.20	*0.80	0.20	0.00	0.00
45a	*0.60	0.20	0.00	0.20	*0.20	0.00	0.60	0.20

* Indicates the answer option that was awarded the most points