

Exploring Perceptions of Second Language Speech Fluency through Developing and Piloting a
Rating Scale for a Paired Conversational Task

by

Kent Williams

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

In

Applied Linguistics and Discourse Studies

Carleton University

Ottawa, Ontario

© 2020

Kent Williams

Abstract

Much research has explored how perceptions of speech fluency are influenced by a variety of temporal speech features (e.g. speech rate). However, less is known about the influence of non-temporal and conversational speech characteristics, as well as listener characteristics such as accent familiarity and conversational style, on fluency perceptions. To address this gap, the present study explored the influence of these characteristics through developing and piloting a fluency rating scale for a paired conversational task for assessment for learning purposes. A two-phase mixed-methods sequential exploratory design (Creswell, 2009) provided the methodological framework for the present study. In the first phase, seven trained English for Academic Purposes (EAP) instructors watched videos of seven-minute conversations, elicited from 14 intermediate-to-advanced EAP learners, who performed the conversational task twice with two different partners. Afterwards, instructors were audio-recorded discussing their observations about learners' fluency. These recordings were coded using in-vivo and pattern coding techniques (Saldaña, 2009). Six themes were identified: smoothness, efficiency, sophistication, clarity, facilitating topics and turns, and supporting the conversation partner. These themes informed the development of a multi-item fluency rating scale, used in the second phase of the study. In this phase, a new group of 35 EAP instructors watched four seven-minute video-recorded conversations between eight learners, and then used the scale to rate the performances. Before watching each video, instructors reported their familiarity with students' accents on a six-point scale. Once rating was completed, instructors completed a conversational style questionnaire. The results were as follows. First, a Principal Component Analysis of scale items produced two separate components - individual fluency and conversational fluency. Second, temporal measures of within-clause pause rate correlated significantly with items

representing individual fluency whereas measures of filled-pause rate correlated significantly with items representing conversational fluency. Third, non-parametric analyses of group differences showed that accent familiarity moderately affected fluency assessments. Finally, correlational analyses revealed significant correlations between fluency assessments and questionnaire items representing certain conversational style characteristics of the listeners. The overall findings showed that speech and listener characteristics affected fluency perceptions to varying degrees, providing implications for the assessment of 'higher-order fluency' (Lennon, 2000).

Acknowledgements

I would like to begin by sending my warm regards to everyone who has helped me throughout this long process. First, I would like to thank my supervisor, Dr. David Wood, for being the invisible hand that guided this process and especially for providing me with much needed support and advice at the end when there was little gas left in my tank. I would also like to thank my committee members, Dr. Janna Fox and Dr. Ron Thomson, for their good humour and their insightful and challenging questions and comments. Also, thank you kindly to the internal and external examiners, Dr. Randall Gess and Dr. Nikolay Slavkov, for their crucial contributions, which challenged me to see the project in ways I would not have imagined otherwise.

I must also thank my colleagues at Carleton University - Shahin Nematizadeh, Alisa Zavalova, and Ana Lúcia Tavares Monteiro - for their continual and unwavering support for me at various stages throughout this process. I could always count on you. Likewise, I would like to express my sincere gratitude to my colleagues at Renison University College at the University of Waterloo, most especially to Julia Williams and Stefan Rehm, for all they have done for me over the years. I also thank Bahattin Altay and Sara Kafashan for their kindness, wisdom, and encouragement at crucially important times in the process. I would also like to thank Dr. Hedy McGarrell, Dr. John Sivell, and Dr. David Hayes at Brock University for inspiring and encouraging me to pursue studies at the doctoral level in the first place. Finally, I would like to thank my friends, my family, and my music for helping me battle my way out of the trenches. I am sincerely grateful.

Table of Contents

Abstract	ii
Acknowledgments	iv
Table of Contents	v
List of Tables	xiv
List of Figures	xvi
List of Abbreviations.....	xvii
List of Appendices.....	xxi
Chapter One: Introduction	1
1.1. Motivation for the Present Study.....	2
1.2. Background.....	3
1.2.1. Core Features.....	3
1.2.2. Peripheral Features.....	4
1.2.3. Conversational Features.....	4
1.2.4. Listener Characteristics.....	5
1.2.4.1. Accent Familiarity.....	5
1.2.4.2. Conversational Style.....	6
1.2.5. Rating Scale Design.....	8
1.3. Research Questions.....	9
1.4. Overview of the Dissertation.....	10
Chapter Two: Literature Review	12
2.1. Overview	12
2.1.1. Defining Fluency	12

2.1.2. Measuring Fluency	14
2.1.3. Outline of the Literature Review	15
2.2. Fluency Development	16
2.2.1. Cognitive Processes.....	16
2.2.1.1. Automaticity	16
2.2.1.2. Attention	19
2.2.1.3. Segalowitz’s (2010) Framework for Fluency	22
2.2.1.4. Willingness to Communicate (WTC)	23
2.2.1.5. First Language (L1) Transfer	24
2.2.2. Formulaic Language (FL)	26
2.2.2.1. Pawley and Syder’s (2000) Two Hypotheses	27
2.2.2.2. FL and Proceduralization	28
2.2.2.3. Functions of FL in Fluent Speech	29
2.2.2.4. FL and Pausing	31
2.2.2.5. FL and Articulation Rate	32
2.2.3. Fluency Pedagogy	33
2.2.3.1. Fluency-focused activities	33
2.2.3.2. Complexity, Accuracy, and Fluency	42
2.2.3.3. Summary	44
2.3. Fluency Perceptions	44
2.3.1. Temporal Features	44
2.3.1.2. Speed Fluency	46
2.3.1.3. Breakdown Fluency	46

2.3.1.4. Repair Fluency	49
2.3.1.5. Composite Measures	50
2.3.1.6. Cross-Sectional Analyses by Proficiency Level	51
2.3.1.7. Self-Perceptions and Task Type	52
2.3.2. Non-Temporal Features	53
2.3.2.1. Analysis of Raters' Observations	53
2.3.2.2. Perceptions of FL and Fluency	58
2.3.2.3. Intonation	61
2.3.2.4. Linking	63
2.3.2.5. Summary	64
2.4. Accent Familiarity	64
2.4.1. Effect of Exposure	65
2.4.2. Exposure and L2 Speech Ratings	66
2.4.3. Teaching Experience/Training and L2 Speech Ratings	67
2.4.4. The Interlanguage Intelligibility Benefit	70
2.4.5. Listener Expectations	70
2.4.6. Exposure Training Programs	71
2.4.7. Intergroup Attitudes	72
2.5. Conversational Fluency	75
2.5.1. Interactional Competence	76
2.5.2. Non-Verbal Communication (NVC)	80
2.5.3. NVC and Fluency	81
2.5.4. Measuring Conversational Fluency	85

2.6. Conversational Style	92
2.6.1. Stereotypes.....	93
2.6.2. L2 Development.....	94
2.6.3. Topic Shifting and Elaborating.....	95
2.6.4. Conversational Parity.....	96
2.6.5. Avoiding Conversational Silence.....	97
2.6.6. Politeness Strategies.....	99
2.6.7. Gender.....	100
2.6.8. WTC in a Conversational Setting.....	101
2.6.9. Listener Preferences.....	102
2.7. Fluency Rating Scale Design	103
2.7.1. Discourse Analysis.....	103
2.7.2. Scaling Descriptors	106
2.7.3. Corpus Analysis	107
2.7.4. Analysis of Raters' Perceptions	109
2.7.5. Assessment for Learning Purposes	110
2.8. Implications of the Literature Review	111
2.8.1. Purpose of the Present Study.....	115
2.8.2. Research Questions	116
Chapter Three: - Method (Phase One)	117
3.1. Overview	117
3.1.1. Research Design	117
3.1.2. Pilot Stages	119

3.1.3. Ethics and Research Site	120
3.2. Participants	120
3.2.1. Sampling Procedure	120
3.2.2. EAP Learners	120
3.2.3. EAP Instructors	124
3.2.4. Expert Raters	124
3.3. Instruments	124
3.3.1. Paired Conversational Task	124
3.3.2. CAEL/OLT Task Rubrics	129
3.3.3. Semi-Structured Interview Guide	130
3.4. Procedures	132
3.4.1. EAP Learners	132
3.4.2. Expert Raters	134
3.4.3. EAP Instructors	134
3.4.4. Analysis	135
Chapter 4 - Results and Discussion (Phase One)	136
4.1. Question One: Instructors' Fluency Perceptions	137
4.1.1. Overview	138
4.1.2. Smoothness	141
4.1.3. Efficiency	151
4.1.4. Sophistication	156
4.1.5. Clarity	161
4.1.6. Facilitating Topics and Turns	165

4.1.7. Supporting the Conversation Partner	172
4.1.8. Summary	178
4.2. Fluency Rating Scale	180
Chapter 5 - Method (Phase Two)	182
5.1. Overview	182
5.2. Ethics and Research Site	182
5.3. Participants	183
5.3.1. EAP Instructors	183
5.4. Instruments	184
5.4.1. Fluency Rating Scale and Descriptors	184
5.4.2. Task Specifications	184
5.4.3. Conversational Style Questionnaire	188
5.5. Procedures	194
5.5.1. Rater Training	194
5.5.2. Rating Procedures	196
5.6. Analysis	197
5.6.1. Data Cleaning	197
5.6.2. Inter-Rater Reliability and Internal Consistency	198
5.6.3. Assessing Data Suitability	199
5.6.4. Temporal Analysis	199
5.6.4.1. Sample Preparation	200
5.6.4.2. Clause Analysis	204
Chapter 6 - Results (Phase Two)	205

6.1. Overview	205
6.2. Question Two: Fluency Scale Items and Construct Relevance	206
6.2.1. Assessing Data Suitability	206
6.2.2. Factor Extraction	209
6.2.3. Factor Rotation and Interpretation	210
6.2.4. Summary	212
6.3. Question Three: Temporal Measures and Fluency Ratings	213
6.3.1. List of Temporal Measures	213
6.3.2. Assessing Data Suitability	214
6.3.3. Correlation Analysis	215
6.3.4. Summary	218
6.4. Question Four: Accent Familiarity and Fluency Ratings	219
6.4.1. Overview	219
6.4.2. Assessing Data Suitability	219
6.4.3. Kruskal-Wallis H Tests	220
6.4.4. Mann-Whitney U Tests	222
6.4.5. Summary	223
6.5. Question Five: Conversational Style and Fluency Ratings	224
6.5.1. Overview	224
6.5.2. Assessing Data Suitability	225
6.5.3. Correlation Analysis	225
6.5.4. The Efficacy of the Questionnaire Items	228
6.5.5. Summary	229

6.6. Overall Summary of Key Results	230
Chapter Seven - Discussion	231
7.1. Question Two	231
7.2. Question Three	234
7.3. Question Four	240
7.4. Question Five	243
Chapter Eight - Classroom Applications	250
8.1. Using the fluency rating scale for assessment for learning purposes.....	251
8.2. Self-study.....	252
8.3. Accent familiarity awareness	252
8.4. Conversational style awareness.....	253
Chapter Nine - Conclusion	257
9.1. Overview of the study.....	258
9.2. Overview of the findings (phase one).....	260
9.3. Overview of the findings (phase two).....	262
9.3.1 Individual fluency and conversational fluency.....	262
9.3.2. Listener characteristics.....	264
9.3.2.1. Accent familiarity.....	265
9.3.2.2. Conversational style.....	266
9.4. Attention to limitations.....	267
9.5. Future areas of research.....	269
9.6. Future applications.....	273
9.6.1. Methodological applications.....	273

9.6.2. Classroom applications.....274

9.6.3. Training programs.....274

References277

Appendix A: Carleton University REB Clearance 307

Appendix B: Renison University College at University of Waterloo REB Clearance308

List of Tables

Table 1.	List of Temporal Measures	45
Table 2.	Participants and Mean Ratings on the CAEL/OLT Task One	122
Table 3.	Data Collection Groups, Conversations, Participants, Levels of Acquaintanceship.....	134
Table 4.	Themes, Categories, and Number of Comments per Performance Group...	140
Table 5.	Smooth: Categories, Groups, Comments, and Descriptors	142
Table 6.	Efficient: Categories, Groups, Comments, and Descriptors	151
Table 7.	Sophisticated: Categories, Groups, Comments, and Descriptors	157
Table 8.	Clarity: Categories, Groups, Comments, and Descriptors	160
Table 9.	Facilitating Topics and Turns. Categories, Groups, Comments, and Descriptors	165
Table 10.	Supporting the Conversation Partner: Categories, Groups, Comments, and Descriptors	172
Table 11.	Tannen's (2005) Criteria for High-Involvement Style and Nine Items on the CSQ	190
Table 12.	Temporal Measures Investigated in the Present Study	201
Table 13.	Correlation Matrix. Inter-Item Correlations between Items on the Fluency Scale.....	208
Table 14.	PCA - Pattern and Structure Matrix with Oblimin Rotation of Fluency Scale Items	212
Table 15.	Temporal Measures Investigated in the Present Study	214
Table 16.	Evans' (1996) Classifications of Correlation Sizes	216
Table 17.	Pearson <i>r</i> Correlations between Temporal Measures and Raters' Assessments.....	217
Table 18.	Sample Descriptives Using Kruskal-Wallis H test (Mean Ratings vs Accent Familiarity).....	220

Table 19.	Kruskal-Wallis H: Mean Ratings x Accent Familiarity Groups	223
Table 20.	Mann-Whitney U: Fluency Ratings vs Accent Familiarity x Group (Gp 1 vs Gp 2).....	224
Table 21	Evans' (1996) Classifications	226
Table 22.	Raters' Fluency Ratings of HI-Style Students	227
Table 23.	Fluency Rating Scale and Suggested Pedagogical Activities	251
Table 24.	Descriptions of Suggested Activities	253

List of Figures

<i>Figure 1.</i>	Levelt's Model of Speech Production	18
<i>Figure 2.</i>	Exploratory Design: Instrument Development Model (Method Phase One)	120
<i>Figure 3.</i>	Paired conversational Task (Instruments)	130
<i>Figure 4.</i>	Semi-Structured Interview Guide (Instruments)	133
<i>Figure 5.</i>	Instructors' Listening Guide	134
<i>Figure 6.</i>	Fluency Rating Scale: Analytic Criterion Questions	180
<i>Figure 7.</i>	Fluency Rating Scale Descriptors	181
<i>Figure 8.</i>	Exploratory Design: Instrument Development Model	186
<i>Figure 9.</i>	Task Specifications	190
<i>Figure 10.</i>	Example Question from the Conversational Style Questionnaire.....	191
<i>Figure 11.</i>	Conversational Style Questionnaire	193
<i>Figure 12.</i>	Fluency Rating Scale Items	195
<i>Figure 13.</i>	Scree Plot from PCA of Fluency Rating Scale Items	211
<i>Figure 14.</i>	Instructors' Mean Assessments of <i>Smooth</i> and Learners' <i>WCpr</i>	212
<i>Figure 15.</i>	Mean Ranks of Fluency Ratings and Accent Familiarity by Group	222

List of Abbreviations

AccFam	Accent Familiarity
APL	Average Pause Length
AR	Articulation Rate
ASU	Analysis of Speech Unit
BCpl	Between-Clause Pause Length
BCpr	Between-Clause Pause Rate
CA	Consonant Attraction
CAEL	Canadian Academic English Language Assessment
CEFR	Common European Framework Reference of Language Benchmarks
Cle	Clear [Fluency Rating Scale Item]
CLT	Communicative Language Teaching
CPE	Cambridge Proficiency of English Test
ConvA	Conversational Style Questionnaire Item (A) [sharing personal information]
ConvB	Conversational Style Questionnaire Item (B) [change topics suddenly]
ConvC	Conversational Style Questionnaire Item (C) [interrupt]
ConvD	Conversational Style Questionnaire Item (D) [speak more quickly than your partner]
ConvE	Conversational Style Questionnaire Item (E) [overlap]
ConvF	Conversational Style Questionnaire Item (F) [speak loudly and with enthusiasm]
ConvG	Conversational Style Questionnaire Item (G) [avoid conversational silence]
ConvH	Conversational Style Questionnaire Item (H) [finish your partner's sentences]
ConvI	Conversational Style Questionnaire Item (I) [persist in getting across what you want to say]

ConvTot5	Conversational Style Questionnaire [Sum of 5 Items]
ConvTot9	Conversational Style Questionnaire [Sum of 9 Items]
CS	Conversational Style
CSQ	Conversational Style Questionnaire
EAP	English for Academic Purposes
ELTS	English Language Testing System
Eff	Efficient [Fluency Rating Scale Item]
FA	Factor Analysis
Fac	Facilitative [Fluency Rating Scale Item]
FCE	First Certificate of English
FL	Formulaic Language
FPR	Filled Pause Rate
FS	Formulaic Sequences
Gp1	Group 1
Gp2	Group 2
Gp3	Group 3
HC	High-Considerateness
HC-Style	High-Considerateness Style
HI	High-Involvement
HI-Style	High-Involvement Style
Hol	Holistic [Fluency Rating Scale Item]
IC	Interactional Competence
IELTS	International English Language Testing System

KMO	Kaiser-Meyer-Olkin Measure of Sampling Adequacy
L1	First Language
L2	Second Language
LPI	Language Proficiency Interview
MLR	Mean Length of Runs
NVC	Non-Verbal Communication
OLT	Oral Language Test
OPI	Oral Proficiency Interview
PCA	Principal Component Analysis
PSR	Pruned Speech Rate
RepetR	Repetition Rate
RepairR	Repair Rate
SLA	Second Language Acquisition
Smo	Smooth [Fluency Rating Scale Item]
Sop	Sophisticated [Fluency Rating Scale Item]
SPR	Silent Pause Rate
SR	Speech Rate
Sup	Supportive [Fluency Rating Scale Item]
TBLT	Task-Based Language Teaching
TOEFL	Test of English as a Foreign Language
TotBC	Total Backchannels
TR	Turn Rate
WCpl	Within-Clause Pause Length

WCpr	Within-Clause Pause Rate
WTC	Willingness to Communicate

List of Appendices

Appendix A: Carleton University REB Clearance.....307

Appendix B: Renison University College at University of Waterloo REB Clearance308

Chapter One: Introduction

Second language (L2) speech fluency (i.e. fluency) has long been considered an integral component of L2 speech (Fulcher, 2003). However, fluency has been famously problematic to define, categorize, analyse, and evaluate, as perceptions of this construct may vary widely, even among trained L2 practitioners (Tavakoli & Hunter, 2018; Koponen & Riggensbach, 2000). According to Lennon (1990), in the everyday use of the term, fluency is equated to general oral proficiency, whereas in the realm of language teaching, testing, and learning, the term is defined more narrowly as the overall speed and flow of speech, characterized by its *core* temporal features (i.e. speed, pauses, and repairs, as per Tavakoli & Skehan's 2005 taxonomy). Yet, "temporal variables are merely the tip of the iceberg as indicators of fluency" (Lennon, 2000, p. 25) as a wide variety of *peripheral*, non-temporal features such as lexical sophistication, comprehensibility, and conversational features of speech (e.g. turn-taking management) seem to have some degree of influence over how fluency is perceived by L2 practitioners. It would seem then that *core* temporal measures are inherently integrated with non-temporal *peripheral* features of the fluency construct, suggesting that temporal measures only constitute one aspect of a larger whole.

To date, a fairly substantial amount of research has investigated the effects of a variety of both temporal and non-temporal speech features on influencing perceptions of fluency (e.g. Magne Suzuki, Suzukida, Ilkan, Tran, & Saito, 2019; Williams, 2018; Préfontaine & Kormos, 2016; Sato, 2014; Bosker, Quené, Sanders, & DeJong, 2014; Götz, 2013; Rossiter, 2009). However, despite this wealth of research, there are a few research gaps, which the present study intends to address.

1.1. Motivation for the present study

Before addressing these gaps however, it is first necessary to expand upon the reasons why research on fluency perceptions is important. A lack of fluency has been shown to negatively affect speech comprehensibility (Suzuki & Kormos, 2019) as it affects how well the listener attends to and engages with the speaker (Derwing, Rossiter, & Munro, 2002; Lennon, 2000). Consequently, a lack of fluency may result in negative judgements not only about speakers' levels of language proficiency, but also about the speakers themselves (Magne et al., 2019). Hincks (2010), for instance, revealed that the amount of information conveyed by academic professionals during lectures and oral presentations may be negatively affected by a lack of fluency, which may result in negative responses about job performance, potentially hindering future career development in academia.

Moreover, as fluency is perceived to be a key component of overall oral proficiency, it is often a key learning outcome in L2 instructional settings. In high-stakes language performance situations (e.g. university admissions examinations; immigration examinations), fluency has long been featured in scoring rubrics for L2 oral proficiency tests such as the 1930 College Entrance Examination Board's English Competence Examination and the 1948 Foreign Service Institute English Examination (Fulcher, 2003). Moreover, fluency continues to be a key component of influential modern-day tests and benchmarks such as the Test of English as a Foreign Language (TOEFL) (TOEFL, 2015), the International English Language Testing System (IELTS) (IELTS, 2015), and the Common European Framework for Languages benchmarks (Council of Europe, 2015). However, as mentioned, fluency is a complex construct, which, even within the relatively narrow realm of L2 research, assessment, and pedagogy, may be differently conceptualized by L2 practitioners (Tavakoli & Hunter, 2018). According to Chambers (1997), these conflicting conceptualizations "may have consequential implications for testing and assessment" as the

"validity of the judgements made by assessors (could be) seriously in question" (p.543). The present study is thus motivated by a desire to gain a fuller understanding of the wide variety of factors that differentially contribute to listeners' inferences and judgements of L2 speech fluency.

1.2. Background

1.2.1. Core (temporal) features

It is generally assumed that increases in speed and decreases in the amount and /or length of pausing lead to higher fluency ratings (Lennon, 1990). Some research seems to suggest that, overall, this may be the case (e.g. Préfontaine, 2016; Bosker et al., 2013; Rossiter, 2009), yet other studies suggest that decreased pause length and pause frequency (e.g. Kormos & Dénes, 2004) may have little to no effect on fluency perceptions. Moreover, variations within individual performances inherently exist, warranting a closer examination of how fluency develops in an individualized manner (Derwing, Rossiter, Munro, & Thomson, 2004). As fluency is a contextually dependent construct (Ezjenberg, 2000), it is influenced, both productively and perceptually, by a wide range of contextual factors, such as speakers' levels of familiarity with the discussion topic (Bui & Huan, 2018). Consequently, results seem to be mixed as to which temporal features of speech meaningfully affect fluency judgments within particular contexts. Segalowitz (2010) attributes these varied results to variations in methodologies employed, including differences in participant group sizes and compositions, instruments used, and variables analysed. Therefore, depending on the study, any one or combination of temporal variables may affect listeners' judgments of perceived fluency. Overall, the question remains as to which temporal variables, individually, or in combination, are reliable indicators of perceived fluency within and across contexts.

1.2.2. Peripheral (non-temporal) features

Research has also shown that listeners may make fluency judgements based on a variety of non-temporal and non-linguistic features of utterances (Magne et al., 2019; Williams, 2018; Préfontaine & Kormos, 2016). In some studies (e.g. Götz, 2013; Freed, 1995), raters were not trained to perceive fluency in the narrow, temporal sense of term (i.e. the speed and flow of speech). However, these peripheral features were also shown to be influential on fluency assessments even in studies with raters who were trained in this manner. Results from a number of studies have uncovered meaningful relationships between raters' Likert scale fluency assessments and assessments of a variety of non-temporal speech features, including the following: vocabulary range and grammatical accuracy (Rossiter, 2009), the use of formulaic sequences (Ezjenberg, 2000) and intonation (Wennerstrom, 2000), among other non-temporal features. Several studies have also investigated relationships between comprehensibility, accentedness, intelligibility, and fluency. Thomson (2015), upon reviewing these studies, concluded that "fluency is most related to comprehensibility, somewhat related to accentedness, and apparently least related to intelligibility" (p. 217). Overall, both temporal and non-temporal features appear to exhibit some degree of influence over raters' perceptions; yet the extent of this influence is relatively unknown.

1.2.3. Conversational speech features

There appears to be a lack of research in regards to how conversational speech features (e.g. turn-taking, conversational pauses, non-verbal communication) affect fluency perceptions. Traditionally, much research has focused on how fluency is perceived in monologic tasks. However, some recent studies have investigated how fluency is perceived on dialogic tasks (e.g. Peltonen, 2017a; Sato, 2014). There are, notably, understandable reasons why monologic tasks

have been traditionally used. As Wood (2010) discusses, fluency researchers typically employ monologic speech elicitation tasks because researchers are generally concerned with analysis of temporal (e.g. speech rate) and not interactive (e.g. turn taking) features of speech production; moreover, as conversation is inherently unpredictable, it is much more feasible to standardize monologic speech data for comparability purposes. Regardless, this gap in research is notable because paired conversational tasks are used frequently in both testing and classroom teaching contexts, and most notably, the majority of real-world speech tasks are dialogic in nature (Sato, 2014). Moreover, future research on fluency, as elicited from interactive tasks, is necessary for enhancing fluency assessment. DeJong (2018) highlights this importance in stating the following:

Disfluencies are not only signals for the listener for upcoming complex speech (as pointed out by research in psycholinguistics) but also function as interactional devices to regulate turn-taking, to hold the floor, and to strategically repair errors (Kormos, 1999); therefore, rubrics of tests that use actual conversation need to incorporate aspects of interactional fluency (p. 248).

1.2.4. Listener characteristics

Similarly, there is little research examining the influence of characteristics inherent to the listeners themselves. The present study seeks to examine the potential effects of two key listener characteristics: accent familiarity and conversational style.

1.2.4.1. Accent familiarity

Research has shown that, on the whole, the more familiar listeners are with speakers' accents, the more favorable they will be in their judgements of overall L2 speech proficiency (e.g. Shintani, Saito, & Koizumi, 2018; Saito & Shintani, 2015; Winke, Gass, & Myford, 2012;

Carey, Mannell, & Dunn, 2011). However, to what extent fluency ratings, specifically, are affected by listeners' accent familiarity is still largely unknown as only a few studies have examined how accent familiarity affects fluency ratings in particular (e.g. Browne & Fulcher, 2017; O'Brien, 2014).

The relationship between accent familiarity and L2 speech in general, and fluency specifically, is not necessarily linear; in other words, just because listeners may be more familiar with a speaker's accent does not mean that listeners will rate the speaker more favorably. Notably, this relationship may be complicated by a range of other factors such as the extent of listeners' age, teacher training, teaching experience, and attitudes towards the speakers' L1 and its speakers.

1.2.4.2. Conversational style

Another listener characteristic potentially affecting perceptions of fluency is listeners' conversational style, which is comprised of the values and beliefs constituting what one considers as a good conversationalist (Tannen, 2005). Therefore, it is important to understand which conversational features listeners prefer, as these preferences may influence their judgments. As Ducasse and Brown (2009) argue, "it is important also to investigate what language experts – teachers and assessors – value while rating pairs, because it is their view of interaction which finds its reflection in the test scores" (p. 426). Fortunately, conducting research on fluency, as elicited from a conversational task, provides an opportunity to examine the role of listeners' conversational style on their perceptions of conversational fluency.

The present study uses Tannen's (2005) concept of conversational style. Tannen theorizes that conversational styles are developed during one's formative years due to extended exposure to interactions with persons within one's micro-culture (e.g. peers, family members), who are

inherently informed by macro-cultural (e.g. national, ethnic) influences. One's conversational style is believed to exist somewhere along a continuum between High-Involvement (HI-style) and High-Considerateness (HC-style). These styles are defined by a set of conversational tendencies that distinguish a High-Involvement (HI-style) speaker from a High-Considerateness (HC-style) speaker. According to Tannen, HI-style speakers are more likely to do the following during conversation: share more personal information; ask more personal questions; interrupt and overlap more frequently; speak more rapidly; and avoid conversational silences more often than HC-style speakers. On the other hand, HC-style speakers are less likely to share personal information and pose personal questions. Moreover, they are more likely to wait until the other speaker's turn has completed to begin speaking and they are more tolerant of conversational silences. One's conversational style is not necessarily fixed however as it is mediated by surrounding contextual influences such as interlocutors' social status.

The present study is the first to investigate whether raters' conversational styles affect their perceptions of learners' conversational fluency. As Tannen (2005) discusses, during conversation, one may be more favorable towards a speaker who shares a similar conversational style; on the other hand, when conversational styles conflict, people may have a more difficult time conversing and therefore, they may have less favourable attitudes towards one another. For instance, the HC-style speaker, who prefers to wait until the other speaker has finished speaking before commencing the turn, may find constant interruptions by the HI-style speaker to be rude. Therefore, it is possible that third-party listeners, such as L2 oral proficiency raters, may also be more favorable towards a speaker whose conversational style matches their own. In this study, the term 'conversational style preferences' is used to describe these kinds of potential biases. Overall, however, little is known about how listeners' conversational style preferences

may influence assessments of conversational fluency. The present study intends to address this gap in research.

1.2.5. Rating scale design and assessment for learning

Exploring the influence of speech and listener characteristics on fluency perceptions is the primary aim of this study. However, the process of developing and piloting a rating scale to explore fluency perceptions may have implications for assessment in general and for rating scale design in particular. As Fulcher (2003) discusses, many rating scales have been developed in accordance with an imagined ‘native speaker’ standard. Thus, one of the strengths of the rating scale, as produced from this study, is that its ceiling (i.e. the highest possible scores) is developed from observing expert-level performances from EAP learners, and not from instructors’ intuition, nor from adhering to an imaginary ideal ‘native-speaker’ model of performance.

The goal of assessment for learning on a performance-based task is to enhance learners’ understanding of their current abilities in order to narrow the gap in achieving target-level abilities (Fulcher, 2010). Notably, one of the key strengths of the proposed scale is that each of the criteria is aligned with suggested activities for targeting specific limitations in students’ abilities to speak fluently in dialogic conversation. With this in mind, this scale is developed upon the premise that fluency is a performance-based skill dependent upon specific linguistic, discourse, and pragmatic competencies. The purpose of this scale, therefore, is to draw learners’ and instructors’ awareness to any gaps in competencies as well as any gaps in executive performance skills that may impede one’s ability to speak fluently. With this raised awareness, ideally, learners and instructors could use the suggested activities, presented in the concluding chapter of this study, and aligned with this scale, to negotiate closure of these gaps and thus enhance fluent performance.

1.2.6. Overview of the study

The overall purpose of this study is to explore perceptions of fluency through developing and piloting a fluency rating scale for a paired conversational task for classroom-based assessment for learning purposes. The potential findings from this study can inform readers about which characteristics of learners' dialogic speeches as well as characteristics of the listeners themselves, such as accent familiarity and conversational style, may influence fluency perceptions. This purpose is achieved through using a two-phase mixed-methods exploratory sequential design, which is commonly used to develop an instrument such as a rating scale or a questionnaire (Creswell, 2009).

The purpose of the first phase of the study was to explore fluency perceptions through the development of the scale. Seven instructors watched videos of seven-minute paired conversations, elicited from 14 EAP learners who performed this conversational task twice with two different conversational partners. After watching the videos, instructors were audio-recorded verbalizing their judgments about learners' fluency levels. These audio-recordings were transcribed and coded, resulting in the identification of themes, categories, and salient codes. These themes were used to create a multi-item rating scale representing the range of speech features comprising the fluency construct. Each item was framed as an information question such as "Is the speech smooth?" The format of this scale was inspired by Fox, von Randow, and Volkov (2016), who developed a diagnostic rating scale, consisting of a series of analytic-criterial questions, to analyse first-year undergraduate engineers' writing competence. This format was chosen because it facilitated the deconstruction of the fluency construct into salient sub-constructs, allowing for further investigations of how these sub-constructs may differentially affect perceptions of fluency as a whole. Furthermore, this format was deemed appropriate as the

scale developed in the present study was designed for classroom-based assessment for learning purposes. Themes were used from the phase one findings to create the analytic-criterial questions and the categories and codes were then used to create the performance descriptors associated with each item on the scale.

The purpose of the second phase of the study was to continue to examine perceptions of fluency through the collection and analysis of quantitative data. In this phase, a new group of 35 EAP instructors used the scale to rate four video-recorded conversations between eight learners. Instructors reported their degree of accent familiarity prior to rating each video. After all videos were rated, instructors completed a questionnaire, informed by Tannen's (2005) findings, to elicit their conversational styles. The data was analysed to examine the degree to which the scale items, developed from the phase one findings, were relevant to assessing fluency. Additionally, the role of accent familiarity and conversational style preferences on influencing fluency assessments was also investigated.

1.3. Research questions

The following research questions were posed to meet the aims of this study:

1. How can EAP instructors' perceptions of EAP learners' speech performances on a paired conversational task inform the development of an analytical criterion rating scale to measure speech fluency?
2. To what degree are the items on the scale relevant to assessing fluency?
3. What relationships exist between temporal measures and rating assessments, according to this scale?
4. In what ways, if any, do raters' levels of accent familiarity affect their fluency ratings?

5. What relationships exist between instructors' conversational styles and their assessments of High Considerateness-style and High Involvement-style students, according to this scale?

1.4. Overview of the dissertation

This dissertation is organized as follows. Chapter Two, the literature review, provides relevant background information essential for informing the readers' understanding of the study's purpose. This section provides a review of research on the development and perception of fluency. Relevant research on interactional competence, non-verbal communication, rating scale design, accent familiarity, and conversational style is also discussed.

Chapter Three outlines the method used in phase one. In this phase, the researcher collected and analysed qualitative data, as elicited from instructors' verbalized observations of learners' video-recorded paired conversations, in order to develop the fluency rating scale. Chapter Four provides a discussion of the results from these phase one findings in relation to the construction of the fluency rating scale. This scale is presented at the end of this chapter.

Chapter Five depicts the method used in phase two of the study. The scale, developed from the phase one findings, was used to examine issues of construct-irrelevant variance. Four research questions were posed to investigate this issue. The first two questions were posed to investigate the construct relevance of question items on the rating scale by examining their relationships with one another and their relationships to temporal measures of utterance fluency (e.g. speech rate). The latter two questions were posed to investigate the role of listener characteristics on affecting fluency judgements, in particular, the potential effects of accent familiarity and conversational style. The phase two results are also depicted and discussed in

Chapter Six and Chapter Seven. Classroom applications of these findings are discussed in Chapter Eight.

Chapter Nine summarizes the study's methods, results, and interpretations and provides further suggestions and implications for forwarding fluency research, assessment, and pedagogy. Overall, the results and implications provided in the present study provide a substantial amount of promise for future research.

Chapter Two - Literature Review

2.1. Overview

Research on fluency has grown exponentially over the past 50 years. The theoretical and methodological foundations for the earliest empirical studies on fluency in the 1970s (Grosjean & Deschamps, 1972, 1975) were established by Lounsbury (1954) and Goldman-Eisler (1968), who were among the first researchers to quantify first language (L1) spontaneous speech production through various measurements of speech rate and hesitation phenomena. In the 1980s, Grosjean (1980) and Deschamps (1980) and their colleagues at the University of Kassel in Germany (e.g. Dechert, 1980; Lennon 2005; Raupach, 1987) pioneered fluency research by focusing primarily, but not solely, on how L1 specific structures and L1 production strategies may affect L2 fluent production. Fluency research expanded in the 1990s as researchers explored how fluency may be perceived (Lennon, 1990; Riggenbach, 1991), developed (Towell, Bazergui, & Hawkins, 1996), and acquired within different learning contexts (Freed, 1995). Research continued to blossom in the 2000s and 2010s and now in the 2020s it has become a key area of interest within the realm of second language acquisition (SLA) research.

2.1.1. Defining fluency

Several key definitions of fluency reoccur frequently throughout the literature. Preceding the growth of L2 fluency research in the 1980s, Fillmore's (1979) influential essay provided researchers with four classifications of L1 fluency: a) the ability to speak at a quick rate; b) the ability to speak in semantically and syntactically complex ways; c) the ability to speak in a variety of contexts; and d) the ability to speak creatively. Brumfit's (2005) distinctions between 'fluency-oriented' and 'accuracy-oriented' activities and Sajaavaara's (1987) definition of fluency as "the communicative acceptability of the speech act, or 'communicative fit' " (p. 62) mirror the rise of Communicative Language Teaching (CLT) approaches in the 1970s and 1980s. Lennon (1990) differentiates between the broad sense of fluency, as a term conflated unequivocally with language proficiency, and the narrow sense of fluency, referring to "one, presumably isolatable, component of oral proficiency" (p. 389), as used increasingly by language teachers and testers. Schmidt (1992) defines fluency as an "automatic procedural skill" (p.359), reflecting the incorporation of theories of automaticity/proceduralization into L2 research in the 1980s and 1990s. Task-based research, which developed substantially in the 1990s and 2000s, has been widely influenced by Tavakoli and Skehan's (2005) distinctions between three types of fluencies: speed (e.g. speech rate), breakdown (e.g. pauses), and repair (e.g. self-corrections/repetitions). Derwing, Munro, Thomson, and Rossiter (2009) distinguish between 'state fluency' (i.e. performance-based fluency), which is highly subject to contextual influences and 'trait fluency' (fluency as a generally fixed attribute) which is inherently mediated but is generally stable across contexts. Segalowitz (2010) also classifies fluency as three distinct types of fluencies: a) cognitive fluency, referring to the speaker's underlying cognitive processing efficiency; b) utterance fluency, referring to the linguistic and temporal properties of the utterance; and c) perceived fluency, referring to how the utterance is received aurally by the listener. Segalowitz's

categorizations of fluency have since influenced a growing number of studies on the relationships between cognitive and utterance fluency (e.g. De Jong, Steinel, Floijn, Schoonen, & Hulstijn, 2013; Kahng, 2014) and between utterance and perceived fluency (e.g. Baker-Smemoe, Dewey, Bown, & Martinsen, 2014; Préfontaine, Kormos & Johnson, 2016). Although other definitions have appeared in the literature over time, these definitions seem to be the most influential.

2.1.2. Measuring fluency

Current research often employs phonetic analysis software such as PRAAT (De Jong & Wempe, 2009) to calculate temporal measures, also known as utterance fluency measures (e.g. speech rate, pause length). Computer scripts are available to calculate fluency measures (De Jong & Wempe, 2009), yet some researchers still use a combination of manual and automated calculations to further ensure reliability (De Jong & Perfetti, 2011). Researchers have used different cut-off points to determine what constitutes a silent pause within speech: a) 0.2 sec. (Lennon, 1990); b) 0.25 sec. (Goldman-Eisler, 1968); c) 0.28 sec. (Towell et al., 1996); d) 0.3 sec. (Raupach, 1980); and e) 0.4 sec. (Derwing et al., 2004), among others. Notably, DeJong and Bosker (2013) recently discovered that, among cut-off points ranging from between 0.25 to 0.5, the strongest correlations between utterance fluency measures and oral proficiency ratings were found between 0.25 and 0.3. PRAAT creates spectrograms for each sample and these spectrograms, along with the corresponding transcripts, enable manual and automated calculations of temporal measures.

A wide variety of temporal measures has been used to measure fluency. Notably an expanded list of measures will be presented in a later chapter on fluency perceptions. For now,

the reader should be familiar with the following measures as they are commonly used in studies on fluency (adapted from Kormos & Dénes, 2004):

1. Speech Rate (SR): the total number of syllables produced (including filled pauses, e.g. “uh”, “um”) divided by the total speech duration (including silent pauses).
2. Articulation Rate (AR): the total number of syllables produced (including filled pauses) divided by the percentage ratio of time spent speaking, which is the total speech duration excluding silent pauses.
3. Phonation-Time Ratio (PTR): the percentage ratio of total time spent speaking divided by the total speech duration (including silent pauses).
4. Mean Length of Runs (MLR): the average number of syllables produced (including filled pauses) between silent pauses.
5. Filled Pause Rate (FPR): the frequency of filled pauses (e.g. “um”) in speech.
6. Silent Pause Rate (SPR): the frequency of silent pauses in speech.
7. Silent Pause Length (SPL): the mean length of pauses.

Additional measures of utterance fluency include measures of pause location within and between clauses (Khang, 2015), pause function (Fulcher, 1996), drawls (e.g. “soooooo”) (Raupach, 1980), falling intonation contours (Lennon, 2005), clause-chains (Pawley & Syder, 1983), tone units (Lintunen, Peltonen, & Webb, 2016), smallwords (e.g. “oh”) (Hasselgren, 2002), and formulaic sequences (Wood, 2010), among others. Perceived fluency is often measured through comparing measures of utterance fluency with listeners’ fluency judgements expressed through Likert-scale assessments (Derwing et al., 2004) and/or verbal reports from raters (Préfontaine & Kormos, 2016) or learners (Williams, 2018).

2.1.3. Outline of the literature review

The literature review consists of seven main sections. The first section provides a review of research on how fluency develops, informing understanding of how fluency may be perceived and assessed in an EAP instructional context. More specifically, this first section provides an overview of the cognitive processes underpinning development, the role of formulaic language (FL), and the influence of pedagogy on fluency attainment. The second section covers the range of temporal and non-temporal features of speech affecting fluency perceptions, excluding conversational features, as they are covered in later sections. The third section discusses the influence of listeners' familiarity with speakers' accents (accent familiarity) on ratings of L2 speech in general, and on ratings of L2 speech fluency more specifically. The fourth section focuses specifically on research relating to fluency as mediated through interaction (i.e. conversational fluency), covering relevant research on interactional competence (IC) and non-verbal communication (NVC). The fifth section discusses Tannen's (2005) notion of conversational style (CS) and its potential influence on fluency perceptions. The sixth section reviews key studies depicting various approaches towards fluency rating scale design, providing implications for how fluency rating scales may be designed for classroom-based assessment for learning purposes. The final section highlights the implications of this review for the present study.

2.2. Fluency development

2.2.1. Cognitive processes

'Automaticity' is central to fluency development. Speaking fluently requires the automatization of certain cognitive processes associated with production (Logan, 1988; Levelt, 1989, McLaughlin, 1990); yet as Segalowitz (2000) argues, speaking fluently at a high level

requires the flexibility to strike a balance between both automatic (fast, effortless, spontaneous) and controlled (slow, effortful, nonspontaneous) processes. This section provides several models illustrating the concept of automaticity as applicable to the development of fluency.

Automaticity and proceduralization are terms that are often used synonymously, although 'automatization' seems to co-occur with 'cognitive processes', whereas 'proceduralization' seems to co-occur with 'knowledge'. Anderson's models of cognition (1983, 1993) illustrate how, through knowledge proceduralization, declarative knowledge (the explicit knowledge of linguistic rules) is converted into procedural knowledge (the implicit knowledge of linguistic production processing), due to interactions between declarative, production, and working-memory stores. According to Anderson (1983), during the proceduralization process, knowledge is compiled into larger and larger chunks. During production, the speaker retrieves these chunks from the production memory store and then assembles them into larger chunks in the working memory store. This process occurs quickly because the speaker does not need to retrieve knowledge from the declarative memory store; thus, this declarative memory store is bypassed in the process.

Levelt's (1989) speech production model, which illustrates how proceduralization occurs, has been widely adopted in fluency research (e.g. Towell, Hawkins, & Bazergui, 1996; DeJong & Perfetti, 2011; Tavakoli, 2016). This model illustrates how speech is produced in three stages: a) *conceptualization* (discourse and situational knowledge is retrieved to form a 'pre-verbal message'); b) *formulation* (lexical, grammatical and phonological knowledge is used to encode this pre-verbal message) and c) *articulation* (the speaker uses relevant physiological mechanisms to produce speech). Towell et al. (1996) contend that proceduralization occurs at the formulation stage where lexical, grammatical, and phonological procedural rules are applied. Changes in

learners' speech over time, as measured by temporal variables, have provided some evidence that proceduralization has occurred at either one of two stages. On the one hand, proceduralization may occur at the *formulation* stage, as evidenced by an increase in learners' mean length of spoken runs (MLR), which is the number of syllables uttered between pauses. On the other hand, proceduralization may occur at the *articulation* stage, as evidenced by an increase in articulation rate, which is the number of syllables uttered/duration of speech, excluding pauses (Towell et al., 1996).

Logan (1988) provides an alternative view of automaticity, arguing that learners store and retrieve bits of knowledge as separate representations (a.k.a. instances). According to this theory, the development of fluency is, therefore, dependent on the accumulation of exemplars (a.k.a. separate instances), which can be efficiently retrieved as single encoded units from memory. A lack of automaticity stems from a deficiency in the number of exemplars available for use.

McLaughlin's (1990) restructuring theory posits that, through practice, fluency develops through a process of knowledge restructuring. According to this theory, less complex internal representations of concepts become replaced by increasingly complex representations. This restructuring process produces a U-shaped pattern of fluency development as speech may become slower or more hesitant as knowledge becomes re-organized, ultimately resulting in more efficient and/or more sophisticated speech production. This theory suggests that automatization is more than just faster retrieval of information; thus, fluency is more than just speaking faster and pausing less.

Segalowitz (2010) provides three categorizations of automaticity in speech production: processing speed, processing stability, and processing flexibility. Processing speed in speech production refers to rate of formulation and articulation of pre-verbal messages, as per Levelt's

(1989) model. Processing stability refers to how consistent the speaker is in producing speech, reflecting the degree to which the underlying knowledge and associated processes have been restructured, as per McLaughlin's (1990) restructuring theory. Processing flexibility refers to the speakers' ability to adapt to different circumstances by allotting attentional resources to performing multiple cognitive functions in an efficient manner.

2.2.1.2. Attention

Fluent speakers must have a high degree of 'attentional control' as speaking requires ongoing 'attentional shifts' (Segalowitz, 2007). For example, attentional shifts occur at different stages in spoken narratives, as the beginning of a new stage of a narrative requires a shift in consciousness (Chafe, 1980). Speakers may also shift attention differentially before producing function words and content words (Segalowitz, 2007) as content words require more of the speaker's attention than function words. Similarly, non-formulaic utterances may require more of the speaker's attention than formulaic utterances (Pawley & Syder, 1983), which consist of at least two words, exhibit a relative degree of fixedness, and are likely acquired and recalled by the speaker as if they were single words (Lin, 2018). A more thorough discussion of formulaic utterances will be presented in the next section. Segalowitz's (2007) findings regarding attentional shifts suggest that speakers who possess a substantial store of productive vocabulary, including formulaic utterances, devote little attention to their production, enabling them to be free to shift attentional resources elsewhere, as need be, to the more pressing demands of the communicative context.

Identifying attentional shifts may be key to understanding the relationships between attention, automaticity, and fluency. Chafe (1980) notes that attentional shifts can be measured, not only through pauses, but also through a combination of related prosodic features (pitch reset,

increased articulation rate at the beginning of an utterance, and final syllable lengthening) that comprise an ‘intonation unit’.

Theoretically, attentional shifts occur when learners attempt to prioritize one aspect of speech – complexity, accuracy, or fluency – over another. As Robinson’s (2001) cognition hypothesis asserts, learners are not able to prioritize fluency, accuracy, and complexity simultaneously. Research has shown that these speech aspects exhibit complex interrelationships complicated by the effects of task characteristics (i.e. planning time, time length, topic familiarity, and interactional demands, among others), which may cause learners to allocate more attention to one aspect of speech over others (Skehan, 2009).

One promising area of research on the relationships between automaticity, attention, and fluency involves Functional Magnetic Resonance Imaging (fMRI) scans of the brain. fMRI brain scans provide images of blood flow movement to different areas of the brain as these areas become more activated. Several studies involving fMRI scans have demonstrated how different areas of the brain may become more activated as the result of speakers’ need to allocate attentional resources differentially to complete certain tasks. Shimada et al. (2015) compared oral performances on two tasks with differing levels of complexity between groups assessed at three different fluency levels: high-beginner-to-low-intermediate; intermediate; and high-intermediate. The fMRI scans revealed that on the less complex task (sentence-rebuilding), more fluent speakers showed less activation in certain areas of the brain compared to the less fluent speakers who showed higher areas of activation. On the more complex task (story retell), more fluent speakers showed higher areas of activation compared to low-level learners. Similarly, Fox and Hirotani (2016) set out to measure potential changes in brain activation in 11 low-intermediate Japanese English-language learners who, for 75 hours across several months, engaged in

automated listening-to-speaking activities included in Rosetta Stone software. These activities are akin to ‘shadowing’ activities (for a description, see Wood, 2010), where learners hear phrases produced by an L1 speaker and then are required to recite them repeatedly. A speech recognition algorithm compares their performance with the intonation contour produced by the L1 speaker and then the algorithm provides them with a score based on the accuracy of these comparisons. Learners repeat these activities in an attempt to attain higher scores. Fox and Hirotsu (2016) compared learners’ performances on the sentence rebuild task and story retell before and after the 75-hour program while also comparing their performances with a control group that was exposed to general L2 language instruction. On the less complex, sentence-rebuild task, learners in the experimental group showed higher activation in the putamen, which is an area of the brain that affects behavioral management, learning processes, and L1-to-L2 translation processes. However, no clear activation changes were found for either group for their performance on the story recall task although human raters noted positive qualitative changes in performances on this task. There are a few possible reasons for this finding. As the sentence-rebuild task was the more complex task, it is possible that the task may have been too complex for learners in either group to have automatized language production over the course of the study. However, raters may have noticed other positive changes not necessarily related to automaticity, such as proper word choice, or general confidence in speech delivery, which may still affect perceptions of fluency, but which would not necessarily be revealed by fMRI scans. Overall, future studies involving fMRI brain scans may provide more information about the role of underlying cognitive processes in producing fluent speech.

2.2.1.3. Segalowitz’s fluency framework

Segalowitz's (2010) description of cognitive fluency explains how speakers' degree of automaticity of relevant cognitive processes, proceduralization of linguistic knowledge, and allocation of attentional resources necessary to speech production interact with social and motivational factors to produce fluent speech. As Segalowitz states, "cognitive fluency features include processing speed, stability, and flexibility in the planning, assembly, and execution of utterances in terms of lexical access, and the use of linguistic resources (linguistic affordances) to express construals, *which are beliefs about how one is perceived by the self and others*, handle sociolinguistic functions, and pursue psychosocial goals" (p. 164, *italics added for clarification*). This definition requires further unpacking to understand it more fully: a) *Planning, assembly, and execution of utterances* directly reference Levelt's (1989) model of speech production: planning (conceptualization), assembly (formulation), and execution (articulation). b) *Expressing construals* refers to being mindful of how others perceive you; and c) *Linguistic affordances* refer how task characteristics influence the type of language produced.

Segalowitz's (2010, p. 164) framework illustrates a dynamic system, subject to change and adaptation over time, which is comprised of two main components: (1) L2 speech production indicated by utterance fluency features such as speech rate; pause phenomena; and (2) cognitive-perceptual processing systems, characterized by features described in the next paragraph. These two main components interact in a dynamic, two-way manner with three other components mediated largely by the surrounding environment: (3) motivation to communicate (i.e. willingness to communicate); (4) the interactive communicative (social) context (e.g. task demands in a L2 learning context); and (5) fluency-relevant perceptual and cognitive experiences (e.g. exposure to and opportunity to communicate within the target-language context).

2.2.1.4. Willingness to communicate (WTC)

Segalowitz's (2010) framework helps to explain the interrelationships between fluent speech and the speakers' WTC as affected by the surrounding environment. WTC is a concept first conceptualized and applied to L2 learning by MacIntyre (1994) and colleagues (MacIntyre, Dörnyei, Clément, & Noels, 1998). As Wood (2016) describes, WTC is "a readiness to engage in communication at a specific time and with specific interlocutors" (p. 11). Some research has examined the effects of WTC on fluency through the position that WTC is an underlying cognitive trait (Derwing et al., 2009) implying that WTC is generally stable across all communicative contexts. Yet, newer research has explored the relationship between fluency and WTC as a two-way dynamic state, subject to the unpredictable online demands inherent to any communicative context (Wood, 2016; Nematizadeh & Wood, 2019; Nematizadeh, 2019). These microanalyses of how fluency and WTC interact on monologic tasks on a moment-by-moment basis provide an indication of the role of the interaction between cognitive-social-affective factors and situational variables (e.g. topic, situation, and interlocutor) in affecting the long-term development of fluency. Overall, these studies demonstrate the interrelationships between WTC and fluency as changes in one variable seem to affect changes in another at any given moment. In Nematizadeh and Wood's (2019) and Nematizadeh's (2019) studies, the majority of these interactions were positive. In other words, the act of speaking fluently encouraged one's willingness to communicate, which encouraged one to speak fluently and so on; however, negative patterns also emerged, yet less frequently, as speaking disfluently discouraged one from communicating, which resulted in less fluent production. Unpredictable and incongruous relationships also emerged between these variables. In Nematizadeh and Wood's (2019) study for instance, learners reported low levels of WTC but were shown to be fluent according to measures of mean utterance length (MLR). To explain this phenomenon, the authors inferred that

learners used pre-fabricated utterances (i.e. formulaic sequences) during this time to sustain speech as learners reported that they were internally scrambling to search for words at those particular moments. As a note to the reader, more research regarding the relationship between WTC and situational variables inherent to conversational settings will be discussed in a later chapter on conversational style.

2.2.1.5. L1 transfer

For most adults, the ability to speak the L1 is generally stable, despite the contextual effects affecting one's WTC. Derwing and Munro et al. (2009) refer to this phenomenon as 'trait fluency'. Thus, the effect of the L1 on the L2, also known as L1 transfer, can be considered to be a trait-like variable. The effect of L1 trait fluency on L2 fluent production was the subject of several research studies in the 1980s (Raupach, 1980; Mohle, 2005; Mohle & Raupach, 1989) and interest in this area seems to have renewed recently (Derwing, Munro, Thomson, & Rossiter, 2009; De Jong et al, 2013; Bergmann, Sprenger & Schmid, 2015; Peltonen, 2018). Raupach (1980), for instance, compared L1 and L2 pause profiles of both French and German speakers and discovered that "all speakers maintained their L1 profile in their L2 performance" (p. 268). Due to this finding, Raupach (1980) suggested that learners should work on improving their L1 fluency in order to improve their L2 fluency. Mohle (2005) observed that "respective language-specific structures" (p.48) between French and German contributed to difficulties that L2 learners of these languages had on picture-description tasks, as evidenced through group differences in SR and the number of pauses. In this study, three French learners of German and three German learners of French described a series of cartoons, first in their L1 and then in their L2. The speech samples were then subject to analysis of temporal, lexical, and grammatical features. Mohle inferred from the results that the German learners of French had more difficulty

describing picture sequences due to difficulties in retrieving French lexical items whereas the French learners of German had more difficulty answering questions because of the grammatical structures of German. Derwing et al. (2009) addressed the question of whether fluency is an L1 trait or an L2 state phenomenon by investigating whether L2 learners who were more fluent in their L1, would be more fluent in their L2 than learners who were less fluent in their L1. The results indicated that although highly fluent L1 learners did not make significant gains over their peers, L1 fluency may, in some ways, predict L2 fluent production, depending on the degree of influence of contextual factors. De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2013) also found that average syllable duration in the L1 and average length of pauses in the L1 could account for 21% and 57% of the respective variance in L1-L2 production differences. Bergmann, Sprenger, and Schmid's (2015) comparative study of L1 attriters, near-native L2 learners, and L1 monolingual speakers of German revealed that the significant number of disfluencies among the L1 attriters was due to the 'co-activation' of lexical items in competing languages and not as a result of incomplete language acquisition. Finally, in a recent study, Peltonen (2018) employed a mixed-methods approach to examining L1-L2 relationships through examining picture-description monologues elicited from 42 L1 Finnish secondary school learners of English. The learners performed each task twice: first in Finnish and then again in English. Their two performances were compared according to measures of speed and pausing as well as measures that Peltonen referred to as 'stalling mechanisms' (repetitions, filled pauses, drawls, repetitions). Correlational analysis revealed strong correlations between L1 and L2 utterance fluency measures, most notably for mean length of between-clause pauses. Regression analysis showed that L2 fluency measures could be predicted from the majority of fluency measures with exception to the stalling mechanisms. The qualitative analysis showed idiosyncratic use of these

stalling mechanisms to compensate for gaps in production. The results indicate the importance of L1 transfer on L2 production as well as the idiosyncratic nature in which learners employ strategies to maintain a smooth flow of speech. Taken together, these results indicate that L1 transfer affects L2 production in a variety of complex ways.

2.2.2. Formulaic language

As Pawley and Syder (1983) claim, a large majority of spoken language is formulaic, and as such, it may be necessary for learners to acquire a vast repertoire of formulas in order to attain a high degree of fluency. Particularly for less fluent learners, formulas may serve as “islands of reliability” (Dechert, 1983) that aid in maintaining a smooth flow of speech during potentially disfluent moments.

However, defining formulaic language (FL) is as difficult as defining fluency. A wide variety of terms have been used synonymously with FL, such as idioms, phrasal verbs, transition phrases, and many more (Wood, 2015). However, it is important to first distinguish between FL and formulaic sequences (FS). A “*formulaic sequence* is generally used to refer to one such item, (whereas), *formulaic* language is the uncountable noun referring to these items as a collective” (Wood, 2015, p. 2). FL is, therefore, an umbrella term for all types of FS. As a working definition for this study, Lin (2018) describes FS as consisting of at least two words, exhibiting a relative degree of fixedness, and are likely acquired and retrieved by the speaker as if they were single words. Whether or not FS are stored and retrieved as individual words is still under debate as not enough evidence has been found to support this theory fully; notably, only a limited amount of research, mostly limited to speech read aloud, can substantiate it (Wood, 2015).

Identifying FS in speech is also problematic. Numerous researchers have provided a list of criteria for defining formulaic sequences in speech. As an example, the following criteria adapted from Wood (2015, p. 89), are listed below:

1. Nattinger and DeCarrico's (1992) taxonomy: syntactic strings (e.g. NP + Aux + VP); collocations; and lexical phrases (i.e. pragmatic collocations such as "how do you do?")
2. Phonological coherence: the formulas exhibit no internal pausing
3. Greater/length complexity than other output: chunks of speech that may be longer in length than novel utterances (e.g. "I would like to...")
4. Semantic irregularity: Formulas may be metaphoric or figurative (e.g. "under my belt")
5. Syntactic irregularity: Formulas may be grammatically unusual (e.g. "beat around the bush"). In this example, the utterance cannot be pluralized.

2.2.2.1. Pawley and Syder's (2000) two hypotheses

Pawley and Syder (2000) highlight the importance of fixed expressions in the production of long stretches of spontaneous speech as "there is considerable evidence that novel lexical combinations cannot be planned across clause boundaries in a single focus of consciousness" (p. 195). To explain how fluent speakers create long runs of fluent speech, the authors propose two hypotheses. The first hypothesis refers to "the one-clause-at-a-time constraint" (p. 163); in other words, speakers are not capable of producing new (i.e. never before produced) utterances larger than an independent clause, "in a single planning act" (p. 163). As Pawley and Syder contend, the presence of pauses and other prosodic discontinuities (e.g. intonation falls) at clause-boundaries provide support for this claim. The second hypothesis refers to "the one-clause-at-a-time capacity (which states that) competent speakers are routinely able to encode the full lexical content of some independent clauses in a single planning act" (p. 164). In other words, fluent

speakers are only capable of producing more than a single clause in a single planning act because they have already acquired a vast repertoire of fixed expressions from which they can readily retrieve; thus, the storage and retrieval of complex multi-clause expressions are essential to producing long utterances.

The implications for the development of L2 learner speech are evident. For instance, as shown in a few studies (e.g. Forsberg & Fant, 2015; Foster, 2001), when faced with time-constrained performance tasks, L1 speakers use more fixed phrases than advanced L2 speakers, who may rely more on their creativity to produce spontaneous speech. As a consequence, drawing attentional resources to producing novel speech may overload L2 speakers' capacity to produce long stretches of speech, likely resulting in more within-utterance planning pauses, a reduced speech rate, and/or other perceived disfluencies.

2.2.2.2. FL and proceduralization

Several studies have revealed how FL proceduralization positively affects fluent speech in classroom settings (e.g. Wood, 2006; 2010). In these studies, Wood investigated the role of FL acquisition on positively affecting the development of fluency over the course of a six-month intensive ESL program. These studies measured learners' development of fluency through examining quantitative changes in learners' speech over time, revealing significant increases in the mean length of spoken runs (MLR), which is the average number of syllables/words produced between silent pauses. Additionally, evidence of proceduralization can be found in significant increases in speakers' formula per run ratio (FPR) (Wood, 2006; 2010), which is the average number of formulas produced within each spoken run. Through fluency/FL focused instruction, these studies have shown that learners may increase their length of spoken runs and reduce their amount/length of pausing, indicating that proceduralization has occurred.

2.2.2.3. Functions of FL in fluent speech

Formulas serve a multitude of functions in the production of fluent speech; however, the ways in which learners use FS to enhance fluency appears, to a great extent, to be idiosyncratic (Wood, 2006). Wray and Perkins (2000) claim that FL serves three main functions in overall language production: signalling identity, marking discourse, and reducing processing time. First, the authors claim that FS have a social function. They can be used to operate on the environment to create increased alignment between the speaker and the hearer. For instance, they may be used to affect how they wish to be perceived by the hearer as certain sequences may be used to signal individual and in-group identity affecting, to some degree, what kind of language the listener uses in response. One example is a speaker's use of slang expressions. Upon hearing these expressions, the listener must then choose whether to align with the speaker by responding with slang sequences to create a sense of shared in-group identity. The second main function concerns discourse marking. A speaker may use sequences to indicate the direction of the information that is to come. For instance, a speaker may use phrases such as "there are three points I want to make" to inform the listener that the speaker wishes to make a long turn. In another example, a speaker may use certain phrases to hedge disagreement with the speaker as in "well, you know, that's one way to put it". The third main function concerns the use of FS to reduce the amount of processing time required for not only the speaker to produce the utterance but for the listener to receive it. Recognizable sequences reduce listeners' processing effort because the listener is likely to recognize these sequences more readily, thus enabling them to have more time to make a prediction about what is to be said. This is why, as Wray and Perkins (2000) explain, common phrases are used so pervasively in speech despite the numerous possibilities available for conveying the same meaning and intention. In L2 speech, the use of formulas help learners to

sound more proficient in interview scenarios; moreover, pedagogy that focuses on helping learners to notice FS positively affects how instructors perceive learners' overall oral proficiency levels (Boers et al., 2006).

Although all three of Wray and Perkins' (2000) listed functions are relevant to the production of fluent speech, the implications of the third (reducing processing time) is the most salient. Nattinger and DeCarrico (1992) identify a specific set of lexical phrases that function as fluency devices, including the following: *you know*; *it seems to me that X*; *I think that X*; *at any rate*; *I mean*; *and so on*; *so to speak*; and *as I was saying*. Notably, these devices perform other functions than maintaining speech flow. For instance, whereas "you know" doubles as a clarifier, which highlights specific information; "as I was saying" also functions as a topic shifter.

For Wray and Perkins (2000), fluency devices are "compensatory devices for memory limitations" (p. 16), sub-categorized as *processing shortcuts* and *time buyers*. Two types of processing shortcuts are *standard phrases (with or without gaps)* such as 'I have known _ for _ years' and *standard ideational labels with agreed meanings*, such as 'the current economic climate'. As for the second main category, the authors identify five types of time buyers: *standard phrases with simple meanings* (e.g. make a decision), *fillers* (e.g. if you want my opinion), *turn-holders* (e.g. "and another thing"), *discourse shape markers* (e.g. "there are three points I want to make"), and *repetition of previous input* (e.g. A: What's the capital of Peru? B: What's the capital of Peru? Lima, isn't it?). As the authors contend, these types of devices are used to keep the pace and rhythm of speech and assist the speaker in maintaining one's turn. Notably however, one of the key disadvantages of Nattinger and DeCarrico's (1992) and Wray and Perkins' (2000) lists is that they were not derived from analyses of learner speech.

To fill this gap, Wood (2006) investigated the role of FS in affecting the fluency development of 11 intermediate-level English language learners over a six-month period. Through a qualitative analysis, the author identified five key functions of sequences in facilitating fluency:

- (1) repeating formulas (“he came back the cat came back”);
- (2) stringing multiple formulas together (e.g. “making music – by himself – in his room”);
- (3) relying on one formula (e.g. “and then...”);
- (4) using self-talk and filler formulas (e.g. “I guess”);
- (5) using formulas as rhetorical devices (e.g. “at the beginning”).

Learners used these formulas in different ways to extend their length of spoken runs and to reduce the amount and length of pausing as evidenced by quantitative measures of MLR and FRR. Overall, the results reveal the individualized manner in which fluency develops, in part due to the differential manner in which formulaic sequences are likely acquired. As Wood (2006) concludes, “one theme arising from these data is the complexity of human speech and the varying routes available to arrive at the same speech goal” (p. 29). These findings have implications for the development of individualized scales to measure and assess the development of fluency.

2.2.2.4. FL and pausing

One of the key benefits of using FS to enhance fluency is that pauses tend to be shorter before sequences (Erman, 2006) and absent within sequences (Lin, 2018). This reduced pausing gives credence to Wray’s (2002) notion that sequences are likely stored and retrieved holistically because the presence of noticeable pausing either before or within sequences indicates a delay in

processing. Through qualitative analysis, Tavakoli (2011) discovered that the frequency of mid-clause pausing significantly differentiated between L1 English and advanced-level L2 English speech. The author attributed this finding to the prevalent use of FS among the L1 English speakers. Lin (2018) also investigated the extent to which FS consist of internal pausing, discovering that 83% of the sequences were not interrupted by internal pausing.

However, as Lin (2018) claims, researchers cannot rely solely on pauses to identify FS within learner speech. As Lin argues, this may be the case in beginning level speech (e.g. Dechert, 1983) as learners may be limited by the one-clause-at-a-time facility (Pawley & Syder, 1983) and they may not yet be able to string together multiple formulas, thus producing more pauses between utterances. In more advanced-level speech however, pauses do not completely align with formulas at sequence boundaries nor at intonation contours. These findings suggest that more advanced speakers are able to string together multiple formulas without noticeable pausing under larger and larger intonation contours.

2.2.2.5. FL and articulation rate

As Lin (2018) describes, increased articulation rate at the beginning of formulaic utterances (i.e. anacrusis) also enhances fluency as these word combinations are often phonologically reduced through blending and linking due to frequent repetition of the sequence over time. The final few words are often lengthened, producing a ‘surge and fade’ type of intonation contour. To investigate these assertions further, Lin analysed a corpus of utterances of the short sequence ‘I don’t know why’ revealing that the first words of the sequence (“I don’t know”) were uttered more quickly with phonological reductions whereas the word ‘why’ was lengthened and uttered slowly, exemplifying the ‘surge and fade phenomenon’. However, contrary to expectations, Lin discovered that the mean rate of articulation of formulaic speech

did not differ significantly from the mean rate of articulation of regular speech, which consisted of both formulaic and non-formulaic language.

2.2.3. Fluency pedagogy

2.2.3.1. Fluency-focused activities

Nation and Newton (2010) contend that fluency will develop if classroom activities are designed to enable the following conditions: (1) a focus on meaning rather than form; (2) learners are able to draw from their own experiences; and (3) learners are encouraged to perform at a higher level than normal. The authors claim that instructors can enable these conditions to be met if they use tasks that include the following characteristics, which have been shown, in various studies, to positively impact fluency development: (1) pre-task planning (Skehan & Foster, 2009); (2) topic familiarity (Bui & Huang, 2018); and (3) topic repetition (De Jong & Perfetti, 2011).

One well-researched activity that meets these conditions by providing such characteristics is Maurice's (1983) 4/3/2 activity. This activity contains three key features assistive to fluency development: time pressure, repetition, and a change of speakers. In this activity, after being provided one minute of planning time, learners are required to speak on a familiar topic three times, at three different lengths (4 min/3 min/2 min), and to three different speakers. The positive effects of this activity on fluency development have been shown by several researchers (Nation, 1989; Arevart & Nation, 1991; DeJong and Perfetti, 2011; Boers, 2014). The results from Nation's (1989) study indicated significant increases of speech rate and reductions in the number of false starts, filled pauses, repetitions, and grammatical errors from speakers' first delivery to the third delivery. Arevart and Nation (1991) replicated this study with a larger participant-group and achieved similar results. Yet they also discovered that several participants increased their

number of hesitations while simultaneously increasing their speech rate, revealing individual differences in development.

Nation (1989) posits that planning time is a key task characteristic that enhances fluency. Ellis (2005) distinguishes between two main types of planning. The first type is pre-task planning, which consists of rehearsal (practicing a task before completing it) and strategic planning (analysing how one wishes to present the content). Strategic planning is further subdivided into detailed and undetailed conditions. In the first condition, learners receive direction on how to plan for the task. In the undetailed condition, students do not receive any direction. The second main type of planning is within-task planning, which may consist of pressured conditions, where students must plan within a certain timeframe, and unpressured conditions, in which students are able to control how much time it takes to complete a task.

Several studies have shown that pre-task strategic planning positively affects fluency as evidenced by an increased speech rate (Tavakoli & Skehan, 2005; Bui, 2014; Bui & Huang, 2018) and mean length of run (Skehan & Foster, 2005) whereas no meaningful effects seem to have been found for repairs (e.g. Bui & Huang, 2018; Bui, 2014; Tavakoli & Skehan, 2005; Mehnert, 1998). Within-task planning, however, does not seem to have much effect on fluency. Tavakoli and Skehan's (2005) found no significant effects for within-task planning on fluency whereas Ellis and Yuan (2005) revealed that within-task planning only marginally reduced repairs.

Given that pre-task planning seems to enhance fluency, a number of studies have investigated how much time is needed to produce meaningful results (e.g. Mehnert, 1998; Wigglesworth, 2000; Elder & Iwashita, 2005). Mehnert (1998) compared the effects of planning time across four groups (no planning, one-minute, five minutes, and ten minutes) and discovered

that fluency significantly increased as planning time increased. However, gains were only marginal and non-significant between the one-minute and five-minute condition. Wigglesworth (2000) also did not reveal any significant differences between the one-minute and five-minute conditions. Similarly, Elder and Iwashita (2005) also discovered no differences between groups who received a minimal amount of planning (1 min, 15 sec.) and groups who received a more substantial amount of planning (4 min, 15 sec.). On the other hand however, Bui and Huang (2018) discovered significant group differences for fluency between a planning condition (five-minutes) and a non-planning condition (30 sec). Taken together, the results from these studies suggest that at least one minute of planning is required to enhance fluent speech yet there may be little difference in performance between providing one minute and five minutes of planning. Providing ten minutes of planning however may result in substantially different results regarding fluency.

Task type may also affect the extent to which pre-task strategic planning is effective. Foster and Skehan (2009) compared performances of 32 high-beginner students, across two planning conditions – no planning and planning (10 minutes), on three different tasks: decision-making, personal exchange, and narrative. The planning condition was further subdivided into detailed, with guidance from the instructor, and undetailed, with no guidance from the instructor. The results indicated that detailed planning was only effective on the narrative task, as it likely enabled students more time to keep track of the order of events in a narrative. The authors also discovered that strategic planning affected fluent production more considerably on the personal exchange task than the narrative and decision-making. The authors inferred that the personal-exchange task also included another fluency-facilitating characteristic, topic familiarity, which, in combination with pre-task planning provides optimal conditions for producing fluent speech.

Topic familiarity is regarded to have a strong influential effect on fluency (Nation 1989; Skehan, 1999; Bui & Huang, 2018). Bui and Huang (2018) set out to investigate which task characteristic, pre-task planning or topic familiarity, is more effective in producing fluent speech. In this study, 58 university-level students, majoring in either nursing or computer science, were required to make paired-presentations about two topics: computer viruses and natural viruses. Students were intentionally paired so that there would be a mismatch of content knowledge. In other words, nursing majors were paired with computer science majors. Students were also equally divided into two groups to determine the effects of pre-task planning: a planning condition and a non-planning condition. The first group was allowed 5 minutes to plan whereas the second group was only allowed 30 seconds to plan. One surprising result is that planning increased pause frequency at the end of dependent clauses, yet not significantly. This result could indicate that pre-task planning may encourage students to attempt to produce more complex utterances, with some detriment to fluent production, as evidenced by the increase of between-clause pauses. Overall, the results revealed that content familiarity and pre-task planning affected speech in similar ways. Both variables affected speech rate, phonation-time ratio, and the frequency of mid-clause pauses with no effects on mean length of runs or the total amount of silence. Notably however, the effects of pre-task planning were stronger than the effects for content familiarity.

The third task characteristic that may enhance fluency is topic repetition. DeJong and Perfetti (2011) examined the 4/3/2 activity at three different testing intervals, across three different learning conditions: (1) *Repetition I*. Learners spoke about the same topic three times; (2) *Repetition II*. Learners spoke for three times on the same topic but after longer testing intervals; and (3) *No-Repetition*. Learners spoke three times about three different topics. The

results highlighted the role of topic repetition on fluency development as learners in the Repetition conditions showed increased or stable gains on mean length of runs (MLR), phonation-time ratio (PTR), which is a percentage ratio of time spent speaking, and average pause length (APL), suggesting evidence for proceduralization of linguistic knowledge. On the other hand, learners in the No-Repetition condition did not show significant gains in development.

The essential contribution of FL to fluency development has been previously discussed; thus, in addition to Nation and Newton's (2010) suggestions for developing fluency-based activities, some research has shown that supplementing such activities with explicit instruction of FL may lead to significant fluency gains.

Wood's (2009) six-week fluency workshop illustrates how explicit instruction of FL, reinforced by fluency-focused activities, can enhance fluency development. As Wood describes, this workshop was organized into four stages, *input*, *automatization*, *practice and production*, and *free talk*, which combined explicit instruction of FS with fluency-building activities in order to enhance fluent speech. During the *input* stage, learners listened to recordings of several L1 English speakers telling stories, providing the 'input text', which consisted of a variety of formulas. The instructor then drew learners' attention to FS and asked the learners to mark pauses according to a transcription of the recording. The *automatization* phase involved a series of activities (*shadowing*, *dictogloss*, *mingle jigsaw*, and *chat circles*) designed to facilitate acquisition and fluent production of FS. The activities used in the automatization phase are described as follows: (1) In a language laboratory, learners, following a transcript, listened and imitated the intonation and pausing patterns of the input text. This *shadowing* activity encouraged learners to notice FS and their inherent prosodic cues (intonation contours and

pauses) through imitation and repeated listening of the input text. (2) The *dictogloss* activity required learners to listen to key sentences from the input text, make notes about key words and phrases, and then collectively reconstruct the text from their notes. This activity enabled learners to notice and reproduce formulas through repeated recognition and production. (3) The *mingle jigsaw* activity required learners to memorize formulas verbatim from the dictogloss text, share them with their peers, and then listen to others share theirs. (4) The *chat circle* activity first required that learners form two concentric circles, with students in the inner circle and students in the outer circle facing one another. Then, students discussed a topic from a list of topics related to the input text, in pairs, for two minutes. Afterwards, students switched partners within their respective circles and discussed a new topic. In between discussions, students commented and reflected on any potentially disfluent moments within the conversations. Following this, the 4/3/2 activity was used during the *practice and production* stage to encourage learners to use formulas in longer speech turns. Finally, learners engaged in a spontaneous *free-talk* phase in random pairings and groups, and then once again, commented and reflected on pausing patterns.

This case study from Wood (2009) examined the efficacy of this workshop on the fluency development of a Japanese L2 English learner over a six-week instructional period. The learner produced two spoken narratives – one before the start of the course and one afterwards. The learner's speech samples were subject to temporal analysis (speech rate, mean length of runs) and the number of FS produced, the number of formulaic sequences produced from L1-speaker models, and the percentage of syllables produced from sequences. The workshop consisted of weekly 90-minute classes; each class progressed through the aforementioned four stages: input, automatization, practice and production, and free talk. The quantitative results revealed increases on all measures and the qualitative results that formulas helped the learner to extend utterance

length, fulfill speech functions in a more complex manner, and approach target language expression.

Thomson (2017) replicated Wood's (2009) fluency workshop with 44 Japanese learners of English over a six-week period. Learners studied a series of dialogues, which were designed to focus on 30 target multiword expressions. Learners also engaged in a series of fluency-building activities: listen with gist questions; mingle jigsaw, marking pauses; role-plays; phrase instruction; decreasing time role-play; shadowing; recording role-play; dictogloss; and free related situation roleplay. The author discovered that learners significantly increased their speech rate, in contrast to learners in the control group ($n = 29$), who were enrolled in an integrative-skills program. Learners in the experimental group also increased their use of multiword expressions and a positive but not significant correlation was found between speech rate and the use of multiword expressions. Of all activities, learners were most favorable towards the shadowing activities and least favorable towards the pause-marking activities.

In a related study, McGuire and Larson-Hall (2017) investigated the influence of explicit instruction of FS on fluency development over a five-week instructional period on two groups: a control group ($n = 8$), which received task-based instruction; and a treatment group ($n = 11$), which received task based instruction in addition to explicit instruction of FS. One of the key differences between instructional approaches is that learners in the treatment group were encouraged to notice FS in texts rather than being encouraged to notice individual words and grammatical structures. Notably, learners were encouraged to notice FS as wholes without analysing how they are constructed grammatically. Another key difference between approaches is that learners in the treatment group were encouraged to use the formulas during speaking activities. Learners in both groups produced oral pre-tests and post-tests, which were assessed by

16 raters. Speeches were subject to cross-sectional analyses of group differences (paired t-tests) of temporal variables (speech rate and mean length of runs) and measures of FS (the average number of syllables/sequence). The results indicated significant changes for the treatment group as measured by speech rate but not by mean length of runs. Moreover, strong correlations with large effect sizes were found between temporal measures and measures of FS, providing implications for the relationship between FS acquisition and the development of fluency.

Instruction that encourages learners to raise their awareness about what is transpiring during disfluent moments has produced some positive results. In addition to the pause-marking and shadowing activities (Wood, 2009; Thomson, 2017) mentioned previously, instruction can encourage learners to notice how fluent speakers use strategies to compensate for disfluent moments in order to maintain a smooth flow of speech. For instance, Tavakoli, Campbell, and McCormack (2016) examined the effects of a pedagogic intervention designed to promote fluency development by tweaking the existing pedagogical approach by increasing the number of fluency activities and by providing explicit instruction of awareness-raising and fluency-developing strategies, which revealed significant increases in SR, AR, MLR, and PTR. Activities included the analysis of transcripts for disfluent moments, as per Wood (2006), encouraging the employment of lexical fillers and chunks to maintain fluency during disfluent moments, as well as the provision of activities that incorporate Nation and Newton's (2009) criteria for designing tasks that enhance fluency development (e.g. repetition, topic familiarity).

There are several other noteworthy activities beneficial to fluency development: (1) *Explicit instruction of pragmatic formulas*. House (1996) revealed the positive effects of explicit instruction of pragmatic formulas as learners successfully acquired content-oriented gambits, which are multi-word sequences used to organize interactions. Speakers used these gambits to

initiate and sustain topics and turns in order to enhance their pragmatic fluency. However, learners were unable to vary their range of response gambits and instead, relied largely upon one-word responses (e.g. “yes”) when responding to other-speakers’ questions or statements. (2) *Online chats*. Blake (2009) examined learners' fluency development across three instructional contexts: a text-based online chat condition, a traditional student interactional condition, and a non-interactional condition. The results indicated that learners made significantly greater gains in mean length of runs and phonation-time ratio (i.e. percentage of time speaking) in the text-based online chat condition, suggesting further research in the role of writing activities in fluency development. (3) *Drama activities*. Galante and Thomson (2016) examined the use of drama activities (e.g. well-rehearsed monologues and roleplays) to enhance fluency development, which resulted in significantly higher fluency ratings for learners enrolled in the experimental condition than in the control condition (a traditional communicative program).

As a note to the reader, additional activities that may aid in developing fluency can be found in Wood (2010), Nation and Newton (2009), and Derwing et al. (2010). Overall, the activities mentioned above may assist fluency development in various ways, either through meeting Nation and Newton’s (2009) conditions, promoting FL acquisition in both monologic (Wood, 2006, 2010) and dialogic (House, 1996) settings, or by raising awareness of compensatory strategies that learners can employ during disfluent moments (Tavakoli et al., 2016).

L2 instructors can also encourage learners to seek out opportunities beyond the classroom to engage in contact activities, which are extra-curricular activities with the target language and culture (Rossiter et al., 2010). First, instructors can raise learners’ awareness of opportunities in the community such as volunteering. Springer and Collins’ (2008) study of two native French-

speaking learners of English in two different volunteering situations revealed that both learners made considerable gains in general oral proficiency while also increasing their linguistic self-confidence and understanding of the target culture. Second, instructors can assign extensive listening and reading tasks, which have been correlated with improvements in turn length and fluidity (Segalowitz & Freed, 2004). Finally, as Wu (2012) concluded from his surveys of Chinese L2 learners of English in Hong Kong on the efficacy of extra-curricular tasks, that where appropriate, extracurricular tasks can be included as part of the curriculum. Several examples of such contact activities include the following: conducting surveys, joining English-speaking clubs, and engaging in online chat groups that have been co-ordinated by the instructors. Therefore, with consideration to learners' relative WTC and contextual limitations, L2 instructors can promote opportunities for learners to engage in contact activities to further fluency development.

2.2.3.2. Complexity, accuracy, fluency

Another concern that practitioners must consider when designing fluency activities is the complex relationships between three components of oral proficiency: complexity, accuracy, and fluency (CAF). Research has shown that CAF constructs exhibit complex interrelationships. As Robinson's (2001) cognition hypothesis asserts, learners are not able to prioritize fluency, accuracy, and complexity simultaneously. Research has provided support for this hypothesis by showing that certain tasks cause learners to allocate more attention to one aspect of speech over others (Skehan, 2009). Since fluency is a performance phenomenon, when instructors use tasks that focus on performance (such as the one in the present study) learners will tend to prioritize fluency over accuracy and complexity; however, tasks that focus on development influence learners to prioritize complexity (through restructuring) over accuracy and fluency (Skehan,

2009). Tavakoli and Skehan's (2005) review of the effects of task characteristics on CAF production shows that planning, topic familiarity, and degree of structure positively affect fluency, whereas outcome complexity and within-task transformations have no effects; moreover, dialogic tasks negatively affect fluent production more than monologic tasks.

Distinctions between complexity, accuracy, and fluency may be useful for language practitioners, such as testers, and teachers, but they may not reflect how speech develops as development in one area is often related to development in another (Nation, 2013). Lennon (2000) asserts that Brumfit's (2005) dichotomy between fluency and accuracy facilitates classroom methodology as it enables instructors to create and use activities aligned with specific learning objectives associated with either accuracy or fluency; moreover, it allows testers and researchers to isolate features of speech to discretely test or examine. However, as Lennon (2000) claims, this dichotomy does not reflect speech development. Lennon's notions reflect McLaughlin's (1990) restructuring theory that, through practice, speech not only becomes more automatized but that it becomes reorganized conceptually into more complex internal representations; as a result, this restructured speech becomes both more efficient and more accurate. For Skehan (2009), restructuring is more akin to the development of speech complexity; yet, he also indicates that restructuring must occur in order for the speaker to achieve higher-order fluency: "effective fluency is achieved when previous restructuring becomes automatized or becomes a (correct) exemplar" (p. 94). It would then seem that a certain threshold of speech complexity, as achieved through restructuring, is necessary for the attainment of higher-order fluency.

Development in one area may thus foster development in another. For instance, as Tavakoli et al. (1996) discovered, learners' acquisition of subordinate clauses may increase

measures of both fluency (e.g. mean length of runs) and complexity (e.g. average clause length). Additionally, Nation (1989) and Arevart and Nation (1991) revealed that the 4/3/2 activity enabled learners not only to make fluency gains, as measured through temporal variables (e.g. speech rate), but also to make accuracy gains, as measured through the reduction of phonological and grammatical errors. As Nation and Newton (2009) state, since these new procedures become both more efficient and accurate due to restructuring through practice, “it is therefore not surprising that developments in fluency are related to developments in accuracy” (p. 152).

Arguably, it is still useful to teachers, testers, researchers, as well as students, to continue to adopt this fluency, accuracy, complexity trichotomy when designing and using teaching and testing materials. As mentioned, it allows instructors to allocate valuable instructional time, testers to operationalize test constructs, and researchers to measure discrete aspects of speech. Students also benefit from this trichotomy as it allows them to prioritize their attentional resources on certain aspects of speech production over others.

2.2.3.3. Summary

Research on fluency development provides insight into understanding research on fluency perceptions. Understanding the impact of speakers’ underlying cognitive processes, the role of formulaic language, the influence of pedagogy, and the interaction between complexity, accuracy, and fluency is important to understanding the extent to which perception reflects production.

2.3. Fluency perceptions

2.3.1. Temporal features

The following table provides a list of measures of speed, breakdown, repair, composite, and dialogue, that have been commonly used to investigate the relationships between utterance

fluency (i.e temporal measures) and perceived fluency (i.e. listeners' judgements) (Tavakoli, 2016).

Table 1

Temporal Measures, Sub-Measures, and Calculations (adapted from Tavakoli, 2016)

<u>Measure</u>	<u>Sub-Measure</u>	<u>Calculation</u>
Speed	Articulation Rate (AR)	Number of syllables/min. (excluding pauses)
Breakdown	Silent Pause Rate (SPR)	Number of silent pauses/min.
	Average Length of Silent Pauses (APL)	Average length of silent pauses/number of pauses
	Within-Clause Pause Rate (WCpr)	Number of silent within-clause pauses/number of clauses
	Within-Clause Pause Length (WCpl)	Average Length of within-clause pauses/closures
	Between-Clause Pause Rate (BCpr)	Number of between-clause pauses/# of clauses
Repair	Between-Clause Pause Length (BCpl)	Average length of between-clause pauses/number of clauses
	Filled Pause Rate (FPR)	Number of filled pauses/min.
	Repetition Rate (RepetR)	Number of repetitions/min.
Composite	Repair Rate (RepairR)	Number of repairs/min.
	Speech Rate (SR)	Number of syllables/min. (including pause time)
	Pruned Speech Rate (PSR)	Number of syllables – filled pauses, repetitions, and repairs/min.
Dialogue	Mean Length of Runs (MLR)	Average number of syllables between silent pauses
	Turn frequency (TF)	Number of turns

Backchannel frequency (BF)	Number of all backchannels/speech duration
----------------------------	--

2.3.1.2. Speed fluency

As Segalowitz (2010) describes, “perceived fluency is the inference that listeners make about the connection between utterance and cognitive fluency” (p. 49). In other words, listeners make observations about utterance fluency features – such as those listed above (speed, breakdown, repair, composite and dialogue) - in order to make inferences about cognitive fluency (i.e. the underlying workings of speakers’ cognitive processes). The following section reviews research related to the kinds of inferences that listeners have made about speakers’ cognitive fluency through observations of utterance fluency.

Perceptions of fluency are often influenced by the speed of speech, as shown by Likert scale correlations between raters’ assessments and measures of speech rate (SR) (Kormos & Denes, 2004). However, according to DeJong (2016b), SR is not a true measure of speed fluency; instead, it is a composite measure of both speed fluency and breakdown fluency as it includes pause time in its calculations, reflecting “speech-pause relationships” (Lennon, 1990). On the other hand, articulation rate (AR) is a truer measure of speed as it does not include pause time in its calculation. Therefore, as DeJong (2016) professes, AR is a more accurate measure of the articulation stage of speech production, as per Levelt’s (1989) speech production model. AR also has strong perceptual saliency. Préfontaine (2013), for instance, discovered that strong correlations were found between AR and self-reported measures of speed, pausing, and overall smoothness. Suzuki and Kormos (2019) also discovered that, among a wide variety of

phonological and temporal variables, AR was the best predictor of listeners' comprehensibility judgments.

2.3.1.3. Breakdown fluency

Pauses are a natural and necessary aspect of speech in any language, occurring for a variety of physiological (Poyatos, 2014) and pragmatic (Schelgoff, 2007) purposes. Yet L2 speech is often characterized by the perceived 'unnaturalness' (in terms of relative frequency, length, and/or location) of the pause (Lennon, 1990). Therefore, the silent pause is at the heart of fluency research. Much research has examined the number, length, location, and to a lesser extent, the functions of silent pauses in L2 speech. L2 learners tend to pause more often and for longer lengths of time (DeJong, 2016a); however, a greater frequency, or length, of pausing does not necessarily lead to perceptions of disfluency, according to a review of studies from Préfontaine and Kormos (2015). The location of pauses may be a more revealing indicator of fluency. Lounsbury (1954) is credited as first distinguishing between 'juncture pauses', which occur at the beginnings and ends of clauses and 'hesitation pauses', which occur within clauses. Although L1 speakers pause for numerous reasons, including attentional shifts (Chafe, 1980), they typically produce juncture pauses to plan the content of their next utterance, whereas L2 learners more frequently produce hesitation pauses to retrieve linguistic knowledge and assemble utterances (Pawley & Syder, 1983). In reference to Levelt's (1989) speech production model, pauses within clauses reflect the formulation stage of speech production whereas pauses at clause boundaries reflect the conceptualization stage of speech production (DeJong, 2018).

Both qualitative and quantitative analyses of pause location have provided support for this claim. For instance, Tavakoli's (2011) qualitative analysis revealed that pauses produced at clause boundaries indicate planning time to produce the content of the utterance while pauses

produced within clause boundaries are more indicative of planning time used to retrieve lexical items and organize the grammatical structure of the utterance.

Quantitative studies have provided substantial support for the claim that pause location discriminates between fluency levels. Kahng (2014), for example, revealed strong correlations between rating assessments and three measures of pause rate, which is a measure that blends pause frequency with pause location. The three measures are as follows:

- (1) Pause rate within a clause, calculated by dividing the number of within-clause pauses by the total number of clauses
- (2) Pause rate at a clause boundary, which is calculated in a similar manner
- (3) Pause rate within an Analysis of Speech Unit (AS-unit) boundary (Foster, Tonkyn, and Wigglesworth, 2000), which is defined as “a main clause plus any other clauses which are dependent upon it” (Foster et al., 2000, as cited in Kahng, 2014, p. 821).

Kahng (2014) discovered that, of these three measures, pause rate within a clause was clearly the strongest determiner of scores on the Speaking Proficiency English Assessment Kit (SPEAK) oral proficiency test. These findings show that measures of pause rate within a clause may reveal important information about raters' judgements of perceived fluency.

Several recent studies have applied Kahng's (2014) approach and have corroborated her results (e.g. De Jong 2016; Suzuki & Kormos, 2019). Most notably, as Shea and Leonard (2019) recently revealed, among all temporal calculations combining pause rate, length, and location, within-clause pause rate was the strongest discriminator between levels of fluency.

2.3.1.4. Repair fluency

Several studies have investigated the workings of 'repair fluency', which includes filled pauses, repetitions, reformulations, false starts and replacements (Tavakoli & Skehan, 2005). An

overview of research indicates that quantitative measures of repairs are associated, to some degree, to measures of perceived fluency (Riggenbach, 1991; Freed, 1995; Freed, 2000; Bosker, Pinget, Quené, Sanders, & De Jong, 2012; Kahng, 2014). These results align with previous research studies showing that the number or type of repairs have moderate (Bosker et al. 2012) effects in some studies but little or no effect in others (Cucchiaroni et al., 2000; Kormos & Dénes, 2004) on fluency judgements. Notably however, Skehan (2016) revealed that certain types of repairs (filled pauses, repetitions, and self-corrections) correlated strongly and significantly with quantitative measures of vocabulary – lexical diversity, represented by the abbreviation, VocD. Skehan concluded that learners with a large vocabulary range are able to draw more readily from their wide range of lexical resources to repair utterances efficiently.

However, it may be more useful to understand how repairs are used functionally, rather than measuring how many repairs are used, in determining their effects on fluency judgments. Advanced speakers may use repairs more strategically to maintain speech flow. As Götz (2013) remarks, “self-corrections seem to be a rather positively evaluated speech management strategy production, but they should not necessarily be considered a disfluency marker” (p. 38). Likewise, the CEFR rubrics for the highest fluency levels state that the speaker “can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it” (Council of Europe, 2001, p. 37). However, as Götz (2013) suggests, there may be “a threshold concerning their frequency” (p. 31) even though quantitative studies on repairs have yet to determine what this threshold may be.

Some studies have revealed that filled pause rate correlates inversely with fluency ratings (e.g. Révész, Ekiert, & Torgersent, 2016; Bosker et al., 2013), indicating that less fluent speakers

use more filled pauses; yet it is also important to understand how filled pauses, as well as repetitions, may positively affect speech production. These devices may be regarded as stalling mechanisms (Peltonen, 2017a) used to signal that the speaker needs more processing time to formulate utterances in order to claim or sustain a turn. As Crible and Pascaul (2019) revealed, repetitions and filled pauses often co-occur with silent pauses in turn-initial position in fluent English, French, and Spanish speech. Similarly, as Carroll (2004) describes, for less-fluent speakers using these devices are “interactional achievements” to claim and maintain a smooth flow of speech. Moreover, as Fox Tree (2001) discovered, different kinds of filled pauses convey different messages to the listener as “uh” signals a short delay in speech whereas “um” signals a long delay; additionally, “um” indicates that the following utterance will be more complex (Watanabe, 2008). Filled pauses, as well as repetitions and lexical fillers, serve additional pragmatic conversational functions, which will be discussed in more detail in the later section on conversational fluency. With this in mind, it may be more pertinent for researchers to investigate the nature of repairs through qualitative analyses rather than quantitative analyses.

2.3.1.5. Composite measures

Composite measures blend speed and breakdown fluencies in order to provide information about speech-pause relationships (Tavakoli, 2016). Speed fluency is measured by syllable/word counts whereas breakdown fluency is measured by pause length, frequency, and/or location. Composite measures, such as speech rate (SR) and mean length of runs (MLR), include both syllables and pauses in calculations, thus blending measures of speed and breakdown fluencies. Composite measures have been shown to be perceptually salient variables in numerous studies, according to correlational analyses between composite measures and rating assessments: a) SR, (e.g. Préfontaine et al., 2016; Préfontaine, 2013; Bosker et al., 2014; Kormos & Dénes,

2004); b) MLR, (e.g. Préfontaine et al., 2016; Rossiter, 2009; Kormos & Dénes, 2004; Cucchiarini, Strik, & Boves, 2000). c) PTR (e.g. Ginther, Dimova, & Yang, 2010; Derwing et. al, 2004); d) PSR (e.g. Derwing et al., 2009; Derwing et al., 2004); and e) decreases in number and/or length of pauses (e.g. Préfontaine et al., 2016; Bosker, Pinget, Quené, Sanders, & De Jong, 2012; Rossiter, 2009). These composite fluency measures are useful in providing information about overall speech-pause relationships but provide only surface-level information about the nature of speech fluency. Speech rate, for instance, “gives us little information about the workings of fluency unless it is viewed in interaction with certain other variables” (Wood, 2010, p.20).

2.3.1.6. Cross-sectional analyses by proficiency level

Some insightful information has come from studies investigating perceptions of fluency through cross-sectional analyses of different proficiency levels. Ginther, Dimova, and Yang (2010), for instance, discovered that Spearman rank-order correlational analyses produced strong correlation coefficients between proficiency scores and SR, AR, PTR, MLR, and the number and length of silent pauses. No significant results were found for any measures involving filled pauses, suggesting that their use is idiosyncratic. At the highest levels, the authors also discovered that speech rate and mean length of runs did not discriminate between groups at the highest levels, indicating that learners reach a threshold of temporal fluency at the near-to-highest levels. The authors suggest that at the highest levels, related linguistic features of speech (e.g. pronunciation, vocabulary), which were not measurable by utterance fluency variables, may be indicators of the highest fluency levels.

Bosker, Pinget, Quené, Sanders and De Jong (2014) examined group differences between listeners’ perceptions of L1 Dutch (n = 10) and L2 Dutch (n = 10) fluency. In two separate

experiments, the authors manipulated the speech rate and pause length for both groups and discovered that these phonetic manipulations affected judgements of both groups relatively equally by a mean reduction of .5 points. Additionally, the results showed that idiosyncratic disfluencies in L1 speech affect fluency ratings of L1 speech providing implications for the use of L1 speech as the highest attainable performance benchmark.

In a study by Saito, Ilkan, Magne, Tran, and Suzuki (2018), 10 raters assessed 90 Japanese L2 English speakers and 10 L1 English speakers on a nine-point fluency scale, who were then divided into four groups of proficiency levels: low ($n = 29$); mid ($n = 30$); high ($n = 31$); and native ($n = 10$). Correlational analysis revealed that certain temporal measures differentiated between different levels of fluency. Whereas between-clause pause frequency differentiated low- and mid-levels, within-clause pause frequency differentiated between mid- and high-levels, while articulation rate differentiated between high- and native-proficiency levels. Taken together, the results from these three studies show that proficiency level is an attribute that needs to be carefully considered when making generalizations about relationships between temporal measures and rating assessments.

2.3.1.7. Self-perceptions and task type

Task type is another impactful variable to consider. Préfontaine (2013) examined relationships between learners' self-assessments, raters' assessments, and utterance fluency measures, across three different tasks. 40 L1 English learners of L2 French produced three different spoken narratives: Task 1 (picture narration); Task 2 (story retell); and Task 3 (another picture narration). Afterwards, learners self-assessed their fluency levels by completing a questionnaire consisting of eight items assessed across a six-point continuum between two semantic opposites (ex. Very X/Not Very X). Each of these items represented various

components of the fluency construct: smoothness, rhythm, efficiency, effortlessness, pauses, speed, self-corrections, and lexical retrieval. Three raters then listened to these audio-recorded performances and then completed the same questionnaire for each participant. The results revealed moderate correlations between raters' and learners' assessments, and that on two of the three tasks, self-assessments on the questionnaire items were most strongly associated with mean length of runs and average pause time. Individually, each questionnaire item (e.g. self-perceptions of smoothness) correlated with utterance fluency measures to varying degrees. For instance, strong correlations were found between AR and self-reported measures of speed, pauses, and smoothness. Also, strong correlations were found between mean length of runs, mean pause length, and pause frequency and perceptions of effortlessness. However, these correlations were mediated by task type as these correlations were found to be strong on Task 3 but only moderate on Task 1. The author attributes differences across tasks to the contention that different tasks require learners to attend to different aspects of speech production in terms of accuracy, complexity, and fluency (Robinson, 2001). In other words, some tasks may elicit more complex or more accurate speech whereas other tasks may elicit more fluent speech depending on the task characteristics. In this study, although all three tasks elicited narratives, Task 2, the story retell, required more attentional resources related to complexity as performance on the task necessitated the knowledge of various verb tenses and topic-specific vocabulary. The results reveal not only the saliency of specific temporal features (MLR and SPL) on self-perceptions but also that these perceptions may be subject to variation as a result of task type.

2.3.2. Non-temporal features

2.3.2.1. Analysis of raters' observations

Through analysing raters' observations, several studies have demonstrated the link between temporal and non-temporal features of fluency. Freed (1995) discovered that raters' perceptions were influenced by a wide range of non-temporal and non-linguistic factors including learners' "richness of vocabulary, accuracy of grammar, accent, clarity of voice, enunciation, rhythm of the phrases, tone of voice, ease, confidence in speech, and comfort in the ability to converse" (Freed, 1995, p.143). In Kormos and Dénes' (2004) study, raters' written comments revealed that accuracy, lexical diversity, and average pause length also affected their judgements to varying degrees, depending on the rater. Brown, Iwashita, and McNamara's (2005) analysis of raters' think-aloud reports revealed that their fluency judgments consistently overlapped with the other measured constructs of oral proficiency (vocabulary/grammar, pronunciation, content), suggesting that fluency is highly interrelated with these constructs. Brown (2007), through the think-aloud procedure, elicited discussions from eight raters who were asked to verbalize their rating processes of four audio-recorded International English Language Testing System (IELTS) interviews. Brown then organized raters' responses into two types (positive and negative) across seven main categories and their associated sub-categories. Brown discovered that raters made inferences not only about temporal features (speed, hesitancy, fillers, and repetitions) but also on non-linguistic aspects such as personality, affective state, and topic interest. Notably, the majority of raters' comments were negative; for instance, 27 out of 30 comments regarding speed and hesitancy were negative. In the following quote, Brown (2007) comments on the problem of making inferences about ability from evidence of disfluency, not from evidence of fluency:

"A major problem with performance tests is the fact that while evidence of a behavior can clearly be taken as an indication of mastery, lack of evidence cannot always be assumed

to indicate non-mastery. So in the case of fluency, the question arises 'Is the lack of fluency evidence of linguistic shortcomings (i.e. a search for words) or simply evidence of cognitive planning, a consequence of the type of task or question?' Whatever the case, it is likely that the inference drawn by the rater as to the cause of hesitation will affect the way the perception of fluency is integrated into the final judgement" (p. 122).

Similarly, Rossiter (2009) provided a qualitative analysis of raters' comments, which were categorized as either positive or negative, revealing that 75% of the negative comments pertained to speech rate, fillers, repetitions, and pauses. Raters' comments were also categorized as temporal and non-temporal. Non-temporal comments included references to grammar (21% and 27% per data collection time), vocabulary (19% and 17%), confidence (6% and 5%), and pronunciation (50% and 51%). Half of the comments under the non-temporal category related to pronunciation. Consequently, this latter category was further sub-categorized as segmental (vowels and consonants) and non-segmental (linking, rhythm, and voice quality). The majority of these comments were negative. Overall, these findings indicate that fluency seems to be identified by its lack of disfluencies and that non-temporal linguistic features affect fluency judgments to some extent, and in a largely negative manner.

In another study, Götz (2013) conducted a contrastive corpus analysis of a corpus containing both L1 English and L2 English speech, revealing differences between L1 and L2 English users in terms of temporal variables (speed, pauses, repairs) but also in terms of the number of FS used, including discourse markers and smallwords. In spite of these differences between L1 and L2 English speech, raters in her study were not trained to perceive fluency in the temporal sense and were therefore not attuned to the influence of temporal variables on fluent speech; instead, for these untrained raters, the non-temporal variables were believed to be more

salient aspects of oral proficiency/fluency. From her overall findings, Götz proposes a framework for fluency consisting of productive fluency (temporal variables) and perceptive fluency (non-temporal variables) and a third type of fluency, which was discussed but not investigated in this study – non-verbal fluency. Through this framework, Götz provides a broad overview of the nature of L2 fluency and of the wide variety of peripheral features that may influence listeners' perceptions of fluent performance. Götz (2013) elicited 50 L1 English raters' perceptions of 5 L2 English speakers' (L1 German) fluency through questionnaires and focus-group interviews with 9 respondents. The results from the questionnaires and interviews revealed that raters reported that their ratings were influenced by a wide range of non-temporal factors, such as accuracy, idiomaticity, intonation, accent, pragmatic features, lexical diversity, and sentence structure. It is not surprising, however, that raters in this study would report that a wide range of non-temporal speech features would affect their ratings because these raters were untrained and uninformed of how fluency is defined commonly by its temporal features. Thus, these findings are a reflection of raters' perceptions of oral proficiency more generally, not of fluency specifically.

Préfontaine and Kormos (2016) conducted a qualitative analysis of three raters' perceptions of 42 L2 French learners' levels of fluency. Salient features of fluent speech included speed, rhythm, pause phenomena, self-correction and efficiency/effortlessness in word choice, and target-like rhythm and prosody. The authors discovered that raters were influenced by a wide range of core features (speed, pause, self-corrections, and ease of lexical retrieval) but also by related measures of rhythm and prosody. The authors attributed these perceptions to cross-linguistic differences between English and French as the two languages exhibit salient

differences in rhythm. The authors emphasize the importance of rhythm in L2 learning and suggest its salience in affecting perceptions of fluency.

Williams (2018) investigated expert raters' perceptions and intermediate-to-advanced learners' self-perceptions of fluency. Learners' speeches were elicited from task one of the Canadian Academic English Language (CAEL) assessment (Fox, 2004), which is a high-stakes test of English proficiency designed for university admission purposes. Raters and learners used the accompanying CAEL oral proficiency rating scale as well as the Common European Framework of References (CEFR) fluency benchmarks coded to a six-point scale to rate their performances and, through a think-aloud protocol, verbalized their rationales for their judgements. Correlational analyses revealed significant correlations between speech rate, phonation-time, and oral proficiency scores (CAEL) and fluency scores (CEFR). Williams' qualitative analyses revealed that listeners' judgements were influenced by the perceived degree of automaticity, comfort, grammatical acceptability, speed, continuity, contextual/cultural familiarity, and receptivity of speech.

Tavakoli and Hunter (2018) investigated the relationships between instructors' definitions of fluency and their pedagogical practices through analysing how 84 L2 English instructors define fluency and use activities to enhance fluency in the classroom. The findings indicated that instructors predominantly define fluency in Lennon's (1990) broad sense of the term, which equates fluency to general oral proficiency; relatedly, instructors used activities, such as free-production activities, to enhance general oral proficiency rather than fluency specifically. The authors suggest 'tweaking' these free-production activities to include task characteristics that enable conditions that promote fluency (as per Nation & Newton, 2009). Relevant to the present study, Tavakoli and Hunter (2018) recommend that "more research is

needed to develop a clearer picture of why teachers adopt a broad approach to defining fluency and in what ways their understanding of fluency relate to their professional practice” (p. 345).

Magne et al. (2019) applied a mixed-methods approach to investigating perceptions of fluency. Ten L2 English raters from a variety of L1 backgrounds rated 100 speech samples from 90 L1 Japanese learners of English on a nine-point scale. Once all ratings were complete, participants discussed speech features that influenced their fluency judgements. Rating scale assessments and utterance fluency measures were subject to regression analyses revealing that between-pause clauses were significant predictors of performance. Raters’ discussions were analysed and coded, revealing three key themes: temporal factors, non-temporal factors, and social factors. This latter factor included comments related to education, socio-economic status, gender, and nativelikeness. The authors infer that the L2 raters refer to a native-speaker model of performance when making their assessments, indicating that raters’ implicit social biases may have a degree of impact on their fluency assessments.

2.3.2.2. Perceptions of FL and fluency

As discussed previously, FL plays a unique role in the production and perception of fluency. FL has relatively unique linguistic properties in terms of pausing, prosody, and pragmatics, which contribute to fluent speech (Wood, 2015). Research has shown that these features work in tandem to affect not only the production, but also the perception of fluent speech as FS have been shown to contribute not only to higher oral proficiency ratings (e.g. Boers et al. 2006) but to fluency ratings as well (Ezjenberg, 2000). In this latter study, Ezjenberg (2000) revealed that, in monologues, the more fluent speakers used FS, and one-word lexical fillers more frequently and more appropriately to organize their speeches and to maintain speech flow. In dialogues, more fluent speakers "imitated and incorporated useful formulaic language

and vocabulary" (p. 308) from interlocutors to keep the dialogue going. Less fluent speakers however, indicated difficulty in retrieving and incorporating FS into their speeches accurately and appropriately.

It would seem, however, that not all of the functions in which learners use FS, as identified by Wood (2006) [see p. 45 for a list of these functions] are equally valued by raters. In a study by Brown (2007), eight raters verbalized their processes of making oral proficiency judgements about learners' performances on the IELTS oral interview. In this study, too much perceived reliance on one formula, including self-talk and filler formulas, was perceived negatively by raters. On the other hand, the use of infrequent and idiomatic formulaic sequences was perceived positively, perhaps because they signal to the listener that the speakers are part of the in-group of proficient language speakers. This finding relates to Wray and Perkins' (2000) argument that one of the key functions of formulaic sequences is to signal in-group identity. Moreover, as Brown (2007) observed, raters make predictions about performance in future real-world contexts based on their observations of learners' speech, predicting that the use of FS that have a high degree of idiomaticity signals that learners may be fluent in more demanding contexts where these phrases may be more frequently used.

Formulaic language may be uttered more quickly (i.e. increased AR) with less pre-sequence pausing (Erman, 2006) and with less within-sequence pausing (Lin, 2018); therefore, FL may, from a purely temporal sense, contribute not only to more fluent speech productively but perceptually as well. Pause phenomena, in particular, may be key to discriminating between perceptions of fluent and non-fluent speech; yet pauses are only one of the prosodic elements that may influence perceptions of discontinuity. Lin (2018, p. 51) lists several prosodic features that may disrupt perceptions of speech continuity: (1) pitch reset (the beginnings of utterances

are often marked by a high pitch, which gradually cascades downwards); (2) anacrusis (increased articulation at the beginning of an utterance); (3) syllable-lengthening (the final syllables of an utterance are often lengthened); and (4) the absence of linking/elision (the phonological blending of phonemes). Discontinuity may be affected by any combination of temporal (pauses, syllable-lengthening, anacrusis), pitch (pitch reset), and segmental (linking/elision) features (Knowles 1991, p. 153, as cited in Lin, 2018). Knowles places these discontinuities into a hierarchy, in terms of their saliency in disrupting speech continuity, rated on a scale from 1 (low saliency) to 5 (high saliency): 1 = segmental run-on cancelled; 2 = segmental separation; 3 = pitch discontinuity; 4 = pause; 5 = pause accompanied by audible breathing. According to this hierarchy, pauses are the most influential indicators of continuity, followed by pitch, and then segments.

To investigate the relationships between FL and prosodic features, Lin (2018) examined whether or not formulaic sequences could be identified based on a combination of prosodic cues (see the list in the preceding paragraph) rather than by pauses alone. Lin claims that whereas silent and filled pauses may be useful to identify sequences in less fluent speech (e.g. Dechert, 1983; Raupach, 2005), more fluent learners string together sequences with less noticeable pausing between them. Contrary to expectations, however, Lin discovered that only 55% of sequences could be identified through using prosodic cues alone, ultimately contending that “each sequence must be assessed on a case-by-case basis” (2018, p. 18).

Relatedly, Lintunen et al. (2016) investigated the efficacy of the tone unit as a measure of fluent speech. The authors compared spoken narratives from 30 L1 Finnish learners of English across three proficiency levels to narratives produced from an L1 English speaker group (n = 10). As mentioned previously, tone units may express a single focus of consciousness (Chafe, 1980),

which has commonly been measured by utterance units, (a.k.a. the average number of syllables produced between pauses, otherwise known as MLR). Although a tone unit is similar to a unit of utterance length (MLR), MLR is delineated by pause boundaries, and tone units are delineated by a combination of features. Each tone unit consisted of the nuclear syllable (i.e. the most prominent vowel at the beginning of the utterance), a pre-head, a head (containing the nuclear syllable) and a tail. Also, as per Lin (2018), anacrusis, final-syllable lengthening, and pitch reset were used to identify tone units. It should be noted that Lintunen et al. (2016) discovered that tone units could be delineated by pauses only 70% of the time, meaning that MLR and tone units do not reliably coincide. Overall, the results revealed that mean tone unit length differentiated proficiency levels. More proficient speakers produced clause-length tone units whereas less advanced learners produced tone units as individual syllables/words. The authors suggest that tone units are viable measures of fluent speech. However, one drawback of using this unit is that utterance units (MLR) are much easier to identify than tone units, which requires arduous and detailed judgment from multiple well-trained researchers about a variety of prosodic features in determining what constitutes a tone unit. Notably, Lintunen et al. (2016) reported that a small percentage of tone units (3.8%) were excluded from the analysis in their study because the researchers ($n = 3$) could not come to an agreement. The results are promising in that tone units may be even more accurate in measuring cognitive fluency, through the analysis of a tone unit as an expression of conceptualized utterance, than MLR. Overall, Lintunen et al.'s (2016) and Lin's (2018) studies show that although FS cannot be identified solely by prosodic features, and therefore, cannot yet be used as a reliable measure of fluent speech, the findings from these studies further reveal the strong interrelationships between FL, prosody, and fluency.

2.3.2.3. Intonation

Relatedly, as discussed previously, several studies have revealed the influence of intonation on fluency perceptions through both quantitative (e.g. Derwing et al., 2004, Rossiter, 2009) and qualitative (Préfontaine & Kormos, 2016; Götz , 2013; Freed, 1995) measures of monologic performances.

To investigate the relationship between intonation and fluency in dialogic settings, Wennerstrom (2000) investigated the role of tone boundaries in dialogic settings in producing fluent interaction between speakers. The author investigated L2 speakers' use of three types of tone boundaries during conversation with L1 speakers, categorized as follows: (1) high-rising pitch to signal an adjacency pair by another speaker; (2) low-falling pitch to signal finality and the end of a turn; and (3) low-rising pitch and level pitches (plateaus) to signal continuity within the speaker's discourse. The author compared the L2 speakers' use of these three types of tone boundaries with their fluency scores on a three-point rating scale derived from the SPEAK oral proficiency interview test. From this analysis, Wennerstrom (2000) revealed the following results:

More fluent speakers used plateaus and low rises on words and used plateaus on pause fillers to signal the intention to continue. Pauses following these tended to be tolerated without interruption. For less fluent speakers, we saw examples where inappropriate boundaries, that is, those that were too frequently placed, cut short, or of the wrong type, all led to interruption by native speakers (p. 124 - 125).

Wennerstrom (2000) concludes from these findings that "these components (tone boundaries) are not mere stylistic features; they are part of a grammatical system that encodes the cohesive relationships of spoken text" (p. 125) proposing that tone boundaries should be included in any future analysis of interactional fluency.

Pike (1972) describes how fluency and intonation, although inherently related to one another, are separate constructs, as they can potentially fulfill separate communicative functions. Pike argues that although intonation contours, created by variations in pitch over time, interact with other supra-segmental features of speech (speech rate, pauses, and rhythm), intonation must ultimately be regarded as a separate component. Although pauses may define intonation contours, in some cases, pauses may occur within contours or may not exist at the conjunction of two or more intonation contours in a 'complex rhythm unit' (Pike, 1972, p. 81). As Celce-Murcia et al. (2010) describe, if this unit is uttered rapidly, without any internal pausing, then it will create relatively few prominent elements. As a result, speech may be perceived as sounding smooth. However, "too many pauses and therefore intonation units can slow speech down to create too many prominent elements" (p. 232). Consequently, the speech may sound choppy.

2.3.2.4. Linking

Phonological linking, which is "the connecting of the final sound of one word or syllable to the initial sound of the next" (Celce-Murcia et al., 2010 p. 165) may also be a perceptually salient feature of fluent speech. The amount that English speakers link speech in any given situation depends largely on speech rate, speech style, and the formality of the situation (Celce-Murcia et al., 2010) as "un-emphasized words become more and more reduced as speech becomes more rapid and more informal" (Chun, 2012, p. 83). Thus, a more rapid speech rate may likely increase the amount of linking, whereas a slower speech rate may decrease the amount of linking; however, this relationship may not be entirely predictable because of variations in speech style and register (Celce-Murcia et al., 2010; Chun, 2012) and to the extent that linking occurs in the speakers' L1 (Hieke, 2005).

Only a few studies seem to have investigated the relationship between fluency and linking. Hieke (2005), for instance, argues that one type of linking, Consonant Attraction (CA), is a distinguishable feature of fluent speech in English. CA refers to the linking of the final consonant of an initial word/syllable to the initial vowel of a subsequent word/syllable resulting in a re-syllabification of the utterance. Hieke's comparisons of 12 L1 English speakers' and 29 advanced L1 German speakers' of English revealed the L1 English group actualized twice as many CA links as the L1 German group, at a rate of about 12 CA links/100 syllables. As Hieke observes, cross-linguistic differences are evident as linking in German is 'restricted' as compared not only to English, but also to French, in which linking is more highly prevalent, highlighting the impact of L1 transfer on perceptions of the smoothness of speech. Moreover, the amount of CA links is likely affected by the prevalence of pauses in the L1 German group's speech as a pause removes the possibility for a CA link, once again potentially affecting perceptions of smoothness.

In another study, Waniek-Klimczak (2014) argues that aspects of rhythm and connected speech are 'fluency factors' as they were shown to correlate with higher fluency ratings. Waniek-Klimczak also discusses how formulaic sequences, intonation, and linking combine to produce smoothly flowing speech as fully acquired formulaic sequences may contain elements of reduced speech that preserve their natural intonation contour. Waniek-Klimczak observes that the positive impact of linked speech, as created through the production of formulaic sequences within intonation units, on raters' judgements noting that the learners' "use of lexicalized chunks (e.g. 'kinda', 'gotcha' etc. which may give the impression of greater fluency for this speaker...(therefore) the strategy of using 'articulatory chunks' proves very successful in

speaking” (p. 179). Overall however, not much is known about the effects of linking on fluency perceptions as research has been limited thus far.

2.3.2.5. Summary

In summary, previous research has shown that a wide variety of temporal (speed, breakdown, and repair phenomena) and non-temporal (lexical resources, FL, intonation, linking) linguistic features may affect fluency perceptions to varying degrees. In upcoming sections, the effects of conversational features, including non-verbal features, on fluency perceptions will be discussed. In addition, certain listener characteristics, such as listeners’ degree of familiarity with speakers’ accents, may also affect fluency judgements to some degree.

2.4. Accent familiarity

As Derwing et al. (2004) state, perceived fluency is a perceptual phenomenon in the ear of the listener. It is therefore reasonable to assume that listeners’ degree of accent familiarity affects fluency perceptions to some extent. Several studies have investigated the relationships between accent familiarity and ratings of different aspects of L2 speech (Shintani, Saito & Koizumi, 2018; Saito & Shintani, 2015; Winke, Gass, & Myford, 2013; Carey, Mannell, & Dunn, 2011; Kennedy & Trofimovich, 2008). Overall, these studies show that listeners’ degree of accent familiarity seems to affect oral proficiency ratings. Recent research has also provided new insights into how accent familiarity affects fluency ratings specifically (e.g. Reid, Trofimovich, & O’Brien, 2019; Browne & Fulcher, 2017). However, research is generally limited in this area thus far.

2.4.1. The effect of exposure

The mere exposure effect phenomenon refers to the likelihood that gradually increasing exposure to a certain stimulus gradually increases liking of that stimulus (Van Dessel et al.,

2017). More specifically, in terms of L2 speech perception, Carey, Mannell, and Dunn (2011) contend that exposure positively affects one's "interlanguage phonology familiarity" (p. 204), explaining how intelligibility may vary from individual to individual and from community to community as "certain features of interlanguage pronunciation may be accepted by one community but may deviate from expectations in another" (p. 204). Kuhn's (1991) 'perceptual magnet effect model' explains how, over time, listeners store speech features associated as "the best instances of phonetic categories, or phonetic prototypes" (Carey et al., 2011, p. 201) which are then used as a reference point to decode similar-sounding speech. Exposure to non-prototypical speech features (e.g. L2 vowel sounds that are not exact replications of L1 vowel sounds, but are within the same acoustic space) cause the prototypical feature to, in a sense, pull the non-prototypical speech features towards it (Carey et al., 2011). In other words, through exposure, listeners become more tolerant of slight deviations to their stored prototypical sounds. As a result, accented speech becomes more comprehensible over time due to length of accent exposure.

This phenomenon has implications for perceptions of fluency. Although listeners have been shown to distinguish clearly between accentedness and fluency when rating L2 speech (O'Brien, 2014) and that heavily-accented speech may still be comprehensible (Munro & Derwing, 1995), fluency and comprehensibility still exhibit strong interrelationships (Saito & Shintani, 2015; Kang, 2012). Therefore, if listeners' degree of accent exposure affects judgements of comprehensibility, then it would likely affect judgements of fluency as well, as suggested by findings from a few recent studies (Reid et al., 2019; Browne & Fulcher, 2017).

2.4.2. Exposure and overall L2 speech ratings

Several studies seem to indicate that the degree of exposure to accented speech facilitates listeners' ease of processing (Bradlow & Bent, 2008), which, in turn, may positively affect L2 speech ratings. Ockey (2014), for instance, discovered that listeners who reported having less familiarity with speakers' accents assigned a higher degree of accentedness than listeners who reported having more familiarity. In a similar study, Ockey et al. (2016) discovered that post-secondary lecturers who spoke a variety of English reported to be markedly different from listeners' local variety, were reported to be much more difficult to comprehend, particularly when multiple words in an utterance were pronounced in a different manner. Winke et al. (2013) discovered that, on the TOEFL test of oral proficiency, L2 Spanish raters assessed L1 Spanish test-takers significantly more leniently whereas L2 Chinese raters assessed L1 Chinese test-takers significantly more leniently. In Carey et al. (2011), 99 IELTS raters assessed the pronunciation three groups of speakers (Chinese, Korean, & Indian) at various international locations. The authors discovered that raters awarded higher pronunciation ratings based on their level of phonological familiarity; moreover, test location was a factor as raters awarded higher scores to speakers from their home country. Similarly, O'Brien (2014) revealed that listeners with higher levels of German proficiency were more favorable in all areas (accentedness, fluency, and comprehensibility) to L2 English learners of German than listeners with lower levels of German proficiency.

Browne and Fulcher (2017) examined the relationship between raters' L1 Japanese familiarity and fluency ratings on the TOEFL iBT test. Multi-faceted Rasch analyses revealed that raters who were more familiar with Japanese were significantly more lenient towards assessing L2 Japanese speech fluency. The authors contended that any construct definition of fluency should consider the potential impact of listener characteristics. One notable point about

this study is that although the authors attempted to examine the effects of accent familiarity on fluency specifically, the TOEFL rubrics for fluency refer to “Delivery”, which includes descriptors referring to phonological and grammatical accuracy, not just fluency.

Taken together, the studies mentioned above would lead one to believe that L2 speakers may be more favorable towards speakers of the same language; however, this may not necessarily be the case. Xi and Moulann (2009) found that bilingual (Hindi, English) speakers did not assign higher ratings to L1 Hindi/L2 English speakers. Similarly, Derwing and Munro (2009) could not claim any instance of L1 bias in their investigations of L1 Japanese speakers’ and L1 Cantonese speakers’ respective ratings of L2 Japanese speech and L2 Cantonese speech. Overall, whether an L1 bias exists or not, these studies provide support for the contention that exposure seems to breed liking (i.e. the mere exposure effect) when it comes to making judgements about L2 speech.

2.4.3. Teaching experience/training and L2 speech ratings

Studies investigating the role of teaching experience and linguistic training on raters’ global assessments and analytic assessments of fluency, accentedness, and comprehensibility have produced mixed results. On the one hand, some studies (e.g. Huang and Jun, 2015) discovered that inexperienced raters were more severe than experienced raters in assessing L2 speech globally. Similarly, Kennedy and Trofimovich (2008) also discovered that experienced listeners assigned higher intelligibility ratings to both the L1 and L2 groups than the inexperienced listeners did; yet there were no differences in terms of comprehensibility of accentedness ratings. Isaacs and Thomson (2013) did not find any significant differences between experienced and inexperienced raters in their assessments of accentedness, comprehensibility, and fluency.

What is more interesting is how the degree of experience/training may affect the relative saliency of certain speech aspects. In Derwing and Munro's (2009) study, persons with no language teaching/testing experience have been shown to prioritize fluency and comprehensibility when it comes to making preferential judgements about certain L2 speech accents. Other studies have revealed similar results. Huang (2013, 2016), for instance, discovered that inexperienced L2 instructors were more attentive to speech content whereas experienced L2 instructors were more attentive to accent, grammar, and vocabulary. In another study, Dujim, Schoonen, and Hulstijn (2018) revealed that raters with linguistic training attended more to accuracy whereas raters without linguistic training attended more to fluency in making oral proficiency judgements about L2 Dutch speech. Listeners' age may also affect which speech aspects receive the most attention. Reid et al. (2019) recently discovered that younger L1 English raters assigned higher ratings of flow and intonation to L1 French/L2 English speakers whereas older raters were less favorable in their ratings of flow. Shintani, Saito & Koizumi (2018) revealed that L2 raters who were more proficient in English at an earlier age were more severe in their judgments of L2 accentedness and comprehensibility. It would seem then, that for a non-language practitioner, who may have a limited degree of exposure to L2 speech, the interrelated constructs of fluency and comprehensibility are the most salient aspects of speech.

Inner-circle/outer-circle speaker status may also affect L2 speech ratings. Speaker status, according to Kang (2012), refers to Krachu's (1988) classifications of English varieties according to inner-circle (e.g. Canadian, Australian) English varieties, and outer-circle (e.g. Japanese, Singapore) English varieties. Kang (2012) discovered that speaker status affected between 7 to 9% of the score variance in oral proficiency ratings, which means that inner-circle raters were slightly less lenient when rating L2 speech than were the outer-circle raters.

Similarly, Saito and Shintani (2015) analysed how two groups of raters, L1 Canadian English raters ($n = 10$) and L1 Singapore English raters ($n = 10$), assessed the comprehensibility of L2 Japanese speech. Comprehensibility ratings were subject to correlational analysis with utterance measures of fluency, pronunciation, grammar, and lexis. The results indicated that Singaporean raters were significantly more lenient than Canadian raters when assessing the comprehensibility of L2 Japanese speech. Moreover, certain measures of fluency (speech rate) and phonology (segmentals) were the most influential in affecting both raters' assessments of comprehensibility, whereas measures of grammatical accuracy affected Canadian raters' judgements and lexical appropriateness affected Singaporean raters' judgments to a lesser degree. The authors speculate that these differences may be attributed to accent familiarity, commenting that "because the Singaporean raters must have accumulated a great deal of experience in decoding and processing such non-native-like speech signals, they may be able to attend to the universal characteristics of non-native speech in Japanese-accented English and understand it with relative ease" (p. 438).

2.4.4. The interlanguage intelligibility benefit

This previous statement from Saito and Shintani (2015) refers to an "interlanguage speech intelligibility benefit" (Bent, & Bradlow, 2008), which according to Kang et al. (2019), may advantage L2 listeners over L1 listeners in terms of comprehending L2 speech. When listening to L2 speech, L2 listeners "may be especially well-tuned to specific acoustic-phonetic features of an L2 that reflect features of their own L1" (p.4), such as the absence of vowel reduction in Chinese, Japanese, and Spanish. Stringer and Iverson (2019) provide some support for this under-researched hypothesis. In their study, 16 English and 16 Spanish raters listened to the speech of Southern British English, Glaswegian English, and Spanish-English speakers in a speech recognition task. The results revealed that both phonetic similarity and accent familiarity

affected raters' general ability to recognize utterances; yet individual variations existed and phonetic similarity could not fully explain rating variance. Munro, Derwing, and Morton (2006) also found inconsistencies in the effect of accent similarity as although Japanese listeners understood Japanese-English slightly better than the L1 English listeners did, the Cantonese listeners did not understand the Cantonese-English speakers better than the L1 English listeners did.

2.4.5. Listener expectations

Listeners who expect to hear accented speech may be more severe in their ratings (Kang & Rubin, 2009). Ockey et al. (2016) discovered that post-secondary lecturers who spoke an unfamiliar variety of English (as reported by students) were much more difficult for the students to comprehend. Notably, listeners in this study reported that they expected that they would not understand the lecturer comfortably. Relatedly, listeners in Derwing's et al (2002) study reported 'zoning out' because they did not expect to understand disfluent L2 speech. Expectations of miscomprehension may affect listeners' willingness to attend to speech because, as Munro, Derwing, and Morton (2006) explain, a "lack of familiarity might make people apprehensive about their own abilities, which might lead to their not paying attention to accented speech because they are convinced that they will not understand it" (p. 128).

2.4.6. Exposure training programs

Rater training programs that include the provision of exposure to a variety of different accents have been widely recommended to help negate the effects of a lack of familiarity (Kraut & Wulff, 2013; Carey et al., 2011; Derwing et al., 2002). In Derwing et al.'s (2002) study, for instance, rater training improved listeners' confidence in making judgements about L2 speech, as well as enhancing their empathy for L2 speakers. Relatedly, Kraut and Wulff (2013) discovered

that L2 exposure and training improved 78 L1 English listeners' attitudes towards L2 speech, as produced by L1 speakers of Chinese, Spanish, and Farsi. The effects of training may be immediate but less is known about their lasting impact. For example, Wittenman et al. (2013) reported that L1 Dutch listeners adapted to L2 Dutch speech within only a few minutes of being exposed to priming tasks that focused listeners' attention on segmental differences in heavily accented words. It is unknown however, whether this adaptation is long lasting.

Jiang, Sandford, and Pell (2018) discovered that, in interview scenarios, L2 speakers who more frequently displayed vocal features associated with confidence (i.e. vocal confidence cues) such as increased volume and a more dynamic pitch range, were more positively perceived despite their level of accentedness. From this, one can infer that training programs should inform instructors about helping students acquire these 'vocal confidence cues' to offset any potential accent familiarity biases.

2.4.7. Intergroup attitudes

Despite all of this research, negative intergroup attitudes may override the positive effects of exposure. As Reid et al. (2019) discovered, social biases may affect raters' fluency assessments, in spite of how familiar listeners are with certain accents. In this study, 60 novice raters were divided into three groups: positive bias, negative bias, and neutral. With the exception of the neutral group, each group was exposed to anecdotes that reinforced positive or negative social biases immediately prior to rating L1 French/L2 English speakers' speech samples for flow (i.e. fluency), intonation, segmental accuracy, accentedness, and comprehensibility. The authors discovered that the social manipulation biasing affected ratings of all five aspects, including flow. These results indicate that accent familiarity may not necessarily lead to enhanced fluency ratings due to the effects of social biasing.

It is therefore reasonable to assume that intergroup attitudes may play a role in affecting fluency perceptions. As Jenkins (2007) claims, attitudes about language may reflect attitudes towards users of that language and vice-versa; thus, language-specific characteristics have an effect on how, not just the speech, but also how the speaker is perceived. For instance, Hindi speakers typically use falling intonation rather than rising intonation when posing questions, which may give the impression that the speaker is rude, which is not actually the case (Celce-Murcia, 2010). Since L2 speech, as measured by temporal fluency features such as speech rate and pause phenomena may be highly influenced by the L1 (Segalowitz, 2016), perceptions of fluency may also be affected by intergroup biases.

Speed of speech, as measured by speech rate or articulation rate, is commonly believed to vary across languages. However, whether language-specific speech rate differences really exist is under debate. There is some research demonstrating the existence of cross-linguistic differences in speech rate, such as between Spanish and English, as group differences in the average range of speech rate has been shown to be higher in Spanish than in English (Pellegrino, Coupe, & Marsico, 2011). Within-language differences may also exist, such as between Northern U.S. (Wisconsin) and Southern U.S. (North Carolina) speakers of English, as collectively, the Wisconsin speakers in the study were shown to produce a higher articulation rate than the North Carolinian speakers did (Ewa Jacewicz, Fox, & Wei, 2010). To complicate matters further, speech rate differences may exist across age groups as younger speakers may speak more quickly than older speakers (Quené, 2008). On the other hand, other studies have revealed no significant differences in temporal fluency measures across languages (Kowal, Eeisse, & O'Donnell, 1983) and have provided mixed results regarding age group differences (Ewa Jacewicz et al., 2010).

However, whether or not some languages are inherently faster or not is subject to debate (Roach 1998). As the author observes, perceptions of speech rate are highly influenced by stereotypes and intergroup biases about speakers of those languages/dialects. For instance, slower speech may be affiliated with less intelligence whereas faster speech may be affiliated with higher intelligence. Languages may be judged as ‘fast’ languages either because they are unfamiliar and less intelligible to the listener and/or because of stereotypes about its speakers. Therefore, speech rate differences across languages may be a myth. Moreover, it is also problematic to compare speech rates by counting syllables. For instance, English syllable structure allows for more consonant clusters than the Spanish syllable structure. Therefore, Spanish speakers may produce more vowels, and thus more syllables. In turn, measures of English speech rate may be lower according to syllable counts.

Furthermore, differences in mean syllable duration exist between stress-timed languages, such as English, and syllable-timed languages, such as Spanish. Wennerstrom (2001) defines a stress-timed language as “a language for which the stressed syllables tend to align rhythmically (and a syllable-timed languages as) “a language for which syllables tend to be rhythmically aligned” (p. 275). As Chun (2012) describes, in syllable-timed languages, syllables are of relatively equal duration. Yet, syllable duration in stress-timed languages is much more variable as syllables on the stressed syllable of content words (e.g. nouns, verbs, adjectives, adverbs) are typically longer than syllables within function words (e.g. prepositions, determiners) which feature the highly reduced vowel, the schwa – notably however, function words may also be stressed for emphatic or discourse cohesion purposes. Overall, comparing measures of syllable counts and syllable duration across languages is somewhat problematic.

Regardless, there is some research indicating that shows cross-linguistic differences in speakers' use of silent and filled pauses. Grosjean and Dechamps (1975) comparisons of French and English revealed that the English speakers pause more frequently but for a shorter duration. These L1 features may influence L2 production as speakers may transfer their 'pause profiles' (i.e. frequency, length, and distribution tendencies) from their L1 to their L2 (Raupach, 1980). Silent pauses may also fulfill meaningful language/culture-specific functions in conversation. Young and Hallebeck's (1998) comparisons of Japanese and Mexican-Spanish speakers of English revealed that the Japanese speakers paused more at initial-turn position, which is believed to be a feature transferred from the L1 to indicate thoughtfulness. On the other hand, the Mexican-Spanish speakers paused less frequently between turns in order to create rapport through conversational cohesiveness. These features may affect L2 oral proficiency ratings, as raters may perceive turn-initial pauses in Japanese L2 English speech more negatively; for instance, Sato, (2013) discovered that raters inferred from these turn-initial pauses that learners were not able to comprehend the previous speaker or that they were disinterested in the conversation. Sato (2013) also discovered however that the Japanese learners, upon reviewing their performances, did not perceive these turn-initial pauses as being problematic.

2.4.8. Summary

To conclude this section, although there is much research on the effect of accent familiarity on oral proficiency ratings in general, there is limited research (with the exception of the limited findings found in Reid et al., 2019; Browne & Fulcher, 2017; O'Brien, 2014) on the effect of accent familiarity on fluency ratings more specifically. Overall, intergroup attitudes may affect fluency ratings in two distinct, but related, ways: First, attitudes towards speakers of a language reflect attitudes towards the language itself and vice versa (Jenkins, 2007). Second,

temporal fluency features may be language-specific, and thus influence listeners' intergroup attitudes. Several studies have shown that attitude change is possible through increased exposure and training (Kraut & Wulff, 2013; Derwing et al., 2002); yet there is no indication that these rater-training programs address the prevalence of fluency features transferred from the speakers' L1.

2.5. Conversational fluency

The present study is unique in that it focuses on fluency as elicited through a conversational, rather than an individual (i.e. monologic) speech task. As Lennon (2000) contends, temporal measures, as indicators of lower-order fluency, represent only the tip of the fluency iceberg; therefore, the attainment of higher-order fluency may be dependent upon lexical sophistication and other factors. Arguably, one's level of interactional competence is requisite to being conversationally fluent as certain interactional skills may be necessary to facilitate and support the mutual flow of speech between speakers. Through this lens, whether fluency-within-conversation be termed 'conversational fluency' (Riggenbach, 1991), 'pragmatic fluency' (House, 1996), 'communicative fluency' (North, 2000), 'interactional fluency' (Sato, 2014), or 'dialogue fluency' (Peltonen, 2017), fluent conversation has been theorized to be inherently co-constructed by all participants to some degree (Sato, 2014). As indicated in the studies listed above, this concept has been discussed for decades in both the second language acquisition and language testing literature.

The following discussion of conversational fluency is divided into several sections. The first section provides a brief discussion of interactional competence as a theoretical construct underpinning the notion of conversational fluency. The second section provides a discussion of non-verbal communication as it relates to perceptions of interactional competence and to the

production and perception of conversational fluency. The third section provides a review of studies that have investigated the production and perception of conversational fluency either through quantitative or qualitative means. The fourth and final section discusses Tannen's (2005/2005) concept of conversational style, discussing its potential role in affecting the production and perception of conversational fluency.

2.5.1. Interactional competence

Kramsch (1986) is credited with first coining the term 'interactional competence', (IC) describing it as " the construction of a shared internal context or 'sphere of inter-subjectivity' that is built through the collaborative efforts of the interactional partners" (p. 367). Notions of IC draw from socio-cultural theories of learning and development (e.g. Lave & Wegner, 1991; Hitchens, 1995), centering on the notion that cognition does not reside within the individual but that it is socially constructed through interaction. Sato (2014) applies this concept to the assessment of fluency by defining "the construct of interactional fluency" as a "joint performance between learners" (p. 88), reflecting He and Young's (1998) perspective on IC that oral performance is "co-constructed by all participants in the interaction" (p. 7). The strong version of IC suggests that the construct to be measured in oral proficiency tests is the interaction itself, which may be problematic to operationalize because "if there are no consistencies in performance, any attempt to generalization is suspect; if there are consistencies, we are unable to explain or interpret them" (Bachman, 2007, p. 62). Alternatively, a more moderate version of IC, as posed by Chalhoub-Deville (2003), posits that the learner's ability rests within the individual itself but is largely mediated by the context; thus, the construct to be measured is "an ability - in language user - in context" (p. 572). Through these perspectives, it is conceivable that fluency

theoretically exists neither solely within the speaker and/or within the listener (Freed, 2000; Derwing et al., 2004), but that it is co-constructed in the space between interlocutors.

Understanding the challenges in assessing conversational fluency through an IC lens requires an understanding of the on-going challenges in assessing IC on tests of general oral proficiency. Galaczi and Taylor (2018) provide an overview of the ongoing challenges underpinning the operationalization of this theory into testing rubrics. The authors provide a definition of IC, a framework of macro- and micro-level strategies to operationalize this definition, and a list of on-going inherent issues and challenges. The authors refer to IC as “a socio-cognitive construct” (p. 225), which can be defined and operationalized as follows:

(Interactional competence) is the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event. This ability is supported by the linguistic and other resources that speakers and listeners leverage at a micro-level of the interaction, namely, aspects of topic management, turn management, interactive listening, breakdown repair and non-verbal or visual behaviours (p. 226).

According to Galaczi and Taylor (2018), these macro-skills are further sub-categorized as follows: (1) *Topic management*: initiating, extending, starting, and closing; (2) *Turn management*: starting, maintaining, ending, pausing/latching/interrupting; (3) *Interactive listening*: providing backchannels (continuors, comprehension checkers); (4) *Breakdown Repair*: joint utterance creation (self/other; recasts); and (5) *Non-Verbal Behaviour* (eye contact facial expressions, laughter, and posture). The authors also recognize that the construct of IC may consist of an observable awareness of genre and politeness strategies. Finally, the authors discuss the fundamental challenges inherent to the concept of IC; more specifically, since assessment of

this construct is highly mediated by contextual variables, concerns about variability are ever-present. Testers seek to reduce variability in order to increase reliability and validity; yet variability in topics, situations, and interlocutors are an inevitable part of real-world interaction, which test tasks are supposed to be designed to emulate.

Ongoing research is required to understand the extent to which test tasks measure IC as a single unitary construct. Batenburg et al (2018) addressed this challenge by investigating the extent to which six structured speaking tasks could measure IC in a valid and reliable manner.

The results are as follows:

1. Strong correlation coefficients were found between interactional performance ratings and analytic measures of two types of interactional skills, as identified by Kormos (2006): *self-supporting strategies* (compensation strategies; and meaning negotiation strategies) and *other-supporting strategies* (response to clarification requests and response to misinterpretation of the message).
2. Strong correlations were discovered between all six task types and ratings of both interactional skills and linguistic performance on both analytic and holistic scales.
3. Holistic and analytic scores were found to be highly correlated, suggesting that the interactional skills (self- and other-supporting strategies) represented by analytic measures are part of the IC construct.

In sum, the results indicate that interactional performance, although subject to a wide variety of contextual constraints, may be measured validly and reliably across a set of speaking tasks.

On the other hand, however, some research has shown that different interactional patterns may be differentially favored by raters. Galaczi (2008) identified four different kinds of

interactional patterns within L2 conversational speech. Then, they analysed raters' judgments of interactional competence according to the following patterns: (1) asymmetric: one participant speaks much more than the other; (2) parallel: each participant provides relatively equal long turns, but they do not interact substantively; (3) collaborative: participants provide relatively equal turns and they interact substantively; (4) blended: a mixture of the latter two patterns. Galaczi discovered that scores were highest for the collaborative group and lowest for the parallel group, whereas the mean scores for the asymmetric and blended groups fell in between. These results indicated that raters react more favourably towards conversations that are balanced but highly interactive, as evidenced through a substantial amount of turn taking.

Another challenge regarding IC assessment is how practitioners can adequately measure and assess a substantive response to initial-speakers' statements. Lam (2018) addressed this issue through conducting a conversational analysis of Hong Kong secondary school English learners' speeches to identify which speech features substantiate an interactionally competent response to initial-speakers' turns. Lam framed her analysis around three conversational actions: formulating, accounting, and extending. She discovered that the following features of speech constitute substantive responses to initial-speakers' turns: (a) formulating previous speaker's contributions through paraphrasing or summarizing; (b) accounting for (dis)agreement with previous speakers' ideas by providing reasons for agreeing/disagreeing; and (c) extending previous speakers' ideas by providing examples or supportive arguments. In all, Lam (2018) contends that her paper "has demonstrated how test-takers discursively construct their IC through producing responses contingent on previous speaker contribution, and has argued for the inclusion of this feature in the construct of IC within the context of paired/group speaking assessments" (p. 395).

2.5.2. Non-verbal communication (NVC)

As Plough, Banerjee, and Iwashita (2018) discuss, one of the most under-researched areas regarding the assessment of IC is the role of NVC in affecting interaction. As Kendon (2004) describes, any speech utterance is an ensemble of both oral and visual activity, which together communicate information. Yet, to what extent NVC expresses one's interactional competence and thus contributes to one's ability to be fluent within an interaction is largely unknown. Most of the focus of L2 pedagogy, research, and assessment has been predominantly centered on verbal aspects of communication.

There are several challenges to assessing NVC. Firstly, one needs to consider which features of NVC are most salient to performance. NVC encompasses a wide variety of actions including gestures, posture, eye contact, facial expressions, among others (Plough et al. 2018). Additionally, there may be considerable differences in the use of NVC within individuals, across contexts, and across cultures. These challenges may therefore explain why little is known about how NVC shapes and reflects perceptions and assessments of IC (Plough et al., 2018). More specifically, little is also known about the role of NVC in shaping and reflecting the perceptions of fluency (Götz, 2013). This section provides a brief review of pertinent studies exploring the relationships between NVC and perceptions of IC and fluency.

Several studies have examined listeners' perceptions of IC in regards to NVC (e.g. Jenkins & Parra, 2003; DuCasse & Brown, 2009). DuCasse and Brown (2009), for instance, grouped raters' observations about learners' performances on paired conversational tasks into three key categories: *non-verbal interpersonal communication*, which included comments regarding gaze and body language; *interactive listening*, which included NVC; and *interactional management*, which referred to the ways in which partners directed the interaction. Jenkins and

Parra's (2003) analysis of examiners' written and oral comments on L2 speaking tests revealed that eye contact, posture, and the use of specific gestures for turn-taking and active listening (head nodding and providing backchannels) aided less L2 proficient students in passing the test despite their lack of language proficiency. However, proficient L2 students who may have not have demonstrated an active use of NVC also passed the test, indicating that NVC may enhance perceptions of OC in spite of a lack of linguistic proficiency. Taken together, these studies reveal the importance of NVC on affecting listeners' perceptions of interactional competence.

2.5.3. NVC and fluency

The relationship between NVC and fluency has received little attention in the literature. Poyatos' (2005) argues that acquiring cultural fluency, which refers to the speakers' ability to interweave between cross-cultural patterns of discourse, requires acquisition of culturally specific non-verbal communicative features. For Poyatos, "fluency is a dimension central to any interactive encounter, whether intercultural or intracultural, conversational or non-conversational" (2005, p. 456) and that "nonverbal behaviors may act positively or negatively during the encounter" (p. 457). In an empirical study on the effects of NVC on fluency, Bavelas (2000) analysed video-recorded conversations among peers and observed the salience of non-verbal features on fluent production concluding that "non-redundant nonverbal acts supplement words so that, taken alone, the words would seem inadequate or even disfluent" (p. 100). Influenced by these findings, Götz (2013) sheds light on its importance by categorizing non-verbal fluency as a distinct kind of fluency, separate from productive (characterized by utterance measures) and perceptual (characterized by rating assessments) fluencies.

L2 gesturing seems to be the most salient feature of NVC (e.g. May, 2011; Nakatusara, 2011; Gullberg, 2008; Stam & McCafferty, 2008). Different classifications of gestures exist

within the literature. For example, whereas Guillberg (2008) broadly classifies gestures as being either functional or symptomatic (e.g. scratching), Stan and McCafferty (2008) identify five types of gestures applicable to L2 communication:

1. Emblems – culturally specific gestures.
2. Illustrators – movements that are directly tied to speech, which [...] [are] divided into six categories: batons, ideographs, deictic movements, spatial movements, kinetographs and pictographs.
3. Affect displays of the face – facial expressions that show emotion.
4. Regulators – movements that are involved in conversation and turn-taking; and
5. Adaptors (self-adaptors, alter-adaptors, and object adaptors), – movements involved in self-grooming, interpersonal contact, and related to an instrumental task (p. 5).

Similarly, Plough et al. (2018) list four key characteristics of gestures relevant to the assessment of L2 speech:

1. Gestures have multiple individual functions - therefore, there is not necessarily an equal correspondence between form and function.
2. Gesture and speech exhibit a complex interrelationship.
3. Gestures serve two overall functions: self-directed (e.g. lexical searching) and other-directed (turn-taking and providing backchannels).
4. Generalizations about gestures can be made but there is also plenty of individual variation.

The third characteristic from Plough et al.'s (2018) list appears to be most relevant to the study of fluency perceptions. Gestures used for self-directed functions may provide visual cues

of online processing whereas gestures used for other-directed functions may provide insights into users' levels of IC and/or conversational fluency.

As for Stan and McCafferty's (2008) list, regulators are most relevant to the study of conversational fluency perceptions. Regulators are "non-verbal habits...which direct the back-and-forth nature of speaking and listening" (Rowe & Levine, 2015, p. 322), which include hand movements, eye gaze, and head nodding, which may be used in a variety of ways and combinations to manage turn taking.

According to Rowe and Levine (2015), eye gaze serves different functions within conversations, one of which is to signal to the speaker to start, continue, or stop talking. "A pattern of gazing and gazing away, as well as the length of a gaze give subconscious cues to the interactants about when it is time to start or stop talking" (p. 325). One's eye gaze may be influenced by whether or not the speaker is discussing concrete or abstract ideas as the speaker may be more likely to turn away or even close their eyes when discussing abstract or cognitively challenging concepts (Rowe and Levine, 2015). However, the use of eye contact is cultural-specific and the appropriateness of its use is determined by one's and the interlocutors' gender, age, and social status. Brown (2008) illustrates these differences in the following example:

In American culture it is permissible, for example, for two participants of unequal status to maintain prolonged eye contact. In fact, an American might interpret lack of eye contact as discourteous lack of attention, while in Japanese culture eye contact might be considered rude. Intercultural interference in this nonverbal category can lead to misunderstanding (p. 239).

Ultimately, what is determined to be supportive body language is inherently affected by the listener's beliefs and values about the use of regulators such as eye gaze within a conversation.

Wolf (2008) discovered the positive effects of head nodding on fluent speech in his study involving 14 Japanese learners of English in three different back-channeling conditions: verbal/non-verbal, including the use of 'uh-huh'; non-verbal, which only included head nodding; and no backchannels. Learners were most fluent in the verbal/non-verbal condition, the second most fluent in the non-verbal (head-nodding) condition, and least fluent in the no-back-channeling condition. Although, levels of significance were not reached between the non-verbal and the no-backchannels condition, the results highlight the emphasis of using head nodding to contribute to the interlocutor's speech, and possibly, the mutual flow of speech between interlocutors.

Cross-cultural differences in the use of head nodding may be problematic however. For instance, Cutrone's (2014) analysis of 30 conversations between L1 English-speaking Americans and L1 Japanese speakers of English revealed that cross-cultural variation in the function, distribution, and frequency of verbal and non-verbal backchannels led to negative cross-cultural perceptions across groups.

Proxemics, the use of space, is another aspect of NVC than may affect interlocutors' willingness to engage in the conversation. Leaning forward in conversation may be perceived positively as it closes the physical distance between interlocutors; however, as Rowe and Wharton (2015) explain, although "the use of space is important in regulating interactions" (p. 332), the proximal distance between interlocutors is culturally-influenced and inherently determined by interlocutor characteristics, including one's familiarity with the interlocutor, and

the formality of the situation. In other words, the appropriateness of closing the space is mediated not only by the interlocutors themselves but by any third-party raters assessing conversational fluency.

2.5.4. Measuring conversational fluency

Typically, in studies on fluency, the majority of elicited speech is monologic for several reasons (Wood, 2010). For one reason, the emphasis in fluency research is primarily on temporal features and not on interactive features, and monologic tasks generally elicit more temporally fluent production than dialogic tasks. For another reason, monologic speech samples enable researchers to better standardize the speech data for comparability purposes. As conversational data may be unpredictable, monologic speech samples allow researchers to examine how learners perform similar tasks under similar conditions. However, researchers have become increasingly interested in measuring the production and perception of fluency on interactive tasks both quantitatively and qualitatively, as depicted in this review of selected studies.

Riggenbach (1991) conducted the earliest known study on perceptions of L2 conversational fluency. The author provides a detailed qualitative and quantitative analysis of conversations produced from four L2 English learners. The author conducted a quantitative analysis of the role of interactive phenomena/features, specific only to dialogic settings: backchannels, echoes, questions, repair initiations, laughter particles, latches, overlaps, gaps, and collaborative completions. Her quantitative analyses revealed that these interactive features did not much influence raters' perceptions of conversational fluency; yet the author was careful to highlight the need to examine these features with greater sample sizes. Her qualitative analysis, however, indicated that the following features may be still be integral to the construct of conversational fluency: a) initiating topic changes; b) using backchannels and substantial

responses to carry conversational weight; c) anticipating turn-endings through latches and overlaps; d) producing a relatively equal amount of speech as the conversation partner. From her findings, the author rank-ordered the saliency of speech features in affecting listeners' perceptions of fluency: "(1) Frequency, placement, and degree of chunking and type of filled and unfilled pauses; (2) rate of speech and (3); frequency and function of repair" (p. 438-439). The author also noted that her analysis is complicated by the contextually dependent nature of conversations.

Sato (2014) investigated the impact of interactive features (e.g. turn taking) on fluent production during different modes of interaction (learner vs. learner/learner vs. interviewer) and whether fluent performance on individual monologic tasks could predict fluent ability on interactive tasks. To answer the first question, the author elicited speech performances from 56 Japanese English language learners on individual picture description tasks and paired (learner-to-learner) decision-making tasks. Four experienced raters then listened to 16 speech samples of individual tasks and eight samples of learner to learner interaction tasks and then, through the think-aloud procedure, discussed the fluent quality of each sample. The author then transcribed and coded the raters' recorded discussions, creating categories reflective of typical fluency measures (e.g. pausing, ease of speech) but also two interactive measures: turn-taking and scaffolding. The author then used this coded data to create band descriptors for two separate four-point fluency rating scales for individual and interactive tasks. The raters then used these scales to rate the speech performances. The rating scores were then correlated with two measures of speech rate: pruned and unpruned. The descriptors included the following descriptive phrases related to the 'turn-taking' category (e.g. 'hesitant to take turns', 'engaged in conversation', 'replies quickly') and scaffolding (e.g. 'scaffolding conversation'). From comparing temporal measures

across the individual and interactive tasks, Sato concluded that individual and interactive fluency are two separate constructs.

Peltonen (2017a) framed certain conversational fluency measures (self-repair, other-repair, repetitions, and filled pauses) as problem-solving mechanisms used to sustain the flow of conversational speech. The author analysed one-minute extracts from monologues and dialogues produced from 42 Finnish L2 English language learners for traditional monologue fluency measures (e.g. speech rate), dialogue fluency measures (e.g. collaborative completions), stalling mechanisms (e.g. repetitions), and communication strategies (e.g. paraphrases). The author also discovered that learners were more fluent on the dialogic than the monologic tasks, according to temporal analyses. Peltonen found that repetitions and fillers were used more strategically by fluent speakers to maintain one's turns in both monologues and dialogues. Moreover, less fluent speakers allocated attention to planning their own turns within dialogues whereas fluent speakers were able to allocate attention to their interactive strategies to maintain dialogic flow. The author concluded that interactive aspects are important aspects of fluency with implications for definitions, measurements, and assessments of fluency.

In another study, Peltonen (2017b) conducted a qualitative analysis of dialogic problem-solving tasks, elicited four L1 Finnish young learners of English, revealing the contributions of collaborative completions and other-repetitions to the joint construction of conversational fluency. These features were found to create cohesiveness between partners; in particular, collaborative completions were found to have a scaffolding effect as these next-speaker completions helped the initial speaker compensate for gaps. The findings highlight the importance of inter-speaker alignment and accommodation in contributing to the mutual flow of speech. It is likely however, that the presence of these features may be characteristic of the task

itself; in other words, collaborative completions and other-repetitions may be better elicited by problem-solving tasks than information-exchange tasks such as the one used in the present study. Therefore, the findings from this study may not be generalizable beyond problem-solving tasks.

Gambits, which are multi-word sequences used to organize interactions by “linking turns to the previous or the next one, or clarify or modify the interactional content of the current turn” (Luoma, 2004, p. 91), may be another indicator of conversational fluency. House (1996) investigated the effects of several types of gambits:

1. *Up-takers* (e.g. “yes”, “oh”, and “go on”), which are typically in turn-initial position to respond to the previous speakers’ turn;
2. *Clarifiers*, which are used to cajole (promote agreement) (e.g. “I mean”) or underscore (highlight importance) (e.g. “the point is”) information and which may occur in any position;
3. *Appealers*, which are used to promote agreement in post-position (e.g. “right”, “okay”)
4. *Starters*, which are used to initiate topics (e.g. “well”, “now”)

House (1996) discovered that explicit instruction of up-takers and starters was able to increase L1 German learners’ ‘pragmatic fluency’ (a.k.a. conversational fluency) by using these devices to initiate and change topics and provide more substantive turns, whereas students who were not exposed to this explicit instruction were unable to do so, and thus perceived as less pragmatically fluent.

Baron and Celya (2010) used the same qualitative measures as House (1996) to analyse the pragmatic fluency development in Spanish-Catalan young learners of English by comparing the use of pragmatic formulas by four groups at different stages in their development: Group 1 (age 10); Group 2 (age 12); Group 3 (age 15); Group 4 (age 17). However, one key difference

between this study and House's (1996) is that gambits were not taught explicitly. The participants learned English in EFL instructional settings but they were not exposed to explicit instruction of pragmatic routines. The author discovered qualitative differences in the use of pragmatic formulas across age groups and attributed these findings to incidental processes of pragmatic development in the L1 and general development in the L2. Although these results indicate that gambit acquisition may occur incidentally without explicit instruction, these results, more pertinently, indicate the importance of using gambits in order to be conversationally fluent.

The frequency, function, and location of discourse markers may be important indicators of conversational fluency. Hasselgren (2002), for instance, investigated the frequency, function, and location of smallwords, which may be regarded as a sub-category of discourse markers, in contributing to fluent speech. Hasselgren defined smallwords as follows: "small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself" (p. 150). Adhering to this definition, the author identified seventeen examples of smallwords: "well, right, all right, okay, you know, you see, I know, I see, oh, ah, I think, I mean, like, sort of/kind of, a bit, just, or something, and everything/and that/and stuff/and things" (p. 151). To investigate the contribution of smallwords to fluent speech, Hasselgren (2002) developed a fluency rating scale from a corpus analysis of young learners' (ages 14 to 15) performances. Raters placed learners into three groups: (1) less fluent; (2) more fluent; and (3) L1 speakers. Hasselgren hypothesized that L1 speakers and more fluent L2 speakers would use more smallwords than less fluent speakers. The author used corpus analysis to identify the frequency and location of smallwords as produced per participant groups. The author also provided a descriptive analysis of two temporal variables (MLR, frequency of filled pauses) per group, as categorized by a three-point fluency

rating scale. The results indicate that the most fluent group produced greater MLR while also using small words more frequently, especially in turn-initial position.

Crible (2017) conducted a corpus analysis investigation of the location and function of discourse markers (DM) in L1 English and L1 French speech in various interactional settings (e.g. conversations, interviews) and transactional settings (e.g. political speeches). Discourse markers include Hasselgren's (2002) criteria mentioned above but they also include connectors (e.g. "and"), subordinators (e.g. "because"), and transition phrases (e.g. "on the other hand") among others. The author discovered that DMs may co-occur with silent pauses, repetitions, and self-corrections depending on the function of the DM. Ideational DMs, used, for instance, to express 'cause' (e.g. "because") do not tend to co-occur with pauses or other disfluencies, and, as such, are generally integrated into the smooth flow of speech. Rhetorical DMs used, for instance, to express 'emphasis' (e.g. "you know") tend to co-occur with repetitions and restarts. Sequential DMs, used, for instance to express 'topic shifting' (e.g. "on the other hand") co-occur frequently with pauses. Interpersonal DMs, used, for instance, to express agreement/disagreement co-occur with false starts and truncations. Overall, these findings indicate that ideational DMs positively enhance the smooth flow of speech whereas the three other kinds of DMs are linked to speech disfluencies. The author also indicated that task characteristics such as planning and interactivity affect DM types and frequencies.

In a similar study, Crible and Pascaul (2019) provided a cross-linguistic corpus analysis of DMs used in English, French, and Spanish and their co-occurrence with repairs and repetitions. The authors discovered that 85% of repairs co-occurred with DMs across all three languages; moreover, in most cases, these repairs/DMs fulfilled a rhetorical speech function such as

emphasis or reformulation. The authors recommend that instructors provide explicit instruction in the use of a variety of DMs to enhance fluency.

Taken together, the results from these studies provide further support for the value of discourse markers to create confluence in conversation as the frequency, function, and location (e.g. turn-initial position) of discourse markers may be important indicators of conversational fluency.

In sum, exploring the nature of conversational fluency requires a review of related research of IC, NVC features, and other features of conversational speech. Understanding conversational fluency is inherently problematic as conversation is affected by a variety of situational variables, which have not received relatively much attention by researchers. For instance, one important, yet under-researched, situational variable is the relationship between conversational style and conversational fluency.

2.6. Conversational style

There does not appear to be a widely adopted definition of conversational style, communication style, nor speaking style across the literature. Tannen's (2005) depiction of conversational style was chosen for this study because of the author's well-researched descriptions of conversational style features, which are diametrically opposed, yet existing upon a continuum, enabling conversion of these features into questionnaire items that can be assessed quite readily across a multi-item semantic-differential questionnaire. A semantic-differential questionnaire contain words that convey opposites such as 'always' and 'never' that are situated on opposing ends of a scale. For example, respondents could choose between 'always' (6) and 'never' (1) on a six-point scale in response to the question prompt: 'when communicating with someone you do not know well (e.g. an acquaintance), do you share personal information

easily?’ Tannen’s concept of ‘conversational style’ posits that one’s style exists somewhere along a continuum between High-Involvement (HI-style) and High-Considerateness (HC-style). HI-style speakers are more likely to, among other characteristics, share personal information more willingly, ask personal questions more readily, interrupt and overlap more frequently, speak more rapidly, and avoid conversational silences more often. On the other hand, the reverse may be generally true for HC-style speakers. HC-style speakers are less likely to do the following: share personal information and pose personal questions; wait until the other speaker’s turn has completed to begin speaking; and are more tolerant of conversational silences. Communication breakdowns may be produced due to a clash of conflicting styles and, as a result, speakers of opposing styles may assign negative attributes towards one another (Tannen, 1986). Therefore, it is pertinent to understand how these potential biases may influence listeners’ judgements of speakers’ conversational fluency.

2.6.1. Stereotypes

It is important to first point out that the present study considers conversational style to be an individualized construct, influenced both by micro-cultural and macro-cultural influences. Pizziconi (2009), in her discussion of Japanese communicative style, is careful to note the dangers of overgeneralizing and stereotyping any collective’s (national, ethnic, gender-based) communicative style:

Overgeneralizations can ‘frame’ the character of a language or a culture, and produce fixed and often crude images which are then hard to shred, hindering, for example, our sensitivity to variability within a culture. Moreover, stereotypes are not neutral representations (p. 222).

With this in mind, the following section presents research on national/linguistic conversational styles (e.g. German, Japanese, Mexican-Spanish, US-English) in order to inform our understanding of the potential relationships between one's individual conversational style and one's oral production; yet, inferences drawn from these generalizations about collective groups need to be taken with a grain of salt.

2.6.2. L2 development

Tannen's (2005) notion of conversational style has not been widely adopted in research on L2 oral development. One notable exception comes from Ziegler, Seals, Ammons, Lake, Hamrick, and Rebuschat (2013) who examined 11 intermediate-level L1 English-speaking American learners' acquisition of L2 German conversational style. According to the authors, German conversational style exhibits relatively more high-involvement style characteristics than American conversational style as German conversational style is typically characterized by overlaps, collaborative completions, avoidance of conversational silence, and an argumentative format (Byrnes, 1986, Kotoff, 1991, 1991; Straehle, 1998, as cited in Ziegler et al., 2013). On the other hand, American conversational style has characteristically more established turn-taking boundaries and speakers are more willing to allow other speakers to continue should they overlap (Berry, 1994; Byrnes, 1986, as cited in Ziegler et al., 2013). In Ziegler et al.'s (2013) study, learners participated in instructor-moderated conversation groups of five and six persons respectively and met for one hour each week over a six-week period. The authors recorded, transcribed, and analysed three of these meetings: one at the beginning, one in the middle (week 3), and one at the end. The authors discovered that about half of the participants exhibited changes in their conversational style characteristics whereas little change in this manner was observed for the other half. In particular, the authors discovered that successful learners

produced more overlaps, interruptions, and greater topic shifts from week one to week six. As a result, successful learners were able to hold more of the conversational floor and participate more in the discussions. Learners were interviewed following completion of the six-week conversation group sessions. The findings garnered mixed results. On the one hand, by the end of the six-week sessions, learners could not identify features of a typified German high-involvement style. On the other hand, learners reported an increased level of communicative confidence and a willingness to communicate with German speakers. From these findings, the authors contend that conversational groups are useful activities to increase learners' willingness to engage in conversations with speakers of an opposing style in a cross-linguistic/cross-cultural context. Notably, it may also be useful to provide learners with explicit explanations of conversational style differences to increase their awareness and perhaps increase their use of conversational strategies associated with L2 conversational styles.

2.6.3. Topic shifting and elaborating

With the exception of Ziegler et al.'s (2013) study, research on the relationship between conversational style and conversation fluency is limited and can only be tangentially linked thus far. Young and Hallebeck (1998) analysed discourse produced between two groups of L2 English learners (L1 Japanese and L1 Mexican-Spanish) and L1 English assessors on the Language Proficiency Interviews (LPI). The findings were analysed descriptively according to proficiency level (superior, advanced, and intermediate-mid) and one interview from each language group was analysed at each level. The authors revealed conversational style differences between groups as the Mexican-Spanish speakers exhibited more increased topic elaboration, number of turns, speech rate, and topic shift frequency than the Japanese speakers at each proficiency level. Consequently, the English interviewers were required to intervene more in the

interviews. From these findings, one can infer that L1 conversational style may affect one's level of conversational fluency, as measured quantitatively by researchers, and as judged qualitatively by assessors. As the authors infer, talkativeness, as a feature of HI-style, may be a conversational quality that is generally more valued by Spanish speakers and English speakers than by Japanese speakers. Notably however, the authors also observe that L2 educational background and not just conversational style differences may account for reduced conversational fluency on part of the Japanese speakers despite being categorized at the same proficiency levels. The authors speculate that there are limited opportunities for speaking English outside of the classroom in Japan and so learners may generally have limited opportunities to develop their conversational fluency.

In a related manner, as Sato (2013) discovered, turn-initial silences and a lack of topic elaboration were judged negatively according to verbal reports of L1 English raters of Japanese learners of L2 English. However, the Japanese learners, according to their self-assessments, did not view these features negatively.

2.6.4. Conversational parity

As discussed in the previous section on interactional competence, Galaczi (2008) and Isaacs (2013) discovered that raters seem to be more favorable towards interactional patterns in which both speakers produce relatively equal amounts of speech. For Morales-López (2000), conversational parity is a marker of a high level of conversational fluency. Her analysis of Spanish-Mexican English language learners' conversational speeches, as categorized by levels of fluency as elicited from an interview, revealed that more-proficient L2 speakers were able to maintain a level of conversational parity with a L1 English interviewer whereas the less proficient speakers were not able to do so as effectively. For Morales-López, conversational

parity is achieved as follows: “in developing the topic initiated by an interlocutor, a certain balance between the participants is usually respected, as is the amount of information communicated” (p. 270). Yet, to what extent conversational parity is a marker of one’s level of conversational fluency is debatable. A lack of conversational parity may be negatively affected not only by linguistic proficiency-dominance but by conversational-style dominance as well (Ziegler et al., 2013). In other words, conversational parity may more likely to be achieved when there is a convergence of conversational styles; when there is divergence however, one speaker may dominate the conversation or the conversation may break down completely (Tannen, 1986).

2.6.5. Avoiding conversational silence

McCarthy (2006) has described conversational fluency as attaining confluent alignment between speakers. Ideally, this confluence could be measured reliably through an analysis of between-speaker pause frequency and/or length. Although some research has indicated that conversational alignment has indicated that a lack of pausing evident in between-speaker turns is evidence of L2 conversational fluency (Galaczi, 2014, Peltonen, 2017a), other research has produced no conclusive results (e.g. Riggenbach, 1991; Tavakoli, 2016). As discussed in previous sections, there are several methodological concerns in attributing the cause of a pause between speakers immersed in dialogue and there are no clear solutions as of yet how to address these concerns adequately (Tavakoli, 2016; Sato, 2014; Michel et al., 2017). As discussed, there are several pragmatic reasons for the conversational pause; for instance, the speaker may attempt to achieve conversational intelligibility by pausing between topic shifts to allow time to process what was previously said and to formulate what will be said next (Pickering, 2006).

Additionally, one may pause in response to dispreferred statements or questions (Schelgoff, 2007). Preferred statements provide alignment with the speaker. Examples of

preferred statements include agreeing with a proposition or accepting a request or invitation, as in the following example adapted from Schegloff (2007, p. 60):

Lot: Well I just thought maybe we'd go over to Richard's for lunch and then after that
uh get my hair fixed.

Emm: Alright.

Lot: Ok.

On the other hand, dispreferred statements do not align with the previous utterances. Examples include express disagreements, refusals, and they are often accompanied by excuses or appreciations for the initial request. Other examples include personal questions that may cause embarrassment, or questions/statements that change the topic suddenly. Dispreferred statements or questions create a sense of distance and lack of alignment between speakers as in the example below, adapted from Schegloff (2007, p. 69):

Emm: Wanna come down and have a bit of lunch with me and get some beer and stuff?
(0.3)

Nan: Well you're real sweet hun. Umm...

Dispreferred responses often result in noticeable between-speaker gaps or turn-initial delays, marked by silent pauses, filled pauses and discourse markers (e.g. well) used to hedge responses to dispreferred questions or statements (Schegloff, 2007). Several researchers have suggested that conversation is fluent if it is confluent (McCarthy, 2006), in which speakers create alignment through latching turns onto one another (Peltonen, 2017a); however, a natural disalignment naturally occurs to hedge dispreferred responses. This phenomenon is therefore important to consider when measuring and assessing conversational fluency.

Previous speakers' dispreferred statements or questions may be perceived as a form of imposition. However, to what degree speakers feel they are imposed upon may be influenced largely by their own conversational styles. In other words, HC-style speakers may feel more imposed upon by personalized questions than HI-style speakers would because HI-style speakers use personalized questions frequently within conversation (Tannen, 1986). Additionally, as Tannen's research shows, HI-style speakers may seek to avoid conversational silences through a variety of measures such as overlaps, interruptions, and collaborative completions to create rapport by indicating a sense of shared experience or by posing 'machine-gun' style questions (i.e. short yes/no questions that often elicit personal information). HC-style speakers, on the other hand, may be more tolerant of shared silence and may prefer it for a variety of reasons. Thus, it is necessary to consider L2 learners' conversational styles when making judgments about their levels of conversational fluency.

2.6.6. Politeness strategies

Between-speaker pauses may also occur as the result of clashes between politeness strategies across speakers; moreover, speakers may use politeness strategies differentially to attain and maintain conversational rapport during conversation (Fiskdal, 2000). Brown and Levinson (1987) identify two types of politeness strategies used to maintain fluent conversation - positive politeness strategies (i.e. strategies used to seek alignment through reinforcing inclusion of the speaker within the group) and negative politeness strategies (i.e. strategies used to provide deference to the other speaker). As Fiskdal (2000) contends, when speakers' use incongruent politeness strategies with one another, disfluent (i.e. awkward) moments may occur. Through her analysis of American L1 English speaker-to Taiwanese L2 English speaker conversations, conflicting politeness strategies can produce disfluent and awkward moments, which not only

result in between-speaker gaps, but the awkwardness that arises can cause disfluencies in the surrounding turns of both speakers, regardless of language proficiency levels. In this study, whereas the American L1 English speakers used positive face strategies to seek alignment through co-repairing each other's utterances, the Taiwanese L2 English speakers used negative face strategies through pausing and avoiding co-repairs as strategies to provide deference to the other speaker.

2.6.7. Gender

Finally, cross-gender differences may have an impact on fluency production and perception. It should first be noted that research in this area presents gross generalizations between men and women despite the considerable individual and contextual variations that exist; moreover, it should be noted that much research in this area considers gender to be a binary construct and not one that is fluid and existing across a continuum. With this in mind, it is still useful to delve into this research as it may provide insights regarding generalizable differences between men's and women's speech in order to inform understanding about the interrelationships between conversational style and conversational fluency.

Holmes (2008), in citing Lakoff's (1975) seminal research on cross-gender differences, provides useful insights relating to the interrelationships between conversational style and conversational fluency in L1 English speech. *Lexical fillers*: Women may tend to use more lexical fillers (e.g. you know, you see) and filled pauses (e.g. um) to impose less upon their conversation partners, while also creating options for them to either continue to listen or to interject. Additionally, lexical fillers such as "you know" may be used more by women as rapport-building strategies to formulate a sense of shared experience between conversation partners. *Backchannels*: According to Holmes (2008), women may tend to use more

backchannels (e.g. uh-hmm) to support the conversation partner while listening, which may result in an enhanced mutual flow of speech between speakers. *Topic expansion*: Women may also be more inclined to expand upon ideas presented by the previous speaker. *Conversational silences*: Men may be more tolerant of between-speaker silences, according to cross-sectional conversational analyses of men and women in various workplace settings. *Changing topics abruptly*: Men may be more likely to change topics abruptly in situations where overt power is to be attained, such as in departmental meetings or doctor-patient interactions. *Interruptions and overlaps*: Similarly, men may also be more inclined to interrupt and/or overlap during these aforementioned power-attaining situations. In sum, although the present study considers conversational style to be an individualized construct, useful generalizations about cross-gender differences may provide insight into the relationships between conversational style and conversational fluency.

2.6.8. WTC in a conversational setting

As discussed previously, WTC is “a readiness to engage in communication at a specific time and with specific interlocutors” (Wood, 2016, p. 11) which is theorized to be a state-like variable influencing moment-to-moment bursts of fluency or disfluency (Wood, 2016; Nematizadeh, 2019). In a related manner, conversational style may be largely, but not entirely, fixed, as it is inherently subject to contextual influences (Tannen, 2005), and likely, one’s WTC. For instance, the interactional context may entice one to interrupt, overlap, and change topics more frequently in power-attaining situations (Holmes, 2008). Similarly, the communicative context (topic, situation, and interlocutor) may entice one to take more responsibility for the conversation than usual (Kang, 2005). Relatedly, as Dornyei and Kormos (2000) show, speakers’ WTC and fluency may be affected to some degree by certain social variables such as speakers’

perceived social status, group cohesiveness, and one's relationship with the interlocutor. Finally, as Ziegler et al. (2013) discovered, exposure to interlocutors with opposing conversational styles over time increased L2 learners' WTC in a conversational setting.

Therefore, certain situational variables, inherent to conversational contexts, interact with one's personal traits in a dynamic manner affecting both one's WTC and one's ability to speak fluently (Wood, 2016; Nematizadeh, 2019). With this in mind, it is important to consider the concept of conversational style as somewhat fixed in terms of its reflection of personality traits, but also somewhat flexible in terms of its malleability to be mediated by surrounding contextual influences. In other words, one's conversational style may be reflective not only of one's personality traits, but to some degree, of one's state of willingness to communicate at any given moment within a communicative context.

2.6.9. Listener preferences

Listeners' conversational style preferences may be of consequence to their judgements of conversational fluency. As Tannen (2005; 1986) and Yule (1997) note, speakers of opposing conversational styles may be inclined to assign negative personality attributes to one another. Thus, a HC-style speaker may have certain conversational style preferences as related to the potential findings from the present study. HC-style speakers may do the following: a) have a higher tolerance for conversational silences; b) wish not to share personal information or ask personal questions to an acquaintance; and c) may value the allowance of topics to reach a state of closure before moving on to a new topic. The HI-style speaker, who may not share these conversational style characteristics, may be perceived as being rude. On the contrary, a HI-style speaker may expect conversation to be at a quick rate and quick topic shifts, full of personal questions and stories, and full of overlaps and interruptions to avoid conversational silence. Yet,

the HC-style speaker, who may not exhibit these stylistic traits, may be considered as unintelligent, uninteresting, or unengaged in the conversation (Yule, 1997). However, as each individual's style is theorized to exist somewhere along a continuum, it is likely that there is a considerable amount of variation regarding the degree of one's stylistic preferences and therefore, the degree to which one assigns negative personality attributes to the other. It is possible then that third-party listeners (i.e. raters) who more frequently share personal information (with an acquaintance), shift topics abruptly, and avoid conversational silence, may be more appreciative of these aspects of oral production and may consider them as valued components of conversational fluency. In sum, conversational style preferences, to some degree, shape and reflect what one values in a conversation, affecting what one values when assessing a conversation as a third-party observer.

2.6.10. Summary

The present study uses Tannen's (2005) notion of conversational style represented as a dichotomy between HI-style and HC-style speakers. These styles are defined by a set of features such as sharing personal information more readily, asking more personal questions, interrupting and overlapping more frequently, speaking more rapidly, and avoiding conversational silences. The present study regards the notion of conversational style as an individualized construct; therefore, caution is warranted when generalizing about any particular group's (national/gender-based) communicative style. Research on conversational style and conversational fluency is tangential yet connections can be drawn in terms of topic shifting and elaborating, conversational parity, between-speaker silences, politeness strategies, gender, and willingness-to-communicate. As Yule (1997) indicates, speakers of opposing styles may possess negative biases towards one another, and therefore, third-party listeners may also possess these negative biases. Overall, it is

important to delve into future research how these potential biases impact assessments of conversational fluency.

2.7. Fluency rating scale design

Understanding which features are most salient to raters within a conversational context has important implications for rating scale design. Fulcher (2003) distinguishes between two main approaches to rating scale design: intuitive and empirical. The intuitive approach relies on the scale developer's subjective judgment of performance levels before performance has occurred, "reflecting the subjective experience of the scale developer" (p. 96). In classroom-based assessment situations, descriptors may align with learner outcomes in a syllabus, but may be developed separately from, and in advance of, an analysis of test-takers' performances. Alternatively, there are empirical approaches to designing fluency rating scales: (1) discourse analysis (Fulcher, 1996) (2) scaling descriptors (North, 2000) (3) corpus analysis (Hasselgren, 2001) and (4) an analysis of raters' perceptions (Sato, 2014). These empirical approaches are data-driven, involving analysis of test-takers' performances or listeners' perceptions of these performances prior to scale development.

2.7.1. Discourse analysis

Fulcher (1996) used a discourse analysis approach to develop a rating scale for fluency, uncovering some of the underlying reasons for L2 learners' hesitations/pauses across proficiency levels. The author conducted a discourse analysis of 21 ELTS interview transcripts deriving the following eight categories to explain instances of pausing:

- 1) End-of-turn pauses: pauses indicating the end of a turn.
- 2) Content planning hesitation: Pauses that appear to allow the student to plan the content of the next utterance.

- 3) Grammatical planning hesitation: Pauses that appear to allow the student to plan the form of the next utterance.
- 4) Addition of examples, counterexamples or reasons to support a point of view: these pauses are used as an oral parenthesis before adding extra information to an argument or point of view, or break up a list of examples.
- 5) Expressing lexical uncertainty: pauses which mark searching for a word or expression.
- 6) Grammatical and/or lexical repair: hesitation phenomena that appear to be associated with self-correction.
- 7) Expressing propositional uncertainty: hesitation phenomena that appear to mark uncertainty in the views that are being expressed.
- 8) Misunderstanding or breakdown in communication. (Fulcher, 1996, p. 217)

Fulcher (1996) then used a discriminant analysis procedure to decipher which of these categories could predict test-takers' scores on the English Language Testing System (ELTS) fluency rating scale. From the results, the author created a new fluency rating scale that included detailed descriptors that emphasised the functions of test-takers' use of pauses across proficiency levels. Notably, rating scales on well-known tests such as the Foreign Service Institute (FSI) do not provide much information about the function of pauses; rather, it is common for rating scale descriptors for fluency to refer mostly to the frequency or length of pauses (Fulcher, 1996).

Most importantly, Fulcher's (1996) findings revealed curvilinear patterns for three types of pauses across five proficiency levels (level 1 = 'lower-intermediate'; level 5 = 'university readiness'). The first pattern showed that the number of content-planning hesitations (category 2) rose gradually across the first four levels until, at level 5, this number suddenly fell below level 1. The second pattern showed that grammatical planning hesitations (category 3) rose from level

1 to level 2, sharply declined between levels 2 and 3, and then diminished almost entirely from level 3 to level 5. The third pattern revealed a U-shape contour for end-of-turn pauses. At level 1, there were a large number of end-of-turn pauses, which indicated communication breakdowns (category 8). From levels 2 to 4, there were few end-of-turn pauses, but at level 5, this number increased sharply, indicating that this pause was functioning as a turn-taking device. To put it another way, the author contends that end-of-turn pauses evolved from signalling communication breakdowns at level 1 to signalling turn-taking devices at level 5, with few instances of end-of-turn pauses between levels 2 and 4. Fulcher infers from these findings that data based empirical scales may provide testers with much more information about the potentially curvilinear nature of fluency development.

Overall, Fulcher's (1996) rating scale provided highly detailed descriptors that explain the likely causes for speech hesitations at different performance levels. On the other hand, as Luoma (2004) notes, this scale, although highly informative, is not user-oriented, and therefore, suffers from impracticality. Additionally, if fluency is partially "a perceptual phenomenon in the listener" (Derwing et al., 2004, p. 565), then these scales, which are based solely on close analysis of the utterance, may not necessarily reflect features that are explicitly salient to the rater. In other words, these scales may not reflect the types of observably measurable features that would influence raters' inferences and judgements.

2.7.2. Scaling descriptors

North's (2000) approach used a scaling descriptors approach to develop a new fluency rating scale from analysis of 30 existing rating scales. This process involved: (1) analysing various rating scales derived from numerous sources; (2) reconfiguring them through qualitative and quantitative analysis of experienced teachers' evaluations of these scales; and then (3)

creating a new rating scale and associated descriptors. This analysis resulted in the development of three categorizations of types of fluency. The first is accessibility, which refers to the ability to access knowledge to create and sustain a flow of speech. The second is pragmatic/discourse fluency, referring to the ability to be flexible, precise, coherent, and logical during speech production. The third category refers to communicative fluency, referring to the ability to use interaction strategies such as turn taking, cooperating, asking for clarification, asking for help, and repairing. North also justified how these scales fit within a larger theoretical framework of L2 production. The author claims that a number of influential communicative language testing models, such as Bachman's (1990) communicative language ability model were "used as a point of departure" (p.74) in developing the revised CEFR framework, comprised of four categories: socio-linguistic competence, strategic competence, linguistic competence, and pragmatic competence. Since Bachman (1990) does not describe where fluency fits within his model, North chose to place fluency within the pragmatic competence framework: North discusses various notions of pragmatic competence in the literature, ultimately preferring Thomas' (1983) conception that pragmatic competence is the knowledge that learners possess about how to use language to achieve a particular purpose. North (2000) explains the decision to position fluency within pragmatic competence is as follows:

The decision to put Fluency under Pragmatic Competence cuts across the traditional competence/performance dichotomy used by linguists since Fluency is clearly related with performance. However...associating Pragmatic Competence with Use and Linguistic Competence with Knowledge or Resources allows one to keep Fluency - in the narrow sense of flow - together with the other elements in teachers' wider interpretation of the term (p. 91)

North's (2000) insights into rating scale development are important because they not only include data-driven analysis of fluency perceptions, but they also recognize features related to 'communicative fluency', providing a 'home' for fluency within models of communicative competence. On the other hand, this method may still be largely intuitive because the resulting scale is based upon analyses of other intuitively based scales, developed separately from an analysis of actual performances.

2.7.3 Corpus analysis

Hasselgren (2002) constructed a fluency rating scale from a corpus analysis of young learners' (ages 14 to 15) performances on the Test of School English (EVA) test of spoken interaction, which consists of three interactive tasks: picture description, giving instructions, and role-play. 62 Norwegian L2 learners of English and 26 L1 English British students participated in the study. Raters familiar with the EVA test placed learners into three groups: (1) less fluent; (2) more fluent; and (3) L1 speakers. Hasselgren hypothesised that temporal measures would reflect the groupings and that L1 speakers and more-fluent L2 speakers would use more smallwords than less fluent speakers. Hasselgren (2002) defines smallwords as follows: "small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself" (p. 150). Adhering to this definition, the author identified seventeen types of smallwords: "well, right, all right, okay, you know, you see, I know, I see, oh, ah, I think, I mean, like, sort of/kind of, a bit, just, or something, and everything/and that/and stuff/and things" (p. 151). The author used corpus analysis to identify the frequency and location of smallwords as produced per group. The author also provided a descriptive analysis of two temporal variables (MLR, frequency of filled pauses) per group. Whereas average utterance length (MLR) increased from level to level, results were

mixed for filled pause frequency. From these findings, the author created a fluency rating scale consisting of three band descriptors per level.

Hasselgren's (2002) results revealed that speakers use smallwords differently at different proficiency levels (L1 speakers; more-fluent L2 speakers; and less fluent L2 speakers) in order to maintain conversational fluency. The results are summarized as follows. *Initiating turns*: L1 speakers used smallwords more frequently to initiate turns whereas less fluent speakers used more filled pauses to initiate turns. *Maintaining turns*: L1 speakers used the widest range of smallwords to maintain turns whereas both more fluent and less fluent speakers used respectively smaller ranges of smallwords. *Using Backchannels*: L1 speakers and more fluent L2 speakers used more smallwords whereas less fluent speakers used more non-verbal backchannels (e.g. "hmm") to maintain the mutual flow of speech. *Filled pauses*: More-fluent speakers used the most filled pauses; however, less fluent speakers used filled pauses to initiate turns whereas L1 speakers used filled pauses to initiate turns with the least amount of frequency. *Range of use*: L1 speakers produced the widest range of discourse markers and used them for the widest range of functions, including expressing nuance. More-fluent L2 speakers used a smaller range of smallwords for a smaller range of functions and the less fluent L2 speakers used the smallest range of smallwords for the smallest range of functions. Overall, the findings reveal the efficacy of using corpora to design rating scales to assess fluency as well as the contribution of smallwords in terms of range, frequency, and location to fluent speech. Hasselgren's use of corpora to create a fluency rating scale is unique and promising, indicating how future studies could incorporate corpus analytical techniques to examine how a wide variety of fluency phenomena (e.g. pauses, repairs) inform the development of fluency rating scales.

2.7.4 Analysis of raters' perceptions

Another empirical approach that has been adapted to develop fluency rating scales is Upshur and Turner's (1995) empirically-derived binary-choice boundary definition (EBB) method. In this rating scale design process, expert judges divide performances into three categories (high, middle, and low) and then design a series of yes/no questions that discriminate performances across levels. Sato (2014) adapted this method to design a rating scale based on listeners' perceptions of fluency on an individual task (picture description) and an interactive task (peer-peer picture description). Sato (2014) elicited learners' discussions of learners' performances on each task, which resulted in four band descriptors. Then, Sato expanded these four descriptors across a seven-point scale by applying Upshur and Turner's (1995) technique for constructing criterial questions to distinguish between performances within levels. For example, the highest level of performance (Band Descriptor 4) was expanded to cover levels 6 and 7 on the rating scale. Criterial questions were then constructed to discriminate between performances. The advantage of the resulting scale is that the descriptors reflect salient features of performance according to specific levels while providing some information for why disfluencies such as a slower speech rate or unnatural pauses may be occurring at different levels. Scales developed from this process may be well suited for diagnostically assessing individual fluency features providing that the scale can supply useable feedback about the likely source of the disfluency for classroom-based assessment for learning purposes.

2.7.5. Assessment for learning purposes

Fulcher (2010) discusses the use of rating scales of speaking ability for assessment for learning purposes. Core to the assessment of learning is the notion that assessments are used prior to or in the early stages of instructional settings to assess relevant strengths and limitations in order to help provide an instructional program to meet these individualized needs. One of the

core principles of assessment for learning is to motivate learners to realize their current state of competence to negotiate the zone of proximal development in order to achieve their potential (Brown, 2008). Thus, the assessment for learning approach centers on formative rather than summative feedback. Providing learners with formative feedback about how learners can improve has been shown to help motivate learners to succeed in a variety of academic contexts (Levesque, Zuehlke, Stanek, & Deci, 2004). In essence, as Fulcher (2010) states, "learners need to know what aspects of their performance can be improved and, critically, how they can make that improvement" (p. 69).

When it comes to fluency, progress is gradual, and not easily tracked in a classroom-based setting. One approach, not yet taken towards resolving this problem, is the division of the fluency construct into a set of sub-constructs depicted upon a rating scale that is designed to target specific aspects of linguistic competence and performance that enhance one's ability to speak fluently. Fox et al.'s (2016) analytic-criterion scale, designed to assess first-year university-level engineering students' writing competence, provides a useful format for achieving this purpose. Fox et al.'s scale compiled information from L2 writing experts and discipline-specific engineering experts to sub-divide the writing-for-engineering purposes construct into a set of analytic-criterion questions. These questions were designed to evaluate strengths and weaknesses of specific competencies in order to provide information regarding how learners can remedy deficiencies in these competencies in order to achieve a higher level of writing in engineering contexts. Potentially, this format could be used to evaluate strengths and weaknesses of specific aspects of linguistic competence and performance related to the development of speech fluency in a conversational setting.

2.8. Implications of the literature review for the present study

This literature review has discussed relevant studies on the following topics. First, the review discussed the effect of underlying cognitive processes, FL, and fluency-focussed pedagogy on affecting fluency development. This background information is necessary for understanding the extent to which production reflects perception. Second, the review discussed the influence of temporal, non-temporal, and conversational features of speech on affecting fluency perceptions. Third, the review discussed the potential influence of listeners' accent familiarity and conversational style on fluency perceptions. Finally, related research on interactional competence, non-verbal communication, and rating scale design was covered.

Before highlighting the relevance of this research, it is useful to review the aims of the present study. To recall, the overall purpose of this two-phase mixed methods study is to examine the influence of a variety of speech characteristics – temporal (core), non-temporal (peripheral), and conversational - as well as certain listener characteristics (accent familiarity and conversational style) on affecting instructors' perceptions of fluency on a conversational task. This purpose is achieved through developing and piloting a scale to assess fluency on a paired conversational task for classroom-based assessment for learning purposes. The first phase of the study explores how instructors' perceptions of fluency translate into an analytic-criterion rating scale (Fox et al., 2016), which is composed of question items representing both core and peripheral fluency features. The second phase examines the following: (1) the degree to which these core and peripheral items relate to one another and (2) the degree to which these items relate to temporal measures of speed and flow (e.g. speech rate). This second phase also focuses on how accent familiarity and conversational style may affect assessments of both core and peripheral fluency features.

Much research on fluency perceptions has involved correlating temporal variables (e.g. speech rate, pause length) with raters' assessments. This research has provided valuable information about how temporal analyses can be used to validate rating scales designed to assess fluency. However, depending on the research context (task, scale, participants) any one of a wide variety of temporal variables may be most salient (Segalowitz, 2010); therefore, which variable most adequately predicts fluent performance is still relatively unknown. With this in mind however, recent research has shown (e.g. Shea & Leonard, 2019; Kahng, 2014) that within-clause pause rate, indicative of processing difficulties in the formulation stage of speech production (Levelt, 1989), has been shown to discriminate between levels of fluency in a fairly predictable manner. Therefore, future research involving analyses of within-clause pause rate may provide validity evidence for the use of rating scales to assess fluency.

From a listener's perspective, temporal speech features cannot be fully isolated from non-temporal speech features (e.g. vocabulary range) as they appear to be inherently interconnected (Williams, 2018). Previous research (e.g. Préfontaine & Kormos, 2016; Rossiter, 2009) has shown that these non-temporal features affect perceptions of fluency to varying degrees. Future research is needed to explore the degree of this influence in order to have a fuller understanding about the degree to which L2 practitioners should recognize the importance of non-temporal features as an integral component of fluency.

Moreover, the extent to which conversational features affect raters' fluency perceptions is under-researched. It is therefore unknown if conversational features should be included in rubrics to assess fluency, despite their potential relevance (McCarthy, 2006; Sato, 2014).

Relatedly, what role interactional competence plays in contributing to perceptions of conversational fluency is under-researched. Chalhoub-Deville (2003) contends that the learner's

ability rests within the individual but it is largely mediated by the surrounding communicative context; thus, the construct to be measured is "an ability - in language user - in context" (Chaloub-Deville, 2003, p. 572). Through this perspective, it is conceivable that fluency neither theoretically exists solely within the speaker and/or within the listener (Freed, 2000; Derwing et al., 2004), but that it is co-constructed in the space between speakers.

Conversation involves a substantial amount of non-verbal communication (NVC); yet it is unknown how NVC may play a role in affecting perceptions of fluency. In particular, self-directed gestures may provide raters with visual clues about online planning while other-directed gestures and regulators (incorporating aspects of eye contact, head nodding, and posture) may provide raters with visual clues about one's ability to be conversationally fluent. As Götz (2013) recommends, using video recordings of learners' performances can provide a more complete understanding of the concept of fluency as a whole. Notably, as discussed in the Method section of this paper, these recommendations were taken into consideration in the design of this study.

The majority of previous research on fluency has focused primarily on monologic tasks, largely because measuring conversational fluency is a challenging task. Notably, there are methodological difficulties concerning temporal analysis of the between-speaker pause, which requires further exploration (Tavakoli, 2016). Moreover, due to the unpredictable nature of conversation, it is difficult to compare results regarding conversational features of speech (e.g. backchannels "um") across studies. Thus, it is not surprising that previous research has provided mixed results regarding the efficacy of correlating raters' assessments with measures of interactive phenomena (e.g. backchannels; overlaps, collaborative completions) (e.g. Tavakoli, 2016; Riggenbach, 1991). Corpus analysis of smallwords (Hasselgren, 2002) and discourse markers (Crible, 2019), as well as House's (1996) analysis of pragmatic formulas have provided

some insight about the nature of conversational fluency; yet these features have been under-researched thus far. Most relevantly, with the exception of Sato's (2014) study, there seems to be a lack of research on how raters perceive the contribution of interactive phenomena, smallwords, and discourse markers, when making fluency judgments.

Conversation is inherently affected by a variety of situational variables. One under-researched situational variable is conversational style. Not much research has revealed how conversational styles affect fluent production, how conversational styles between interlocutors affect interaction, and how this interaction between styles is perceived by a third-party listener, as in the present study. Notably, the present study is the first to investigate the impact of raters' conversational style on their fluency assessments. As Ducasse and Brown (2009) point out, raters may hold widely different beliefs and values about conversational practices, thus it is worthwhile to investigate how these different beliefs and values may differentially affect raters' perceptions of conversational fluency.

In addition to conversational style, accent familiarity is another listener characteristic that is likely to affect perceptions of fluency as raters' degree of accent familiarity has been shown to affect oral proficiency ratings in several studies (e.g. Winke et al., 2013; Carey et al., 2011). However, less is known about its influence on fluency ratings specifically, except for a few studies (e.g. Browne & Fulcher, 2017; Reid et al., 2019).

Finally, more research is needed to understand further how data based approaches to creating fluency rating scales may inform research and practice regarding rating scale design. Several studies have used a variety of data based approaches including discourse analysis (Fulcher, 1996), scaling descriptors (North, 2000), corpus analysis (Hasselgren, 2002), and analysis of raters' perceptions (Sato, 2014). Contributing to Sato's area of research, the proposed

scale is also drawn from an analysis of raters' perceptions. However, the proposed scale is unique in its format, which is adapted from Fox et al., (2016). This format enables the deconstruction of the fluency construct into analytic items representing various fluency features. Therefore, this scale, constructed in phase one, allows for phase two investigations into how these different fluency features relate to one another and to holistic assessments of fluency. Furthermore, this format allows for investigations of the relationships between raters' assessments of each of these scale items and the following: temporal measures of speech; listeners' self-ratings of accent familiarity; and questionnaire items representing listeners' conversational style characteristics. Ultimately, these findings may inform future research and practice regarding the development of data based scales to assess fluency on a conversational task.

2.8.1. Purpose of the present study

The purpose of the present study is to explore the potential influence of speech characteristics (temporal, non-temporal, and conversational) and listener characteristics (accent familiarity and conversational style) on influencing perceptions of fluency. This purpose is achieved through developing and piloting a scale to assess fluency on a paired conversational task for assessment for learning purposes in a university-level EAP classroom. The following research questions are posed to meet the aims of the present study.

2.8.2 Research questions

1. How can EAP instructors' perceptions of EAP learners' speech performances on a paired conversational task inform the development of an analytic-criterion rating scale to measure speech fluency?
2. To what degree are the items on the scale relevant to assessing fluency?

3. What relationships exist between temporal measures and rating assessments, according to this scale?
4. In what ways, if any, do raters' levels of accent familiarity affect their fluency ratings?
5. What relationships exist between instructors' conversational styles and their assessments of High Considerateness-style and High Involvement-style students, according to this scale?

Chapter Three - Method (Phase One)

3.1. Overview

3.1.1. Research design

The overall purpose of this two-phase mixed-methods study is to explore perceptions of fluency through the development and piloting of a rating scale to assess fluency on a paired conversational task. A two-phase mixed-methods sequential exploratory design (instrument development model) provides the underlying methodological framework for this study (see Figure 2 below)

Figure 2 is adapted *from Designing and conducting mixed-methods research* (Creswell & Plano Clark, 2011, p. 76).

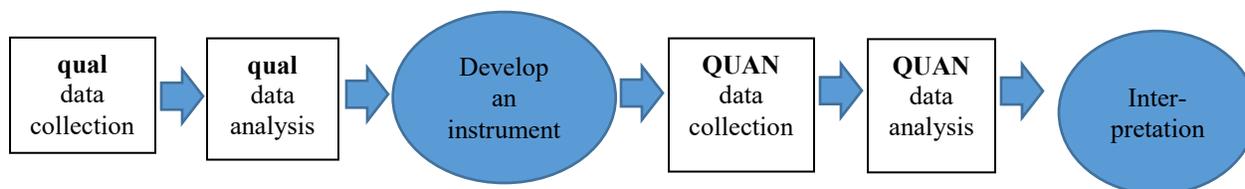


Figure 2: Exploratory Design: Instrument Development Model.

As shown in Figure 2 above, the first phase involves the collection and analysis of qualitative data, resulting in the development of the scale. The second phase involves the collection and analysis of quantitative data through the use of the scale. The results from both

phases provide insight into how temporal, non-temporal, and conversational characteristics of speech, as well as characteristics of the listeners themselves, influence fluency perceptions.

In phase one, seven EAP instructors watched videos of seven-minute paired conversations, elicited from 14 EAP learners who performed this conversational task twice with two different conversational partners. In this task, students were provided with a list of three topics: (1) Learning English; (2) Living in Canada; and (3) Future academic/professional goals. Learners were allowed to discuss any or all of the topics. Learners were given one minute of preparation. After watching the videos, instructors were audio-recorded discussing their observations about learners' fluency levels. Once all videos were viewed, instructors then placed each learner within one of three categories: high, middle, and low. These audio recordings were transcribed and coded using in-vivo and pattern coding techniques (Saldaña, 2009). The codes resulted in the development of themes with corresponding categories and codes. These themes were used to create items in a multi-item fluency rating scale. Categories and codes informed the scale descriptors associated with each item.

In phase two, the scale was used to examine how relevant these scale items are to assessing fluency; additionally, the scale was used to examine how two key listener characteristics, accent familiarity and conversational style, affected fluency perceptions.

3.1.2. Pilot stages

Before data collection for phase one could commence, it was first necessary to choose the most appropriate task for data collection. In order to achieve this purpose, three paired conversational tasks were piloted with a small group of four participants prior to data collection. These four participants were representative of the participant-group in this study because they were L2 English graduate students enrolled in English language programs, which they were

required to take as a condition of acceptance into their regular programs. Each of these tasks contained different task prompts with different topical information. Students were asked to discuss, after one minute of preparation, the topics provided for seven minutes. The most suitable task was chosen based on the familiarity of the topic, which according to participants influenced the amount of interaction and fluent production elicited by the topic. The chosen task included more general topic prompts regarding learning English and living in Canada (See Instruments) whereas the discarded tasks included more topic-specific prompts that guided participants to discuss the use of technology in society (DuCasse and Brown, 2009) and the growth of world population (Fulcher, 2003).

As Wainer (1994) discusses, there are concerns regarding issues of validity related to item/task difficulty when allowing test-takers to choose the task or item that they wish to be tested on; in this study, one of these concerns is the degree to which the task reflects what participants would be asked to do in an authentic EAP setting. More specifically, in this study, two issues exist surrounding the issue of task difficulty as it relates to validity. First, arguably, the topics are much easier to discuss than topics typically discussed in EAP settings represented by the discarded topics presented when piloting these tasks. Second, discussion tasks, although widely used during classroom proceedings – and are the norm in L2 classroom settings (Wood, 2010) - are likely to be the easiest and least consequential type of oral task in a real-world EAP setting.

3.1.3. Ethics and research site

Ethics clearance was provided by two medium-sized Canadian universities in order to recruit participants for this study. All participants were informed verbally and through written informed consent forms about their roles in the study and about any potential risks associated

with their involvement. All personal identifiers were coded numerically and no information that would identify the participants was revealed. All audio recordings and transcripts were stored on a personal password-protected computer.

3.2. Participants

3.2.1. Sampling procedure

A purposive sampling procedure (Teddlie & Tashakkori, 2009) was used to recruit participants. The purpose of this procedure was to enable the researcher only to recruit participants who represented specific groups within the overall population. In order to do this, only participants who met specific requirements were invited to volunteer for the study. These requirements are depicted in each section below.

3.2.2. EAP learners ($n = 14$)

The purpose of the sampling procedure was to recruit students who could represent a small sample of the population of university-level students in EAP classes at Canadian universities as the fluency rating scale is designed for classroom-based assessment in university-level intermediate-to-advanced EAP classes only. Thus, the scale was not designed for a more general population of English language learners across instructional contexts.

With this information in mind, participants were required to have met three key requirements. First, participants must have been enrolled in university-level EAP classes at the time of data collection. Second, participants must have been judged to be at least a low-intermediate level of general oral proficiency in order to have been reasonably capable of meeting the challenge of conversing with a partner for seven minutes. Third, it was necessary to recruit participants of as wide a range of proficiency levels as possible in order for the scale to reflect a broad range of abilities as possible (Upshur & Turner, 1995). Evidently however, due to

the challenging nature of the seven-minute paired conversational task, beginning-level learners were excluded from this participant-group and thus, beginning-level fluency was not covered by the range of the scale.

To meet the first requirement, participation from learners was requested by means of recruitment speeches in EAP classes and through recruitment emails. The resulting participant group consisted of EAP learners who, at the time, were enrolled in regular graduate-level programs at a Canadian university. To meet the second requirement, participants' general oral proficiency levels were judged by performance on a monologic task, which was external to the findings of this study. In other words, performances on this monologic task were not analysed by the researcher and were not compared with the dialogic samples collected from analysis of the paired conversational task. They were only used to ensure that participants met the necessary requirement of speaking English at a low-intermediate level. This task was a simulation of the older version of Task One of the Oral Language Test (OLT), which was the speaking component of the Canadian Academic English Language (CAEL) Assessment (Fox, 2004). For this task, EAP learners were required to speak for one minute into a computer about their English language learning experiences after being given one minute of planning time. Learners were provided with both written and computer-mediated audio-recorded instructions.

Beyond this study, the OLT consisted of five tasks, which enables assessors to gain a much fuller sense of learners' oral proficiency levels. However, due to time constraints during the data collection sessions and due to the limited purpose of the CAEL in this study, the first task was deemed sufficient enough to attain a glimpse at their oral proficiency levels.

Using the CAEL/OLT nine-point (10 – 90) task one holistic rating scale (see Instruments), two expert CAEL/OLT raters, who were well experienced with this scale (See

participant description – Expert raters) would later judge their general oral proficiency levels ranging from intermediate (40) to high-advanced (80 out of 90). To clarify, the CAEL/OLT scores correspond to the following proficiency levels: Beginner (10 – 30); Intermediate (40); High-Intermediate (50); Advanced (60); Adept (70); Expert-fluent (80-90). Table 2 below lists the mean scores provided by two raters for each of the participants in this study.

Table 2

Participants and mean ratings on the CAEL/OLT Task One

<u>Participant #</u>	<u>Mean Rating (CAEL/OLT Task One)</u>
1	60
2	40
3	50
4	45
5	65
6	70
7	40
8	50
9	80
10	45
11	80
12	65
13	50
14	50

It is important to note that participants were, at the time of data collection, graduate students who were enrolled in supplementary EAP courses as a condition of their enrollment. Some explanation may be required regarding how graduate students could be admitted to universities despite exhibiting limited oral proficiency skills. For many of these graduate students, at one of the two universities where data collection transpired, certain departments have alternative admission criterial requirements granting admission to students below a certain overall language proficiency threshold (e.g. 70 on the CAEL). As such, their admission was conditional and required enrollment in a series of supplementary EAP courses. The resulting

fluency rating scale is therefore designed for classroom-based assessment in university-level intermediate-to-advanced EAP classes for students who fit this description only. The scale is not designed for a more general population of English language learners across instructional contexts.

Since the participants in this study were already fluent enough to participate in a seven-minute conversational task, no beginner-level learners participated in the study. Therefore, the 'low' group represented on the scale, constructed from the phase one findings, does not represent performances at the beginning level, but rather, as to be discussed later, represents ability at the intermediate level. Thus, this fluency scale is task-based, designed for a specific context, and not designed to reflect beginning-level performances.

The participant group was diverse in terms of their linguistic, cultural, and educational backgrounds. Students provided the following information at the commencement of the data collection sessions regarding the following:

1. L1 (Farsi = 4; Mandarin = 2; Spanish = 2; Turkish = 2; Arabic = 1; Brazilian Portuguese = 1; Russian = 1; and Thai = 1)
2. Program of study (Mathematics, Engineering, Natural Sciences, Geography, and Applied Linguistics)
3. Age (between 24 and 45)
4. Reported gender (7 males and 7 females)
5. Time spent living in Canada (4 months to 4 years)
6. Degree of familiarity with their conversation partner (friends, acquaintances, strangers)

3.2.3. EAP instructors ($n = 7$)

Participants in this group were also required to meet certain requirements; specifically, they were required to have at least two years of prior teaching experience and to possess relevant qualifications, such as a graduate degree in Applied Linguistics, Education, or Modern Languages, in order to provide informed responses about L2 speech production. Selected participants were known to the researcher and were invited to participate through emails.

3.2.4. Expert raters ($n = 2$)

Participants in this group were selected because they needed to be familiar with the CAEL/OLT Task one rating scale. Participants reported having over 25 years of rating experience in general, and over 15 years of experience with the CAEL/OLT rating scales. Participants were invited to participate through email.

3.3. Instruments

3.3.1. Paired conversational task

This task was designed by the researcher. Decisions made regarding task specifications were informed by synthesising insights from various sources (e.g. Ducasse & Brown, 2009; Fulcher, 2003). Task characteristics such as topic familiarity (Bui & Huang, 2018) and pre-task planning (e.g. Ellis, 2005; Foster & Skehan, 2009), as discussed in the fluency-focussed activities section of the literature review (p. 33), were carefully considered. Fulcher's (2003, p. 57) framework for describing speech tasks, as shown below, was used to provide the initial task specifications. It should be noted that these task specifications would later be developed following the phase one data collection and analysis; the final task specifications provided to phase two participants will be described in the phase two Method section.

1. Task orientation: After one minute of preparation, students had seven minutes to exchange personal information based on the following prompts: (1) Learning English as an additional language; (2) Living in Canada; and (3) Future academic/career goals.

2. Interactional relationship: Two-way interaction. Communication is co-constructed.
3. Goal orientation: Convergent. Test-takers work together to exchange information.
4. Interlocutor status and familiarity: Variable. Participants may have varying levels of familiarity with one another.
5. Topic(s): Familiar. Test-takers should be able to draw from personal experience to discuss the topics.
6. Situation: Exchanging information with a peer.

Traditionally, one of the key reasons why researchers interested in fluency prefer using monologic tasks is because procedures can be more easily standardized as conversation can be much more unpredictable (Wood, 2010). Therefore, several key task characteristics needed to be considered in the design of this task. Justifications for the choices are outlined in the following discussion:

It was deemed that one minute would be sufficient for planning time. Research has shown that providing students with pre-task planning time to strategically prepare their responses results in more fluent speech than providing no or little (e.g. 30 seconds) of planning time (Bui & Huang, 2018; Bui, 2014; Skehan & Foster, 2009; Mehnert, 1998). However, a review of studies (Mehnert, 1998; Wigglesworth, 2000; Elder & Iwashita, 2005) has shown no to marginal (Mehnert, 1998) differences between one minute and five minutes of planning time. Ten minutes, on the other hand, may result in significantly more fluent speech (Mehnert, 1998). With consideration of these findings, one minute was deemed an appropriate length of time for students to plan. Moreover, the decision to provide learners with only one minute of planning time ensures that the methods used in this study align with methods used in similar studies on fluency generated from interactive tasks (e.g. Sato, 2014; Tavakoli, 2016).

The task was chosen to be seven minutes in length. As Weir (2005) discusses, certain factors need to be taken into consideration regarding time constraints for speech tasks in relation to the purpose of the test task. In this study, it was deemed necessary that the task length be long enough for both speakers to “claim speaking rights” (Weir, 2005, p. 66) and to share responsibility for the conversation in order to understand more fully the relationships between fluency and interactional skills. In other words, the task length needed to be long enough for both speakers to have the opportunity to do the following: provide both long and short turns; initiate, close, extend, or transition between multiple topics; and to actively listen to support the speech. Ten minutes appears to be a common length of time stipulated for interactional test tasks (Weir, 2005; DuCasse & Brown, 2009) in order for raters to have sufficient time to rate two speakers simultaneously. However, ten minutes was deemed too long for this study because, as will be discussed in the phase two Method section, EAP instructor-participants would be required to watch multiple videos during the subsequent data collection sessions. As described further in phase two, there were several reasons why the data collection sessions were limited to 90 minutes, including a) enabling busy teachers to allot time to the study; b) allowing time for developing rapport; c) allowing time to introduce instructors to the scale; d) allowing time for assessment; e) allowing time to provide additional information as necessary.

Recent studies exploring conversational fluency have used only six-minute tasks (Tavakoli, 2016; Pletonen, 2017a) in their examinations of fluency in interaction; yet these studies did not involve an additional phase requiring participants to use a multiple-criterion scale to assess two speakers simultaneously, as in phase two of this study, which would likely require additional time. In the end, a seven-minute task was chosen as an appropriate compromise among all of these conflicting factors.

Topic familiarity has been shown to affect the quality of fluent production in a variety of studies (e.g. Bui, 2014; Bui & Huang, 2018). Therefore, it was necessary to reduce this topic effect by prompting speakers with topics that would be likely be familiar to all participants based on their background characteristics: learning English, living in Canada, achieving future academic/professional goals. As discussed earlier, the pilot participants preferred these familiar topics because they could elicit the greatest amount of fluency and interaction, which was essential in meeting the purpose of this study.

Evidently, the topic effect could be reduced even further by providing learners with only one topic; yet there would be one considerable drawback in doing so. Topic management is regarded as being a key characteristic of interactional competence (Galcazi & Taylor, 2018), which, intertwined with turn taking skills, is theorized to be a component of interactional fluency (Sato, 2014; Tavakoli, 2016, Pletonen, 2017). It was hypothesized that learners of different levels of fluency would exhibit different topic management skills; thus, it was deemed necessary to include a variety of topic prompts from which to choose.

One of the drawbacks of using peer-to-peer tasks is the effect of interlocutor familiarity (Fulcher, 2003), which refers to how well the conversation partners know one another and their social status relations. Interlocutor familiarity has been shown to affect the production and perception of speech performances. O'Sullivan (2002), for instance, revealed that participants attained higher oral proficiency scores when paired with a more familiar partner.

Interlocutor familiarity is a variable that may not easily be controlled in a research setting. In this study, as participants responded individually to participate in-group meetings for data collection, it was not possible to know or foresee participants' levels of familiarity with one

another. During the data collection sessions, participants reported having varying degrees of familiarity with one another: (1) friends, (2) acquaintances, and (3) strangers.

Paired Conversational Task

Instructions

With a partner, discuss the following topics:

- Learning English as an additional language
- Living, studying, and/or working in Canada
- Your future academic and/or career goals

Take as much time as you need to discuss each topic. It is okay if you do not discuss all topics. If time remains after discussing all topics, start a new topic.

Preparation Time: 1 minute. You may make notes in the space below

Speaking Time: 7 minutes

Figure 3. Paired Conversational Task

The purpose of this study is to examine perceptions of fluency as mediated by an interactional context; therefore, it was necessary to choose an interactive task. A peer-to-peer task was chosen because it reflects the target language domain (e.g. discussing topics with a peer in an academic context) and, according to DuCasse and Brown (2009), peer-to-peer tasks provide additional benefits to both test-takers and to researchers investigating interaction:

Peer-to-peer interaction has been found to be more balanced (Eygud & Glover, 2001) and interactive (French, 1999), with candidates producing a greater range of functions (Kormos, 1999; Lazaraton, 2002) and interactional patterns being more varied (Saville & Hargreaves, 1999)...In short, peer-to-peer tasks have been found to provide the potential for a wider range of functional and interactional moves than is generally possible in the more traditional interviewer-led oral interview (p. 425).

Moreover, as Weir (2005) claims, peer-to-peer tasks enhance the potential for ‘reciprocity’ conditions, referring to who can claim speaking rights and responsibility for the conversation. In this study, as the goal of the task was to exchange information about a familiar topic, both speakers have speaking rights and share responsibility for the conversation. In contrast, speaking rights and level of responsibility for the conversation are often unequal in a teacher-to-student conversation task (Weir, 2005).

3.3.2. CAEL/OLT task one & rubrics

In this study, the sole purpose of this task is to gauge participants' general levels of oral proficiency. Performances on this task would later be assessed by two expert raters to gauge their overall oral proficiency levels. Beyond this study, the purpose of the OLT, which is the task-based, computer-mediated speaking component of the CAEL assessment, is to predict speech performance on five academic speech tasks, such as making short presentations (Fox, 2004).

Only task one was used in this study. For this task, EAP learners were required to speak for one minute into a computer about their English language learning experiences after being given one minute of planning time. Learners were provided with both written and computer-mediated audio-recorded instructions. This task was chosen because it requires speakers to provide a one-minute speech turn, which enables raters to use rubrics available to the general public to assess the performance; moreover, samples longer than one minute could increase the potential of rater fatigue; therefore, the CAEL/OLT task one was deemed to be the most suitable task available.

3.3.3. Semi-structured interview guide

The purpose of this interview guide was to prompt EAP instructor-participants to verbalize, through a think-aloud technique (Gass & Mackey, 2007), salient features of speech

related to the construct of fluency in interaction. Notably, similar procedures were used by DuCasse and Brown (2009) on raters' perceptions of the construct of interaction and by Sato (2014) in his study of raters' perceptions of interactional oral fluency. This interview guide (Figure 3) was supplemented by the Raters' Listening Guide (Figure 4), which helped the EAP instructors document their responses as they analysed the speeches.

Semi-structured interview guide

Thank you once again for your participation in this study on the assessment of second language speech fluency. I would like to remind you that I will be recording this session. With your permission, I would like to turn on the audio-recording device.

The purpose of this interview is to develop a rating scale for fluency on an interactive task. You will be asked to listen to four seven-minute paired conversations. In this speech task, they are required to exchange information about the following topics as provided through written prompts. (1) Learning English as an additional language; (2) Living in Canada; and (3) Future academic or career goals.

1. Speech fluency, in its most narrow definition, refers to the flow, fluidity, or continuity of speech (Koponen & Riggensbach, 2000, p. 6). Fluency in an interactional context however, is theorized to be “a joint performance between learners” (Sato, 2014, p. 8); in other words, each speaker contributes to the mutual flow of speech between speakers. With this in mind, for this study, speech fluency is presently defined as “the flow of speech within or between speakers”. Therefore, while watching the video, make comments about any linguistic features that contribute to the flow of speech within or between speakers.

2. Watch the speeches and make notes in the table provided. More specifically, please look for significant moments of fluency or disfluency and be prepared to discuss why these moments may be occurring. Be sure to indicate the significant moment with a time marking. You may ask to stop the video to write down the time marking. Afterwards, we will review these moments and you will be asked to discuss why these moments might be occurring.

3. Please assign individuals to a high, middle, or low group in the table provided.

4. Once we have watched all four videos, please review your decisions, with the goal of identifying features that distinguish between performances.

Figure 4. Semi-Structured Interview Guide

Performance # _____
Speaker (Left): _____
Performance Notes: Write time markings for significant moments
Overall Level of Fluency (High, Middle, Low)
Speaker (Right): _____
Performance Notes: Write time markings for significant moments.
Group (High, Middle, Low)

Figure 4. Raters' Listening Guide

3.4. Procedures

3.4.1. EAP learners ($n = 14$)

Participants were asked to meet in groups of four so that they could perform this task with three different partners. Although some participants convened in groups of four, due to sudden and unexpected absences at the data collection sessions, some groups only convened in groups of three, which still provided the researcher with the ability to record participants perform this task with two different partners. In the end, the following groups were formed: Group 1 (4 participants); Group 2 (4 participants); Group 3 (3 participants); and Group 4 (3 participants). First, after providing informed consent, participants were asked preliminary questions about their first language backgrounds, areas of study, gender, age, years living in Canada, and their familiarity with the other participants.

Then, participants individually performed the one-minute CAEL/OLT task, which would later be assessed by two expert raters to gauge their overall oral proficiency levels. These performances were audio-recorded using a small condenser microphone.

Afterwards, in pairs, participants performed the seven-minute paired conversational task. Participants from each group performed the task twice with different conversation partners. Participants were given one minute before each task to plan their speeches. The performances on the paired conversational tasks were video-recorded and two small condenser microphones were directed towards each participant in order to isolate individual speeches within this interactive task. The audio recordings from the microphones would later be edited in Ableton 9.2 and synced with the video footage through Microsoft Movie Maker 8.1 software to provide audio enhancement.

In total, 14 minutes of audio recordings from the CAEL/OLT task and 98 minutes (14 paired conversations x 7 minutes/conversation) of video footage/audio recordings from the paired conversational task performances were collected. The table below provides information regarding the groups, participants, and levels of acquaintanceship with one another (friends, acquaintances, and strangers).

Table 3

Data Collection Groups, Conversations, Participants, Levels of Acquaintanceship

<u>Group</u>	<u>Conversation #</u>	<u>Participant #</u> (Left)	<u>Participant #</u> (Right)	<u>Acquaintanceship</u>
A	1	1	2	Strangers
A	2	3	4	Acquaintances
A	3	1	3	Strangers
A	4	2	4	Strangers
B	5	5	6	Friends
B	6	7	8	Strangers
B	7	5	7	Strangers
B	8	6	8	Strangers
C	9	9	10	Acquaintances
C	10	9	11	Friends
C	11	10	11	Strangers
D	12	12	13	Acquaintances
D	13	12	14	Strangers
D	14	13	14	Strangers

3.4.2. Expert raters (n = 2)

As mentioned, raters individually assessed learners' performances on the CAEL/OLT task to provide a holistic assessment of their overall oral proficiency levels. The 14-minute speeches from the CAEL/OLT task one were compiled into a digital folder on the researcher's personal computer and were randomized. Each rater, who reported having over ten years of experience using the CAEL/OLT rating scale, listened to each speech through a set of headphones. Each rater then listened to each speech and assessed each speech accordingly.

3.4.3. EAP instructors (n = 7)

Interviews were conducted with instructors individually. Each individual interview session lasted about two hours. After establishing rapport, instructors were reminded of their overall purpose in the study and the data collection procedures. Instructors asked questions as necessary for further clarification. Following this introduction, instructors provided signed informed consent forms and the audio-recording device was activated.

The instructors were provided with the semi-structured interview guide (see instruments) to assist them throughout the data collection process. Using this guide, participants followed these procedures. Notably, similar procedures were followed by participants in DuCasse and Brown's (2009) study regarding the development of a scale to measure interactional competence on a paired conversational task and Sato's (2014) study regarding the development of a fluency rating scale for an interactive task.

1. Participants reviewed the topics that the learners would discuss in the videos.
2. Participants reviewed the working construct definition for fluency.
3. While watching the videos, participants made comments about any linguistic features that contribute to the flow of speech within or between speakers.
4. Participants provided approximate time-markings in the video regarding significant moments of fluency or disfluency. Participants stopped the video if necessary to record the time markings.
5. Once the video was over, participants discussed their overall observations and the significant moments in the video.
6. Once all four videos were watched, participants consulted their notes and then placed each of the eight speakers into high, middle, and low groups, relative to one another.

7. Participants once again reviewed their decisions with the goal of identifying features that distinguish between performance groups.

3.4.4. Analysis of EAP instructor interviews

EAP instructors' discussions were transcribed in Microsoft Word and then later analysed in NVivo 12, by using in-vivo and pattern coding techniques (Saldaña, 2009), which resulted in the creation of relevant themes and categories. Notably, Creswell and Plano Clark (2007), whose two-phase sequential exploratory design (instrument development model) provides the methodological framework for this study, justifies the use of in-vivo and pattern coding to develop an instrument. Creswell and Plano Clark state that "a researcher might use the significant statements or quotes to help write specific items for the instrument; the codes would be major variables and the themes would be constructs or scales on the instruments" (p. 145, as quoted in Saldaña, 2009, p. 49).

The themes developed from the analysis were then used to create analytical criterion questions on the rating scale; as mentioned in the introduction to this paper, the format for this scale was inspired by a scale used to assess first-year engineering students' writing competence (Fox et al., 2016). The categories and sub-categories were then used to create the rubric descriptors. Within these themes, instructors' comments were further categorized into three groups: high, middle, and low. It is important to point out that, since speakers were previously judged to be at least at a low-intermediate level of oral proficiency, the low group does not refer to a beginning-level group. The low group, in this case, refers to students at a low-intermediate to intermediate level of fluency; the middle group refers to speakers at a high-intermediate level of fluency; and the high group refers to speakers at an advanced level of fluency. Thus, as no beginner-level learners participated in the study, and the least-proficient learners were gauged as

being at least at a low-intermediate level, the “low” group does not represent performances at the beginning level. Selected comments were used and modified to create the rubric descriptors for each performance level: High = Yes (n = 4); Somewhat = Middle (n = 4); Low = No (n = 6).

The three band descriptors were then used to create a six-point fluency rating scale: “Yes” (6 and 5); “Somewhat” (4 and 3); and “No” (2 and 1). There are several justifications for this decision. For one, Sato (2014), in a similar study, created a seven-point fluency rating scale from four band descriptors. He justified this choice by citing Preston and Colman (2000) who “compared validity, reliability, and discriminating power of scales with different response categories (from 2 to 10) and concluded that those indices were significantly higher with response categories of 7 than those of 4. In other words, they found that 7 response categories (up to 10) were more reliable and valid than 4 response categories (below 4) as rating scales” (Sato, 2014, p. 84). Moreover, similar to the nine-point scale presented in Fox et al. (2016), in which each category is divided into three sub-categories (indicating a ‘strong yes’, a ‘medium yes’, and a ‘weak yes’), in this study, each category is divided into two sub-categories. Finally, the format of this scale is further inspired by the Cambridge English Proficiency scale used in the CEP oral proficiency exam (CEP, 2015). Like the fluency scale in the present study, this six-point scale (5 – 4 – 3 – 2 – 1 – 0) consists of three descriptors bands (5); (3); and (1). For levels 4 and 2, the descriptors indicate that level 4 shares features of levels 5 and 3 and level 2 shares features of levels 2 and 1. The number 0 indicates that the performance is below band 1. This format was adopted for the fluency scale in the present study. Please see the fluency scale at the end of the following section.

Chapter Four - Results and Discussion (Phase One)

This section illustrates how an analysis of EAP instructors' fluency perceptions informed the development of a rating scale to measure fluency on a paired conversational task. In particular, this section includes a depiction and discussion of the results related to the first research question:

How can EAP instructors' perceptions of EAP learners' speech performances on a paired conversational task inform the development of an analytical-criterion rating scale to measure speech fluency?

The rating scale, developed from these findings, is presented at the conclusion of this chapter on pages 180 to 181.

4.1. Overview

As mentioned previously, instructors were asked to make comments about any linguistic features related to the flow of speech within or between speakers. Six themes and their associated categories were created from analysing instructors' comments: (1) smoothness; (2) efficiency; (3) sophistication; (4) clarity; (5) facilitating topics and turns, and (6) supporting the conversation partner.

Notably, the naming for these themes was inspired by Lennon's (2000) definition of higher-order fluency: "the rapid, smooth, accurate, lucid and efficient translation of thought into language" (p. 40). 'Smooth' and 'efficient' were taken verbatim from this definition. Lucid was translated into a more frequently used synonym ('clear'), which corresponded to some of the comments that instructors made. 'Accurate' was not included in this definition because as mentioned, only comments that showed how peripheral elements of speech could be evidently related to the flow of speech either within or between participants. Only one comment referred to the value of accuracy has being a crucial component of speech - "accuracy and comprehensibility

- you need them for fluency” (Instructor 4) – which did not explicitly reference how accuracy affects the flow of speech. Notably however, the interrelationship between accuracy and fluency is discussed in the ‘Clarity’ section of this chapter.

What is evidently absent from Lennon’s (2000) definition, but present in instructors’ discussions, are themes related to the flow of speech between participants: facilitating topics and turns and supporting the conversation partner. These themes relate to components of interactional competence (see Galaczi & Taylor, 2018, p.227, for a clearly illustrated schematic diagram of interactional competence).

Only comments about features of participants’ speeches that were evidently connected to the flow of speech were coded. For instance, in the following example, the instructor makes an inference about a speaker’s limited range of vocabulary and connects it to her observation about the speaker’s apparent lack of conversational fluency:

That’s often the trouble with people who are not that fluent is they don’t have enough lexis to keep a conversation going. So they go back and start to say the same thing over again - repeat themselves.” (Instructor 3)

It should therefore be noted that any references to aspects of oral proficiency peripheral to the fluency construct that were not explicitly shown to be connected to the speed or flow of speech within or between participants were not coded, and therefore, excluded from the results. For example, references to lexical, grammatical, or phonological accuracy (e.g. “good pronunciation, good use of grammar”...etc.) or references to the use of body language (e.g. “nice eye contact”) that did not appear to contribute to the flow of individual or conversational speech were ignored and not included in the results.

The six themes and their associated categories and codes containing relevant participant quotes are provided in the table below. These themes are arranged so that core features of

fluency (smoothness, efficiency), which include references to length of runs, silent and filled pauses, repairs and repetitions, are followed by features that become more and more peripheral to the construct (sophistication, clarity, facilitating topics/turns, and supporting the conversation partner) which include references to lexical range, comprehensibility, and interactional competence. The numbers represent the frequency of instructors' responses related to this theme/category per performance group. As mentioned, instructors were asked to compare speakers' performances with one another and then categorize them into three groups: high, middle, and low. It is important to remind the reader that, since speakers were previously judged to be at least at a low-intermediate level of oral proficiency, the low group does not refer to a beginning-level group. The low group, in this case, refers to students at a low-intermediate to intermediate level of fluency; the middle group refers to speakers at a high-intermediate level of fluency; and the high group refers to speakers at an advanced level of fluency.

Table 4

Themes, Categories, Sub-Categories, and Comments per Group (High, Middle, and Low)

<u>Theme</u>	<u>Categories</u>	<u>High</u> <u>(n = 4)</u>	<u>Mid</u> <u>(n = 4)</u>	<u>Low</u> <u>(n = 6)</u>	<u>Total</u> <u>(n = 14)</u>
Smooth	Length of runs	3	6	5	14
	Intonation contour	3	2	2	7
	Linked speech	2	1	4	7
	Discourse cohesion	2	1	2	5
	Total	9	10	11	30
Efficient	Ease of lexical retrieval	4	6	5	15
	Pause frequency	0	3	4	7
	Filled pauses	0	4	3	7
	Pause length	0	2	3	5
	Repetitions	0	2	3	5
	Pause location	3	0	0	3
	Repairs	2	3	0	4
	Total	9	20	16	42
Sophisticated	Lexical resources	2	6	2	10
	Topic knowledge	4	3	2	9
	Total	6	9	4	19
Clear	Clarity (speed)	2	4	2	3
	Clarity (flow)	3	4	2	4
	Total	5	8	4	17
Facilitative	Extending topics and turns	5	5	5	15

	Sustaining turns	2	2	2	6
	Filling gaps	4	0	2	6
	Initiating/transitioning between topics	2	3	1	6
	Total	13	10	10	33
Supportive	Using supportive body language	4	4	1	9
	Responding with short lexical phrases	5	2	0	7
	Providing non-lexical backchannels	3	1	2	6
	Providing one-word responses	2	1	1	4
	Total	14	8	4	6
TOTAL		63	75	60	198

It should also be noted that there existed a seventh theme, “rapid”, which was originally created from analysing the results. However, this category was removed for several key reasons. First, as De Jong (2018) noted in her state-of-the-art review of fluency and language testing, current well-known rubrics and performance benchmarks do not include references to overall speed, as it may be more indicative of one’s personal speech style than of one’s own level of fluency. Second, as to be discussed in the “Phase Two – Method” section of this paper, during initial piloting of the rating scale, the participants had great difficulty assessing speakers according to seven criteria per speaker. Furthermore, they requested that of all criterion, the ‘rapid’ criteria should be removed, as references to speech rate within the ‘rapid’ category were believed to overlap with the “finds words quickly” descriptor in the “Efficient” category. Finally, since the rating scale is intended, beyond this study, for future use in classroom settings, encouraging learners to speak faster is not helpful and may in fact be detrimental to learners by potentially increasing their anxiety and/or decreasing their level of comprehensibility (Chambers, 1997). For these reasons mentioned, the ‘rapid’ category was removed from the results.

The following discussion presents results for each of the six themes and includes tables that illustrate how specific comments from instructors’ were adapted by the researcher to become scale descriptors for each performance group (high, middle, and low). Additionally, quotes from

instructor-participants are discussed to highlight how peripheral fluency features interact with core fluency features, and therefore, likely belong in a fluency rating scale.

4.2. Smoothness

Fluency has long been characterized by the overall smoothness of speech (Koponen & Riegenbach, 2000). According to the results of this study, ‘smoothness’ is comprised of the following categories: (1) length of runs; (2) intonation contour; (3) linked speech; and (4) discourse cohesion. The table below illustrates how specific comments from instructors were adapted by the researcher to become scale descriptors for each performance group (high, middle, and low) as categorized by length of runs, intonation contour, linked speech, and discourse cohesion. No descriptors were created for categories/performance groups where no comment existed; for instance, in the final category (discourse cohesion) and performance group (low), no comments exist so no descriptors were created.

Table 5

Smooth: Categories, Groups, Comments, and Descriptors

<u>Category</u>	<u>Group</u>	<u>Comments</u>	<u>Descriptors</u>
Length of Runs	High	<i>“(She) speaks in long sentences - long flowing runs” (Instructor 4)</i>	produces long, smooth, and cohesive utterances between pauses
	Middle	<i>“There were a few long utterances but there were a lot of shorter utterances with pauses in between” (Instructor 5)</i>	produces some long utterances between pauses but overall the speech may lack smoothness and cohesion
	Low	<i>“(He) struggled to keep a fluent run going or utterance going and a lot of difficulty articulating himself” (Instructor 5)</i>	Struggles to keep a fluent run going; may be content to produce short runs just to get the meaning across.

		<i>“(He) was just happy with shorter sentences and just getting the meaning across” (Instructor 1)</i>	
Intonation Contour	High	<i>“(She) very much imitates the intonation patterns and the flow of an English speaker...natural cadence, nice chunking” (Instructor 2)</i>	See length of runs (high)
	Middle	<i>“intonation is quite smooth, flows nicely” (Instructor 7)</i>	See length of runs (middle)
		<i>“the way the groups of words are pronounced -where the pauses go and where the inflections go so it doesn’t flow smoothly” (Instructor 2)</i>	
	Low	<i>“both of them choppy, no chunking” (Instructor 4)</i>	See length of runs (low)
Linking	High	<i>“(He) was a little more fluent cause he linked his speech” (Instructor 7)</i>	See length of runs (high)
	Middle	<i>“very smooth, linked his speech together” (Instructor 3)</i>	See length of runs (middle)
		<i>“(she) doesn’t link at all” (Instructor 7)</i>	
	Low	<i>“I found (him) very hesitant in his speech, very slow, not linking words together” (Instructor 3)</i>	See length of runs (low)
		<i>“Neither one of them is able to link language together. No linking.” (Instructor 4)</i>	
Discourse Cohesion	High	<i>“good connecting strings of utterances together” (Instructor 4)</i>	See length of runs (high)

	<i>“cohesive use of discourse markers” (Instructor 6)</i>	
Middle	<i>“he would make longer sentences....he would make compound sentences or complex sentences - putting two simple sentences together” (Instructor 1)</i>	See length of runs (middle)
Low	<i>No comment</i>	See length of runs (low)

Notably, of these four categories, only ‘length of runs’ is widely agreed to be considered as a ‘core’ fluency feature. Thus, it is surprising that although fluent speech is often equated with smoothness, the aesthetic quality of smoothness seems to be perceptually constructed by peripheral features such as intonation, linking, and discourse cohesion. However, these three features can still be subsumed under the core feature – length of runs. The discussion below explains how these features are embedded in the creation of long runs of speech.

It should also be noted that, in order to maintain ‘smoothness’ as a core feature of speech, only ‘length of runs’ is explicitly mentioned in the fluency rubric descriptors. The other features are implied through references to ‘smooth’ and ‘cohesive’.

4.2.1. Length of runs

The mean length of spoken runs (MLR) produced between pauses has long been perceived as a core feature of fluent speech (Wood, 2010). MLR has been shown to be linked to the use of formulaic language (e.g. Raupach, 1980; Towell et al., 1996; Wood, 2006; 2010). Notably, formulas may enhance utterance length, as evidenced through comparative or longitudinal measures of MLR (e.g. Towell et al., 1996), formula-per run- ratio (FRR) (Wood, 2010), and through qualitative analysis of their functions in monologic (Wood, 2006) and dialogic (House, 2015) speech. Moreover, as shown in the discussion that follows, MLR is

inherently connected to other three sub-categories (intonation contour, discourse cohesion, and linking), affecting one another in interrelated ways to influence perceptions of smoothly/flowing or choppy/halting speech.

Also, notably, in the realm of high-stakes language testing, references to length of runs can be found in well-known fluency rubrics and performance benchmarks. For instance, the oral fluency scoring criteria for the Pearson Test of English Academic (PTE Academic, 2001) for level 2 on a six-point (0 to 5) states: “Speech (if ≥ 6 words) has at least one smooth three-word run”. Additionally, the Common European Framework for References (CEFR, 2001) performance benchmarks for fluency describe speech rated at the lowest levels as containing “very short isolated utterances” (A1) and “very short utterances” (A2). Similarly, as professed in the CEFR rubrics, in this study, instructors associated a speaker’s short length of runs with lower levels of fluency:

“(He) struggled to keep a fluent run going or utterance going and a lot of difficulty articulating himself” (Instructor 5)

“(He) was just happy with shorter sentences and just getting the meaning across” (Instructor 1)

Instructors observed that speakers at the mid-range level produced some longer runs:

“There were a few long utterances but there were a lot of shorter utterances with pauses in between” (Instructor 5)

However, at the highest level, speeches were characterized by long flowing runs:

“(She) speaks in long sentences - long flowing runs” (Instructor 4)

These comments illustrate how instructors are able to identify spoken runs of varying lengths and to use these observations to make inferences about speakers’ abilities to produce smoothly flowing speech. Surprisingly, in a similar study, Préfontaine and Kormos’ (2016) study investigating raters’ qualitative perceptions of fluency, despite the wide variety of linguistic

features mentioned to be associated with fluency, length of runs was not among them. Instead, raters in that study identified related measures of speech rate and pause phenomena as the primary indicators of fluent speech, along with a variety of other non-temporal features (as in the present study). Thus, the present study contributes to this realm of research in that regard.

In several quantitative studies however, speakers' MLR has revealed that listeners, despite not verbalizing their observations about utterance length, tacitly relate longer length of runs with more fluent speech (e.g. Towel et al., 1996; Kormos & Dénes, 2004; Bosker et al., 2013). In this study, instructors' explicit references to learners' length of runs is reflected in their implicit categorizations of learners by performance levels: (1) high, middle, and low. As a reminder to the reader, raters categorized learners' performances relative to one another: high = mid-to-high advanced; middle = high-intermediate to low-advanced; low = low-intermediate to intermediate.

4.2.2. Intonation contour

Like fluency, second language intonation has often been described, as evidenced in the participant-quotes in this study, as being 'smooth' or 'choppy'. As Wennerstrom (2001) argues, fluency and intonation are strongly intertwined conceptually, illustrating her point through her critique of the Foreign Service Institute's (FSI) four-point fluency scale, "the terms "halting" and "fragmentary" refer to length and frequency of pauses...in contrast; terms such as 'smooth' and 'flow' refer to a lack of interruption in the intonation contour (p. 244). Hence, the decision was made to group instructors' references to intonation under the core fluency category of 'Smooth', despite disagreements in the literature and in high-stakes testing rubrics about whether or not intonation contours comprise 'core' features of fluent speech.

Length of runs and intonation contours are inherently related in producing smoothly flowing speech. Long lengths of runs may be composed of a series of FS linked together, each of which exhibits unique prosodic features comprising its own intonation unit (Lin, 2018; Lintunen et al., 2016). As Celce-Murcia et al. (2010) describe, if produced at a reasonable rate, without any internal pausing, sequencing these intonation units produces a smooth speech stream that combines multiple intonation units into a “complex rhythmic unit” (Pike, 1972). This complex rhythmic unit produces only relatively few prominent elements; however, “too many pauses and therefore intonation units can slow speech down to create too many prominent elements” (p. 232). In other words, too many prominent units, punctuated by pauses, may sound choppy to the listener. In the present study, instructors made connections between intonation and the speed and flow of speech. For more fluent speakers, their use of intonation is perceivably smooth:

“(She) very much imitates the stress patterns and the flow of an English speaker...natural cadence, nice chunking” (Instructor 2)

“Intonation is quite smooth, flows nicely” (Instructor 7)

For less fluent speakers, their use of intonation is perceivably choppy:

“The way the groups of words are pronounced, where the pauses go and where the inflections go, it doesn’t flow smoothly” (Instructor 4)

“Both of them choppy, no chunking” (Instructor 2)

As these comments suggest, pauses are only one of several prosodic features that may influence perceptions of speech discontinuity. Although pauses and other temporal prosodic features (syllable-lengthening and anacrusis) are the most salient, perceptions of speech discontinuity may be affected by prosodic features relating to pitch reset and segmental linking (Knowles 1991, p. 153, as cited in Lin, 2018). Since these features (pauses, pitch, segments) are integral to the formation of intonation units (Lin, 2018), it is not surprising then, that listeners

often identify intonation as one of the key linguistic features affecting perceptions of fluency. Both quantitative studies correlating raters' assessments of fluency with assessments of intonation (e.g. Derwing et al., 2004, Rossiter, 2009) and qualitative studies involving analysis of raters' verbal reports (Préfontaine & Kormos, 2016; Götz, 2013; Freed, 1995) have revealed the salience of intonation on perceptions of fluency.

In high-stakes testing rubrics, intonation and fluency are grouped together under the 'Delivery' category of the TOEFL speaking rubrics (Educational Testing Services, 2015) in contrast to how it is presented in the FCE speaking rubrics (FCE, 2015). However, despite these associations, contrary to the TOEFL speaking rubrics, fluency and intonation are often defined separately in the speaking rubrics for the First Certificate of English (FCE) – “very little hesitation” (FCE, 2015, p. 58) – references to hesitations and speech rate are categorized under 'Discourse Management (FCE, 2015) while references to intonation are depicted under 'Pronunciation' – “intonation is appropriate” (FCE, 2015, p. 58).

According to Pike (1972), intonation is its own isolatable speech component. Therefore, despite instructors' observations of the connections between fluency and intonation, to be conservative, explicit references to intonation were not included in the fluency rubrics for core elements of fluency. Instead, as per Wennerstrom (2000), intonation was subtly referred to in the descriptors by the aesthetic quality of appearing 'smooth' or 'choppy'.

4.2.3. Discourse cohesion

Lexical features of speech, comprising discourse markers (Crible, 2019) (a.k.a. 'small words', Hasselgren, 2002) may also affect perceptions of smoothly flowing speech, as shown in comments from instructors below about more fluent speakers' performances:

good connecting strings of utterances together (Instructor 4)

cohesive use of discourse markers (Instructor 6)

This link between fluency and cohesion in discourse is also depicted in several well-known high-stakes fluency testing rubrics such as the International English Language Testing System (IELTS), First Certificate of English (FCE), and Cambridge English Proficiency (CEP). For instance, the rubrics for the IELTS speaking test are currently comprised of four categories, one of which is “Fluency and Coherence” (British Council, 2015). This category includes descriptors connecting temporal features with knowledge of cohesive devices and rules of coherence, as per Canale’s (1983) amendment of Canale and Swain’s (1980) operationalization of communicative competence, by illustrating that “Discourse Management” should supplement operationalization of the construct. The IELTS, FCE, and CEP all seem to align with this framework for communicative competence, and fluency, as a performance-based skill, through this lens, seems to fit best under ‘Discourse Competence’. Thus, in the IELTS, FCE, and CEP rubrics, descriptors for speech temporality are interwoven with descriptors of cohesion and coherence.

In relation to instructors’ comments about discourse markers, several studies have demonstrated that discourse competence, as evidenced through speakers’ use of discourse markers/small words (e.g. oh, well, okay, yeah), may be indicative of fluent speech. Notably, more fluent speakers have been shown to use these devices more frequently than less fluent speakers (Crible, 2019; Crible & Pascaul 2018; House, 2015; Götz , 2013; Carroll, 2004; Hasselgren, 2002; Ezjenberg, 2000; House, 1996) for varying functions both within-turns to express relationships between utterances or between-turns to create “smooth continuity in on-going talk” (House, 2015, p. 67).

4.2.4. Linking

Learners' use of phonological linking which refers to "the connecting of the final sound of one word or syllable to the initial sound of the next" (Celce-Murcia et al., 2010 p. 165) may perceivably create smooth connections within intonation units embedded within long runs of speech. As Lin (2018) describes, formulas exhibit unique prosodic properties such as phonological reductions through linking, which tend to occur within the first few syllables uttered; relatedly, these first few syllables are uttered more quickly than the rest of the formula (a.k.a. anacrusis). Since the acquisition of FS is a requisite to producing fluent speech (Pawley & Syder, 2000), the presence of linking, most notably within formulas, may be an indicator of a higher-order fluency. Raters in Waniek-Klimczak's (2014) study, for instance, noted that learners' use of formulas in which linking was evident (e.g. 'kinda', 'gotcha'), was perceived favorably by raters in their fluency judgments. Relatedly, Hieke (2005) discovered that one type of linking, consonant attraction (CA), should be a salient indicator of speech as more fluent learners produced a higher rate of CA links.

In the present study, instructors made connections between levels of fluency and the amount of linking. Perceivably, more fluent speakers linked their speech effectively:

(He) was a little more fluent cause he linked his speech (Instructor 7).

Very smooth. Linked his speech together (Instructor 3).

On the other hand, less fluent speakers perceivably did not:

I found (him) very hesitant in his speech, very slow, not linking words together (Instructor 3)

Neither one of them is able to link language together. No linking. (Instructor 4)

Although Hieke (2005) and Waniek-Klimczak (2014) posed the argument that linking should be considered to be a marker of fluent speech, it would appear that not much research has since investigated this area; moreover, the amount that English speakers link speech in any given

situation is unpredictable as it depends largely on speech rate, speech style, and the formality of the situation (Celce-Murcia et al., 2010) as “un-emphasized words become more and more reduced as speech becomes more rapid and more informal” (Chun, 2012, p. 83). Thus, references to ‘linking’ or ‘elision’ do not appear explicitly in fluency rubrics in the present study; instead, as discussed earlier regarding intonation, and as per Wennerstrom’s (2000) observations, perceptions of linking are implied under the descriptor, *Smooth*.

In sum, as observed by instructors in this study, speakers’ length of runs, intonation contours, discourse markers, and the use of linking interact with speech rate, pause phenomena, and formulaic sequences in a variety of ways to affect the perception of smoothly flowing speech. The qualitative results from this study thus seem to align with previous quantitative research indicating that speakers’ length of spoken runs is a key component of fluent speech.

4.3. Efficiency

Similar to the *Smoothness* category, fluent speech is often qualified by descriptors such as ‘effortless’ and ‘efficient’ (Lennon, 2000; Préfontaine & Kormos, 2016). These two categories - *Smoothness* and *Efficiency* - are similar in nature but there is a key difference to how they are conceptualized in this study. While *Smoothness* reflects observations about the aesthetic quality of speakers’ utterances, the *Efficiency* item reflects inferences made about the automatization of the cognitive processes involved in producing these utterances. In the present study, instructors made observations about the number and length of silent pauses, filled pauses, repetitions, and repairs, often relating these breakdowns to a lack of efficiency in retrieving vocabulary or organizing utterances.

The table below depicts selected responses from instructors and their associated descriptors for each performance group (high, middle, and low) for each of the categories: ease

of linguistic retrieval, pause frequency, pause length, pause location, filled pauses, repetitions, and repairs. Notably, not all categories contain comments from instructors for several categories (e.g. silent pauses, filled pauses, and repetitions); in these cases, scale descriptors were not created from instructors' comments.

Table 6

Efficient: Groups, Categories, Comments, and Descriptors

<u>Groups</u>	<u>Categories</u>	<u>Comments</u>	<u>Descriptors</u>
Ease of Linguistic Retrieval	High	<p><i>“slowing down here, finding the right technical language to explain what she’s doing” (Instructor 1)</i></p> <p><i>“she’s looking for those rarely used words, words I would have to search for too when discussing complex things and ideas (Instructor 2)”</i></p>	finds words quickly but may hesitate or slow down slightly to retrieve infrequent words (e.g. technical jargon) or when discussing complex ideas
	Middle	<p><i>“I wouldn’t say that he was a low fluency because he certainly was able to sustain speech. Like not big gaps. Like if someone is low fluency there’s going to be big gaps in the speech because they’re really searching for words” (Instructor 4)</i></p> <p><i>“It’s a pretty lower-intermediate word but he’s still struggling with those which means he’s thinking....it may not be necessarily vocabulary. It may be ideas...could be retrieving ideas. So it’s not only vocabulary”</i></p>	may pause, hesitate or slow down to retrieve both frequent and infrequently-used words/phrases or when discussing complex ideas

		(Instructor 1)	
	Low	<i>“he’s often trying to retrieve something that is easily retrievable but he’s just struggling” (Instructor 1)</i>	pauses, hesitates, or slows down noticeably when retrieving frequently-used words or phrases
		<i>“(He) had much more hesitation - what I would call searching for vocabulary - so there was more disfluency in his hesitations” (Instructor 3)</i>	
Pause Frequency	High	No comments	See ‘Ease of linguistic retrieval’ (High)
	Middle	<i>“a few pauses” (Instructor 6)</i>	See ‘Ease of linguistic retrieval’ (Middle)
	Low	<i>“pauses noticeably” (Instructor 7)</i>	See ‘Ease of lexical retrieval’ (Low)
Pause Length	High	No comments	See ‘Ease of linguistic retrieval’ (High)
	Middle	<i>“a bit of an awkward long pause right there” (Instructor 6)</i>	See ‘Ease of linguistic retrieval’ (Middle)
	Low	<i>“couple of long pauses” (Instructor 7)</i>	See ‘Ease of linguistic retrieval’ (Low)
Pause Location	High	<i>“there were those pauses in between but where in the sentence those pauses were happening was optimal” (Instructor 5)</i>	Pauses naturally and in the appropriate places
		<i>“Her hesitations were natural - just like you and I”. (Instructor 4)</i>	
	Middle	No comment	No comment
	Low	No comment	No comment

Filled Pauses	High	<i>No comment</i>	No comment
	Middle	<i>Quite a few 'ums' and 'ahhs' (Instructor 6)</i>	See 'Ease of linguistic retrieval' (Middle)
	Low	<i>A lot of filled pauses (Instructor 5)</i>	See Ease of linguistic retrieval' (Low)
Repetitions	High	No comments	No comment
	Middle	<i>"a certain amount of repetition" (Instructor 4)</i>	See 'Ease of linguistic retrieval' (Middle)
	Low	<i>"a lot of repetitions" (Instructor 5)</i>	See 'Ease of linguistic retrieval' (Low)
Repairs	High	<i>"some things that (she) exhibited were immediate self-corrections, especially around 5:19 - plural to singular noun form that she made a quick self-correction" (Instructor 6)</i> <i>"she even corrects herself in the right way. She would be talking along and all of a sudden, her mind has gone faster than her mouth. But she kept going and fixed it along the way" (Instructor 3)</i>	makes quick self-corrections
	Middle	<i>"a lot of self-corrections and false starts" (Instructor 2)</i>	See 'Ease of linguistic retrieval' (Middle)
	Low	<i>No comment</i>	No descriptor

4.3.1. Ease of linguistic retrieval

Fluent speech is believed to be largely dependent on one's ability to retrieve lexis efficiently, as revealed in studies analysing reaction time to lexical categorization tasks

(Segalowitz, 2007) and as reflected in the CEFR performance benchmarks for fluency for high-intermediate learners: “he/she can be hesitant as he/she searches for patterns and expressions” (Council of Europe, 2001, p. 28). Notably, in Préfontaine and Kormos’ (2016) study of raters’ perceptions, ease of lexical retrieval was also shown to be a salient feature of fluent speech. In the comment below, the instructor inferred that this speaker was struggling at the formulation stage of speech development (Levelt, 1989) as he needed to devote a high level of attention and effort through controlled processing to recall language that was not fully automatized.

Here, he’s pausing, repeating, struggling. It’s definitely vocabulary and he’s trying to retrieve something that is easily retrievable but he’s just struggling because he’s thinking about two things at the same time - trying to construct and vocabulary. (Instructor 1)

4.3.2. Pause location

Several studies have shown how pause location is a marker of fluency (e.g. Shea & Leonard, 2019; DeJong, 2016; Kahng, 2015; Kahng, 2014). De Jong (2016) notes that pauses produced at clause boundaries reveal that speakers are using planning time to produce the content of the utterance, reflecting delays at the conceptualization stage of speech development. On the other hand, pauses produced within clause boundaries are more indicative of planning time used to retrieve lexical items and organize the grammatical structure of the utterance, reflecting delays at the formulation stage of speech development. Notably, in previous studies eliciting raters’ verbal reports on fluency judgments (e.g. Préfontaine & Kormos, 2016), no references were made to pause location. In the present study however, some instructors made references to pause location either explicitly...

There were those pauses in between but where in the sentence those pauses were happening was optimal. (Instructor 5)

...or tacitly, by referring to the naturalness of the pause:

Her hesitations were natural - just like you and I. (Instructor 4)

4.3.3. Repairs

As exhibited in the following quotes, instructors noted that more fluent learners were able to make quick and natural-seeming repairs, allowing them to maintain the flow of speech:

Some things that (she) exhibited were immediate self-corrections, especially around 5:19 - plural to singular noun form that she made a quick self-correction (Instructor 6).

She even corrects herself in the right way. She would be talking along and all of a sudden, her mind has gone faster than her mouth. But she kept going and fixed it along the way (Instructor 3).

When speakers make immediate self-corrections, they seem to be exhibiting a high degree of ‘attentional control’ (Segalowitz, 2007) as they have automatized lower-order processes of formulation and articulation so that they can devote attentional resources to higher-order processes of speech monitoring, as per Levelt’s (1989) model. With this in mind, it is not surprising that self-corrections are often classified as advanced communication strategies used to maintain speech flow (Peltonen, 2017b; Götz, 2013, Lennon, 1990; Riggensbach, 1991). The way that speakers use repairs to maintain fluency is largely believed to be idiosyncratic (Tavakoli, 2016) suggesting that learners need to have a high degree of ‘processing flexibility’ (Segalowitz, 2010) to be able to adapt efficiently to temporary breakdowns in speech production. Likewise, one’s repertoire of repairs may be related to one’s ability to use a wide variety of repairs in various ways to maintain fluent speech. Skehan et al. (2016), for instance, discovered that quantitative measures of repairs (filled pauses, self-corrections, repetitions) correlated strongly and significantly with quantitative measures of speakers’ range of lexical diversity (VocD). As a result, Skehan concluded that less fluent speakers “are not able to access such a wide range of words, and are more often in trouble with those that they can access” (p.110). The inherent relationship between lexical range and fluency is discussed in the next section.

4.4. Sophistication

The results from this section provide some information about how instructors make inferences about the relationships between vocabulary range, topic knowledge, and willingness to communicate (WTC), and their contributions to fluent speech. These relationships have been alluded to previously in the literature and in performance benchmarks and testing rubrics.

Fillmore (1979), in his seminal paper on L1 fluency, described one of the four kinds of fluency as the ability to speak appropriately, effectively, and confidently in a wide variety of contexts.

The Canadian Language Benchmarks (CLB) (Pawlikowska-Smith, 2002) performance benchmarks for speaking proficiency differentiate between high-intermediate levels of fluency and advanced levels of fluency by speakers' abilities to speak fluently on increasingly demanding topics in increasingly demanding contexts. Similarly, the IELTS 'Fluency and Coherence' rubrics indicate that, at the highest levels, speakers are able to develop topics fully/coherently and appropriately (IELTS, 2015). As shown in the table below, instructors in this study perceived that a speaker's range of linguistic resources enable the speaker to speak with ease and willingness on a wider range of topics whereas a speaker with a limited range of resources may speak with more difficulty and less willingness on a narrower range of topics.

Table 7

Sophisticated: Categories, Groups, Comments, and Descriptors

<u>Categories</u>	<u>Groups</u>	<u>Comments</u>	<u>Descriptors</u>
Lexical Resources	High	<i>"she had lots to draw from...the vocabulary items and stock phrases being used to buy time, I think there was a little bit more variance there and also how they were appropriately used" (Instructor 5)</i>	Draws from a wide range of lexical resources and ideas

Topic Knowledge	Middle	<i>“they were both very slow but I think John had more resources at his disposal. Either way they were able to discuss the topics ok” (Instructor 2)</i>	The range of lexical resources is sufficient for this task
		<i>“keeps using the same formulaic phrases – ‘I mean, I mean’” (Instructor 4)</i>	May overuse some formulaic phrases
	Low	<i>“not a big vocabulary range, relies a lot on practiced structures” (Instructor 3)</i>	shows a limited range of lexical resources
	High	<i>“she didn’t even need to stop and think when she was talking about her research...she could explain complicated things, using complex language, in depth” (Instructor 2)</i>	discusses complex ideas in depth
	Middle	<i>I thought his fluency was low until he started talking about his research and using words like ‘purification’ (Instructor 1)</i>	may use some infrequently-used words and technical jargon
	Low	<i>“I think he ran out of things to say about it” (Instructor 7)</i>	Seems to run out of things to say
		<i>“he keeps using the same phrases over and over, like, if he were talking about something else, something more challenging, he couldn’t use those anymore and I think he’d struggle a lot to find the phrases he needs to keep it all going. (Instructor 7).</i>	Shows a limited range of resources by repeating words/phrases during the conversation

4.4.1. Linguistic resources

Several of these comments are worth discussing in more detail. For instance, as one of the instructors observed, repetitions - not immediate repetitions of words/phrases - but repetitions

of words/phrases, ideas, or topics discussed over the course of conversation, demonstrate a limited range of resources and, relatedly, a lack of fluency:

That's often the trouble with people who are not that fluent is they don't have enough lexis to keep a conversation going. So they go back and start to say the same thing over again - repeat themselves." (Instructor 3)

On the other hand, more fluent speakers could rely on their relatively larger lexical resources to sustain the flow of speech:

She had lots to draw from...the vocabulary items and stock phrases being used to buy time, I think there was a little bit more variance there and also how they were appropriately used" (Instructor 5)

As the instructor observed, different types of sequences may be used in various ways to maintain fluent speech. Different types of sequences include 'time buyers' (e.g. fillers, turn-holders, discourse shape markers) used to signal delays in production or 'processing shortcuts' to aid both the speaker and the listener exchange information more efficiently (Nattinger & DeCarrico, 1992). Wood (2006) showed that immediate-level learners use formulas in individualized, but specific, ways to extend their utterance length and to reduce the amount and length of pausing: repeating formulas; stringing formulas together; relying on one formula; using self-talk and filler formulas; and using formulas as rhetorical devices. As the instructor's preceding comment illustrates, fluent speakers need to have a sufficient range of formulas and a high degree of attentional control in order to use these formulas in varied and communicatively appropriate ways that meet the online demands of any communicative situation (Pawley & Syder, 2000). From a listener's point of view, speakers who use a wide range of formulas in a sophisticated manner impress a high degree of proficiency upon the listener (Boers et al, 2006).

4.4.2. Topic knowledge

Relatedly, instructors commented that learners appeared to be more fluent when they used sophisticated language (e.g. jargon) to express more complex ideas on familiar topics such as their research areas. The following examples illustrate comments from instructors about speakers at high, mid, and low levels of fluency, respectively:

She didn't even need to stop and think when she was talking about her research...she could explain complicated things, using complex language, in depth (Instructor 2)

I thought his fluency was low until he started talking about his research and using words like 'purification'. (Instructor 2)

I think he ran out of things to say about it (Instructor 7)

Additionally, as learners may possess more topic-specific lexis, then topic familiarity may produce a positive affective response within learners that causes them to speak more fluently:

The words are flowing and he's a little bit more articulate, isn't he? Even his vocabulary is a little bit upgraded. I think he's talking about something he feels strongly about. (Instructor 2)

The act of speaking fluently on familiar topics may also encourage one to speak more:

Her personal experience is a familiar topic about which she feels really comfortable. She becomes more fluent and then her fluency actually encourages her to continue talking. (Instructor 1)

In the quotes above, instructors made inferences about the effect of learners' levels of topic familiarity based on perceived increases in fluency and emotional investment in the speech. Relatedly, Nematizadeh and Wood (2019) also noted the effects of topic familiarity on the two-way dynamic relationship between speakers' WTC and fluency. In one particular example, the learner-participant, upon reviewing her performance, discussed that recalling a personal experience increased her WTC, which resulted in the production of the speakers' longest runs of speech. The authors observed that, in addition to discussing personal experiences, other factors

that affected one's level of topic familiarity, and thus affected one's levels of WTC and fluency, were the amount of task preparedness and the number of ideas and supporting arguments from which to draw.

Finally, instructors in this study made predictions that, due to a limited range of resources, learners would likely be disfluent in discussing more demanding topics in more demanding situations:

He keeps using the same phrases over and over, like, if he were talking about something else, something more challenging, he couldn't use those anymore and I think he'd struggle a lot to find the phrases he needs to keep it all going. (Instructor 7).

As shown in the preceding comment, despite the inherent contribution of FL to fluent speech (e.g. Raupach, 1980; Pawley & Syder, 2000; Wood, 2006, 2010), there are different perspectives on how well learners' use of formulas are perceived by raters. On the one hand, Boers et al. (2007) discovered that the number of formulas produced correlated significantly with oral proficiency ratings use of formulaic language enhanced oral proficiency ratings. Similarly, Hasselgren (2002) also found that learners who used more smallwords (e.g. "I mean") were perceived to be more fluent than those who used less and integrated these observations into a fluency rating scale. On the other hand, Filmore (1979/2000) argues that overuse of a limited number of formulaic phrases indicates that the person is "unable to respond creatively to small differences and novelties in situations" (p. 53) arguing that limitations in one's range of formulaic language likely lead to limitations in one's ability to speak appropriately in varying contexts. Similarly, Lennon (2000) states that the overuse of certain formulas is a kind of false fluency, which, "while enhancing temporal fluency results in triteness, and, for this reason, threatens to lose listener attention, thus violating a higher-order fluency" (p. 33-34). In the pilot study that informed this doctoral research project (Williams, 2018), some raters made inferences

and predictions that if learners were over-reliant on certain formulas to discuss a familiar topic, then they would likely struggle to discuss more demanding topics in more demanding situations. Therefore, it would seem that not all the ways in which learners use formulas are equally valued by raters.

4.5. Clarity

This category illustrates relationships between instructors' perceptions of fluency as related to perceptions of intelligibility, which refers to the ease to which listeners can decode individual words, and comprehensibility, which is "a measure of the processability of speech" (Thomson, 2017, p. 23). These comments are included in the table below:

Table 8

Clarity: Categories, Groups, Comments, and Descriptors

<u>Categories</u>	<u>Groups</u>	<u>Comments</u>	<u>Descriptors</u>
Clarity (speed)	High	<i>Sometimes you had to really concentrate because she was really speaking fast (Instructor 3)</i>	A quick speech rate may sometimes distract the listener from focusing on the content
	Middle	<i>"I was pretty comfortable with the speed for the most part. Just a few moments where he could have slowed down for better clarity" (Instructor 2)</i>	The listener (you) feels generally comfortable but it may require some effort to listen to the speech.
	Low	<i>"He was speaking too fast. There was a total patch of I don't know what he said. More than once. I have patches where I didn't know what he said" (Instructor 4)</i> <i>"He was really slow and plodding and rambling. By the end, I couldn't stay focused" (Instructor 2)</i>	A quick speech rate...often interferes with intelligibility A slow speech rate may make it difficult to sustain attention over the course of the conversation

Clarity (flow)	High	<i>“(Her speech) was effortless. I could just relax engage in what she was saying” (Instructor 3).</i>	The listener (you) feels at ease and engaged in the content
	Middle	<i>Some rough incoherent bits of speech, lots of pauses and fillers and what not, not clear what was said (Instructor 6)</i>	Disruptions in the flow of speech may occasionally distract the listener from focusing on content.
	Low	<i>It was really choppy. I had to listen really hard to piece the bits together (Instructor 4)</i>	Disruptions in the flow of speech often interferes with intelligibility

The term clarity subsumes instructors’ perceptions of comprehensibility and intelligibility in relation to the speed and flow of speech. Several studies have investigated the relationships between comprehensibility, accentedness, intelligibility, and fluency (e.g. Saito & Shintani, 2015; Kang, 2012; Derwing et al., 2004). Thomson (2015), upon reviewing a range of studies, concluded that “fluency is most related to comprehensibility, somewhat related to accentedness, and apparently least related to intelligibility” (p. 217). Notably, these terms are often conflated, but in this study, intelligibility refers to the listeners’ ability to decode individual words, and comprehensibility refers to the procesability of speech (Thomson, 2017). In this study, instructors made connections between comprehensibility, intelligibility, and the speed and flow of speech, as in the quotes below:

His speed was too quick. There was a total patch of I don’t know what he said (Instructor 4).

She speaks with a lot more clarity than Marvin but still stumbles over words (Instructor 7).

Much research investigating the relationships between complexity, accuracy, and fluency has provided support for Robinson’s (2001) cognition hypothesis, stating that prioritizing one aspect of speech (e.g. ‘fluency’) may negatively affect another (e.g. accuracy). In other words,

trying to speak faster may affect phonological accuracy as a hurried speech rate may cause phonological simplifications within individual words and/or the omission of weak forms altogether (Celce-Murcia et al., 2010). On the other hand, a slower speech rate can create “too many unintentional prominent units” (Celce-Murcia et al., 2010, p. 222) which can make it more difficult for the listener to attend to the speech (Lennon, 2000). Listeners in Derwing et al.’s (2006) study, for example, stated that they had difficulty listening to disfluent speech and admitting to ‘zoning out’ from time to time.

Likewise, in this study, instructors also made connections between their own ease of listening and speed:

*(Her speech) was effortless. I could just relax and engage in what she was saying”
(Instructor 3)*

Choppy - I had to listen really hard to piece the bits together (Instructor 4)

Clarity of speech in relation to the listener’s level of comfort may be a clue that listeners draw upon to make inferences about fluency. In the pilot study that informed the development of this doctoral research project (Williams, 2018), one of the expert raters, who had extensive testing experience, elucidated upon this inference made by relationships between speech flow and speech clarity, in relation to one’s personal level of comfort. :

At the advanced level, you’re able to relax as an interlocutor because you’re actually communicating. You forget that you’re in a testing situation because it’s (the learner’s speech) is just floating. So if I had to take one big holistic charge (about categorizing fluency), I would say it’s flow...When assessing fluency, there’s a point in the process where you have to think about what you’re doing, how comfortable you are as the interlocutor (p. 39).

Clarity of speech, although influenced by fluency, cannot be divorced from lexical, grammatical, and phonological accuracy. The extent to which accuracy determines fluency is a point of debate. Lennon (2000) argues that the accuracy-fluency polarity appeals to L2 practitioners because it allows for the creation of activities, test tasks, and scoring rubrics that isolate either aspect of speech; yet, in production, the relationship between fluency and accuracy is dynamic and not easily isolated:

Fluency will be impaired when the learner is unable to render thought or communicative intention into language accurately. In other words, the more accurate the translation of thought into language, the more fluent is the speaker...delaying articulation may mean that a lexical item being searched for will be found, whereas maintaining the speech flow may mean perhaps omitting the item, choosing a near synonym, or abandoning or altering the concept. Fossilized learners may be very fluent in their interlanguage, which may be distinct from the target language in various ways (Lennon, 2000, p. 30-31).

In sum, for this category raters make inferences about speech clarity (comprehensibility and intelligibility) to inform their judgements about speech fluency. These inferences are further informed by raters' feelings of personal comfort when listening to the speakers.

4.6. Facilitating topics and turns

As the instructors in the present study have observed, the attainment of higher-order fluency may be predicated on the acquisition of certain interactional skills necessary to “master smooth continuity in on-going talk” (House, 2015, p. 65). In the present study, instructors observed that being perceived as conversationally fluent requires speakers to facilitate the mutual flow of speech (see the table below). These findings are related to previous work on conversational fluency, interactional competence, and non-verbal communication.

Table 9

Facilitating topics and turns: Categories, Groups, Comments, and Descriptors

<u>Categories</u>	<u>Groups</u>	<u>Comments</u>	<u>Descriptors</u>
Extending topics and turns	High	<i>“the other thing you could tell was very fluent was it was very coherent so when one would say something, the other person wouldn’t just nod and then go on with the next question but the person would kind of draw on her experience and then share it” (Instructor 6)</i>	Expands on what the interlocutor has said
		<i>“another thing is the little questions that are asked without slowing down or without interrupting - to ask the question at the right timing to spur the speaker on to speak some more” (Instructor 6).</i>	Asks follow-up questions to keep the conversation going
	Middle	<i>“she was kind of struggling to facilitate the interaction, like articulate the question...not really sure how formulate the question” (Instructor 5)</i>	May show some hesitancy in formulating questions
	Low	<i>“a lot of times when she says something, he doesn’t really comment on it. They just move on to the next topic” (Instructor 7)</i>	Provides mostly short responses to questions or statements
<i>“he’s trying to ask the question but all those pauses and hesitations – having a lot of difficulty articulating the question” (Instructor 5)</i>		May show difficulty in asking questions	
Sustaining speech	High	<i>“She did most of the talking for the last two minutes” (Instructor 6)</i>	May speak for the majority of the conversation

	Middle	<i>“she doesn’t let him speak I feel. Maybe she feels she needs to carry it” (Instructor 7)</i>	may or may not rely on the other speaker to sustain conversation
		<i>“he actually needed to be prompted a few times” (Instructor 3)</i>	
	Low	<i>“I think he was content to let her carry the conversation” (Instructor 2)</i>	Seems to rely on the other speaker to sustain the conversation
Filling Gaps	High	<i>there was a bit of a lull in the conversation around 1:42. And she asked him a question about what his research was on, what applied math means, and she elaborated on it and waited. But then she took back her question and said you don’t have to answer that question (Instructor 6).</i>	May compensate for gaps in the conversation by elaborating or rephrasing questions to give the interlocutor time to plan
	Middle	<i>No comment</i>	No descriptor
	Low	<i>“there were lots of awkward lulls because he wasn’t saying anything so she needed to fill them in to keep it all going” (Instructor 4)</i>	May initiate gaps in the conversation
Initiating /Transitioning Between Topics	High	<i>“she just integrates the prompts so seamlessly without needing to stop” (Instructor 7)</i>	Integrates the task prompts into the conversation without hesitation
	Middle	<i>“it was kind of an abrupt start. It wasn’t smooth or gradual or anything like that” (Instructor 6)</i> <i>“looking down at the sheet again” (Instructor 6)</i>	Shows some hesitancy in integrating the task prompts
	Low	<i>“he starts looking down at the sheet, wondering what to ask her and so there were these awkward silences” (Instructor 6)</i>	May hesitate when integrating the prompts into the conversation

4.6.1. Extending topics and turns

Instructors commented that more fluent speakers were more capable of extending topics to sustain conversation:

She wouldn't just nod and go on to the next question but she would draw on her experience and then share it to kind of show that she feels the same way (Instructor 6).

On the other hand, less fluent speakers seemed to have more difficulty:

A lot of times when she says something, he doesn't really comment on it. They just move on to the next topic (Instructor 7).

Moreover, more fluent speakers were perceived to ask more follow-up questions to facilitate the other speaker's turn:

Another thing is the little questions that are asked without slowing down or without interrupting - to ask the question at the right timing to spur the speaker on to speak some more (Instructor 6).

However, less fluent speakers struggled to ask follow-up questions:

For the purpose of asking a question, it was a very long drawn out question - asking a question, hesitation, reformulation, not really sure about how to formulate the question (Instructor 5)

Topic and turn extension, either through asking follow-up questions or by providing substantive responses to statements, has been argued to be a key component of higher levels of interactional competence (Lam, 2018; Galaczi & Taylor, 2018) which, relatedly, may be indicative of higher levels of conversational fluency (House, 1996; Young & Hallebeck, 1998; Morales-Lopez, 2000). Young and Hallebeck (1998), for instance, discovered that learners with lower speech rates provided shorter and factual answers to questions whereas learners who exhibited higher speech rates expanded and elaborated more on topics. For Morales-Lopez (2000), being conversationally fluent involves achieving conversational parity: “in developing

the topic initiated by an interlocutor, a certain balance between the participants is usually respected, as is the amount of information communicated” (p. 270).

In a testing situation, Galaczi (2008) and Isaacs (2013) have shown that the amount of conversational parity affects raters’ perceptions of interactional performance. Galaczi (2008) analysed raters’ judgments of four different kinds of interactional patterns: a) asymmetric (i.e. one participant speaks much more than the other); (b) parallel (i.e. each participant provides relatively equally long turns, but do not interact substantively); (c) collaborative (i.e. participants provide relatively equal turns and interact substantively); and (d); a mixture of the latter two patterns. Galaczi (2008) discovered that scores were highest for the collaborative group and the lowest for the parallel group, whereas the mean scores for the asymmetric and blended fell in between. Isaacs (2013) revealed similar results. Taken together, these results indicate that conversational parity, potentially enhanced by topic expansion, may be more favorably salient to raters if it results in a collaborative conversation rather than a parallel conversation.

4.6.2. Sustaining turns

Instructors also commented that more fluent speakers were more capable of extending topics to sustain conversation:

She wouldn't just nod and go on to the next question but she would draw on her experience and then share it to kind of show that she feels the same way. (Instructor 6)

Instructors also perceived that speakers who did the majority of the speaking were regarded as being more fluent and those who spoke much less were deemed to be disfluent:

He's letting his interlocutor do much more of the talking so that's definitely a sign of reduced fluency in the conversation (Instructor 7)

As instructors observed, the ability to sustain individual speech for a substantial period of time within a conversation may be a hallmark of higher fluency. House (1996) for instance, notes

that providing ‘substantive turns’ is a key indicator of pragmatic fluency. In her study, through explicit instruction of conversational routines, students increased their range of content-focused gambits allowing them to provide longer and more substantiative turns, thus developing their pragmatic fluency. In high-stakes testing rubrics, the ability to “speak at length without noticeable hesitation” (IELTS, 2015) and to produce “extended stretches of language with flexibility and ease and very little hesitation” (CPE, 2015) is a distinguisher of higher-level fluency.

On the other hand however, to what extent a speaker is able to sustain a turn within a conversation depends inevitably on each speaker’s conversational style and their effect on the nature of the conversation type itself (collaborative, parallel, asymmetric, or blended) as discussed in the previous section (Galaczi, 2008). Speakers’ conversational styles invariably affect the type of conversation as a more high-involvement style speaker may interject themselves more into the conversation through interrupting or overlapping with the purpose of showing interest or creating rapport (Tannen, 1986), thus reducing the size of the turn. Thus, turn size may reveal more about the nature of conversation, influenced by the nature of the speakers’ conversational style, rather than the nature of the speaker’s fluency (this subject will be discussed in greater detail in phase two of this study). It is not surprising then that quantitative measurements of the average number of turns and average turn length has so far not produced any meaningful results (Tavakoli, 2016; Peltonen, 2017). Regardless, instructors in this study still identified the ability to sustain one’s own speech as an indicator of conversational fluency.

4.6.3. Filling gaps

More fluent speakers would also take the initiative to compensate for gaps:

There was a bit of a lull in the conversation around 1:42. And she asked him a question about what his research was on, what applied math means, and she took back her question and said you don’t have to answer that question (Instructor 6).

However, some instructors attributed between-speaker gaps to the less fluent speaker:

There were lots of awkward lulls because he wasn't saying anything so she needed to fill them in to keep it all going (Instructor 4).

Several researchers have suggested that more fluent speakers take the initiative to compensate for gaps in speech, thus scaffolding the other speaker to create a mutual flow of conversational speech. McCarthy (2006) has argued that three aspects of conversational speech comprise the notion of conversational fluency: a) temporal variables; b) formulaic sequences to organize interactions (e.g. gambits, discourse markers, smallwords) and c) scaffolding. Regarding this latter feature, McCarthy (2006) notes that, for conversational fluency to occur, “the conversation itself is fluent. Speakers contribute to each other’s fluency; they scaffold each other’s performance and make the whole conversation flow” (p. 4). Michel et al. (2007) also attribute the effects of scaffolding to enhancing fluency on dialogic tasks. The authors compared performances on monologic and dialogic tasks, revealing a significantly higher speech rate for speakers on the dialogic task and attributed this finding to the observation that “as soon as a participant paused in the dialogic condition, the interlocutor tried to help and immediately started speaking” (p. 255). In a study similar to the present one, Sato’s (2014) verbal report analysis of raters’ perceptions revealed three categories, including temporal measures, turn taking, and scaffolding. Sato discovered that the most conversationally fluent speakers were perceived to scaffold less fluent speakers by taking the initiative to fill between-speaker gaps.

4.6.4. Initiating/transitioning between topics

Instructors commented that the ability to initiate or transition between topics was a marker of fluency. More fluent speakers integrated the task prompts into the conversation without hesitation and in natural manner:

She just integrates the prompts so seamlessly without needing to stop (Instructor 7)

On the other hand, less fluent speakers struggled to do so:

He starts looking down at the sheet, wondering what to ask her and so there were these awkward silences (Instructor 6).

Turn-beginnings have been identified as key moments in conversation where speakers facilitate the direction of not only their own turn but the negotiation of turns between speakers (Schegloff, 2007); as such, turn-beginnings are salient moments that may affect listeners' perceptions of fluency. For instance, L1 English raters in Sato's (2013) study perceived turn-initial silences produced by Japanese L2 English speakers in a negative manner whereas the Japanese speakers did not.

It is not clear from instructors' verbal reports which of these speech features contributed to their perceptions; yet, insights from previous research may provide a few clues. To begin with, some research has shown that discourse markers (e.g. oh, well, okay, yeah), also known as 'smallwords' (e.g. Hasselgren, 2002), were used more frequently by more fluent speakers (House, 2013; Carroll, 2004; Hasselgren, 2002; Ezjenberg, 2000; House, 1996). As Hasselgren (2002) discovered, these devices often occur at the beginning of turns as "premessage starters" (House, 1996) at the beginning of turns to initiate speech and hold the floor. In addition to discourse markers, as House (1996) discovered, more fluent speakers used certain types of gambits, such as up-takers, which are used to encourage the conversation partner to speak, and starters, which are topic initiators. Finally, Young and Hallebeck (1998) also revealed links between rate of topic shifting and speech rate, suggesting that more fluent speakers initiate and transition between topics more smoothly.

4.7. Supporting the conversation partner

The results indicate that some instructors seem to adhere to the notion that conversational fluency may be co-constructed through interaction (Sato, 2014) as they highlighted the importance of the listener's contributions during the interlocutor's turn to support the mutual flow of speech between participants. The 'Facilitative' and 'Supportive' categories are similar in that they show how fluency may be mediated by interaction; however, there are notable differences. Comments within the 'Supportive' category refer to the quality and quantity of contributions that the listener makes to support the interlocutor's turns, which are categorized as follows: (1) Non-lexical backchannels ("uh-hmm"); (2) Providing one-word lexical responses; (3) Responding with short lexical phrases; (4) Using supportive body language.

Table 10

Supporting the Conversation Partner: Categories, Groups, Comments, and Descriptors

<u>Categories</u>	<u>Groups</u>	<u>Comments</u>	<u>Descriptors</u>
Non-lexical back-channeling (e.g. "uh-hmm")	High	<i>"She has some good verbal cues like going 'um-hmm uh-hmm' to show him she's listening"</i> (Instructor 3)	Encourages the interlocutor through a variety of ways such as backchannels...
	Middle	<i>"a lot more 'uh-hmm' 'uh-hmm'"</i> (Instructor 6)	Uses these devices (back-channeling) with less variety and frequency (than the high group)
	Low	<i>He doesn't seem totally engaged. He's just sitting there going 'uh-hmm'. Doesn't really want to keep this thing going.</i> (Instructor 4)	Encourages the interlocutor mostly through backchannels or one-word responses May seem disinterested in keeping the flow of conversation going
Providing one-word responses	High	<i>"just a very brief 'yes' and not taking the turn away from him"</i> (Instructor 6)	Encourages the interlocutor through a

			variety of ways such as...one-word responses
	Middle	<i>“these little responses like saying ‘cool’ or like or saying ‘ugh’ like these were marks of higher fluency” (Instructor 5)</i>	Uses these devices (back-channeling) with less variety and frequency (than the high group)
	Low	<i>“He’s saying ‘yeah’ a lot in response to what she says. But I’m not sure he totally understands her. She has to carry it” (Instructor 4)</i>	Encourages the interlocutor mostly through backchannels or one-word responses; may have difficulty understanding more fluent speakers
Responding with short lexical phrases	High	<i>“The way she responds. It sounds more natural, like ‘oh that’s great’ Keeps things going” (Instructor 6)</i>	Encourages the interlocutor through a variety of ways such as...short lexical phrases
	Middle	<i>“They’re not like the longest runs but saying ‘oh I know’ I found the strategies he was using to work in a way that is familiar to like a native English speaker in terms of like strategies” (Instructor 5)</i>	Uses these devices (back-channeling) with less variety and frequency (than the high group)
	Low	<i>“He doesn’t really say anything to what she said like ‘oh that’s interesting’ or ‘oh that’s great’ or something like that. There’s no response really, just ‘uh-hmm’ (Instructor 7)</i>	Encourages the interlocutor mostly through backchannels or one-word responses
Using supportive body language	High	<i>“Quite a lot of hand gestures from her - gave him encouragement” (Instructor 6)</i>	Uses supportive body language to keep the mutual flow of speech going
	Middle	<i>“Approving what the other person was saying in terms of nodding and keeping that eye contact, trying to keep the conversation going” (Instructor 1)</i>	Uses supportive body language to show engagement in the conversation but the body language may also be static when concentrating on listening

“Her body language is fairly static but I think she’s concentrating hard on listening” (Instructor 2)

Low

“Even though it’s culturally influenced you can tell by the body language if somebody feels comfortable and is excited and able to kind of participate. Cause sometimes yeah, the lack of body language, the lack of eye contact gives it away that they’re not really able to at least to be fluent” (Instructor 7)

Body language is generally less supportive to the other speaker

4.7.1. Backchannels

The findings indicate that there does not seem to be much difference between high-level speakers and mid-level speakers in their use of backchannels to support the conversational partner. The only key difference is the perceived range of lexical phrases, such as “oh, that’s great”, as identified by an instructor in the quote below, that the higher-fluency level group employed. According to the instructors in this study, more fluent speakers, in addition to using non-lexical and one-word lexical backchannels, used more substantive lexical phrases to backchannel as in the example below:

The way she responds. It sounds more natural, like ‘oh that’s great’ Keeps things going” (Instructor 6)

Whereas, speakers in the lowest levels were perceived to not use these phrases with much frequency:

A lot of times when she says something, he doesn’t really comment on it. I mean he does share his own experience but more awkward because he doesn’t really say anything to what she said like ‘oh that’s interesting’ or ‘oh that’s great’ or something like that. There’s no response really, just ‘uh-hmm’ (Instructor 7).

However, previous research in this area has provided mixed results. Previous qualitative analyses have indicated the importance of substantive use of lexical phrases as an indicator of conversational fluency yet quantitative correlational analyses have revealed opposing results. Likewise, Riggenschach (2001) revealed similar results indicating that substantive (content) backchannels were perceived to be more indicative of fluency than non-lexical backchannels, which, in her study constituted over fifty percent of the total number of backchannels produced. However, those results contradict the findings from an earlier study by Riggenschach (1991), who concluded that more backchannels “could indicate genuine interest in the conversation or topic, and thus might serve to move the talk forward” (p. 434). However, they also “may reveal a lack of attendance to or comprehension of the talk and actually disrupt the flow of the conversation. Thus, the frequent occurrence of backchannels does not necessarily indicate fluency or conversational skill-in fact, the contrary may be true” (p. 435). Riggenschach’s statement is reflected in a comment from one of the instructors:

He doesn't seem totally engaged. He's just sitting there going 'uh-hmm'. Doesn't really want to keep this thing going. (Instructor 4)

House (1996) noted that despite explicit instruction of responding routines used to enhance pragmatic fluency, “the all-purpose token yes (was) being indiscriminately overused to fill various interactional slots around turn-taking” (p. 240). House infers that students’ overuse of one-word lexical responses such as “yes” and underuse of lexical phrases indicates a gap in students’ pragmatic fluency. Moreover, the use of the “yes” backchannel may not be a clear indicator of fluency levels. As Liao (2009) indicates, the use of “yes” backchannels is more typical to the informal interview style genre, as a fair amount of information questions are posed and answered, and the goal of interviewing and small talk is to forge a positive social

relationship, largely through mutual agreement. In the present study, the use of ‘yes/yeah’ was perceived both positively...

“Just a very brief ‘yes’ and not taking the turn away from him” (Instructor 6)

...and negatively:

“He’s saying ‘yeah’ a lot in response to what she says. But I’m not sure he totally understands her. She has to carry it” (Instructor 4)

4.7.2. Body language

Instructors also observed that whereas more fluent speakers were generally more active in their use of body language, less fluent speakers were generally more static in their use of body language during others’ turns. Overall, in the present study, instructors commented on learners’ use of eye contact, head nodding, and body position to support the conversation partner. Wolf (2008) discovered the positive effects of head nodding on fluent speech as speakers were most fluent in the non-verbal (head-nodding) condition, and least fluent in the no-backchanneling condition. In the present study, active use of nodding and eye gaze was perceived positively:

“Approving what the other person was saying in terms of nodding and keeping that eye contact, trying to keep the conversation going” (Instructor 1)

On the other hand, inactive use of nodding and eye gaze was perceived negatively, as instructors inferred that the perceived lack of a willingness to communicate related to a lack of fluency:

“Even though it’s culturally influenced you can tell by the body language if somebody feels comfortable and is excited and able to kind of participate. Cause sometimes yeah, the lack of body language, the lack of eye contact gives it away that they’re not really able to at least to be fluent” (Instructor 7)

In particular, instructors noted the influence of gestures on supporting the conversation partner, thus contributing to the mutual flow of speech between speakers:

“Quite a lot of hand gestures from her - gave him encouragement” (Instructor 6)

The instructor may be referring to a particular type of gesture - ‘regulators’ - which direct the back-and-forth nature of speaking and listening” (Rowe & Wharton, 2015, p. 322), which include hand movements, possibly supplemented by prolonged eye gaze and the use of head nodding, to manage turn taking. On the other hand, a lack of gesturing, and thus an absence of regulators, was perceived negatively:

“Her body language is fairly static but I think she’s concentrating hard on listening” (Instructor 2)

Relatedly, maintaining prolonged eye gaze but not providing expressive body language was perceived negatively. This instructor inferred that the speaker may have been struggling to comprehend what the other speaker is saying:

He doesn’t have any body language but he is concentrating on the conversation and trying to understand what’s going on. I felt he was a bit worried about retrieving ideas sometimes if he was in a position to answer he was, it was his turn to start talking. (Instructor 7).

It should be noted that, due to cultural differences in interpreting the appropriateness of the use of body language, the descriptors are necessarily general rather than specific.

Consequently, this lack of specificity potentially affects the reliability of this assessment, if raters have different opinions of what constitutes ‘supportive body language’. Much more research is needed to understand the relationship between speakers’ use of regulators and proficiency levels in contributing to the mutual flow of speech between interlocutors.

Regarding this *Supportive* category, instructors’ comments point to a wider debate regarding how fluency is conceptualized. Freed (2000) questions whether fluency exists within the mouth of the speaker or within the ear of the listener; however, through an interactional perspective on fluency, perhaps fluency is a product of a joint performance between speaker and

listener. It is unclear to what extent non-verbal, non-lexical and non-lexical backchanneling as indicators of 'active listening', according to Galaczi and Taylor's (2018) interactional competence taxonomy, are components of the fluency construct. If fluency is co-constructed through dialogue and the surrounding context, reflecting strong notions of interactional competence (Chaloub-Deville, 2003), then "the notion of oral fluency must be understood not only in terms of production but also in terms of receptive comprehension skills...both factors make it possible for the conversation to progress" (Morales-López, 2000, p. 284). Therefore, in this sense, 'active listening' should contribute to fluency. Additionally, if fluency resides within the individual as an "automatic procedural skill" (Schmidt, 1992), then the speed to which one reacts to an interlocutor's initial contribution and the ability to interweave the interlocutor's input with one's subsequent output may also be considered as an automatic procedural skill.

4.8. Summary

The purpose of the first phase of the study is to explore how EAP instructors' perceptions of fluency translate into an analytic-criterion scale (Fox et al. 2016) which is designed to measure core and peripheral components of fluency on a paired conversational task. Six themes were developed from these discussions, which comprise the six criterial-questions on the rating scale. These criterial questions are arranged on the scale according to their perceived relevance in assessing fluency; thus, core measures of fluency (smooth, efficient) are followed by measures that become more and more peripheral to the fluency construct (sophisticated, clear, facilitating topics and turns; supporting the conversation partner). Within these themes, instructors' comments were further categorized into three groups: high, middle, and low. Selected comments were used and modified to create the rubric descriptors for each performance level: High = Yes (n = 4); Somewhat = Middle (n = 4); Low = No (n = 6).

As a result of these findings, the following construct definition for this task required redevelopment to ensure that the construct definition for fluency clearly aligns with the criteria used to assess fluency. The redefined construct definition is as follows: *on this task, at the highest level, the speaker produces smooth, efficient, sophisticated, and clear speech (adapted from Lennon, 2000), while facilitating topics and turns and supporting the other speaker to keep the conversation going.*

The scale, presented on pages 180 and 181, consists of two parts: the scale items and the descriptors. The first part includes a six-point scale for each of the analytic-criterial question items. Selected comments were used and modified to create the rubric descriptors for each performance level: High = Yes (n = 4); Somewhat = Middle (n = 4); Low = No (n = 6), which comprises the descriptors.

The analytic questions are followed by a seventh question item that was added to the scale so that instructors could provide global judgements about fluency. This seventh holistic item was also included because, notably, in phase two of this study, relationships between analytic items and the holistic item are investigated to further understand how each of the analytic items is relevant to the fluency construct. The second part includes the scale descriptors for each category arranged per performance group.

The scale also includes information about student's first languages and a question eliciting instructors' familiarity with the students' first languages because phase two of this study investigates the potential impact of instructors' accent familiarity on influencing their perceptions and assessments of fluency on a paired conversational task. The scale and associated descriptors are presented on page 180 and 181.

Fluency Rating Scale: 7-minute paired conversational task

Speaker # 1 (Left) _____

Speaker's L1: _____

How familiar are you with the speaker's first language? [Very] 6 – 5 – 4 – 3 – 2 – 1 [Not at all]

Analytic Criteria	Yes	Some-what	No	Comments
<i>Is the speech...</i>				
1. ...smooth?	6 5	4 3	2 1	
2. ...efficient?	6 5	4 3	2 1	
3. ...sophisticated?	6 5	4 3	2 1	
4. ...clear?	6 5	4 3	2 1	
<i>Does the speaker...to keep the flow of conversation going?</i>				
5. ...facilitate topics and turns...	6 5	4 3	2 1	
6.support the conversation partner...	6 5	4 3	2 1	

7. Overall, is the speech fluent? [Yes] 6 – 5 – 4 – 3 – 2 – 1 [No]

Comments about the role of context (topic, situation, partner, culture) on affecting the speaker's fluency:

Speaker # 2 (Right) _____

Speaker's L1: _____

How familiar are you with the speaker's first language? [Very] 6 – 5 – 4 – 3 – 2 – 1 [Not at all]

Analytic Criteria	Yes	Some-what	No	Comments
<i>Is the speech...</i>				
1. ...smooth?	6 5	4 3	2 1	
2. ...efficient?	6 5	4 3	2 1	
3. ...sophisticated?	6 5	4 3	2 1	
4. ...clear?	6 5	4 3	2 1	
<i>Does the speaker...to keep the flow of conversation going?</i>				
5. ...facilitate topics and turns...	6 5	4 3	2 1	
6.support the conversation partner...	6 5	4 3	2 1	

7. Overall, is the speech fluent? [Yes] 6 – 5 – 4 – 3 – 2 – 1 [No]

Comments about the role of context (topic, situation, partner, culture) on affecting the speaker's fluency:

Fluency Rating Scale Descriptors

Yes [6]

- 1. Smooth produces long and cohesive utterances between pauses
- 2. Efficient finds words quickly but may pause, hesitate (e.g. “um”), or slow down slightly to retrieve infrequent words (e.g. technical jargon) or when discussing complex ideas; pauses naturally and in the appropriate places; makes quick self-corrections
- 3. Sophisticated draws from a wide range of lexical resources and ideas; seems comfortable with all topics; predicted to be fluent on more demanding topics and contexts; discusses complex ideas in depth
- 4. Clear the listener (you) feels at ease and engaged in the content as a quick speech rate or disruptions in the flow of speech do not interfere with comprehensibility and intelligibility
- 5. Facilitates topics and turns integrates the task prompts into the conversation without hesitation; expands on what the interlocutor has said or asks short follow-up questions to keep the conversation going; may need to speak for the majority of the conversation to keep it going; may compensate for gaps by elaborating or rephrasing questions to give the interlocutor time to plan
- 6. Supports the interlocutor encourages the interlocutor through a variety of ways such as backchannels, one-word responses, and short lexical phrases (e.g. “oh that’s great), or uses supportive body language to keep the mutual flow of speech going

[5] shares some features of both [6] and [4]

Somewhat [4]

- 1. Smooth produces some long utterances between pauses but overall the speech may lack cohesion
- 2. Efficient may pause, hesitate, repeat, or slow down to retrieve both frequent and infrequently-used words/phrases or when discussing complex ideas
- 3. Sophisticated shows some limitations in lexical resources (e.g. by overusing some formulaic phrases); may use some infrequently-used words and technical jargon; the range of lexical resources and ideas is sufficient for this task but may struggle with more demanding topics and in more demanding contexts
- 4. Clear the listener (you) generally feels comfortable but it may require some effort to listen to the speech; a quick speech rate or disruptions in the flow of speech may occasionally distract the learner from focusing on the content
- 5. Facilitates topics and turns may show hesitancy in integrating the task prompts or formulating questions/responding to statements to keep the conversation going; may rely on the other speaker to sustain conversation
- 6. Supports the interlocutor encourages the interlocutor through backchannels, one-word responses, and short lexical phrases to keep the flow of conversation going but uses lexical phrases with less frequency than the high group; uses supportive body language to show engagement in the conversation but the body language may also be static when concentrating on listening

[3] shares some features of both [4] and [2]

No [2]

- 1. Smooth struggles to keep a fluent run going; may be content to produce short runs just to get the meaning across
- 2. Efficient pauses, hesitates, repeats, or slows down noticeably when retrieving frequently-used words or phrases or when discussing complex ideas
- 3. Sophisticated shows a limited range of lexical resources by repeating words/phrases or topics during the conversation; seems to run out of things to say; uses some technical jargon when discussing familiar topics; shows difficulty with some topics on the task; would likely have difficulty with more demanding topics and in more demanding contexts due to a limited range of resources
- 4. Clear the listener (you) may feel uncomfortable as it often takes effort to listen to the speaker as a quick speech rate or disruptions in the flow of speech often interferes with comprehensibility and intelligibility; on the other hand, a slow speech rate may make it difficult to sustain attention on the speaker over the course of the conversation
- 5. Facilitates topics and turns seems to rely on the other speaker to facilitate or sustain the conversation; may show difficulty asking questions; may hesitate when integrating the prompts into the conversation; provides mostly short responses to questions or statements; may initiate gaps in the conversation; the interlocutor may need to elaborate/rephrase questions to give the speaker time to speak
- 6. Supports the interlocutor encourages the interlocutor through backchannels or one-word responses; body language may be generally less supportive to the other speaker and it may be static while listening as the speaker may have difficulty comprehending more fluent speakers; may seem disinterested in what the other person is saying or in keeping the flow of conversation going

[1] does not meet criteria for [2]

Chapter Five: Method (Phase Two)

5.1. Overview

To recall, Creswell and Plano Clark's (2011) two-phase mixed-methods sequential exploratory design (instrument model) provided the underlying methodological framework for the overall study. In this second phase, a new group of 35 instructors were recruited to use the fluency rating scale, constructed from the phase one findings, to assess four video-taped paired conversations involving eight learners. These four video-taped conversations were among the 14 videos analysed in phase one. Before watching each video, instructors, having been provided with information about learners' first languages (e.g. Spanish), were then required to indicate their levels of familiarity with learners' accents on a six-point scale to investigate whether or not instructors' accent familiarity would affect their fluency assessments. After all videos were viewed and the paired conversations were rated, instructors completed a conversational style questionnaire as it was hypothesized that instructors' conversational styles would influence their fluency assessments of learners engaged in a paired-conversation task. Since speakers of opposing conversational styles may possess preferential biases towards one another, it is therefore likely that third-party listeners may also possess these preferential biases (Yule, 1997). Figure 8 is adapted from *Designing and conducting mixed-methods research* (Creswell & Plano Clark, 2011, p. 76).

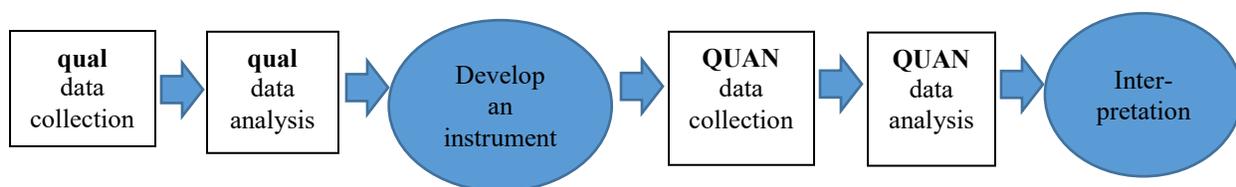


Figure 8: Exploratory Design: Instrument Development Model.

5.2. Ethics and research site

As mentioned in the phase one Method section, ethics clearance was provided for both phases of the study. Ethics clearance was requested from two medium-sized Canadian universities in order to obtain the broadest possible group of instructors from which to recruit. All participants were informed verbally and through written informed consent forms about their roles in the study and about any potential risks associated with their involvement. All relevant information was stored on a personal password-protected computer.

5.3. Participants

5.3.1. EAP instructors ($n = 35$)

Participation was requested through emails and through face-to-face communication. The purposive sampling technique (Teddlie & Tashakkori, 2009) was used to acquire a participant-group of qualified and experienced instructors who could provide informed responses relating to the study's research questions. The resulting participant group (29 females and 6 males) reported possessing the following: (1) at least two years of English language teaching experience in an academic setting; and (2) a graduate degree in Applied Linguistics, Education, or Modern Languages. It should be noted that none of these 35 instructors participated in phase one of the study.

5.4. Instruments

Each of the instruments (fluency rating scale and associated scale descriptors; fluency task specifications; and conversational style questionnaire) used in phase two of this study were piloted with five of the seven instructors whom had participated in phase one.

5.4.1. Fluency rating scale and rubric descriptors.

Please refer to page 180 and 181 to see the scale and descriptors. During the pilot stages, participants aided the researcher in modifying the choice of words/phrases used in the rubric descriptors, and most importantly, participants aided in solidifying the rating procedures.

As for word choice, not all participants in the pilot stage were familiar with some of the terminology. For example, some participants were not familiar with the term ‘backchannels’; thus, participants had suggested that a brief definition of ‘backchannels’ accompany the term.

As for the rating procedures, three key revelations emerged from this pilot stage. First, participants requested the opportunity to practice rating one of the videos in advance to get more accustomed to the scale and to rating two speakers simultaneously. Second, as the task was designed to emulate a real-life testing situation without the use of video-recordings where two students would be talking in front of an assessor, it was necessary for each video to be played only once without stopping. Third, instructors reported being inclined to provide half-points; yet half-points would have distorted the researcher’s ability to analyse ordinal data (Vogt, 2004); thus it was necessary to instruct raters not to provide half-points.

A Cronbach alpha analysis of inter-rater reliability between the five participants was conducted, revealing strong inter-rater reliability coefficients for all seven items on the scale: *Smooth*, *Efficient*, *Sophisticated*, *Clear*, *Facilitative*, *Supportive*, and *Holistic*. For instance, the inter-rater reliability coefficient for *Smooth* was $\alpha = .927$, providing the researcher with confidence in the reliability of the fluency rating scale.

5.4.2. Fluency task specifications

The task specifications were outlined on a one-page document detailing the following: (1) a general description of the task, which included the purpose and construct definition; (2) prompt attributes, including information about the task instructions and relevant background information

about the participants (e.g. interlocutor status and familiarity); and (3) the rating procedures. The task specifications evolved in three key stages: (1) initial design prior to phase one data collection; (2) first revision following the phase one pilot stage; and (3) second revision, following the phase two pilot stages. As a note to the reader, the finalized specifications are provided in this paper following brief descriptions of these three stages. These task specifications developed over time and their development was documented because, as Fulcher and Davidson (2008) contend, it is crucially important to document this process as it contributes to a “validity narrative” (p. 9).

5.4.2.1. Initial design prior to phase one data collection

Davidson and Lynch’s (2002) format for organizing specifications was used in this study. Their framework consists of five main components:

1. A general description, which includes the purpose, construct definition, and a brief summary of the test.
2. Prompt attributes, which includes information about the necessary requirements for constructing the test and information about how the test should interact with the test-taker.
3. Response attributes, which includes information about what is required from the test-taker in order to complete the task sufficiently.
4. A sample task, which provides an illustration of a task suitable for replication.
5. A specification supplement, which provides test developers with further resources to enrich their understanding about the information presented in the specifications.

Initially, the specifications included all of Davidson and Lynch’s five parts: general description, prompt attributes, response attributes, sample task, and specification supplements.

The *general description* section described the purpose of the task and provides a working construct definition, which would later be rewritten following the phase one findings. The *prompt attributes* included relevant background information in regards to assessing performances on speaking tasks such as goal orientation, topics, situations, participants, and interlocutor status and familiarity (adapted from Fulcher, 2003). The *response attributes* included predictions about the kinds of speech functions participants would likely be required to use in order to provide a response, such as describing, narrating, arguing, and so forth. Prior to the phase one data collection, a rating scale had not yet been created, but it was pre-determined that it would resemble the format of a scale used by Fox et al. (2016) in order to isolate specific attributes of the fluency construct for assessment for learning purposes. Finally, the *specification supplements* section provided sources used to construct the initial design of these specifications.

5.4.2.2. First revision following the phase one findings

The working construct definition in the general description section was revised to include the construct definition created from the phase one findings. A “Participants” description was added to the prompt attributes section to inform raters about relevant background information concerning the participants.

5.4.2.3. Second revision following the phase two pilot stages

Several changes were made at this stage. As Fulcher and Davidson (2008) claim, test specifications need to be well-detailed so that they provide substantial validity evidence; however, in this study, due to time limitations, they also needed to be concise and user-oriented so that instructors could digest the information readily. Therefore, the first major change made was that the specifications were restricted to a one-page document. It was then deemed necessary to remove the latter two sections from the task specifications - sample task and prompt attributes

– and to remove information regarding the types of speech functions from the response attributes section. This section was then renamed “Rating procedures” because it focused solely on the three key rating procedures obtained from the phase two pilot stage, described previously.

The rating procedures were also informed by Luoma’s (2004) suggestions for rater training: introducing raters to the test and the criteria; providing examples of test-takers performing at high, medium, and low levels; and discussing why test-takers performed at this perceived to perform at different levels according to the rating scale provided.

The final version used in the phase two data collection process is displayed on the following page:

Fluency Assessment: Task Specifications

(1) General Description

Purpose of the Task:

- This task is designed to enable raters to identify strengths and weaknesses in English as an additional language speakers' abilities to speak fluently on a paired conversational task

Construct Definition of Fluency:

- At the core of the fluency construct is the speed and flow of speech; yet these core, temporal features are believed to be inherently connected to peripheral, non-temporal features of speech. On this task, at the highest level, the speaker produces smooth, efficient, sophisticated, and clear speech (adapted from Lennon, 2000), while facilitating and supporting the flow of conversation between participants.

(2) Prompt Attributes

- Students were presented with a one-page test paper that lists the following: (a) the purpose; (b) the instructions; (c) the speech prompts; and (d) the time restrictions.
- Students were given 7 minutes to exchange personal information based on the following prompts: (1) Learning English as an additional language; (2) Living/studying/working in Canada; (3) Future academic/career goals.
- Students were given one-minute to prepare their discussions. They could choose any topic and they didn't have to discuss all topics. Students could initiate new topics if they discussed all three topics. Students negotiated who went first.
- Goal Orientation: Convergent. Test-takers work together to achieve a mutual communicative goal.
- Topics: Familiar. Test-takers should be able to draw from personal experience to discuss the topics.
- Situations: Academic. Exchanging information with a peer in an academic context.
- Participants: Graduate students who speak English as an additional language. At the time of data collection, participants were enrolled in regular programs at a Canadian University but many of the students, whose admission test scores (e.g. IELTS) were below the standard cut-off scores for admission, were also required to take several supplementary English language courses as a condition of their acceptance to the university. Participants were gauged by their performance on an external measure of oral proficiency (CAEL/OLT – Task one) to range in general oral proficiency from intermediate to advanced. Participants had been in Canada for four months.

(3) Rating Procedures

- Review the rubric descriptors and the rating scale.
- Watch and discuss the training videos to help differentiate between performance levels.
- Practice rating one of the videos to get accustomed to the scale and to rating two speakers simultaneously.
- Before watching each video, indicate your perceived familiarity with the speakers' first languages.
- Rate both performances while watching the video only once. While watching, make short comments in the boxes provided. You may wish to wait until the video ends to consult your comments and rate the performances.
- Do not circle between numbers. There are no half-points (e.g. 3.5).
- Make comments in the space provided about how the surrounding context might affect the speaker's fluency.

Figure 8: Task Specifications

5.4.3. Conversational Style Questionnaire (CSQ)

This questionnaire was designed to elicit responses from instructors in order to answer the fifth research question:

What relationships exist between instructors' conversational styles and their assessments of High Considerateness-style and High Involvement-style students, according to this scale?

This questionnaire was developed by the researcher and it was based on Tannen's (2005) dichotomy of conversational styles: High Involvement (HI) and High Considerateness (HC). Tannen's (p.40-41) list of features that differentiate between styles was used to inform the items on the questionnaire:

1. Topic

- a. Prefer personal topics
- b. Shift topics abruptly
- c. Introduce topics without hesitation
- d. Persist (if a new topic is not immediately picked up, reintroduce it, repeatedly if necessary)

2. Pacing

- a. Faster rate of speech
- b. Faster turn taking
- c. Avoiding inter-turn pauses (silence shows lack of rapport)
- d. Cooperative overlap
- e. Participatory listenership

3. Narrative strategies

- a. Tell more stories
- b. Tell stories in rounds
- c. Prefer internal evaluation (i.e. the point of a story is dramatized rather than lexicalized)

4. Expressive paralinguistics

- a. Expressive phonology
- b. Marked pitch and amplitude shifts
- c. Marked voice quality
- d. Strategic within-turn pauses

The criteria outlined in the first two categories listed above (Topics and Pacing) comprised eight out of nine of the items in the questionnaire. Some of these criteria were adapted to be mindful of Dornyei and Csizer's suggestion for using "simple and natural language" (p. 78) that respondents could readily understand. For instance, the sub-category "Participatory Listenership" under "Pacing" would likely not be understood by participants. Upon reviewing Tannen's description of this category however, one could find references to interruptions and collaborative completions, so two items ("Interrupt" and "Finish your partner's sentences") were created from this sub-category. Finally, it was judged that the criteria outlined in the fourth category (Expressive paralinguistics) may not be easily understood by respondents; thus the criteria in this category were reduced to one item on the questionnaire (speak loudly and with enthusiasm). The table below displays Tannen's criteria in the left-hand column and the finalized items on the questionnaire are displayed correspondingly in the right-hand column:

Table 11

Tannen's (2005) Criteria for High-Involvement Style and Nine Items on the CSQ

<u>Tannen's (2005, p. 40-41) criteria</u>	<u>Conversational Style Questionnaire (CSQ)</u>
Prefer personal topics	A. Share personal information about yourself
Shift topics abruptly	B. Change topics suddenly
Persist (if a new topic is not immediately picked up, reintroduce it, repeatedly if necessary)	I. Persist in getting across what you want to say
Faster rate of speech	D. speak more quickly than your partner
Cooperative overlap	E. Overlap (talk at the same time as your partner)
Participatory listenership	C. Interrupt
	H. Finish your partner's sentences
Avoiding inter-turn pauses (silence shows lack of rapport)	G. Avoid conversational silence
Expressive paralinguistics	F. Speak loudly and with enthusiasm

The finalized questionnaire consisted of nine items assessed across a six-point semantic differential scale, referring to words that convey polar opposites, between the adverbs “always” and “never” in response to a question prompt, as in figure 10 below:

For each question below, please circle a number between 1 and 6. During an informal conversation with an acquaintance (e.g. classmate, colleague), do you typically...					
A. share personal information about yourself					
6	5	4	3	2	1
(always)					(never)

Figure 10: Example Question from the Conversational Style Questionnaire

This particular prompt was chosen because it reflects the situation occurring within the video (i.e. two acquaintances discussing topics) and since conversational style may be mediated by the context surrounding the conversation (Tannen, 2005), the prompt should reflect the context pertaining to the particular test situation.

In addition to Dornyei and Csizer's (2012) suggestion for using "simple and natural language" (p. 78) for writing items and formatting the document informed the design of this questionnaire:

1. Writing items that work

- Aim for short and simple items
- Use simple and natural language
- Avoid ambiguous or loaded words and sentences
- Avoid negative constructions
- Avoid double-barreled questions

2. Formatting the Questionnaire

- Have an appropriate length
- Economize space
- Mix up the scales and items

In summary, Tannen's (2005/2005) conversational style dichotomy and Dornyei and Csizer (2012)'s design recommendations resulted in the production of the questionnaire as seen below:

Conversational Style Questionnaire						
For each question below, please circle a number between 1 and 6. During an informal conversation with an acquaintance (e.g. classmate, colleague), do you typically...						
A. share personal information about yourself	6	5	4	3	2	1
	(Always)					(never)
B. change topics suddenly	6	5	4	3	2	1
	(Always)					(never)
C. interrupt	6	5	4	3	2	1
	(Always)					(never)
D. speak more quickly than your partner	6	5	4	3	2	1
	(Always)					(never)
E. overlap (talk at the same time as your partner)	6	5	4	3	2	1
	(Always)					(never)
F. speak loudly and with enthusiasm	6	5	4	3	2	1
	(Always)					(never)
G. avoid conversational silence	6	5	4	3	2	1
	(Always)					(never)
H. finish your partner's sentences	6	5	4	3	2	1
	(Always)					(never)
I. persist in getting across what you want to say	6	5	4	3	2	1
	(Always)					(never)

Figure 11. Conversational Style Questionnaire

As with the fluency rating scale, this questionnaire was piloted with five instructors who participated in the first phase of the study. The researcher met with each participant individually. Participants were encouraged to identify any potential problematic issues such as the formatting and administration of the questionnaire (Dornyei & Csizer, 2012). Several items were revised and removed from the questionnaire due to feedback from the participants. For instance, the first item (Share personal information about yourself) initially read “Prefer personal topics” which was taken directly from Tannen’s (2005, p. 40-41) criteria for High-Involvement Style. However, several participants required further explanation regarding what “prefer personal topics” means; thus, the item needed to be modified accordingly.

In terms of administration, participants understood the directions readily and they were observed to complete the questionnaire in a relatively short amount of time (less than one minute).

Once all five instructors completed the questionnaire, responses were entered into Statistical Packages for the Social Sciences (SPSS) in order to conduct a Cronbach alpha analysis of internal consistency. This analysis produced a strong reliability coefficient ($\alpha = .876$), providing confidence that the questionnaire would be suitable for data collection.

5.5. Procedures

The researcher met with 35 instructors in small group settings. The meetings were held in private locations at a medium-sized Canadian university. Each of these meetings lasted approximately an hour and a half.

Each instructor was provided with two booklets of information. The first booklet included the task specifications and the fluency rubric descriptors. The second booklet contained fluency rating scales for each of the four of the conversations and the conversational style

questionnaire. The rationale for providing separate booklets was to enable instructors to refer to the rubric descriptors in the first booklet in order to inform their assessments on the rating scales in the second booklet.

5.5.1. Rater training

Issues of rater training needed to be considered carefully. As instructor-participants are naturally busy in their daily lives, it was decided that the data collection sessions needed to be time-set for around 90 minutes. Otherwise, extensive data collection sessions could have limited the number of qualified and experienced instructors willing to participate in the study as well as inducing fatigue over the course of the sessions, thus affecting the reliability of the assessments. The videos are relatively long (seven minutes each), totalling 28 minutes of viewing. Time was also needed for participants to finalize assessments using the rating scales. Additionally, time needed to be allotted to develop initial rapport with participants and introduce them to the scale. Finally, time needed to be allotted to allow participants to complete the conversational style questionnaire following data collection. Thus, time constraints played a key role in determining the amount and extent to which raters should be trained.

Relatedly, to what extent raters should be trained to use a scale constructed by the researcher is open to debate. Taylor and Wigglesworth (2009), in their introduction to a special issue of the *Language Testing* journal, note that this concern is uniquely addressed and justified by the authors of each of the journal articles. Due to time constraints, instructors received only a brief amount of training with the rating scale (approximately 30 minutes). Although relatively brief, 30 minutes was believed to be a sufficient amount of time for several reasons. First, as mentioned previously, a high level of inter-rater reliability (e.g. $\alpha = .876$) was achieved for all seven items on the scale after piloting the scale with five participants. Second, as Fulcher (2003)

contends, depending on the circumstances, it is possible that raters' training should be limited while a test is still in the infant stages of development, stating the following:

If raters are trained, 'socialised' or 'cloned' before the validity argument is constructed, the training itself becomes a facet of the test that cannot be separated from the construct.

This fusion contaminates any validity evidence that uses scores from these raters (p. 146).

A limited amount of training was deemed necessary to standardize rating procedures and to familiarize raters with the rubric descriptors; yet raters were not trained extensively due to the reasons listed above. Taylor and Wigglesworth (2009) also address the need for researchers to justify the extensiveness of rater training, with consideration to the unique demands of each study:

Absence of rater training – whether by design or by default – raises questions of reliability that need to be addressed. Deliberately avoiding rater training in order to achieve 'naïve rating' may promise fresh insights into the experience of assessing proficiency, but it may beg questions about how far a study's findings can generalize to a more formal assessment context. The challenge to the researcher is to determine what role rater selection and training should play in their study design and to justify this in appropriate ways (p. 11).

The rater training procedures were derived from Luoma's (2004) suggestions for rater training, which are described as follows:

1. Rater training sessions often begin with an introduction to the test and the criteria.
2. Different levels on the scale are then illustrated, usually through taped performances that have rated by experienced raters before the training.

3. After this, the participants practise rating by viewing more taped performances (p. 177).

Luoma's (2004) suggestions informed the rater training procedures used in this study in the following ways:

1. Instructors reviewed booklet one, containing the task specifications and the rubric descriptors.
2. Instructors were shown a six-minute compilation video. This video was comprised of three-minute segments of two paired conversations featuring four speakers from the high, middle, and low groups, as identified by the instructor participant-group in phase one. More specifically, the first video featured two high-level performing speakers and the second video featured one mid-level and one low-level performing speaker. After each video, the researcher discussed features of speech related to the rubric descriptors at each level.
3. Instructors practiced rating one of the videos in order to get accustomed to the challenge of watching an entire video without stopping while rating two speakers simultaneously. As mentioned, these rating procedures were informed by findings from the pilot stage. Instructors were asked to write observations in the space provided while watching the videos. Once each video was finished, instructors were asked to consult their notes and circle the numbers associated with each item. Instructors were reminded not to provide half-points.

5.5.2. Rating procedures

Once training was finished, instructors watched four seven-minute videos featuring eight learners conversing in pairs. To recall, in phase one, learners were videotaped performing this

task multiple times with multiple partners; however, the four videos used in this second phase featured paired conversations from only the first time this task was performed. Each group of raters watched these performances in the exact same order.

A specific procedure was followed for each of the videos being shown. Instructors watched each of these videos only once without stopping. During this time, as recommended, instructors made notes in relation to the scale descriptors. Some instructors were observed to provide scores for a few of the items while the videos were still playing. Once the video ended, instructors consulted their notes and finished providing scores for each of the learners. Then, instructors provided comments about how contextual factors might affect each speaker's fluency. Instructors did not share scores nor did they discuss their observations. Instructors were allowed as much time as they needed to complete their ratings. Once finished, instructors were asked to review their scales to ensure that all of the numbers associated with each item had been circled.

Once all four videos had been viewed and all of the rating forms were completed, instructors then filled out the conversational style questionnaire. Once finished, instructors submitted both booklets, containing the task specifications, fluency rubric descriptors, and fluency rating scales to the researcher.

5.6. Analysis

5.6.1. Data cleaning

Once all data was collected and subsequently inputted into SPSS software, the entire database was checked against the rating scales and questionnaires to ensure that the data was entered correctly. Memos were kept with regards to any missing values stemming from either instrument (i.e. the rating scale and the conversation style questionnaire). In the end, 11 values were found to be missing from the database.

Larson-Hall's (2013) insights into handling missing data were consulted in order to solve this problem. Larson-Hall suggests examining the data to see if there are any perceivable patterns to the missing data. In other words, one should examine whether or not the participants are repeatedly not responding to the same item; if there is a pattern, then the item should be removed. However, a pattern did not appear to exist, and the missing values appeared to occur at random. Therefore, as suggested by Larson-Hall (2013), it was appropriate to impute the missing data by calculating the means for each respective item and then entering the means into the missing values as necessary.

5.6.2. Inter-rater reliability and internal consistency

It was necessary to check for the reliability of the instruments (the fluency rating scale and the conversation style questionnaire), in terms of inter-rater reliability and internal consistency, in order to pursue any further analyses involving this data.

5.6.2.1. Rating scale

A Cronbach alpha analysis of inter-rater reliability was conducted for all 35 raters' responses to each of the seven items on the fluency rating scale. The following alpha coefficients were uncovered: *Smooth* ($\alpha = .971$); *Efficient* ($\alpha = .974$); *Sophisticated* ($\alpha = .977$); *Clear* ($\alpha = .975$); *Facilitative* ($\alpha = .972$); *Supportive* ($\alpha = .967$); and *Holistic* ($\alpha = .977$). In sum, the findings indicated strong levels of inter-rater reliability for all items.

Additionally, a Cronbach alpha analysis of the internal consistency of raters' mean assessments was conducted to investigate how well the seven items on the scale grouped together. This analysis produced a strong alpha coefficient ($\alpha = .920$) providing further confidence in the reliability, and therefore, the use of this scale.

5.6.2.2. Conversational style questionnaire

Likewise, a Cronbach alpha analysis of the internal consistency of the nine items on the conversation style questionnaire was conducted, producing the following strong alpha coefficient: $\alpha = .880$. These results from the Cronbach analyses of inter-rater reliability and internal consistency demonstrated that these instruments could be deemed reliable for collecting data relevant to this study.

5.6.3. Assessing data suitability

It was necessary to determine the suitability of the data in order to answer each research question accordingly. This procedure “involves making changes in the dataset prior to the analyses, in order to make it more appropriate for certain statistical procedures” (Dornyei & Csizer, 2012; p. 84). The upcoming phase two results section is divided into four parts based on its associated research question. In light of Dornyei and Csizer’s (2012) suggestions, specific analytical procedures used to prepare the data appropriately to answer each question will be discussed in each section as necessary. This information may include, but is not limited to, the inspection of the data for information related to linearity, distribution, and potential outliers.

5.6.4. Temporal analysis

5.6.4.1. Sample preparation

The third research question investigates how well temporal analysis of learner speech samples correlate with the fluency scale items. To prepare these samples, all 8 seven-minute paired conversations were transcribed with pauses in Microsoft Word. Repetitions, repairs, and clause boundaries were also marked in the transcripts.

The decision to extract and analyse one-minute individual speeches from the seven-minute dialogues derived from methodological choices first initiated by Tavakoli (2016) and then later replicated by Peltonen (2017a). In these studies, both researchers extracted one-minute

individual speeches from six-minute dialogues to analyse them for temporal variables present within individual speech whereas the full dialogues were analysed for features of conversational speech (e.g. backchannel rate). The following table is based on Tavakoli's (2016) list of fluency measures, categorized by speed, breakdown, repair, composite, and dialogue.

Table 12

List of Temporal Measures Investigated in the Study

<u>Measure</u>	<u>Sub-Measure</u>	<u>Calculation</u>
Speed	Articulation Rate (AR)	Number of syllables/min. (excluding pauses)
Breakdown	Silent Pause Rate (SPR)	Number of silent pauses/min.
	Average Length of Silent Pauses (APL)	Average length of silent pauses/number of pauses
	Within-Clause Pause Rate (WCpr)	Number of silent within-clause pauses/number of clauses
	Within-Clause Pause Length (WCpl)	Average Length of within-clause pauses/clauses
	Between-Clause Pause Rate (BCpr)	Number of between-clause pauses/# of clauses
Repair	Between-Clause Pause Length (BCpl)	Average length of between-clause pauses/number of clauses
	Filled Pause Rate (FPR)	Number of filled pauses/min.
	Repetition Rate (RepetR)	Number of repetitions/min.
Composite	Repair Rate (RepairR)	Number of repairs/min.
	Speech Rate (SR)	Number of syllables/min. (including pause time)
	Pruned Speech Rate (PSR)	Number of syllables – filled pauses, repetitions, and repairs/min.

	Mean Length of Runs (MLR)	Average number of syllables between silent pauses
Dialogue	Non-Lexical Backchannel rate (NLBR)	Number of non-lexical backchannels/speech duration
	# of One-Word Lexical backchannels (OWLBR)	Number of one-word lexical backchannels/speech duration
	# of Lexical Phrase backchannels (LPBR)	Number of lexical phrase backchannels/speech duration
	Total # of backchannels (TotBR)	Number of all backchannels/speech duration

As per Tavakoli (2016), each one-minute speech was selected because the participant spoke for at least 70% of the time in the sample and produced at least two turns within the sample. Each sample included backchannels and lexical interruptions that “were considered as overlapping speaking time which was considered to belong to both participants, and as such were included in the measurement of fluency for both speakers” (Tavakoli, 2016, p. 141). Following this approach enabled the researcher of this present study to analyse each speaker’s longest turns for temporal variables while also enabling analysis of interactive speech features over the course of the conversation.

In previous studies, speech segments were extracted directly from the opening section of the dialogue in order to more aptly standardize approaches towards analysing dialogic speech (Derwing et al., 2006). However, two issues may arise as a result of that approach. First, in the opening sections of speech, learners, depending on their proficiency levels, often rely on reproducing memorized stretches of speech planned during the planning time allotted for the task (Ellis, 2005). Raters are often trained to be aware of this possibility when rating speeches so they often disregard the opening sections of speeches as mere warm-up time. Moreover, as

conversation is unpredictable, one speaker may dominate the opening of the conversation, compromising the ability of the researcher to collect substantial information about the interlocutor. Since it is not entirely clear when during a seven-minute conversation speakers will be able to produce their most substantial runs of speech, abiding by Tavakoli's (2016) guidelines enables that the researcher is able to analyse the best possible data suitable for temporal analysis provided by each speaker.

One of the inherent difficulties in analysing dialogic speech, as discussed by Tavakoli (2016), is the attribution of pauses between speakers because, in a sense, the pause belongs to both speakers. To properly attribute the pause, a researcher must be able to adequately infer the cause for the pause. A speaker could wish to continue their turn but may pause within their turn for a variety of pragmatic reasons. For instance, as Schegloff (2007) indicates, perhaps the speaker chooses to pause to relinquish the floor for a varied set of reasons:

1. Providing an adequate response to a previously-posed question
2. Posing a question to the interlocutor
3. Inviting the interlocutor to provide a collaborate completion of one's own turn by using silent pauses either on their own or within a cluster of other perceived disfluencies such as filled pauses, repetitions, or lexical fillers.

Without understanding the cause of the pause, however, it is not possible to determine to whom it should be attributed.

As per Schegloff (2007), a qualitative examination could provide more insightful information. Through analysing transcripts and audio files for where turn-relevance places (TRPs) may occur, this analysis would enhance researchers' understanding about how to attribute pauses accordingly. A variety of perceptual clues such as syntactic completions and/or

final (turn-completion) or high-rising (information question posing) intonation contours may indicate that a TRP is occurring (Wennerstrom & Siegel, 2003). However, Schegloff (2007) highlights that these perceptual clues provide potentialities but not determinants of the possible occurrence of a TRP, stating that a “transition to a next speaker becomes *possibly relevant* there” (p. 4, *italicized* emphasis included in the original statement). Therefore, even a qualitative analysis that considers all the contextual influences of the pause cannot possibly determine to whom the pause should be attributed.

Tavakoli (2016) proposes two solutions to address this problem for fluency researchers needing to calculate mutually-created gaps within collective speech: (1) removing the between-pauses, referred to as ‘gaps’ in this study, as per Riggensbach (1991); (2) including these gaps but attributing them to both speakers but dividing the length of gaps in half. Tavakoli revealed that attributing gaps to both speakers causes the speakers to appear to be more fluent. Thus, Tavakoli prefers this latter approach yet is careful to recognize its imperfections.

Thus, in line with Tavakoli (2016) and Peltonen (2017a), gaps were calculated by dividing their lengths in half. Similarly, in order to reduce the potential for measurement error, it was deemed more sensible to analyse only one minute of individual speech within conversation for each speaker rather than to analyse the entire seven minutes. Notably, analysing even one minute of speech for a wide variety of temporal variables is an arduous task.

Thus, one-minute samples from each learner were extracted from each of the paired conversations. Each of these one-minute speech samples was first edited by using audio-production software, Ableton 9.2. Then, each speech sample was inserted into phonetic analysis software (PRAAT) for temporal analysis.

The following procedure from DeJong and Perfetti (2011) was adapted to analyse the speech samples for temporal variables. In PRAAT, the parameters were set to “text-grid silences” and the cut-off point for the pause was set for > 0.25 , as this cut-off point has been widely used in various studies (e.g. Goldman-Eisler, 1968; De Jong et al., 2013; Préfontaine & Kormos, 2015; Tavakolli et al., 2016).

PRAAT then created spectrograms for each sample, segmenting the speech into 'sounding' and 'not sounding'. The 'sounding' text segments were replaced with the transcribed text from learners' speeches. Filled pauses (e.g. “um”) were marked as 'fp'. The 'not sounding' segments were replaced by 'sp' to indicate silent pauses. These spectrograms and corresponding transcripts were then used as a basis to calculate the temporal variables manually.

5.6.4.2. Clause analysis

The frequency and length of within- and between-clause pauses is one of the aforementioned measures. Finding the location of pauses within learners' speech streams required identifying clauses within speech beforehand. In accordance with recent research exploring pause location (e.g. DeJong, 2016a; Skehan, Foster & Skum, 2016, Tavakoli, 2017), Foster, Tonkyn, and Wigglesworth's (2000) definition of the Analysis of Speech Unit (ASU) as “a single speaker's utterance consisting of an independent clause, or a sub-clausal unit, together with any subordinate clause(s) associated with either” (p. 365) and their definition of a clause as consisting “minimally of a finite or non-finite verb element with at least one other clause element (Subject, Object, Complement, or Adverbial)” (p. 366) was used in the present study.

Chapter Six - Results (Phase Two)

6.1. Overview

The following results section for phase two is divided into four main parts, corresponding to the four key research questions (RQ) posed in the second phase of the study.

RQ2: To what degree are the items on the scale relevant to assessing speech fluency?

RQ3: What relationships exist between temporal measures and rating assessments, according to this scale?

RQ4: In what ways, if any, do EAP instructors' levels of students' accent familiarity affect their fluency ratings?

RQ5: What relationships exist between EAP instructors' conversation styles and their assessments of High Considerateness-style and High Involvement-style students, according to this scale?

6.2. Fluency scale items and construct relevance

This first section depicts results relating to the construct relevance of the six analytic items (*Smooth*, *Efficient*, *Sophisticated*, *Clear*, *Facilitative*, and *Supportive*) in relation to the seventh holistic item (*Holistic*), which comprise the fluency rating scale. The six analytic items on the scale were arranged according to how well they theoretically represent the fluency construct. The *Smooth* and *Efficient* items were followed by items that represent features that are deemed to be more and more peripheral to the fluency construct – *Sophisticated*, *Clear*, *Facilitative*, and *Supportive*. The seventh item, *Holistic*, was added to allow the rater to provide a global assessment of fluency in order to analyse the relationships between holistic and analytic measures of fluency.

Notably, it may be possible that some of the items on the scale represent features that are far too peripheral to the fluency construct. Therefore, the following research question is posed to provide a check on the findings from the first phase:

RQ2: To what degree are the items on the scale relevant to assessing speech fluency?

To answer this question, a principal component analysis (PCA) of all seven items was conducted. A PCA groups the data into components by examining the inter-item correlations in a multitude of ways. As Pallant (2007) describes, “it (the PCA) does this by looking for ‘clumps’ or groups among the inter-correlations of a set of variables” (p. 179). For this study, a PCA enables a fuller understanding of how well the seven items on the scale group together to form the fluency construct and, more specifically, to explore any potential relationships between analytical and holistic scale items. In sum, the PCA provides evidence regarding the degree to which the items on the scale are construct-relevant.

As Pallant (2007) describes, conducting a PCA is a three-step procedure: (1) assessing data suitability; (2) factor extraction; and (3) factor rotation and interpretation.

6.2.1. Assessing data suitability

First, the data must be deemed to be suitable for the PCA approach in terms of sample size and the strength of the inter-item correlations between variables.

Pallant (2007) states “there is little agreement among authors regarding how large a sample size should be” (p. 180). PCAs are typically conducted with sample sizes with at least 150 cases with a subject-to-variable ratio of 10:1; yet it has also been proposed that a subject-to-variable ratio of 5:1 is suitable (Tabachnik & Fidell, 2007 as cited in Pallant, 2007). In this study, the data reflect a 5:1 subject-to-variable ratio as there are 35 raters corresponding to the seven items on the fluency rating scale, thus matching the 5:1 subject-to-variable ratio.

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy also assisted in determining the appropriateness of the sample size. The KMO measure produced a value above

.853, which exceeds the minimum requirement ($> .6$) for a suitable PCA (Tabachnick & Fidell, 2007 as cited in Pallant, 2007). In sum, the sample size was deemed large enough for a PCA.

Determining the suitability of the data also required analysing the strength of inter-item correlations. Three modes of analysis were used to examine these relationships. First, an examination of the correlation matrix (see Table 13 below) indicates that all variables correlated above the .3 threshold, indicating that the items correlated well enough together to load onto individual factors (Tabachnik & Fidell, 2007, as cited in Pallant, 2007). The lowest correlation coefficient was found between the *Efficiency* item and the *Supportive* item ($r(35) = .376$). Bartlett's Test of Sphericity was used to further examine this correlation matrix to uncover how well these variables correlated with one another. The results produced a significant value ($p < .000$) indicating that a PCA was appropriate for this data set.

Table 13

Correlation Matrix. Inter-Item Correlations between Items on the Fluency Scale

	<u>Smooth</u>	<u>Efficient</u>	<u>Sophisticated</u>	<u>Clear</u>	<u>Facilitative</u>	<u>Supportive</u>	<u>Holistic</u>
Smooth	1.000	.797	.762	.668	.522	.471	.905
Efficient	.797	1.000	.693	.632	.427	.376	.814
Sophisticated	.762	.693	1.000	.626	.578	.575	.795
Clear	.668	.632	.626	1.000	.610	.598	.679
Facilitative	.522	.427	.578	.610	1.000	.887	.610
Supportive	.471	.376	.575	.598	.887	1.000	.595
Holistic	.905	.814	.795	.679	.610	.595	1.000

The correlation matrix also provides a key finding regarding the degree to which these fluency scale items are construct-relevant. In Table 13 above, the right-most column displays the relationship between the *Holistic* item and the six analytic items. As mentioned in phase one, the

scale was designed so that the items on the scale were arranged according to how well they were theorized to represent the fluency construct. The correlation matrix (see Table 13 above) supports this arrangement as the correlations between the analytic items and the *Holistic* item weaken in descending order (*Smooth*, $r = .905$; *Efficient*, $r = .814$; *Sophisticated*, $r = .795$; *Clear*, $r = .679$; *Facilitative*, $r = .610$; and *Supportive*, $r = .595$). Similar results are found for the left-most items (*Smooth* and *Efficient*) as the correlations weaken as the items become more and more peripheral to the fluency construct.

Analytic Criteria	Yes	Some -what	No	Comments
<i>Is the speech...</i>				
1. ...smooth?	6 5	4 3	2 1	
2. ...efficient?	6 5	4 3	2 1	
3. ...sophisticated?	6 5	4 3	2 1	
4. ...clear?	6 5	4 3	2 1	
<i>Does the speaker...to keep the flow of conversation going?</i>				
5. ...facilitate topics and turns...	6 5	4 3	2 1	
6.support the conversation partner...	6 5	4 3	2 1	
<i>Overall, is the speech fluent?</i> [Yes] 6 – 5 – 4 – 3 – 2 – 1 [No]				

Figure 12. Fluency Rating Scale Items

6.2.2. Factor extraction

The second step, factor extraction, “involves determining the smallest number of factors that can be used to best represent the interrelations among the set of variables” (Pallant, 2007, p. 181). Two analytical techniques were used to identify how many components should remain: Kaiser’s criterion and scree plot.

The first technique, Kaiser’s criterion, examined the eigenvalue of each component. The eigenvalue “represents the total amount of variance explained by each factor” (Pallant, 2007, p

182). The Kaiser criterion states that only components with eigenvalues of above 1.0 should be retained. As shown in the table below, two components contain eigenvalues that exceed 1.0: Component 1 = 4.911, % of variance = 70.152; Component 2 = 1.031, % of variance = 14.733. These two components explain 84.885 % of the total variance.

The second technique, scree plot, provides a visual representation of the eigenvalues of the components (Pallant, 2007). As seen in the figure below, the interval between component 2 and 3 represents the elbow, and as Catell (1966, as cited in Pallant, 2007), contends, all factors above the elbow should be retained as they explain the most variance in the data whereas factors below the elbow explain the least amount of variance, and as such, should be excluded. In sum, the Kaiser's criterion and scree plot determined that the scale contains two components, which explain the majority of the variance in the rating assessments.

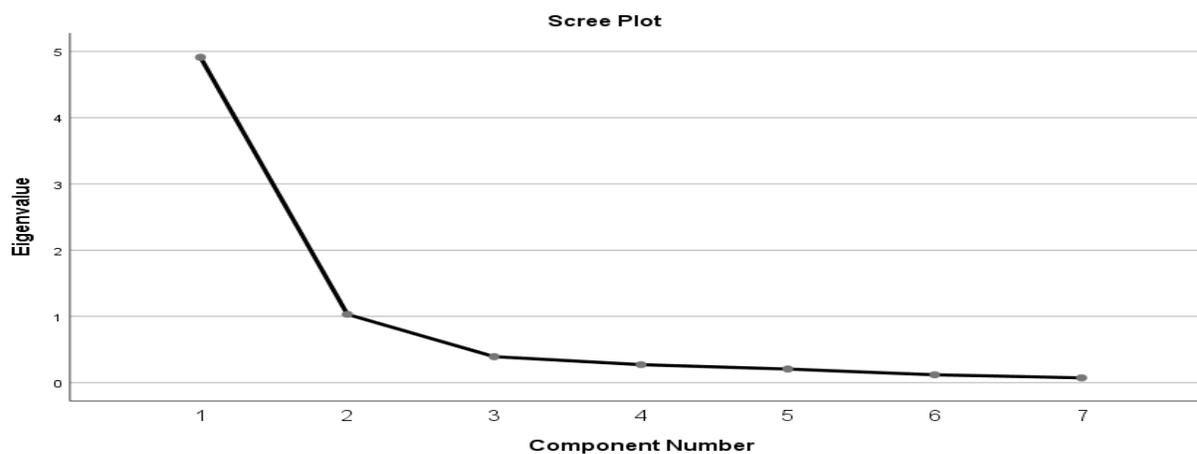


Figure 13. Scree Plot from PCA of Fluency Rating Scale Items

6.2.3. Factor rotation and interpretation

The third step was factor rotation. This step involved interpreting the two components produced by the preceding analyses through factor rotation. As Pallant (2007) describes, two possible approaches to factor rotation exist: orthogonal (uncorrelated) and oblique (correlated). The oblique rotation was chosen since the items on the scale were assumed to be correlated since

the phase one findings concluded that each item comprises the same overall construct – speech fluency. The Oblimin rotation provides several kinds of analyses to examine the degree of correlations within and between the two components: (1) Component correlation matrix; (2) Pattern matrix; and (3) Structure matrix.

The Component Correlation Matrix reveals the degree to which the two components relate to one another. The results show that the two components are moderately correlated ($r = .547$).

In Table 14 below, the pattern matrix displays how well each item loads onto each component whereas the structure matrix displays the correlations for each of the items within each component. In other words, both matrices provide an indication of how well each item belongs to each component.

The pattern matrix shows marked differences in terms of how well each item loads onto each component. Core fluency items - *Efficient* (1.006), *Smooth* (.960) and *Holistic* (.880) - load strongly onto the first component. However, peripheral fluency items load onto this component in various ways. *Sophisticated* (.756) loads somewhat strongly onto this component whereas *Clear* (.564) loads moderately onto it. However, the loadings for *Facilitative* (.051) and *Supportive* (-.009) are quite low. As Pallant (2007) advises, the variables with the highest loadings should be used to label the component. Therefore, as *Smooth* and *Efficient* are core fluency variables, and they produce the highest loadings, the component is labeled *Individual Fluency*.

Inverse results are shown for component two, however. *Facilitative* (.931) and *Supportive* (.974) load highly onto this component whereas other items load poorly: *Clear* (.364), *Sophisticated* (.187), *Holistic* (.121), *Smooth* (-.032), and *Efficient* (-.174). Since *Facilitative* and

Supportive are peripheral fluency variables, and since they are representative of a paired conversation, the component is thus labeled *Conversational Fluency*.

Table 14

PCA - Pattern and Structure Matrix with Oblimin Rotation of Fluency Scale Items

	<u>Pattern Matrix</u>		<u>Structure Matrix</u>	
	Component 1 (Individual Fluency)	Component 2 (Conversational Fluency)	Component 1 (Individual Fluency)	Component 2 (Conversational Fluency)
Smooth	.960	-.032	.942	.492
Efficient	1.006	-.174	.911	.376
Sophisticated	.756	.187	.858	.601
Clear	.564	.364	.763	.672
Facilitative	.051	.931	.560	.959
Supportive	-.009	.974	.524	.969
Holistic	.880	.121	.946	.602

The structure matrix shows similar results (see Table 14 above). In component one, for *Individual Fluency*, strong correlations are found for four items - *Holistic* (.946), *Smooth* (.942), *Efficient* (.911), and *Sophisticated* (.858) – and moderately strong correlations are found for *Clear* (.763). Moderate correlations are found for *Facilitative* (.560) and *Supportive* (.524). As for component two, *Conversational Fluency*, strong correlations are found for *Facilitative* (.959) and *Supportive* (.969). Moderately strong correlations are found for *Holistic* (.602), *Sophisticated* (.601), and *Clear* (.672) whereas the correlations for *Smooth* (.492) and *Efficient* (.376) are weak.

6.2.4. Summary

Overall, the PCA has revealed two key results, which are relevant to answering the question regarding the degree to which the items on the scale are construct relevant. First, in analysing the suitability of the data for analysis, the correlation matrix supports the descending-order arrangement of core to peripheral fluency items on the fluency rating scale as the

relationships between the holistic measure of fluency and its analytic measures weaken as the items become more peripheral to the construct. Moreover, the relationships between core fluency items (*Smooth* and *Efficient*) and the peripheral fluency items also weaken as the items become more peripheral to the construct. The second key result is that the PCA has produced two components, which, according to the component correlation matrix, are moderately correlated with one another. In other words, the scale appears to measure two distinct but moderately related constructs: *Individual Fluency* and *Conversational Fluency*. *Individual Fluency* is comprised of core fluency items (*Smooth*, *Efficient*, and *Holistic*) but it is also comprised to a lesser degree of more peripheral items (*Sophisticated* and *Clear*). On the other hand, the *Conversational Fluency* construct is comprised of fluency features involving interaction (*Facilitative* and *Supportive*). In sum, the fluency scale designed for a paired conversational task seems to measure two distinct, but associated, constructs: *Individual Fluency* and *Conversational Fluency*.

6.3. Temporal Measures and Fluency Ratings

The second section depicts the results regarding the relationships between utterance fluency, as measured by temporal measures of fluency (e.g. speech rate), and perceived fluency, as measured by raters' ($n = 35$) mean assessments of EAP learners ($n = 8$) according to the seven items on the fluency rating scale. In particular, results for this section are provided in response to the research question:

RQ3: What relationships exist between utterance fluency and perceived fluency, according to this scale?

6.3.1. Temporal measures

The following set of features used in this study is based off Tavakoli's (2016) list of fluency measures: speed, breakdown, repair, composite, and dialogue.

Table 15

List of Temporal Measures Investigated in the Study

<u>Measure</u>	<u>Sub-Measure</u>	<u>Calculation</u>
Speed	Articulation Rate (AR)	Number of syllables/min. (excluding pauses)
Breakdown	Silent Pause Rate (SPR)	Number of silent pauses/min.
	Average Length of Silent Pauses (APL)	Average length of silent pauses/number of pauses
	Within-Clause Pause Rate (WCpr)	Number of silent within-clause pauses/number of clauses
	Within-Clause Pause Length (WCpl)	Average Length of within-clause pauses/clauses
	Between-Clause Pause Rate (BCpr)	Number of between-clause pauses/# of clauses
Repair	Between-Clause Pause Length (BCpl)	Average length of between-clause pauses/number of clauses
	Filled Pause Rate (FPR)	Number of filled pauses/min.
	Repetition Rate (RepetR)	Number of repetitions/min.
Composite	Repair Rate (RepairR)	Number of repairs/min.
	Speech Rate (SR)	Number of syllables/min. (including pause time)
	Pruned Speech Rate (PSR)	Number of syllables – filled pauses, repetitions, and repairs/min.
	Mean Length of Runs (MLR)	Average number of syllables between silent pauses

Dialogue	Non-Lexical Backchannel rate (NLBR)	Number of non-lexical backchannels/speech duration
	# of One-Word Lexical backchannels (OWLBR)	Number of one-word lexical backchannels/speech duration
	# of Lexical Phrase backchannels (LPBR)	Number of lexical phrase backchannels/speech duration
	Total # of backchannels (TotBR)	Number of all backchannels/speech duration

6.3.2. Assessing data suitability

As suggested by Pallant (2007), in order to ensure that these temporal measures and raters' mean assessments were suitable for correlation analysis, certain procedures were used to check for the degree of linearity, the type of distribution, and the presence of outliers.

First, several scatterplots were created to assess whether or not there was a perceived degree of linearity between fluency rating assessments and temporal measures. Since many temporal measures were subjected to correlational analysis, only a "spot check" (Pallant, 2007, p. 185), of a small sample of scatterplots was examined. The figure below illustrates the scatterplot for raters' mean assessments of *Smooth* (x-axis) and Within-Clause Pause Rate (*WCpr*) (y-axis) indicating that an inverse correlation may exist between these variables.

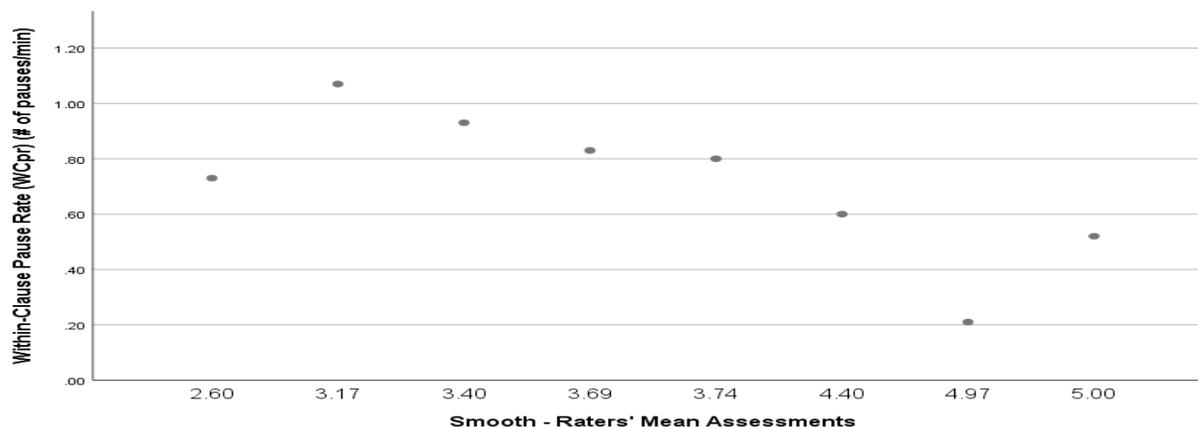


Figure 14. Raters' (n = 35) Mean Assessments of *Smooth* and Learners' (n = 8) *WCpr*

Shapiro-Wilk tests of normality revealed that none of the variables reached significance levels, indicating that they exhibit a normal distribution pattern. Since this spot-check of variables indicates a degree of linearity and a normal distribution pattern, Pearson product-moment correlation analyses were deemed to be appropriate techniques for analysing the relationships between fluency ratings and temporal measures of fluency (Pallant, 2007). Finally, a boxplot analysis did not reveal the presence of outliers for these variables.

Evans' (1996) classifications of correlational strength were chosen for use in this study. These classifications were chosen because they are a more conservative, yet more informative, set of classifications as opposed to Cohen's (1998) commonly-used set: $>.1$ = small; $>.3$ = medium; and $>.5$ = large. Evans' (1996) classifications are shown in the table below. *As an important note to the reader, considering the relatively small sample sizes analysed in the present study, it was deemed important to choose the most conservative set of correlation classifications available.*

Table 16

Evans' (1996) Classifications of Correlation Sizes

<u>Strength</u>	<u>R</u>	<u>r²</u>
Very weak	.00 - .19	(0% – 4%)
Weak	.20 - .39	(4% – 16%)
Moderate	.40 - .59	(16% – 36%)
Strong	.60 - .79	(36% to 64%)
Very strong	.80 – 1.00	(64% to 100%)

The table below illustrates the Pearson-product moment correlation coefficients between temporal measures and the items on the fluency rating scale.

Table 17

Pearson r Correlations between Temporal Measures of One-Minute Excerpts of Learners' Speeches (n = 8) and Raters' (n = 35) Assessments on the Fluency Rating Scale

		<u>Smooth</u>	<u>Efficient</u>	<u>Sophis- ticated</u>	<u>Clear</u>	<u>Facili- tative</u>	<u>Support- tive</u>	<u>Holistic</u>
Speed	AR	.663	.779*	.748*	.647	.648	.693	.674
Breakdown	SPR	-.164	-.178	-.162	-.129	.000	.054	-.190
	APL	-.254	-.311	-.369	-.297	-.153	-.243	-.242
	WCpr	-.772*	-.832*	-.770*	-.707*	-.696	-.742*	-.775*
	WCpl	-.544	-.553	-.541	-.566	-.542	-.506	-.559
	BCpr	-.215	-.178	-.122	-.079	-.014	-.084	-.172
	BCpl	-.255	-.299	-.344	-.371	-.190	-.265	-.230
Repair	FPR	-.709*	-.718*	-.747*	-.794	-.815*	-.786*	-.753*
	RepetR	-.249	-.091	-.189	-.271	-.259	-.234	-.209
	RepairR	-.084	.058	.069	.095	-.003	.003	.107
Composite	SR	.572	.669	.652	.559	.486	.549	.588
	PSR	.608	.750*	.689	.581	.599	.650	.625
	MLR	.392	.462	.442	.379	.270	.331	.418
Dialogue	NLBR	.388	.573	.477	.422	.527	.532	.451
	OWLBR	.118	.266	.182	.078	.105	.120	.150
	LPBR	.350	.358	.259	.203	.362	.391	.291
	TotBR	.352	.549	.433	.331	.427	.444	.400

Note: * = $p < .05$; ** = $p < .01$; AR = Articulation Rate; SPR = Silent Pause Rate; APL = Average Pause Length; WCpr = Within Clause Pause Rate; WCpl = Within Clause Pause Length; BCpr = Between Clause Pause Rate; BCpl = Between Clause Pause Length; FPR = Filled Pause Rate; RepetR = Repetition Rate; RepairR = Repair Rate; SR = Speech Rate; PSR = Pruned Speech Rate; MLR = Mean Length of Runs; NLBR = Non-Lexical Backchannel Rate; OWLBR = One-Word Lexical Backchannel Rate; LPBR = Lexical Phrase Backchannel Rate; TotBR = Total # of Backchannels Rate.

A summary of strong and significant results from the preceding table are as follows:

(1) *Smooth*

- WCpr ($r = -.772, p = .025$)
- FPR ($r = -.709, p = .049$)

(2) *Efficient*

- AR ($r = .779, p = .023$)
- WCpr ($r = -.832, p = -.010$)
- FPR ($r = -.718, p = .045$)
- PSR ($r = .750, p = .032$)

(3) *Sophisticated*

- AR ($r = .748, p = .033$)
- WCpr ($r = -.770, p = .026$)
- FPR ($r = -.747, p = .033$)

- (4) *Clear*
- WCpr ($r = -.707, p = .05$)

- (5) *Facilitative*
- FPR ($r = -.815, p = .014$)

- (6) *Supportive*
- FPR ($r = -.786, p = .021$)

- (7) *Holistic*
- WCpr ($r = -.775, p = .024$)
 - FPR ($r = -.753, p = .031$)

These results were then further subjected to analyses of statistical power in order to increase confidence in these findings. Post-hoc power analyses were conducted through G*Power software (Faul, Erdfelder, Buchner, & Lang, 2009) revealing that several of the aforementioned strong and significant correlations produced suitable statistical power, which according to (Evans, 2008), should be greater than .80. The correlations listed below have statistical power greater than .80. *As another note to the reader, considering the relatively small sample sizes analysed in this study, only the correlations listed below, which were found to be not only significant, but also produced substantial statistical power, are deemed worthy of discussing in the present study.*

- (1) *Smooth:*
- WCpr (power = .81);

- (2) *Efficient:*
- AR (power = .82);
 - WCpr (power = .90);

- (3) *Sophisticated:*
- WCpr (power = .80),

- (4) *Clear:*
- WCpr (power = .84);

- (5) *Facilitative:*
- FPR (power = .88),

(6) *Supportive*:

- FPR (power = .83),

(7) *Holistic*:

- WCpr (power = .81).

Overall, the results suggest that, among all temporal measures investigated, the strongest relationships exist between Within-Clause Pause Rate (*WCpr*), as a measure of Breakdown Fluency (Tavakoli 2016), and items on the scale that comprise individual fluency: *Smooth*, *Efficient*, *Sophisticated*, *Clear*, and *Holistic*. In addition, articulation rate (AR), as a measure of speed (Tavakoli, 2016), is shown to be strongly associated with the *Efficient* item. On the other hand, Filled Pause Rate (*FPR*) provides the strongest correlation coefficients for items related to interactional fluency: *Facilitative* and *Supportive*.

6.4. Accent familiarity and fluency ratings

6.4.1. Overview

The following two sections depict results concerning how listener characteristics, in terms of accent familiarity and conversation style, may affect instructors' fluency assessments in this study. The results from this section provide insight into answering the research question:

RQ4: In what ways, if any, do instructors' levels of accent familiarity affect their assessments on the fluency rating scale?

6.4.2. Assessing data suitability

As mentioned in the Method Phase Two section of this paper, before viewing the video-recorded paired conversations of 8 learners, 35 instructors were each asked to report their amount of exposure to each learner's accents on a six-point scale (1 = Not at all familiar, 6 = Very familiar), referred to in this study as accent familiarity (*AccFam*). Mean scores for *AccFam* were calculated from all instructors' assessments of all learners. A boxplot analysis of *AccFam*

was then conducted to search for the presence of outliers, revealing three outliers, which were then removed from any further analyses involving the *AccFam* variable; thus, the following analyses examine 32, not 35, instructors' assessments of 8 learners' fluency levels.

To answer the research question, the data were first subjected to Spearman Rho correlation analyses, which revealed little to no relationship between *AccFam* and the seven items on the fluency scale. It was therefore necessary to use a different statistical approach; thus, Kruskal-Wallis H and Mann-Whitney tests were conducted to analyse non-parametric group differences in terms of instructors' reported accent familiarity.

6.4.3. Kruskal-Wallis H tests

Instructors' responses to the six-point accent familiarity scale were divided into three groups: Gp1 = Not Familiar (1, 2, 3) ($n = 9$); Gp2 = Somewhat Familiar (4) ($n = 14$); and Gp 3 = Familiar (5, 6) ($n = 9$). The table below depicts the mean, standard deviation, and the median for each of the accent familiarity groups categorized by items on the fluency rating scale.

Table 18

Sample Descriptives Using Kruskal-Wallis H test (Mean Ratings vs Accent familiarity)

		<u>N</u>	<u>M</u>	<u>SD</u>	<u>Md</u>
Smooth	Not Familiar	9	3.76	.51	3.75
	Somewhat Familiar	14	3.99	.32	4.00
	Familiar	9	3.78	.29	3.75
	Total	32	3.87	.38	3.88
Efficient	Not Familiar	9	3.52	.36	3.50
	Somewhat Familiar	14	3.94	.43	3.88
	Familiar	9	3.64	.46	3.75
	Total	32	3.74	.45	3.75
Sophisticated	Not Familiar	9	3.65	.51	3.88
	Somewhat Familiar	14	4.04	.52	3.94

	Familiar	9	3.76	.34	3.75
	Total	32	3.86	.49	3.88
Clear	Not Familiar	9	3.91	.36	3.88
	Somewhat Familiar	14	4.32	.38	4.25
	Familiar	9	4.04	.31	4.13
	Total	32	4.12	.39	4.13
Facilitative	Not Familiar	9	4.24	.46	4.38
	Somewhat Familiar	14	4.37	.66	4.32
	Familiar	9	4.32	.63	4.32
	Total	32	4.32	.58	4.31
Supportive	Not Familiar	9	4.42	.58	4.63
	Somewhat Familiar	14	4.41	.66	4.31
	Familiar	9	4.44	.63	4.50
	Total	32	4.42	.61	4.38
Holistic	Not Familiar	9	3.83	.46	3.75
	Somewhat Familiar	14	4.13	.36	4.06
	Familiar	9	4.00	.47	4.13
	Total	32	4.01	.43	4.00

As seen in the table above, a Kruskal-Wallis H test revealed observable differences between groups. The mean and median were the highest for the Somewhat Familiar group (Grp2) for all categories, with the exception of *Supportive*, whereas the mean and median for the Familiar group (Grp3) were the second highest. The mean and median were lowest for the Not Familiar (Grp1) group for all categories, also with exception to *Supportive*. For *Supportive*, the mean and median were approximately equal across all groups.

The Kruskal-Wallis H test also produced mean ranks for each of the groups corresponding to the items on the fluency rating scale (see Figure 15 below). The mean ranks were the highest for Grp2 (Somewhat Familiar) for all categories with exception to *Supportive*.

The mean ranks were the second highest for Grp3 (Familiar) with exception to *Smooth*. The mean ranks for Grp1 (Not Familiar) were the lowest for all categories, except for *Smooth*, where the mean rank is the second highest, and also except for *Supportive*, where the mean rank is the highest.

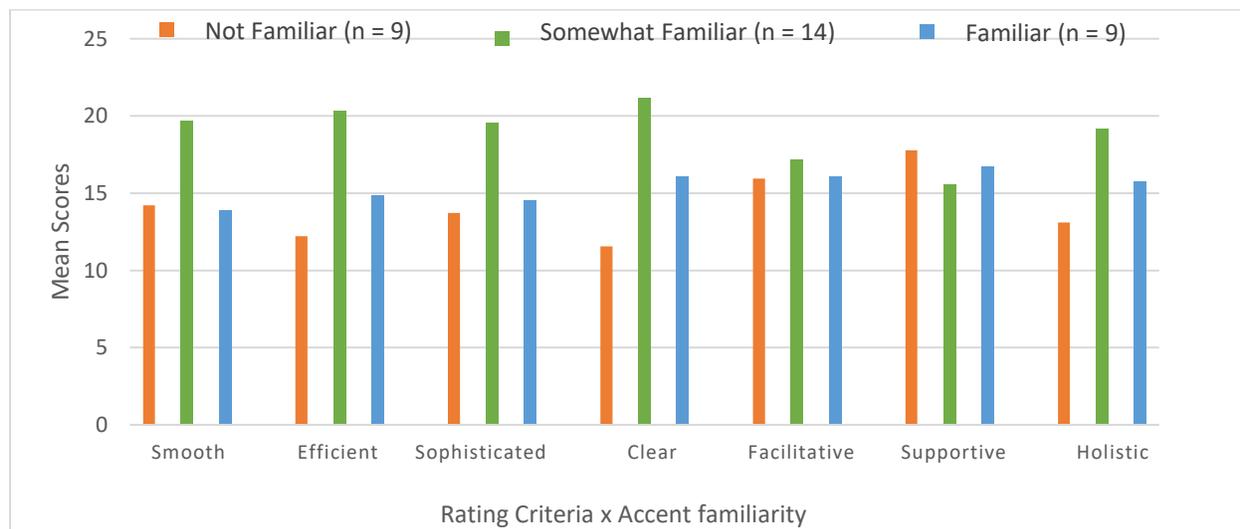


Figure 15. Mean Ranks of Fluency Ratings vs. Accent familiarity x Group

As seen in Table 19 below, the Kruskal-Wallis H test revealed a statistically significant difference ($X^2(2) = 6.633, p = .036$) in instructors' mean assessments of learners on the *Clear* item according to group differences in instructors' reported accent familiarity: (a) Not Familiar; (Gp1, $n = 9$); (b) Somewhat Familiar (Gp 2, $n = 14$); and (c) Familiar (Gp 3, $n = 9$). The Somewhat Familiar group recorded the highest median ($Mdn = 4.25$), followed by the Familiar group ($Mdn = 4.13$), whereas the Not Familiar group recorded the lowest median ($Mdn = 3.88$).

Table 19

Kruskal-Wallis H: Mean Ratings x Accent Familiarity

	<u>Smooth</u>	<u>Efficient</u>	<u>Sophisticated</u>	<u>Clear</u>	<u>Facilitative</u>	<u>Supportive</u>	<u>Holistic</u>
Kruskal-Wallis H	2.838	4.541	2.730	6.633	.118	.323	2.364

Df	2	2	2	2	2	2	2
Sig	.242	.103	.255	.036*	.943	.851	.307

Note: * = $p < .05$

Post-hoc tests using the Bonferroni approach were then conducted to investigate where pairwise differences may be occurring between the three groups for the *Clear* item. Significant differences were found ($X^2(2) = -9.623, p = .047$) between the Somewhat Familiar group (Gp2) and the Not Familiar group (Gp3). Significant results were not found between the Not Familiar group (Gp1) and the Familiar group (Gp3), nor between the Somewhat Familiar group (Gp2) and the Familiar group (Gp3).

6.4.4. Mann-Whitney U tests

Having indicated apparent group differences for *Clear* across the three groups, a series of Mann-Whitney U tests with Bonferroni corrections were conducted to further investigate where additional group differences may exist. First, a Mann-Whitney U test with the Bonferroni correction was conducted to investigate group differences between the Not Familiar group (Gp1, $n = 9$) and the Somewhat Familiar group (Gp2, $n = 14$). Statistically significant results were discovered for *Efficient* ($U = 31.0, z = -1.241, p = .046, r = 0.41$) with a medium effect size (Gp1, $Mdn = 3.5$; Gp2, $Mdn = 3.88$) and *Clear* ($U = 28.0, z = -2.216, p = .028, r = 0.59$) with a large effect size (Gp 1, $Mdn = 3.88$; Gp2, $Mdn = 4.25$). Table 20 below depicts these results.

Table 20

Mann-Whitney U: Fluency Ratings vs Accent Familiarity x Group (Gp 1 vs Gp 2)

	<u>Smooth</u>	<u>Efficient</u>	<u>Sophisticated</u>	<u>Clear</u>	<u>Facilitative</u>	<u>Supportive</u>	<u>Holistic</u>
Mann-Whitney U	43.5	31.0	42.0	28.0	57.0	54.5	38.0
Wilcoxon W	88.5	76.0	87.0	73.0	102.0	159.5	83.0
Z	-1.241	-2.027	-1.332	-2.216	-.380	-.537	-1.587

Sig.	.224	.046*	.201	.028*	.734	.600	.124
------	------	-------	------	-------	------	------	------

Note: $p < .05$. Sig. value with Boneforri correction; Gp 1 = Not Familiar; Gp 2 = Somewhat Familiar.

Additional Mann-Whitney U tests were conducted for the same purpose to investigate group differences between the Not Familiar group (Gp1, $n = 9$) and the Familiar group (Gp3, $n = 9$) and between the Somewhat Familiar group (Gp2, $n = 14$) and the Familiar group (Gp3, $n = 9$). However, no significant results were obtained.

6.4.5. Summary

In sum, Spearman Rho correlational analyses exploring potential relationships between *AccFam* and the seven items (*Smooth, Efficient, Sophisticated, Clear, Facilitative, Supportive, and Holistic*) on the fluency rating scale did not produce any meaningful results. However, by examining group differences using Kruskal-Wallis H and Mann-Whitney U tests, the results demonstrate, with exception to the *Supportive* item on the fluency rating scale, observable differences between groups for all items on the fluency rating scale. Except for *Supportive*, which was the highest for Gp1 (Not Familiar), the mean ratings were highest on all categories for Grp2 (Somewhat Familiar). These differences were shown to be statistically significant between Gp1 and Gp2 for one item (*Clear*), according to Kruskal-Wallis H post-hoc tests, and between two items (*Efficient* and *Clear*), according to Mann-Whitney U tests, indicating medium and large effect sizes respectively.

6.5. Conversation style and fluency ratings

6.5.1. Overview

The previous section illustrated how instructors' level of accent familiarity may affect their assessments of EAP speakers on the fluency rating scale. Similarly, this section explores the

influence of another potentially influential listener characteristic - the relationship between instructors' conversation styles and their fluency assessments. In particular, the results relate to the research question:

RQ5: What relationships exist between instructors' conversation styles and their assessments of High-Considerateness-style and High-Involvement-style students, according to this scale?

Seven out of the eight students completed the questionnaire. Scores for each of the students were tallied for all questionnaire items, producing one final score. The median score (35), from a range of 22 to 46, was used to divide students into two categories: High Considerateness (*HC-style*), (≤ 35), ($n = 4$) and High Involvement (*HI-style*), (> 35), ($n = 3$). Raters' mean assessments were calculated for each group (*HI-Style*, *HC-Style*) for each of the items, resulting in the creation of the following variables: *HC_Smo/HI_Smo* (Smooth), *HC_Eff/HI_Eff* (Efficient), *HC_Sop/HI_Sop* (Sophisticated), *HC_Sop/HI_Cle* (Clear), *HC_Fac/HI_Fac* (Facilitative), and *HC_Sup/HI_Sup* (Supportive).

Following instructors' assessments of 8 students on the paired-conversation tasks, 35 instructors completed the conversation style questionnaire, resulting in the creation of ten variables on the questionnaire: a) each of the nine questionnaire items (*ConvA* to *ConvI*); and b) an additional item, *ConvTot9*, which is the sum of all nine questionnaire items combined. The questionnaire items are as follows: *ConvA* (share personal information); *ConvB* (change topics suddenly and without hesitation); *ConvC* (interrupt); *ConvD* (speak more quickly than your partner); *ConvE* (overlap); *ConvF* (speak loudly and with enthusiasm); *ConvG* (avoid conversational silence); *ConvH* (finish your partner's sentences); and *ConvI* (persist in getting across what you want to say).

6.5.2. Assessing data suitability

As suggested by Pallant (2007), the data was assessed to determine its suitability for correlation analysis; thus, the ten variables (the nine individual questionnaire items and the one summated item) were examined for their perceived degree of linearity, type of distribution, and the potential for outliers. A “spot-check” (Pallant, 2007, p. 185) of several scatterplots were created and analysed, revealing the potential linearity of these items. Shapiro-Wilk tests of normality revealed that all items reached levels of significance, signaling that each item follows a non-normal distribution pattern; thus, non-parametric Spearman Rho correlation analyses were deemed to be more appropriate than parametric statistical techniques (e.g. Pearson r) for analyses of these variables (Pallant, 2007). A boxplot analysis revealed the presence of four outliers across multiple questionnaire items. These outliers were removed; thus, the analyses for this section involve 31 raters, not 35.

6.5.3. Correlation analysis

Students’ mean scores, as categorized by group (e.g. *HC_Smo*, *HI_Smo*) were then correlated with each of the items on the Conversation Style questionnaire, including the summated score from all items, *ConvTot9*. Once again, Evans’ (1996) classifications of correlational strength were used to qualify correlation size (see Table 21 below).

Table 21

Evans’ (1996) Classifications of Correlation Sizes

<u>Strength</u>	<u>R</u>	<u>r²</u>
Very weak	.00 - .19	(0% – 4%)
Weak	.20 - .39	(4% – 16%)
Moderate	.40 - .59	(16% – 36%)
Strong	.60 - .79	(36% to 64%)
Very strong	.80 – 1.00	(64% to 100%)

Spearman Rho analyses were conducted between the questionnaire items and raters' mean HC-Style assessments on the fluency rating scale, producing only slightly meaningful results. Weak correlation coefficients, which did not reach levels of significance, were found between *ConvA* (share personal information) and *HC_Sop* ($r = .342$).

More meaningful results were discovered regarding the HI-Style group of students. The table below illustrates the correlation coefficients between the Conversation Style questionnaire items and raters' ($n = 31$) mean assessments of HI-Style students ($n = 3$) on the fluency rating scale. Spearman Rho correlation analyses reveal moderate correlation coefficients between several rating scale items and several questionnaire items: *ConvA* (share personal information); *ConvB* (change topics suddenly); and *ConvG* (avoid conversational silence).

Table 22

Raters' Fluency Ratings of HI-Style Students

	<u>HI Smo</u>	<u>HI Eff</u>	<u>HI Sop</u>	<u>HI Cle</u>	<u>HI Fac</u>	<u>HI Sup</u>	<u>HI Hol</u>
ConvA	.203	.362*	.440*	.065	.184	.252	.440*
ConvB	.459**	.366*	.248	.125	.283	.230	.439*
ConvC	.076	.305	.131	.024	-.068	.015	.063
ConvD	.122	.297	.007	-.056	.265	.340	.144
ConvE	.265	.324	.322	.061	.120	.248	.321
ConvF	.037	.185	.189	.202	-.001	.128	.222
ConvG	.360*	.010	.252	.224	.298	.258	.337
ConvH	.157	.062	-.062	-.076	-.179	.006	-.017
ConvI	.031	.252	-.140	-.242	-.152	-.123	.023
ConvTot9	.275	.322	.154	.052	.099	.198	.203
ConvTot6	.386*	.336	.334	.176	.279	.292	.488**

Note: * = $p < .05$; ** = $p < .01$; ConvA = share personal information; ConvB = change topics suddenly; ConvC = interrupt; ConvD = speak more quickly than your partner; ConvE = overlap; ConvF = speak loudly and with enthusiasm; ConvG = avoid conversational silence; ConvH = finish your partner's sentences; ConvI = persist in getting across what you want to say; ConvTot9 = sum of all questions combined; ConvTot6 = sum of ConvA, ConvB, ConvC, ConvD, ConvE, and ConvG.

6.5.3.1. Sharing personal information (ConvA)

Table 22 above indicates that moderate ($> .4$) (Evans, 1996) but significant correlations were found regarding raters' reported frequency of sharing personal information during their own conversations and their assessments of HI-style students on the fluency rating scale. First, significant effects were found for *ConvA* and *HI_Eff* ($r < .362, p = .048$), indicating that this aspect of raters' conversational style (sharing personal information) explains 13% of the variance in raters' assessments of HI students on the *Efficient* item. Second, significant effects were also found for *ConvA* and *HI_Smo* ($r = .448, p = .013$) explaining 20% of the shared variance in ratings of *Sophisticated*. Finally, significance levels were reached for *ConvA* and *HI_Hol* ($r = .440, p = .013$), explaining 19% of the shared variance in the *Holistic* rating. Post-hoc power analyses conducted through G*power (Faul et al., 2009) revealed suitable statistical power ($> .80$) for *ConvA* and *HI_Sop* (power = .83) and *ConvA* and *HI_Hol* (power = .82), indicating that these latter two findings provide the greatest amount of confidence the reader can have concerning the relationship between this HI-style characteristic and fluency assessments on a conversation task.

6.5.3.2. Changing topics suddenly (ConvB).

Significant results were found for *ConvB* and *HI_Smo* ($r = .459, p = .009$) indicating 21% of the variance explained regarding the *Smooth* item; *ConvB* and *HI_Eff* ($r = .366, p = .046$), explaining 13% of the variance explained on the *Efficient* item and *ConvB* and *HI_Hol* ($r = .439, p = .014$), indicating 19% of the variance explained on the *Holistic* item. Post-hoc power analyses reveal suitable statistical power for *ConvB* and *HI_Smo* (power = .85) and *ConvB* and *HI_Hol* (power = .82).

6.5.3.3. Avoid conversational silence (ConvG)

Significant results were found for *ConvG* and *HI_Smo* ($r = .360, p = .046$), revealing 13% of the variance explained on the *Smooth* item. Post-hoc power analyses did not produce results with suitable statistical power for this correlation.

The results also demonstrated weak correlations ($> .03$) (Evans, 1996) between certain questionnaire items and certain fluency scale items, which did not reach levels of significance: *ConvC* (interrupt) and *HI_Eff* ($r = .305$); *ConvD* (speak more quickly than your partner) and *Supportive* ($r = .340$); *ConvE* (overlap) and *HI_Eff* ($r = .324$); *ConvE* and *HI_Sop* ($r = .322$); *ConvE* and *HI_Hol* ($r = .321$); *ConvG* and *HI_Hol* ($r = .337$); and *ConvTot9* and *HI_Eff* ($r = .322$). Considering that this research area is exploratory in nature, the moderate-to-weak correlations found in this study provide information regarding the potential and suitability of this questionnaire in measuring the relationship between raters' reported conversation style and their fluency ratings.

6.5.4. The efficacy of the questionnaire items

The results indicate that a six-item questionnaire may be more useful than a nine-item questionnaire in analysing participants' conversational styles. On the one hand, six out of the nine items produced inter-item correlations greater than the $>.3$ threshold for questionnaire analysis (Pallant, 2007) warranting further use of these items in the investigation of the relationship between conversation style and fluency ratings in future studies. On the other hand however, correlation analyses regarding *ConvF* (speak loudly and with enthusiasm), *ConvH* (finish your partner's sentences), and *ConvI* (persist in getting across what you want to say) produced correlation coefficient sizes less than 0.3 across all fluency ratings. Removing these three items resulted in a six-item questionnaire; totalling the results for each of the six items produced a new item, *ConvTot6*.

ConvTot6 was then correlated with the fluency ratings for HI-Style students, producing meaningful results. A weak but significant correlations between *ConvTot6* and *HI_Eff* ($r = .396$, $p = .027$), and a moderate but significant correlation between *ConvTot6* and *HI_Hol* ($r = .403$, $p = .02$). However, post-hoc power analyses did not find sufficient power for these correlations.

As items *ConvF*, *ConvH*, and *ConvI* did not produce any meaningful associations, it may be wise to remove these items from future use of this questionnaire regarding fluency assessments. A Cronbach alpha analysis of this newly created six-item questionnaire produces a marginally similar coefficient ($\alpha = .863$) as the original nine-item questionnaire ($\alpha = .880$). The results indicate that a six-item questionnaire may provide more meaningful results than a nine-item questionnaire in analysing conversation style in future studies.

6.5.5. Summary

In sum, these results suggest that weak-to-moderate but significant relationships may exist between instructors' conversation styles and their assessment of HI-style students on several items of the fluency rating scale (*Smooth*, *Efficient*, *Sophisticated*, *Supportive*, and *Holistic*). Instructors who reported exhibiting certain characteristics associated with a HI style (e.g. sharing personal information, changing topics suddenly, and avoiding conversational silence) were weakly to moderately more favorable towards HI-style students on several items. On the contrary, instructors who reported having more of an HC-style were weakly-to-moderately less favorable towards students with a HI-style. The results show only weak and not significant relationships regarding instructors' assessments of HC-style students. Finally, the results also indicate that a six-item questionnaire may be more useful for investigating participants' conversational styles than the nine-item questionnaire that was originally constructed.

6.6. Overall summary of key results

The first two sections displayed results relating to the following: (1) the construct relevance of the scale, consisting of core fluency items (*Smooth* and *Efficient*) and peripheral fluency items (*Sophisticated*, *Clear*, *Faciliative*, and *Supportive*), in its measure of speech fluency on a paired conversational task; and (2) the potential relationships existing between the seven items on the fluency rating scale and temporal measures of fluency. The results suggest that the scale identifies and measures two separate constructs - individual fluency and conversational fluency- which are moderately associated but distinct from one another. Moreover, by investigating the relationships between temporal measures and items on the fluency rating scale, the results reveal that strong correlations exist between fluency ratings and a number of temporal measures for all items, but most strongly for within-clause pause rate, supporting previous theoretical discussions and empirical evidence (e.g. DeJong, 2016) that underpins the importance of pause location in differentiating between levels of fluency.

The latter two sections provided an examination of how two kinds of listener characteristics – accent familiarity and conversational style – may affect assessments according to this scale. The results indicate that instructors’ levels of accent familiarity may have a moderate impact on their assessments of EAP students on two of the items on the scale: *Efficient* and *Clear*. The results from the final section reveals weak-to-moderate and significant relationships between certain aspects of instructors’ conversational styles (e.g. sharing personal information, changing topics suddenly, and avoiding conversational silence) and their assessments of HI-style students, according to this scale.

Chapter Seven -- Discussion (Phase Two)

The purpose of this second phase of the study is twofold: (1) to continue examining the role of core and peripheral fluency features on influencing fluency perceptions through analysing the construct relevance of the rating scale items; and (2) to investigate the potential effects of listener characteristics such as accent familiarity and conversational style on fluency judgements. This discussion is arranged according to the four research questions comprising this phase of the study.

RQ2: To what degree are the items on the scale relevant to assessing speech fluency?

Principal Component Analysis (PCA), or similar forms of Factor Analysis (FA), is a commonly-used statistical approach to analyse scores from instruments such as questionnaires (Dornyei & Csizer, 2003) or rating scales (Fulcher, 2003) by examining which items, if any, cluster together to form larger components/factors. Researchers may then interpret these individual components/factors as individual constructs or sub-constructs.

The following example illustrates how Hinofotis (1983), as cited in Fulcher (2003), used FA in a manner similar to the present study. Hinofotis used FA to analyse twelve analytic criterion on a rating scale designed to analyse international teaching assistants' oral performances elicited from videotaped replays. The twelve criterion used in Hinofotis' study were as follows: vocabulary, grammar, pronunciation, flow of speech, eye contact, non-verbal aspects, confidence in manner, presence, development of explanation, use of supporting evidence, clarity of expression, and ability to relate to students. FA revealed that these twelve items grouped together to create five factors, which the author later labelled to reflect the conceptual connections between items within each factor. For example, the first factor contained five criteria: development of explanation, use of supporting evidence, clarity of explanation, and ability to relate to students. Hinofotis later labelled this factor 'Communication of Information'.

In the present study, a PCA produced two separate components, moderately correlated with one another. The first, ‘individual fluency’, is composed of the *Smooth, Efficient, Sophisticated, Clear, and Holistic* items, and the second is labelled ‘conversational fluency’, composed of the *Facilitative and Supportive* items. Overall, the results seem to indicate that, within a conversation task, instructors seem to perceive individual fluency and conversational fluency as separate constructs, or perhaps, as separate sub-constructs of the larger fluency construct, depending on one’s interpretation. It would seem however, that according to this scale, the items representing conversational features of speech, *Facilitative and Supportive*, may be so peripheral to the fluency core, that these features may comprise a separate construct altogether. These findings provide support for Sato’s (2014) contention that individual fluency and conversational fluency are separate constructs due to significant differences in speech rates across individual and conversational tasks.

Additionally, the findings from the present study indicate that not only are instructors able to reliably assess two speakers simultaneously, but that they can aptly distinguish between features related to individual fluency and features related to conversational fluency. These findings have implications for one of the on-going challenges in assessing interactional competence – the extent to which test scores reflect characteristics of the interaction itself (Galaczi & Taylor, 2018).

How to assess an individual’s ability within a co-constructed performance has long been a topic of discussion within research on interactional competence (Galaczi & Taylor, 2018). One on-going debate concerns the extent to which the characteristics of the interaction (e.g. the interlocutor’s contributions; the task characteristics, and the rater’s characteristics) should be accounted for in the definition and operationalization of constructs and in the assignment of test

scores; in other words, should testers assign individual scores or joint scores? On the one hand, assigning only individual scores may not fully account for the joint-construction of performance; yet on the other hand, assigning joint scores may not fully reflect the individual's contribution to the interaction. Galaczi and Taylor (2018) discuss a potential solution to this issue in the following quote:

It is interesting that some researchers have since then argued for the awarding of shared scores for interactional competence in paired tasks (May, 2009; Taylor & Wigglesworth, 2009). They argue that we may have to design and use different assessment scales and criteria, some aimed at the assessment of individual performance, and some aimed at joint performance. This is a question that future research endeavours and academic discussions would need to shed light on (p. 231).

The findings from this study perhaps contribute to answering this question. It is conceivable that the criteria on this scale differentially measure both individual performance (individual fluency) and, to some extent, joint performance (conversational fluency), although the latter scores are not shared between participants. If the conversational fluency scores more accurately account for the characteristics of the interaction, yet are still reflective of one's ability-within-interaction, then it is conceivable that these conversational fluency scores would be much more subject to change across conversations with different conversation partners than the individual fluency scores, which likely would be much more stable. If this hypothesis is accurate, then instructors in this scale perceived individual fluency and conversational fluency differently because the former is more reflective of an individual's ability whereas the latter is much more reflective of the co-constructed interaction. Overall, the findings suggest that, to some degree, this analytic-criterion scale enables raters, to separate individual performance

(individual fluency) from co-constructed interactional performance (conversational fluency), enabling raters, to some extent, to separate the influence of the interaction in making their judgements of individual fluency. However, future research involving the use of this scale with multiple conversation partners is necessary to have further confidence in these findings.

RQ3: What relationships exist between temporal measures of fluency and items on the fluency rating scale?

The correlational analysis between temporal measures and fluency items on the rating scale produced several strong and significant correlation coefficients with suitable statistical power. Within-clause pause rate (WCpr) correlated strongly, and significantly, with all individual fluency items (*Smooth, Efficient, Sophisticated, Clear, and Holistic*) whereas filled pause rate (FPR) correlated with the conversational fluency measures (*Facilitative* and *Supportive*). Correlations between articulation rate (AR) and the *Efficient* item were also strong, significant, and statistically powerful. Other measures (e.g. pruned speech rate, PSR) also produced statistically significant correlation coefficients but they were lacking in statistical power. Since the sample size of the speaker-participants is small ($n = 8$), it is worthwhile to only discuss those correlations that crossed the threshold for meaningful statistical power ($> .80$).

It is widely agreed that pause phenomena remain at the core of the fluency construct (Wood, 2010). Even though silent pauses and other related phenomena (e.g. filled pauses, repairs, and repetitions) are a natural and necessary aspect of speech, perceivably unnatural pauses often occur within-utterances in L2 speech because of the additional processing time required for L2 learners to retrieve and organize language at the formulation stage (Levelt, 1989) of speech production. These types of pauses tend to occur more frequently within clauses rather than at clause boundaries whereas frequent pauses at the conceptualization phase indicate that

speakers are pausing to conceptualize their pre-verbal message (DeJong, 2016; Tavakoli, 2011). The findings from the present study contribute to findings from recent research (Shea & Leonard; DeJong, 2016; Khang, 2015) demonstrating that within-clause pauses influence raters' judgements about fluency as evidenced by significant inverse correlations between within-clause pause rate and holistic fluency scores. Thus, the qualitative and quantitative results align as, in phase one, instructors in the present study verbalized their observations about the importance of pause location in differentiating between more fluent and less fluent speakers.

Considering that the present study's findings are supported by related research (Shea & Leonard, 2019; DeJong, 2016; Khang, 2015), it is not surprising that a higher WCpr would disrupt the perceived smoothness and efficiency of speech; yet how WCpr disrupted the perceived sophistication and clarity of speech may require some discussion. To recall, the *Sophisticated* item reflects speakers' linguistic and topic knowledge. In phase one, instructors identified key moments in speech production where gaps in this knowledge became apparent. For instance, two instructors identified a moment in one of the learners' speech where, following the word 'jobs', the learner struggled to produce potential collocations such as 'positions' or 'openings' and eventually settled on the word 'places', which is an unusual collocation. The following excerpt illustrates this key moment:

it's also um (0.84) very friendly country in terms of receiving immigrants (0.51) to fulfill their uh (0.66) uh job (0.47) uh like uh (0.37) places (0.69) so (0.4) uh [Participant 1, conversation 3]

In the above example, the learner produces several silent pauses, in association with several filled pauses and discourse markers, not at a clause boundary, but within a clause. The instructors inferred that the learner had not fully automatized these collocations and was perhaps

hesitant to produce the word ‘places’ because, through self-monitoring, he may have known that it was incorrect and that a more suitable word was available. From this observation, the instructors inferred that the learner has not fully acquired these collocations and inferred that these gaps in linguistic knowledge resulted in the production of several within-clause pauses. These findings give credence to Pawley and Syder’s (2000) notion that fluent speakers can readily draw from a vast repertoire of automatized clausal and multi-clausal sequences to produce long stretches of speech. On the other hand, less fluent speakers may tend to formulate one-word-at-a-time sequences in order to compensate for lexical resource limitations during time-constrained performance situations (e.g. Forsberg & Fant, 2015; Foster, 2001).

WCpr also correlated meaningfully with the *Clear* item, which, in this study, refers to how the third-party listener (e.g. instructor) comprehends and attends to the speaker. Several studies have revealed relationships between fluency and comprehensibility (See Thomson, 2015 for a review); thus, if WCpr is a reliable indicator of fluent speech, then WCpr likely affects speech clarity. As Celce-Murcia (2010) describes, pauses divide larger intonation units into shorter intonation units, creating “too many unintentional prominent units. This can cause the listener to experience difficulty in processing and comprehending the overall message and may even lead to a misunderstanding of the speaker’s intent” (p. 222). Therefore, with this in mind, a higher WCpr may reduce the learners’ ability to focus the speakers’ attention on message (Lennon, 2000), reminiscent of how some listeners in Derwing et al’s (2006) study reported ‘zoning out’ when listening to disfluent speech.

It was expected, and discovered, that Articulation Rate (AR), as a measure of speech automaticity, would correlate well with the *Efficient* item on the task. This item includes references to lexical retrieval (e.g. ‘finds words quickly’) across proficiency levels. Mean length

of runs (MLR), a well-established indicator of fluency levels in monologic performances (Towell et al, 1998; Kormos & Dénes, 2004), was also expected to correlate well with the *Smooth* item on the scale, which references variations in utterance length across proficiency levels (e.g. produces long utterances between pauses). However, despite phase one instructors' observations regarding utterance length, very small to no correlations were found between MLR and any item on the fluency rating scale. There are several possible reasons for this discrepancy between the qualitative and quantitative findings in the present study. Composite measures of MLR and speech rate (SR) tend to be higher on dialogic tasks than on monologic tasks (e.g. Tavakoli, 2016; Sato, 2014; Michel et al., 2007). However, as the authors of these aforementioned studies profess, these discrepancies may be attributed to either: a) the scaffolding nature of dialogic speech; or b) the interference of between-speaker pauses on calculations of speech-pause relationships. Since raters in the present study reported that utterance length was a salient factor in discriminating between fluency levels, it is more likely that composite measures, such as MLR, are perhaps not well suited to measuring dialogic speech, which had been suggested by Tavakoli (2016) and Sato (2014). On the other hand, measures such as WCpr, which do not include between-speaker pauses in their calculations, may be a more appropriate measure of dialogic speech.

Another revealing finding is that none of the dialogue fluency measures correlated with any items, including those representing conversational fluency. This is not particularly surprising however, as so far, correlational analyses between quantitative measurements of conversational speech features and fluency ratings have not produced substantive results (e.g. Riggensbach, 1991; Tavakoli, 2016).

One interesting finding is the strong inverse correlational coefficients found between filled-pause rate (FPR) and the two conversational fluency items: *Facilitative* and *Supportive*. Several possible explanations exist for this finding. First, as Carroll (2001) claims, although filled pauses, in association with repeats and repetitions, are often used as useful compensatory strategies within L1 conversations, they are often perceived as disfluencies within L2 speech; however, instead of recognizing these compensatory strategies as ‘interactional achievements’ (Carroll, 2001), assessors may perceive them as signs of disfluent communicative speech. It is possible then that less fluent speakers used these types of compensatory strategies to claim and sustain turns more frequently than more fluent speakers. However, this would require future research with this data, using discourse analysis techniques, as in House’s (2013) analysis of the contribution of discourse markers to fluent speech.

Additionally, as McCarthy (2006) discusses, how speakers initiate turns is vital to creating confluence in conversational fluency. As McCarthy contends, beginning one’s turn is a crucial aspect of claiming and sustaining the floor as well as signalling the communicative intent of the turn. Filled pauses aid the speaker in this regard. For instance, as Fox Tree (2001) revealed, listeners may respond differently to different kinds of pauses as “um” signals a short delay whereas “uh” signals a longer delay. Moreover, as Watanabe (2008) uncovered, “um” signals that a more complex utterance is to follow; therefore “ums” may draw the listener’s attention to the utterance more effectively than “uhs”. Notably, drawing listeners’ attention to message, as Lennon (2000) describes, is an essential component of higher-order fluency. In this study, these two types of filled pauses were grouped together and analysed as one variable; yet it would be worthwhile to uncover if more fluent speakers used “um” more frequently than less fluent speakers in future analysis of this data.

Filled pauses are also devices that speakers can employ not only to claim turns but to negotiate turns as well (Schelgoff, 2007). On this task, speakers received a specific set of topic prompts and were required to negotiate topic shifting together over the course of the conversation. More proficient speakers were perceived to incorporate the topic prompts more smoothly into the conversation whereas less fluent speakers struggled to do so. It is possible that more fluent speakers used more interactional gambits for topic shifts, as per House (1996) and Bardov-Harlig (2012), whereas less fluent speakers may have used more filled pauses to negotiate topic shifts. However, this cannot be known without a detailed discourse analysis of the transcripts, which is suggested for future qualitative analysis of this data.

Overall, correlations between temporal measures of speech and rating assessments have provided insight into the construct irrelevance of the items in measuring speech fluency in a conversational setting. WCpr, as an indicator of the degree of automaticity in speech production, correlated highly with all four analytic items comprising the individual fluency construct, along with the holistic item. These results align with findings from previous studies (Shea & Leonard, 2019; Khang, 2015; De Jong, 2016) as well as the phase one findings in the present study. AR also, expectedly, correlated strongly with the *Efficient* item on the scale. FPR, on the other hand, correlated highly with the conversational fluency measures, revealing potential implications for future analysis of conversational fluency. In sum, examining how the core and peripheral items on the fluency rating scale relate to temporal measures of fluency provides further insight regarding this study's exploration of fluency perceptions.

RQ4: In what ways, if any, do raters' levels of accent familiarity affect their fluency ratings?

Thus far, the present study has discussed how a wide variety of temporal and non-temporal features of speech elicited from a paired conversational task may affect listeners' perceptions of fluency, raising questions about the construct relevance of these features. In a similar manner, this section investigates how third-party listener characteristics may also contribute to varying degrees of construct-irrelevant variance (Huang, 2013). As the following findings seem to indicate, these listener characteristics may be somewhat influenced by the 'mere exposure effect', which, according to Van Dessel et al. (2017), refers to how the amount of exposure to a stimuli positively relates to one's preference for that stimuli. To apply this concept to this study, the amount of exposure instructors have to speakers of a particular language may affect their implicit liking of a particular language, and therefore, biasing their judgements of speech fluency.

Such potential findings have implications for assessment. For instance, as Browne and Fulcher (2017) contend, if fluency is both a productive and perceptual phenomenon (Freed, 2000), then listener characteristics such as one's accent familiarity may need to be accounted for when developing construct definitions for fluency. More specifically, the authors argue that "any definition of the construct of fluency must include *familiarity of the listener with the entire context of an utterance* [emphasis added]...as fluency is as much about perception as it is performance" (Browne & Fulcher, 2017, p. 37). Additionally, rater training procedures may need to enhance raters' levels of accent familiarity, as suggested by several researchers (e.g. Derwing et al., 2002; Carey et al., 2011; Winke et al., 2013) in order to ensure a greater degree of fairness in awarding scores, thus reducing the amount of construct-irrelevant variance (Huang, 2013).

The results in the present study indicate that, on this scale, instructors who reported being somewhat familiar with speakers' accents, as a whole, were significantly more lenient towards

speakers on the *Efficient* and *Clear* items on the scale with moderate (*Efficient*) to large (*Clear*) effect sizes than instructors who reported being not familiar with the speakers' accents.

However, the descriptive results also indicate that instructors who reported being Somewhat Familiar with speakers' accents were, as a whole, more severe in their ratings than the Very Familiar group, although not significantly. Moreover, correlational analyses did not reveal any discernable pattern, neither linear nor curvilinear. Together, these results suggest that although there are clear distinctions between groups, most meaningfully regarding the *Efficient* and *Clear* items on the scale, there exists a considerable amount of variation between individual listeners, which indicates that the relationship between listeners' accent familiarity and their assessments of L2 speech is fundamentally complex and not necessarily linear.

Similarly, previous research has provided mixed results regarding this relationship, highlighting its apparent complexity. Whereas several studies (Shintani et al., 2018; Browne & Fulcher, 2017; Saito & Shintani, 2015; Winke et al., 2013; Carey et al., 2011) have revealed that the amount of reported accent familiarity has positively affected oral proficiency ratings, other studies have found conflicting results (Kennedy & Trofimovich, 2008; Derwing et al., 2006; Munro et al., 2006). Divergence may also exist within studies between the qualitative and quantitative analyses of this relationship. For instance, in Huang et al. (2016), researchers discovered that qualitative analysis of raters' verbal reports revealed that accent familiarity affected their judgements; however, the quantitative analyses revealed no significant group differences.

It has been suggested that rater training would increase raters' exposure and awareness of students' L2s (Kraut & Wulff, 2013; Winke et al., 2013; Wittenman et al., 2013; Carey et al., 2011; Derwing et al., 2002) as this enhanced familiarity with speakers' accents may increase the

‘scoring validity’ of oral proficiency assessments (Weir, 2005). In a notable study, Derwing et al. (2002) discovered that although explicit instruction in accent awareness did not improve listeners’ comprehensibility of particular accents, it did enhance their confidence in making judgments and their empathy towards speakers of that accent. In turn, Derwing and colleagues argue that this increased self-confidence and empathy could somewhat counter any implicit biases or any potential intergroup attitudes towards unfamiliar accents when making fluency judgements, as discovered by Reid et al. (2019).

If ratings are affected by listeners’ levels of accent familiarity through the extent of one’s exposure, then these findings ultimately provide some support for the prevalence of the ‘mere exposure effect’, which, as Van Dessel et al. (2017) describe, is that the amount of exposure to a stimulus relates to one’s preference for that stimulus. Yet, as the findings from the present study indicate, this mere exposure effect is not necessarily linear as correlational analyses showed otherwise as the ‘Somewhat familiar’ group, although significantly more lenient than the ‘Not familiar’ group, was also descriptively more lenient than the ‘Very familiar’ group.

RQ5: What relationships exist, if any, between raters’ conversational styles and their assessments of High Considerateness-style and High Involvement-style students, according to this scale?

In recent studies on conversational fluency (e.g. Tavakoli, 2016; Peltonen, 2017a), the importance of conversational style has been mentioned within their respective literature reviews; yet its potential relationship with fluency has not been examined exclusively. The findings from the present study, although limited in their generalizability due to the small participant-sample sizes, point towards a greater need to examine the potential influence of raters’ own conversational styles on their assessments of L2 speakers’ fluency on a conversational task.

This study's findings revealed significant relationships between instructors' HI-style characteristics (sharing personal information, changing topics suddenly, and avoiding conversational silence) and their ratings of HI-style students on several core items on the fluency scale (*Smooth, Efficient*) and the nearest peripheral item, *Sophisticated*. The results suggest that HI-style instructors may be more lenient towards HI-style students in assessing fluency as elicited from a conversation task, according to this scale.

Tannen (2005) theorizes that conversational styles are developed during one's formative years due to extended exposure to interactions with persons within one's micro-culture (e.g. peers, family members), who are inherently informed by macro-cultural (e.g. national, ethnic, linguistic) influences. With consideration of the mere exposure effect, in which exposure breeds liking, it is reasonable to assume that one's conversational style, developed from exposure to a multitude of conversational interactions over time, constitutes one's conversational style preferences. In this sense, it is vitally important to understand which conversational features listeners prefer, as these preferences may influence their judgments. To reiterate a quote presented earlier in this study: "it is important also to investigate what language experts – teachers and assessors – *value* while rating pairs, because it is their view of interaction which finds its reflection in the test scores" (Ducasse and Brown, 2009, p. 426, *italics added for emphasis*). In sum, in this study, although sharing personal information (with an acquaintance), changing topics suddenly, and avoiding conversational silence may be preferred features of conversation for HI-style instructors, the reverse may be true for HC-style instructors.

It has been often argued that avoiding conversational silence is a key indicator of higher levels of conversational fluency (McCarthy, 2006; Peltonen, 2017b). Similarly, within the present study, the phase one findings indicate that conversational silences were perceived

negatively so instructors attributed the cause of these mutual gaps to the perceivably less fluent speaker; as a result, these attributions are reflected differentially in the fluency rating scale within the *Facilitative* category.

Several researchers (e.g. McCarthy, 2006; Peltonen, 2017b) have argued that avoiding conversational silence through the use of latches, collaborative completions, overlaps, and/or interruptions may be indicators of conversational fluency yet quantitative measures of these features have not been promising (e.g. Riegenbach, 1991; Tavakoli, 2016). Riegenbach (1991) for instance, indicated that these interactive features of speech were not particularly salient to raters.

Qualitative measures have produced different results. On the one hand, Peltonen's (2017b) qualitative analysis of collaborative completions and other-repairs highlighted significant moments in learners' speeches where conversational alignment occurred due to these features, creating a co-construction of conversational flow between speakers. On the other hand, Fiskdal (2000) revealed that the relationship between other-repairs and conversational silences are affected by one's culturally informed use of positive/negative politeness strategies (Brown & Levinson, 1987). In Fiskdal's study, the Taiwanese L2 learners of English avoided other-repairs through purposeful use of the between-speaker pause to defer to the L1 English interviewer. In light of Fiskdal's (2000) contentions and as the results from this study seem to indicate, the degree to which these features of speech contribute to the avoidance of silence between speakers may be influenced by listeners' own conversational style preferences about what should transpire during a conversation.

HI-style speakers may tend to change topics abruptly, which, especially to HI-style listeners, may be a valued quality of conversational speech. Notably, perceivably long between-

speaker gaps naturally tend to occur at topic boundaries. Even in fluent speech, gaps greater than 0.8 seconds typically occur at topic boundaries (Kang, 2012), allowing both speakers time to allocate attentional resources as necessary (Pickering, 2001). It is likely that these topic-boundary gaps may be even longer within L2 speech. Therefore, more conversationally fluent L2 speakers may be able to change topics more abruptly. The qualitative results from phase one showed that this feature was salient to some instructors and the results from this phase of the study show that this feature may even be more salient to instructors who also change topics abruptly.

There is some research to support that this HI-style characteristic may be indicative of fluent speech. Young and Hallebeck (1998), for instance, observed the link between rapid topic shifts and rapid speech rates among more proficient speakers during Oral Proficiency Interviews (OPI). More specifically, the authors stated that “differences in rate of speaking translate into more rapid topic shifts (p. 374)... and perhaps as a consequence of this faster speed, higher proficiency speakers tend to shift topics more frequently than lower proficiency speakers” (p. 375).

From this standpoint, it would be reasonable to assume that quick topic shifting should be a feature of conversational speech valued by all listeners; however, the perceived appropriateness, and not just the rate of these shifts may be key to influencing their judgements. As such, whether a shift is appropriate or not may be determined by one’s own conversational style preferences. For example, Flowerdew and Miller (2005), in their text on the nature of second language listening pedagogy and assessment, comment:

Although it is not exactly known when a topic shift is appropriate, we may easily recognize when it is not appropriate. This explains why some people are accused of

‘changing the subject’ when speakers consider that the current topic has not been completed (p. 54).

However, it is possible that Flowerdew and Miller’s comments, which do not reference any research on the subject, seem to be informed by their own intuition on the matter; and if that is the case, then these comments are likely informed by their own conversational style preferences.

According to Tannen (1986), a HC-style person may prefer topics to reach closure whereas it may be perfectly acceptable to a HI-style person to change the topic abruptly without closure. With this in mind, it is not altogether surprising that instructors who themselves change topics abruptly would be more favorable towards HI-style students who may do the same. Yet, it is important to consider how situational variables may influence L2 speakers to take more responsibility for the conversation than usual (Kang, 2005) which require shifting topics more when necessary. On the other hand, frequent topic shifting, in addition to overlapping and interrupting, may be used as techniques for asserting power within a conversation (Holmes, 2008), which may or may not be relevant to one’s ability to speak fluently in an additional language.

Instructors who themselves reported being more inclined to share personal information with an acquaintance were more lenient towards assessing HI-style students. It is possible then, that instructors who exhibit this HI-style characteristic, along with ‘avoiding conversational silence’ and ‘changing topics suddenly’ perceive these characteristics favorably when assessing learners’ conversational fluency.

Asking personal questions may be perceived as a form of imposition (Tannen, 1986) upon the receiver. Therefore, the receiver may naturally respond to this dispreferred question

with a cluster of silent pauses, filled pauses, and/or discourse markers (Schelgoff, 2007). However, to what degree speakers feel they are imposed upon may be influenced largely by their own conversational styles. In other words, HC-style speakers may feel more imposed upon by personal questions than HI-style speakers would because HI-style speakers use personal questions frequently within conversation (Tannen, 1986). Additionally, as Tannen's research shows, HI-style speakers may seek to avoid conversational silences through a variety of measures such as overlaps, interruptions, and collaborative completions to create rapport by indicating a sense of shared experience or by posing 'machine-gun' style questions (i.e. short yes/no questions that often elicit personal information). HC-style speakers, on the other hand, may be more tolerant of shared silence and may prefer it. Thus, it is necessary to consider L2 learners' conversational styles when making judgments about their levels of conversational fluency.

Additionally, although instructors should be mindful of conversational style differences, there is some research indicating that adopting this HI-style characteristic, the willingness to share personal information, may help to enhance one's fluency. Nematizadeh (2019) examined L2 English learners' self-reports of fluctuations in their willingness to communicate (WTC) and fluency over the course of a three-minute picture-description task. Among his many results, Nematizadeh discovered that learners' willingness to share personal beliefs and experiences, including their personal accomplishments, contributed to their overall WTC, which enhanced their retrieval of language and ideas, ultimately contributing to enhanced fluent speech.

One's inclination to share personal information is believed to be indicative of one's degree of extroversion (Ockey, 2011) and several studies have demonstrated the effects of extroversion on fluency. Dewaele and Furnham (2000), for instance, discovered that, in their

study, extroverts exhibited faster speech rates whereas introverts hesitated more under pressure, which would seem to indicate that extroversion enhances fluency. However, in terms of ‘lexical richness’ (i.e. *Sophisticated* in this study), introverts used more infrequent and more sophisticated words than extroverts, who used more frequent and less sophisticated words, which, due to the extra cognitive processing demands, may have reduced fluent production. This study indicates the complex nature of extroversion and fluency.

In another study, Ockey (2011) discovered that assertiveness, which is a component of extroversion, and which may also be a personality trait that contributes to an HI-style, has been shown to be effective in facilitating fluent speech. Ockey claims that more extroverted speakers are more likely to seek out and capitalize on more opportunities to speak with members of the target language. From these studies, it can be inferred that certain characteristics associated with HI-style speakers may naturally contribute to enhancing their fluency, and, in reference to the results from this study, their importance may be more salient to instructors who exhibit similar conversational style characteristics.

It is somewhat surprising however that relationships were found to exist among these three aspects of HI-style features (avoiding conversational silence, changing topics suddenly, and sharing personal information) and components of the individual fluency construct (*Smooth*, *Efficient*, and *Sophisticated*), rather than components of the conversational fluency construct (*Facilitative* and *Supportive*). It was expected that any meaningful results found would be in regards to scale items that draw attention to speakers’ conversational fluency features; however, this was not the case.

These findings once again provide some implications for the development of implicit biases resulting from the mere exposure effect (Van Dessel et al., 2017), which posits that

exposure influences preferences. Since conversational styles are developed during one's formative years as a by-product of extended exposure to those around us, then it is possible that implicit conversational style biases develop as a result. It also seems likely that these conversational style biases influence raters' judgements about fluency, not unlike the well-researched implicit biases raters may possess concerning speakers' accents (Carey et al., 2011; Winke et al., 2013; the present study). However, the results from this study are far from conclusive considering the relative small sample size of participants involved in this study. Further research in this area is therefore necessary.

Chapter Eight - Classroom Applications

Although the main purpose of this study is to explore fluency perceptions through the development and piloting of a rating scale, the scale itself provides a useful assessment for learning tool that is directly applicable to classroom-based fluency pedagogy. The scale is designed to raise awareness of any gaps in linguistic competencies as well as any gaps in executive performance skills that may impede one's ability to speak fluently. With this raised awareness, ideally, learners and instructors could use the suggested activities provided below, aligned with each of the criteria in the scale, to negotiate closure of these gaps and thus enhance fluent performance. Additional pedagogical implications, beyond the use of this scale, will be provided in the conclusion section of the dissertation.

8.1. Using the fluency rating scale for assessment for learning purposes

The fluency rating scale is designed to provide information about how learners can attain a level of 'higher-order fluency' (Lennon, 2000). Whereas attaining lower-order fluency requires learners to speak at a steady pace and pause less often, attaining higher-order fluency requires learners to produce sophisticated and clear speech to speak fluently in a range of contexts, as

well as using interactional skills effectively to maintain a mutual flow of speech between speakers in a conversational setting. This scale is designed to enable learners to attain higher-order fluency through: (1) developing instructors' and learners' awareness of different strengths and weaknesses in learners' abilities to speak fluently; and (2), with instructors' guidance, use this information along with the suggested activities to meet pedagogical goals that are related to fluency development.

The scale is underpinned by theoretical concepts of fluency development, particularly McLaughlin's (1990) 'restructuring theory'. This theory posits that developing fluency not only involves retrieving information faster but through a process of conceptual re-organization, speech becomes more sophisticated and efficient. This restructuring process may result in a U-shaped pattern of fluency development, as speech becomes not only more fluent, but more complex as well (Skehan, 2009). This scale is therefore designed to reflect this pattern of development.

The table below lists the criteria provided by the fluency rating scale, development goals, and suggested fluency-building activities for achieving these goals. The following activities are suggested:

1. Activities that include task characteristics essential for fluency-building such as planning, repetition, and a change of partners (Nation & Newton, 2010). As Tavakoli and Hunter (2018) suggest, free-production speaking activities can be tweaked to become fluency-building activities by including these aforementioned task characteristics.
2. Activities that raise learners' awareness about speech characteristics that contribute to fluent speech (i.e. fluency awareness-raising activities) such as identifying pause

locations (Wood, 2009) and developing strategies for compensating for speech breakdowns (Tavakoli, 2016).

3. Explicit instruction of formulaic language (FL) (Wood, 2009; 2015) or in the use of ‘smallwords’ (Hasselgren, 2002). Short descriptions for each activity are provided in the subsequent table. Sources are provided for further information and suggestions are provided for how learners can adapt these activities for self-study.

Table 23

Fluency Rating Scale, Development Goals, and Suggested Pedagogical Activities

Fluency Scale Criteria	Development Goals	Suggested Activities
Smooth	<ul style="list-style-type: none"> • Produce long, smooth, and cohesive speech between pauses 	shadowing; formulaic language instruction; clause-chaining instruction; linking activities; drama activities (rehearsed monologues and roleplays); awareness-raising instruction;
Efficient	<ul style="list-style-type: none"> • Find words quickly • Compensate for breakdowns in speech effectively 	4/3/2; read and retell; mingle jigsaw; drama activities (rehearsed monologues and roleplays); speed dating; awareness-raising instruction; smallword instruction
Sophisticated	<ul style="list-style-type: none"> • Use a wide range of formulaic language (i.e. multi-word expressions) • Speak comfortably on a wide range of topics using topic-specific vocabulary 	formulaic language instruction; dictogloss; mingle jigsaw;
Clear	<ul style="list-style-type: none"> • Speak at a comprehensible pace (i.e. not too fast; not too slow) 	shadowing; pause-marking activities; thought groups; utterance boundary instruction

	<ul style="list-style-type: none"> • Pause in the appropriate places to help the listener understand more clearly 	
Facilitative	<ul style="list-style-type: none"> • Initiate and transition between topics smoothly • Extend topics by asking follow up questions or by providing sustained responses • Compensate for conversational pauses between speakers 	pragmatic formula instruction; problem-solving tasks; speed dating; L1-L2 discussion groups; Q> SA + EI; utterance boundary instruction
Supportive	<ul style="list-style-type: none"> • While listening, support the other speaker effectively by using a variety of the following: backchannels (e.g. “uh-hmm”); one-word responses, (e.g. “yeah”); supportive body language (e.g. head nodding); and multi-word responses (e.g. “oh that’s great”) 	pragmatic formula instruction; smallword instruction; L1-L2 discussion groups

Table 24

Descriptions of Suggested Activities

#	Descriptions of Activities (in alphabetical order)	Source(s)
1	4/3/2: Learners are required to speak on a familiar topic three times, at different lengths (4 min/3 min/2 min), and to three different speakers.	(Nation, 1989; DeJong & Perfetti, 2011)
2	Chat Circles: This activity first requires that learners form two concentric circles, with students in the inner circle and students in the outer circle facing one another. Then, students discuss a topic from a list of topics related to the input text, in pairs, for two minutes. Afterwards, students switch partners within their respective circles and then discuss a new topic. In between discussions, students comment and reflect on any potentially disfluent moments within the conversations.	Wood (2009); Thomson (2017)

- 3 ***Dictogloss:*** Learners are required to listen to key sentences from the input text, make notes about key words and phrases, and then collectively reconstruct the text from their notes. This activity enables learners to notice and reproduce formulas through repeated recognition and production. Wood (2009); Thomson (2017)
- 4 ***Drama Activities (well-rehearsed monologues and roleplays):*** Learners study, rehearse, and dramatize monologues and roleplays in a performance setting, encouraging learners to exaggerate range of volume, dramatic silences, and stress patterns, while automatizing language output. Galante & Thomson (2017)
- 5 ***Explicit Formulaic Language (FL) Instruction:*** In addition to the number of activities which enhance both FL and fluency acquisition (e.g. dictogloss, mingle jigsaw), instructors can encourage learners to notice FL in input texts and tweak free-production speaking activities by encouraging learners to use target formulas. Wood (2009); Thomson (2017); McGuire & Larson-Hall (2017)
- 6 ***Linking Activities:*** Gilbert suggests drawing learners' attention to the phenomenon of consonant attraction (i.e. re-syllabification) as it pertains to formulaic expressions (e.g. "how's it going?" > howz it going?) and structure words such as "and" (e.g. "cream and sugar" > creamən sugar). Gilbert suggests paired-reading of individual sentences and dialogues accompanied by peer feedback to enhance noticing and automatization of this phenomenon. Gilbert (2012)
- 7 ***L1-L2 Discussion Groups:*** In this activity, L1 speakers and L2 learners engage in a free-production speaking exercise prompted by a list of topics. Ziegler et al (2013) found that over time, successful L2 learners adopted conversational style characteristics of the L1 group such as overlapping and collaboratively completing. As a whole, all learners reported a higher willingness-to-communicate with the target L1 group. It is likely that this activity may enhance one's conversational fluency. Moreover, incorporating Nation and Newton's (2010) fluency task characteristics of repetition and a change of partners could turn this free-production activity into a fluency-building activity. Ziegler et al. (2013)
- 8 ***Mingle Jigsaw:*** This activity requires learners to memorize formulas verbatim from the dictogloss text, share them with their peers, and then listen to others share theirs. Wood (2009); Thomson (2017)
- 9 ***Pause-Marking Activities:*** Learners listen to recordings of several L1 English speakers telling stories, which consist of a variety of formulas. Wood (2009); Thomson (2017)

The instructor then draws learners' attention to FL and asks the learners to mark pauses according to a transcription of the recording.

- 10 **Pragmatic Formula Instruction:** House (1996) revealed the positive effects of explicit instruction of pragmatic formulas learners used these formulas to initiate and sustain topics and turns in order to enhance their pragmatic fluency. However, House did not describe specifically which activities were used in her study. Foreseeably, Dictogloss and Mingle Jigsaw can be used to help learners notice and reproduce pragmatic formulas. Additionally, as Larson and McGuire (2017) discovered, instructors can tweak their instructional practices by encouraging learners to notice and reproduce formulas.

House (1996);
Wood (2009);
Larson and
McGuire-Hall
(2017)
- 11 **Problem-Solving Tasks:** These encourage learners to negotiate an outcome through interaction have been shown to enhance learners' conversational fluency. For example, in Peltonen's (2017a) study, learners were required to choose and rank a list of items that they would want to have with them should they be stranded on a desert island.

Peltonen
(2017a)
- 12 **Q > SA + EI:** Nation and Newton (2009) suggest using an interview sequence called Q > SA + EI (Q = Question; SA = Short Answer; EI = Extra information) to help learners keep a conversation going. The focus of the activity is to give learners practice in providing extra information, which could be a fact, feeling, question, and which learners could use to initiate new topics, extend upon topics, or sustain one's own turn.

Nation &
Newton (2010)
- 13 **Shadowing:** Learners listen and imitate the intonation and pausing patterns of the input text. This activity encourages learners to notice FS and their inherent prosodic cues (intonation contours and pauses) through imitation and repeated listening of the input text. Chun (2012) suggests that providing learners with visual information of the intonation contours enhances acquisition. Fox and Hirotoni (2017) discovered that computer-mediated versions of the shadowing activity enhance language automatization.

Wood (2009);
Chun (2012);
Fox & Hirotoni
(2017);
Thomson (2017)
- 14 **Smallword/Discourse Marker Instruction:** Hasselgren (2002) discovered that more fluent speakers use more smallwords (e.g. oh, well, I mean), most notably in turn-initial position. Any of the activities listed here which focus on identifying FL in input texts could also be adapted to encourage learners to notice small words.

Hasselgren
(2002)
- 15 **Speed Dating:** The speed dating activity incorporates Nation's task characteristics of repetition, time pressure, a focus on meaning, and a change of partners. In this activity, students introduce themselves to

East (2012)

one another in a time-pressure situation. Then, students change partners. As an instructor in East's (2012) study recounted, this activity requires students to build interactional skills through negotiating meaning. Moreover, feedback between each repetition allows learners to reflect on strengths and weaknesses within each interactional scenario.

- 16 ***Thought Groups:*** Chafe (1988) describes how the intonation unit is a unit of consciousness. For pedagogical purposes, Gilbert (2012) presents the notion of the intonation contour as a thought-group. Although, as indicated by Lin (2018), not all intonation units are marked by pause boundaries in advanced learner speech, less fluent speech may be marked more precisely by pause boundaries. Therefore, for developing learners, intonation units are more likely to coincide with pause boundaries. With this in mind, Gilbert suggests drawing learners' awareness to thought-groups by marking pause boundaries in transcripts and then reciting them accordingly. This procedure is similar to the Shadowing activity. Gilbert (2012)
- 17 ***Utterance-Boundary Intonation Instruction:*** Wennerstrom (2000) discovered that L1 speakers often interrupted L2 learners who produced falling intonation contours mid-utterance. Producing plateau (flat and even) intonation while searching for words can signal to the other speaker that one wishes to continue. Chun (2012) recommends using software that can provide learners with visual information about utterance-boundary intonation. Wennerstrom (2000); Chun (2012)

8.2. Self-study

Many of these activities may likely be adapted for self-study purposes, especially if students are able to rehearse, record, and repeat their speeches. Students are also encouraged to seek out opportunities to use English in a conversational setting through joining extra-curricular clubs (Wu, 2012), volunteering (Springer & Collins, 2008), and through participating in online chats (Blake, 2009). Reading and listening extensively for substantial periods of time may also help to build speech fluency (Segalowitz & Freed, 2004). Notably, more research is needed to develop and understand the effects of self-study activities on building fluency.

8.3. Accent familiarity awareness

The findings from the present study revealed how instructors' level of familiarity with students' accents affects their judgments on the *Clear* and *Efficient* question items to a moderate degree. Exposing one's self to different accents has been shown to counter this bias (Derwing et al., 2002). Moreover, encouraging students to project 'vocal confidence cues' such as an increased volume and a more dynamic pitch range, may potentially offset listeners' biases. As Jiang, Sandford, and Pell, 2018 have shown, using these vocal confidence cues may help to counter any potential biases from the general public in real-life communicative scenarios.

8.4. Conversational style awareness

Tannen's (2005) concept of 'conversational style' posits that one's style exists somewhere along a continuum between High-Involvement (HI) and High-Considerateness (HC). According to Tannen, HI-style speakers are more likely to share personal information more willingly, ask personal questions more readily, interrupt and overlap more frequently, speak more rapidly, and avoid conversational silences more often whereas the reverse may be generally true for HC-style speakers. The results indicated weak-to-moderate but statistically significant relationships between instructors who report exhibiting certain HI-style characteristics (sharing personal information more readily, changing topics suddenly, and avoiding conversational silence) and their fluency judgments on the *Smooth*, *Efficient*, and *Holistic* question items. In other words, in this study, HI-style instructors were shown to be more favorable towards HI-style students in their fluency assessments. Becoming aware of conversational style differences and reflecting upon what one values as a good conversationalist may enable more valid judgements about conversational fluency.

Chapter Nine - Conclusion

9.1. Overview of the study

The overall purpose of this study was to explore EAP instructors' perceptions of fluency through developing and piloting a fluency rating scale for a paired conversational task. This study was motivated by a desire to inform research about which characteristics of learners' speeches, as elicited by a conversational task, influence fluency perceptions and how characteristics of the listeners themselves, such as accent familiarity and conversational style, also influence fluency perceptions. A two-phase mixed-methods sequential exploratory instrument model design (Creswell, 2009) provided the underlying methodological framework for this study. The first phase involved the collection and analysis of qualitative data. Seven instructors watched videos of seven-minute paired conversations, elicited from 14 EAP learners who performed this conversational task twice with two different conversational partners. After watching the videos, instructors were audio-recorded verbalizing their judgments about learners' fluency levels. Once all videos were viewed, instructors then placed each learner within one of three categories: high, middle, and low. These audio recordings were transcribed and coded using in-vivo and pattern coding techniques (Saldaña, 2009). The codes resulted in the development of six themes with corresponding categories and codes. The six themes were as follows: smoothness, efficiency, sophistication, clarity, facilitating topics and turns, and supporting the conversational partner. These themes were used to create a multi-item analytical-criterion fluency rating scale, inspired by Fox et al.'s (2016) scale which was used to diagnostically assess university-level engineering majors' writing competence. Categories and codes informed the scale descriptors for the fluency rating scale.

The purpose of this second phase was to continue to examine perceptions of fluency through quantitative means. Four research questions were posed in this phase. The first two questions were posed to examine the degree to which these scale items were relevant to

measuring fluency through analysing inter-item relationships and through analysing relationships between scale items and temporal measures of fluency. Additionally, this second phase investigated the potential influence of two key listener characteristics, accent familiarity and conversational style preferences, on fluency assessments.

In this second phase of the study, a new group of 35 EAP instructors were recruited to watch videos of four paired conversations between eight learners, and then use the newly created scale to rate the performances. Prior to rating, instructors underwent a brief training session, involving a review of the task specifications, construct definition, and scale descriptors. As part of the training session, instructors watched one video to get accustomed to the scale and to practice rating two speakers simultaneously. Prior to rating each video, instructors were informed of the students' L1s and then instructors assessed their own familiarity with students' L1s on a six-point scale. Once all the instructors rated all eight learners according to the scale, instructors then completed a questionnaire that was designed to elicit their conversational style preferences according to Tannen's (2005) High-Involvement/High-Considerateness dichotomy.

As indicated below, the findings from both phases of the study were informed by theoretical notions of the following: a) how speech is produced with regards to automaticity (Levelt, 1989; McLaughlin, 1990; Segalowitz, 2010) and attention (Segalowitz, 2007); b) discussions about the role of interactional competence (Galaczi & Taylor, 2018) in co-constructing conversational fluency between speakers (Sato, 2014); and c) the effects of exposure on perception, such as the mere exposure effect (e.g. VanDessel, 2017).

9.2. Overview of the Findings (Phase One)

The findings revealed that a wide range of temporal, non-temporal, and non-verbal features of communication informed instructors' perceptions of fluency. The following is a list of the most revealing findings:

1. Instructors reported that learners' utterance length was a perceivably salient fluency feature, which was not revealed in Préfontaine and Kormos' (2016) qualitative analysis of fluency perceptions.
2. Instructors highlighted the saliency of phonological linking, which has been under-investigated thus far, with exception to a few studies (Hieke, 2005; Waniek-Klimczak, 2014).
3. Instructors inferred that learners who used a wide range of lexical resources, including a wide range of formulas would be fluent in more demanding contexts; on the other hand, learners who were perceived to have a reduced range of resources, as demonstrated through repeating words, phrases, or ideas were perceived to be less fluent. These findings suggest that raters do not equally value the individualized ways, as identified by Wood (2006), in which learners use formulas to maintain fluency.
4. Instructors commented on the inherent interrelationship between fluency and comprehensibility. Instructors reported that speakers who spoke at a high speech rate but were less comprehensible were perceived as being less fluent than speakers who spoke at a slower speech rate but were more comprehensible. These findings provide support for the notion that fluent speakers have greater attentional control (Segalowitz, 2007) and processing flexibility (Seglaowitz, 2010) as more fluent speakers have automatized processes of formulation and articulation so that they can shift their attentional resources

more efficiently towards monitoring their speech rate and pause placement to make themselves more comprehensible.

5. Instructors observed that more fluent learners integrated the task prompts into the conversation without hesitation and in natural manner, also indicating that more fluent learners have the attentional capacity to process and integrate information from the surrounding communicative context in an efficient manner.

6. Instructors observed that fluent speakers would take the initiative to compensate for gaps in conversation, whereas instructors attributed between-speaker gaps to the less fluent speaker. Similarly, Sato (2014) discovered that fluent speakers ‘scaffolded conversation’ through overlaps, collaborative completions, and other-repairs to keep the mutual flow of speech going.

7. Instructors observed that fluent speakers were more capable of extending topics and turns to sustain conversation, either by expanding upon what the initial speaker said, or by asking follow-up questions to encourage the other speaker to continue discussing the topic. House (1996) discovered similar results in her investigation on the development of pragmatic fluency.

8. Instructors observed that more fluent speakers used a wider range of non-lexical (“uh-hmm”), non-verbal (e.g. eye contact, head nodding), one-word (“yeah”) backchannels, as well as, most notably, substantive lexical phrases to keep the flow of conversation going.

9. Overall, instructors’ fluency perceptions were influenced by core and peripheral language features to varying degrees: Smooth > Efficient > Sophisticated > Clear > Facilitative > Supportive. These findings provide implications regarding the extent to which peripheral features should be included in fluency testing rubrics.

9.3. Overview of findings (phase two)

The phase two findings revealed the following. First, within a conversational task, EAP instructors rated individual fluency and conversational fluency differentially. Second, within-clause pause rate correlated meaningfully with items representative of the individual fluency construct whereas filled-pause rate correlated meaningfully with items representative of the conversational fluency construct. Third, instructors' accent familiarity moderately affected assessments of certain items (*Efficient* and *Clear*). Finally, significant correlation coefficients were found between certain instructors' conversational style characteristics (sharing personal information, changing topics suddenly, avoiding conversational silence) and their assessments of fluency on a conversational task.

9.3.1. Individual fluency and conversational fluency

The PCA analyses yielded two key findings. The rating scale consisted of six analytic items and one holistic item. In formatting the rating scale, the analytic items were arranged so that items reflecting core fluency features (*Smooth*, *Efficient*) would be followed by items that, theoretically, reflect speech features that become more and more peripheral to the construct (*Sophisticated*, *Clear*, *Facilitative*, and *Supportive*). The correlational matrix shows support for this arrangement as correlations between each of the six analytic items and the seventh item, a holistic measure of fluency, gradually weakens from the first item (*Smooth*) to the sixth item (*Supportive*). The second key finding is that the PCA analyses, which grouped the items together into components based on their correlational strength, produced two key components. The first component consists of both core (*Smooth*, *Efficient*) and peripheral (*Sophisticated*, *Clear*) speech features and the second component consists of peripheral variables related to conversational interaction (*Facilitative*, *Supportive*). As a result, these components were labeled 'Individual

Fluency’ and ‘Conversational Fluency’, respectively. These results suggest that the scale measures two distinct, but associated, constructs. However, the question remains as to whether these individual fluency and conversational fluency are part of the same whole or if they constitute two separate wholes as per Sato’s (2014) conclusion that individual fluency and conversational fluency are two separate constructs.

Arguably, the findings show that conversational features of speech may be so peripheral to the fluency core, that these features may comprise a separate construct altogether. Additionally, the findings from the present study indicate that not only are instructors able to reliably assess two speakers simultaneously, but that instructors can capably distinguish between features related to individual fluency and features related to conversational fluency. These findings have implications for one of the on-going challenges in assessing interactional competence – the extent to which test scores reflect characteristics of the interaction itself (Galaczi & Taylor, 2018). This analytic-criterion scale enables raters, to some degree, to separate individual performance (individual fluency) from co-constructed interactional performance (conversational fluency). In other words, this scale may help raters to separate the influence of the interaction in making their judgements of individual fluency. For assessment for learning purposes, distinguishing between fluency features attributable to individual fluency and to those features attributable to conversational fluency enables instructors and learners to better target weaknesses in those areas in order to enhance overall ability to speak fluently on a conversational task.

The present study included an analysis of relationships between a wide variety of temporal variables and rating scale items representing individual fluency (*Smooth, Efficient, Sophisticated, Clear*) and conversational fluency (*Facilitative, Supportive*). As for individual

fluency, the results revealed significant inverse correlations with strong statistical power for between within-clause pause rate and all items representing individual fluency, providing further support for the construct relevance of all the items on the scale in measuring individual fluency. This finding contributes to informing research about the use of within-clause pause rate to measure utterance fluency (Shea & Leonard, 2019; Khang 2018; 2014; DeJong, 2016).

As for conversational fluency, significant inverse correlations with strong statistical power were found between filled-pause rate and measures of conversational fluency. Two possible reasons may account for this finding. The use of filled pauses may be idiosyncratic, particularly in how they are employed as a repair strategy (Tavakoli et al., 2016) or, similarly, as a stalling mechanism (Peltonen 2017a). Since the sample size of the present study was relatively small, then it would not be surprising that idiosyncrasies were found to be more salient than they may actually be beyond this study. On the other hand, more fluent speakers may have used a wider range of discourse markers, particularly in turn-initial position, to claim and sustain turns, as shown in several studies (Crible & Pascaul, 2019; Crible, 2016; House, 2013; Hasselgren, 2002). Thus, less fluent speakers may have relied more on filled pauses in turn-initial position to sustain the flow of conversational speech (e.g. Revesz et al., 2016; Bosker et al., 2013). However, it is not possible to support this inference without a detailed conversational analysis of the function of filled pauses in turn-initial position within conversational speech.

9.3.2. Listener characteristics

This study also examined the effects of two key listener characteristics, accent familiarity and conversational style preferences. Although much research has investigated the effect of accent familiarity on overall oral proficiency, not much research has investigated its effects on

fluency ratings specifically. Moreover, the present study may be the first to investigate the effects of conversational style preferences on fluency judgements.

9.3.2.1. Accent familiarity

The effect of accent familiarity on fluency judgments was examined by using non-parametric tests of group differences (Mann-Whitney U and Kruskal-Wallis). Instructors' self-assessments of accent familiarity on a six-point scale were grouped as follows: Very Familiar (6, 5); Somewhat Familiar (4); and Not Familiar (3, 2, 1). The results indicated that instructors who were somewhat familiar with the speakers' accents, overall, were significantly more favorable towards speakers on the Efficient and Clear items, with moderate to large effect sizes respectively, than instructors who reported being not familiar with speakers accents. However, no significant results were found regarding the instructor-group who reported being Very Familiar with speakers' accents. Moreover, correlational analyses did not reveal any meaningful relationships between accent familiarity and any of the items on the fluency rating scale. Taken together, these findings indicate that, although accent familiarity may influence listeners' fluency judgements, the relationship is not necessarily linear, as it may be complicated by a number of variables not investigated in this study such as the amount of teaching experience (Kennedy & Trofimovich, 2008) inner-circle/outer-circle speaker status (Saito & Shintani, 2015), and intergroup attitudes (Reid et al., 2019). Broadly, these findings can be explained in part by the mere exposure effect phenomenon, referring to the likelihood that exposure to a stimulus, such as another speaker's accent, may foster liking for that stimulus. For instance, exposure to a range of accents may positively affect one's "interlanguage phonological familiarity" (Carey et al., 2011, p. 104), as, according to Kuhn's (1991) 'perceptual magnet effect model', listeners become more tolerant of phonetic deviations from their storage of prototypical sounds. Intergroup exposure

may also breed more liking for the out-group, yet as Reid et al. (2010) showed, negative exposure, even in the short term, can also breed negative judgements. In sum, although the present study's findings are promising in that they inform research about how accent familiarity affects fluency ratings more specifically, and not just oral proficiency ratings more generally, much more research would be needed to uncover the cause of these effects.

9.3.2.2. Conversational style characteristics

The relationships between instructors' conversational style characteristics and their fluency judgements were examined through correlating instructors' responses to the nine-item conversational style questionnaire with mean scores of high-involvement style (HI-style) students and mean scores of high-considerateness style (HC-style) students. The questionnaire consisted of nine items, which were informed by certain features that are characteristic of HI-style speakers, adapted by Tannen (2005): sharing personal information, asking personal questions, interrupting, speaking rapidly, overlapping, speaking loudly and with enthusiasm, finishing the other person's sentences, persisting in getting across what one wants to say, and avoiding conversational silence.

The present study is the first to investigate relationships between conversational style and conversational fluency directly, although indirect links between these two phenomena can be inferred from research on speech features such as topic shift rate (Young & Hallebeck, 1998), conversational parity (Morales-López, 2000), overlaps and collaborative completions (Peltonen, 2017a); and politeness strategies (Fiskdal, 2000). The results from the present study indicated weak-to-moderate but significant correlations between several items on the scale (sharing personal information, changing topics suddenly, and avoiding conversational silence) and mean scores of HI-style students. These findings suggest that instructors who exhibit these HI-style

characteristics may be more favorable towards HI-style students in their judgements of core features of fluency (smooth, efficient, and holistic) according to the rating scale. No meaningful correlations were found regarding HC-style students. As Tannen (2005, 1986) contends, conversational style differences may result in conversationally disfluent moments; moreover, speakers may assign negative attributes to a speaker of an opposing style (Yule, 1996). These findings suggest that third-party listeners (L2 instructors) may also possess conversational style preferences, affecting their judgments of fluency on a conversational task.

9.4. Attention to limitations

One of the key drawbacks is that this scale is designed for a relatively small proportion of the English language learner population – intermediate-to-advanced learners in an EAP setting. Therefore, this study's findings regarding fluency perceptions are somewhat limited in terms of how well they can be generalized to a broader population of learners across proficiency levels in different context.

Another limitation is that quantitative techniques were performed with modest sample sizes: EAP instructors ($n = 35$) and EAP learners ($n = 8$). However, certain procedures were taken to account for these modest sizes. For instance, regarding research question two, quantitative techniques such as the Principal Component Analysis (PCA) are typically conducted using larger sample sizes. However, a series of statistical procedures, such as the KMO measure of sampling adequacy (Tabachnick & Fidell, 2007, as cited in Pallant, 2007) were performed to ensure that the data was suitable for analysis in terms of sample size and the strength of the inter-item correlations. Also, regarding research question three, only correlation coefficients between fluency ratings and temporal measures (i.e. within-clause pause rate), which were not only significant but met the threshold for suitable statistical power ($>.80$), were deemed to be worthy

of much discussion. Finally, Evans' (1996) classifications of effect size and correlational strength were used in this study. Since, Evans' set of of classifications provides more conservative estimates of correlational strength and effect sizes than Cohen's (1988), like with statistical power, only the most promising results regarding effect size were discussed at length in the present study.

Regarding the findings about accent familiarity, non-parametric tests of group differences were used in light of the small sample sizes. Additionally, a Multi-Faceted Rasch Analysis of individual rater characteristics could have been used to provide more information about individual differences, as used in previous studies of this nature (e.g. Browne and Fulcher; 2017; Winke et al., 2013). However, the sample size for the present study is too small to warrant use of this procedure. Ultimately, further analyses with larger groups of participants are necessary to have further confidence in these findings.

Another limitation requiring attention concerns the extent to which the scale measures 'trait fluency'. The study used a performance-based assessment task – a paired conversational task - to elicit instructors' perceptions of fluency from a small group of graduate students in an EAP setting. Thus, the findings may be limited to discussions of "state fluency" (i.e. performance-based fluency), which is highly subject to contextual influences - rather than "trait fluency" (fluency as a general fixed attribute) which is inherently mediated but is generally stable across contexts (Derwing, Munro, Thomson, & Rossiter, 2009). As Segalowitz (2016) argues, without an investigation of the relationships between L1 and L2 fluency, and/or without evaluating performances elicited through a variety of tasks amassed in an oral language portfolio (Riggenbach, 1998), it is not possible to provide generalizations regarding learners' trait fluency from the present study's findings.

However, to argue for the relevance of assessing ‘state fluency’, in task-based assessment, raters are required to make observations about learners’ performances on a particular task and then make inferences from these observations about what the learner is experiencing during the task. This leads raters to make predictions about how the learner would likely perform in similar settings in the future (Brindley, 1994). Therefore, although the findings may be limited in generalizability as they cannot clearly reflect learners’ trait fluency - and therefore the resulting scale cannot adequately assess learners’ trait fluency - the findings are in line with the more qualitative nature of this study by providing a rich and thick examination of a relatively small sample of participants through an examination of listeners’ perceptions of learners’ state fluency as relevant to this particular context. Moreover, since the overall purpose of this study was to examine perceptions of fluency, the true subjects of investigation in this study were the listeners, not the speakers.

9.5. Future areas for research

This study covered a broad range of areas related to perceptions of fluency on a conversational task including fluency development and the role of pedagogy, interactional competence and non-verbal fluency, construct validity and rating scale design, and listener characteristics and rater bias. The findings from the present study provide the potential for future research in these areas.

More research is needed to understand the degree to which perceptions of conversational fluency reflect how conversational fluency develops. A wide range of factors affect the degree to which conversation is fluent at any given moment within any given task, which is why quantitative measures of conversational fluency features (e.g. turn rate) have yielded less than promising results overall (e.g. e.g. Riggenbach, 1991; Tavakoli, 2016). On the other hand,

studies providing qualitative analyses of conversational features have provided more insight (Wennerstrom, 2000; Sato, 2014; Peltonen, 2017b), to which the qualitative findings from the present study regarding conversational fluency features have contributed. Future analysis of this data could include a discourse analysis of linguistic features, occurring at turn-relevant places with a particular focus on discourse markers and pragmatic formulas, in order to attain a fuller understanding of their influence on perceptions of conversational fluency. Relatedly, in the present study, raters perceived formulaic language to be a salient feature of fluent speech; yet the findings also suggest that the idiosyncratic ways in which learners use formulas to maintain fluent speech (Wood, 2006) were not equally valued by raters. Therefore, future research is needed to investigate, more fully, how raters differentially value learners' use of formulas when making judgments about fluent speech.

Future research using this data could investigate the efficacy of this scale in enhancing pedagogical purposes. This scale was designed to aid instructors to help learners achieve a higher level of fluency over the course of an instructional program. More specifically, the scale was designed to help identify learners' weaknesses in certain areas so that instructors could provide instructional activities to help remedy these weaknesses. Future research is required to investigate the efficacy of this scale in conjunction with these suggested activities in developing learners' fluency. Moreover, future research is required regarding the provision and use of self-study activities, which foreseeably, learners could use in tandem with this fluency scale to self-assess progress as they develop fluent speech beyond the classroom.

The influence of non-verbal communication poses another challenge to assessing both interactional competence (Plough et al., 2018) and fluency (Götz, 2013). More research is needed to investigate the effects of non-verbal features on both the production and perception of

fluency; in particular, future research is recommended into how regulators, which are non-verbal communicative features that “direct the back-and-forth nature of speaking and listening” (Rowe & Levine, 2015, p. 322), are used to create conversational confluence.

Further investigations are recommended in terms of using this scale to investigate the effects of context (topic, situation, and interlocutor) on affecting perceptions of fluency. For example, it is worth examining the use of this scale with multiple participants to investigate whether or not the individual fluency measures remain relatively more stable while the conversational fluency measures vary much more widely. However, there would be an inherent problem with such research. Task repetition, and in particular, topic repetition, has been shown to increase utterance fluency measures (DeJong and Perferti, 2011), which therefore may increase perceived fluency measures, as reflected in the items on the scale. Thus, task characteristics may likely interfere with such an investigation; future research of this nature would have to take such characteristics into account.

In addition, these findings may also provide methodological implications for research on conversational fluency. Previous research has shown that, according to analysis of temporal variables, learners perform more fluently on dialogic tasks than on monologic tasks (e.g. Tavakoli, 2016). This may be the case for two reasons. For one, analysis of pause phenomena is complicated by questions of pause attribution, as noted by several researchers (Michel, 2007; Sato, 2014; Tavakoli, 2016; Peltonen, 2017a). Without a detailed discourse analysis of the pause then it not possible to know to whom the pause shall be attributed. Removing these between-turn pauses or dividing these pauses in half may result in widely different results (Tavakoli, 2016), neither of which can be directly comparable to pause analysis of speeches elicited from monologic tasks. The second reason, evidently, is the interlocutor effect. Dialogic tasks allow for

the opportunity to scaffold one another's performances to create a greater sense of confluence in the mutual flow of speech (McCarthy, 2006).

Although the first issue remains problematic, the second issue may be helped somewhat by this study's findings. Potentially, if raters are capable of separating fluency features of individual speech within a conversation (i.e. individual fluency) from fluency features of speech mediated by the conversation (i.e. conversational fluency), then perhaps the interlocutor effect on fluency ratings could be reduced somewhat. If such is the case, it is possible that use of this scale, or one similar, would allow for more stable comparisons of performances elicited from monologic and dialogic tasks by only comparing items reflected by the individual fluency construct.

Finally, more research is required to understand more about the relationships between listener characteristics and fluency judgements. The potential relationship between accent familiarity and fluency ratings demonstrates promise. Previous research has investigated how accent familiarity has affected general oral proficiency ratings but not fluency ratings in particular. Browne and Fulcher (2017) reported the influence of raters' familiarity with L1 Japanese on Japanese L2 English speakers' fluency ratings; yet fluency was defined in that study according to the TOEFL rubrics for 'Delivery', which include references not only to fluency, but also to features of pronunciation and to listeners' interpretation of comprehensibility. As a result, it is difficult to understand from Browne and Fulcher's (2017) article how well accent familiarity affected ratings according to perceptions of core fluency features specifically. Future research with larger groups using Multi-Faceted Rasch Analysis techniques is recommended in order to assess individual, rather than group, differences in assessments.

Further research on conversational style and conversational fluency assessments are recommended using the conversational style questionnaire created and validated in the present study. As explained in the results section, as three of the nine items did not produce any meaningful associations, these items should be removed; thus the questionnaire should only consist of six items: share personal information about yourself; change topics suddenly, interrupt, speak at a faster rate than your partner, avoid conversational silence, and overlap. A Cronbach alpha analysis of this newly created six-item questionnaire produces a marginally similar coefficient ($\alpha = .863$) as the original nine-item questionnaire ($\alpha = .880$). The results indicate that a six-item questionnaire may provide more meaningful results than a nine-item questionnaire in analysing relationships between conversation style and conversational fluency in future studies.

9.6. Future applications

9.6.1. Methodological applications

As discussed previously, two of the key challenges to assessing interactional competence are authenticity and variability (Galaczi & Taylor, 2018). Variability is an inherent aspect of interaction; yet attempts to control variability for comparability purposes, by imposing highly structured task demands, negatively affects the interaction's degree of authenticity. Rating procedures also contribute to the authenticity of the test-taking situation. Sato (2014), whose study on interactional fluency perceptions inspired the present study, had raters listen to each speaker twice through headphones. In Sato's study, the speaker in the left headphone was rated first followed by the speaker on the right. However, in a real-life testing situation, instructors may be required to rate both speakers simultaneously, while considering the effects of non-verbal features of communication. Therefore, the rating procedures of the present study are more

authentic in their reflection of a real-life testing scenario and could be applied to future research on interactional competence and conversational fluency.

9.6.2. Classroom applications

Speaking fluently within a conversation requires the acquisition of an additional set of skills not measured by traditional measurements of fluency as elicited through monologic performances. The main purpose of this study was to examine perceptions of fluency on a conversational task through scale development, not to build a validity argument for the use of this scale. Nevertheless, this scale is a positive outcome of this research and some validity evidence has been provided to support its use in low-stakes classroom-based assessment for learning purposes in intermediate-to-advanced classrooms. As Fulcher (2010) claims, the goal of assessment for learning on a performance-based task is to enhance learners' understanding of their current abilities in order to narrow the gap in achieving target-level abilities. With this raised awareness, ideally, learners and instructors could use the suggested activities, aligned with this scale, to negotiate closure of these gaps and thus a level of 'higher-order fluency' (Lennon, 2000) through acquisition of both core and peripheral fluency features. Acquisition of these peripheral language features are vital to developing one's fluency, since as Chambers (1997) concludes, it is simply not helpful to tell students to speak faster and to pause less. The previous chapter outlined some of the activities that instructors can employ in tandem with this scale such as *4/3/2* (Nation, 1989); *Dictogloss* (Wood, 2009); and *Speed Dating* (East, 2012), among others.

9.6.3. Training programs

The findings from this study may be applicable to informing L2 instructor-training pedagogy. These findings highlight the importance of enabling instructors to self-examine one's degree of accent familiarity, one's own perceptions of fluency, and what one values in a

conversation. Raising one's awareness in these areas would likely positively affect one's instructional practices.

Several studies have shown how training programs that expose listeners to different accents have revealed positive results (Kraut & Wulff, 2013; Derwing et al., 2002). Most notably, these programs positively affected listeners' attitudes towards L2 speech; yet it is unknown if the effects of training are long lasting. Moreover, as Jiang et al. (2018) discovered, L2 speakers who project confidence through an increase volume and a more dynamic pitch range were perceived more positively than speakers who did not project these confidence cues. Jiang et al.'s (2018) findings suggest that instructors should train students to project confidence through voice quality in order to offset any negative judgements caused by listeners' lack of accent familiarity.

The present study's findings suggest that it would be worthwhile for L2 instructors to raise awareness about conversational style differences through self-reflection and through encouraging learners to engage in self-reflection. The conversational style questionnaire presented in this study could be used in these self-reflections. Ziegler et al. (2013) discovered that discussion groups enabled L1 English speaking American learners' acquisition of L2 German conversational style, which is reflective of a high-involvement style. Successful learners adopted features of this style over a six-week period and all learners reported a higher degree of willingness to communicate with the German speakers. However, learners were still unable to identify conversational style preferences. Combining these findings with the findings from the present study, one could infer that explicit instruction of conversational style differences in tandem with these discussion groups could potentially yield even more positive results.

As Tavakoli and Hunter (2018) profess, instructors should be better trained to define fluency in its narrow sense by characterizing fluency only by its temporal features. However, temporal features only represent the tip of the iceberg (Lennon, 2000), and restricting research on fluency to monologic tasks further restricts our knowledge of “what fluency really is” (McCarthy, 2006). Therefore, it would be worthwhile to inform instructors that their judgements about the effects of peripheral features should be valued, rather than discarded, because peripheral fluency features, to varying degrees, comprise essential components of higher-order fluency. It is simply not helpful to tell students to speak faster and pause less (Chambers, 1997); however, it is helpful to target students’ deficiencies in certain linguistic competencies (lexical, phonological, and pragmatic) in tandem with delays in executing certain speech functions, as expressed through speed, pause, and repair phenomena, in order to assist students in attaining higher-order fluency.

Therefore, in conclusion, instructors, and L2 practitioners in general, should recognize and appreciate the relative influence of peripheral features on affecting how fluency is defined, categorized, analysed, and evaluated on a conversational task. As well, training programs that enable trainees to self-examine their own levels of accent familiarity, their own conversational styles, and their own fluency perceptions would positively affect the development and assessment of speech fluency in the L2 classroom.

References

- Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, and C. E. Turner (Eds.), *Language testing reconsidered* (pp. 21-39). Ottawa, ON: University of Ottawa Press.
- Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Arevart, S. & Nation, P. (1991). Fluency improvement in a second language. *RELC Journal*, 22(1), 84-94.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.
- Bachman, L. F. & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Bachman, L. F. (2007). What is the construct: The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language Testing Reconsidered* (pp. 41-72). Ottawa: University of Ottawa Press.
- Baker-Smemoe, W., Dewey, D. P., Bown, J. and Martinsen, R. A. (2014), Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, 47(4), 707–728.
- Bardovi-Harlig, K. (2012). Formulas, routines, and conventional expressions in pragmatics research. *Annual Review of Applied Linguistics*, 32, 296-227.
- Barón, J. & Celaya, M. L. (2010). Developing pragmatic fluency in an EFL context. *EUROSLA*

- Yearbook*. 10(1), 38-61.
- Bavelas, J. B. (2000). Nonverbal aspects of fluency. In H. Riggensbach (Ed.), *Perspectives on fluency* (pp. 91-101). Ann Arbor: University of Michigan Press.
- Bergmann, C., Sprenger, S. A., & Schmid, M.S. (2015). The impact of language co-activation on L1 and L2 speech fluency. *Acta Psychologica*, 161(1), 25-35.
- Blake, C. (2009). Potential of text-based internet chats for improving oral fluency in a second language, *Modern Language Journal*, 93(2), 227-240.
- Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Testing Research*, 19(3), 245-261.
- Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal*. 45(3), 221-235.
- Boersma, P., & Weenink, D. (2013). PRAAT: Doing phonetics by computer (Version 5.3.51). Retrieved from <http://www.praat.org/>
- Bortfield, H. Leone, S.D, Bloom, J. E., Schober, M. F., Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2), 123-147.
- Bosker, H. R., Pinget, A.F., Quené, H., Sanders, T. & De Jong, N. H. (2012). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, 30(2), 159-175.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729.

- British Council (2015). *IELTS Speaking band descriptors*. Retrieved from:
http://takeielts.britishcouncil.org/sites/default/files/Speaking%20Band%20descriptors_0.pdf.
- Brindley, G. (1994). Task-centred assessment in language learning: The promise and the challenge. In N. Bird, P. Falvey, A.B. M. Tsui, D. Allison & A. McNeill (Eds.), *Language and learning* (pp. 73-94). Hong Kong: Institute of Language in Education, Hong Kong Department of Education.
- Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment*. (pp. 98-141). Cambridge: Cambridge University Press.
- Brown, J. (2008). *Principles of language learning and teaching: a course in second language acquisition*. White Plains, NY: Pearson Education.
- Brown, P. & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.
- Browne, K. & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In T. Isaacs & P. Trofimovich (Eds.), *Second Language Pronunciation Assessment: Interdisciplinary Perspectives*. Bristol: Multilingual Matters.
- Brumfit, C. (2005). *Communicative methodology in language teaching*. Cambridge: Cambridge University Press.
- Bui, H.Y.G. (2014). Task readiness: Theoretical framework and empirical evidence from topic familiarity, strategic planning, and proficiency levels. In: P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 63–94). Amsterdam: John Benjamins.
- Bui, G. & Huang, Z. (2018). L2 fluency as influenced by content familiarity and planning:

- Performance, measurement, and pedagogy. *Language Teaching Research*, 22(1), 94-114.
- Cambridge English Language Assessment (CELA). (2015a). *Certificate of Proficiency in English (CPE)*. Retrieved from: <http://www.cambridgeenglish.org/images/168194-cambridge-english-proficiency-teachers-handbook.pdf>
- Cambridge English Language Assessment (CELA). (2015b). *First Certificate in English (FCE)*. Retrieved from: <http://www.cambridgeenglish.org/images/168194-cambridge-english-proficiency-teachers-handbook.pdf>
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 147.
- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), 201-219.
- Carroll, D. (2004). Restarts in novice turn beginnings: disfluencies or interactional achievements? In R. Gardner & J. Wagner (Eds.) *Second language Conversations* (pp. 201-220). Continuum: London.
- Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide. (2nd Ed.)*. Cambridge: Cambridge University Press.
- Chafe, W. L. (1980). Some reasons for hesitating. In H.W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 169-182). The Hague: Moulton.
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20(4), 369-383.
- Chambers, F. (1997). What do we mean by fluency? *System*, 25(4), 535-544.
- Chun, D. M. (2012). *Discourse intonation in L2: From theory and research to practice*.

Amsterdam/Philadelphia: John Benjamins.

- Clahsen, H. (1987). Natural language development: Acquisitional processes leading to fluency in speech production In H. W. Dechert & M. Raupach (Eds.), *Psycholinguistic models of*
- Davies, A. (2007). Assessing academic English language proficiency: 40+ years of U.K. language tests. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language Testing Reconsidered* (pp. 73-86). Ottawa: University of Ottawa Press.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.
- Creswell, J. & Plano Clark, V. (2011). *Designing and conducting mixed-methods research*. (2nd Ed.). Thousand Oaks, CA: SAGE Publications.
- Crible, L. (2017). Discourse markers and (dis)fluencies in English and French: Variation and combination in the DisFrEN corpus. *International Journal of Corpus Linguistics*, 22(2), 242-269.
- Crible L. & Pascaul, E. (2019). Combinations of discourse markers with repairs and repetitions in English, French and Spanish. *Journal of Pragmatics*, 145(2), 1 – 16.
- Cucchiari, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2), 989-999.
- Cutrone, P. (2014). A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners. *Intercultural Pragmatics*, 11(1), 83-120.
- Davidson, F., and Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, USA: Yale University Press.

- De Jong, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385-390.
- De Jong, N. H. & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 61(2), 533-568.
- De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34(1), 893-916.
- DeJong, N. H. (2016a). Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 54(2), 113–132.
- De Jong, N. H. (2016b). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 203–218). Boston/Berlin, Massachusetts/Germany: Mouton de Gruyter.
- De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, 15(3), 237-254.
- Derwing, T. M., Rossiter, M. & Munro, M. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development*. 23(4), 245-259.
- Derwing, T. M., Rossiter, M., Munro, M., & Thomson, R. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 655-679.
- Derwing, T. M., Thomson, R., & Munro, M. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, 34(1), 183-193.

- Derwing, T. M., Munro, M., & Thomson, R.I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, 29(3), 359-380.
- Derwing, T. M., Munro, M., Thomson, R.I., & Rossiter, M. (2009). The relationship between L1 fluency and L2 fluency development. *SSLA*, 31(1), 533-557.
- Derwing, T. M., Munro, M. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *CMLR*, 66(2), 181-202.
- Dewaele, J. M. & Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences*, 28,355-365
- Diepenbrock. L. G. & Derwing, T. M. (2013). To what extent do popular ESL textbooks incorporate oral fluency and pragmatic development? *TESL Canada Journal*, 30(7), 1-20.
- Dornyei, Z. & Kormos, J. (2000). The role of individual and social variables in oral task performance. *Language Testing*, 4(3), 275-300.
- Dornyei, Z. & Csizer, K. (2012). How to design and analyze surveys in second language acquisition research. In A. Mackey & S. Gass. (Eds.) *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 74-94). Hoboken, N.J.: Blackwell Publishing Ltd.
- Doutrich, D. (2000). Cultural fluency, marginality, and the sense of self. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 141-160). Ann Arbor: University of Michigan Press.
- Ducasse, A.M. & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423-443.
- Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal*, 97(1), 131-143.
- Dudley, L. (2007). Connecting L2 learners to the larger community, *Canadian Modern*

- Language Review* 63(4), 539–561.
- Dujim, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing*, 35(4), 501-527.
- East, M. (2012). *Task-based language teaching from the teachers' perspective*. John Benjamins: Amsterdam.
- Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287-314). Ann Arbor: University of Michigan Press.
- Elder, C. & Iwashita, N. (2005). Planning for test performance. Does it make a difference? In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219 - 238). Amsterdam: John Benjamins.
- Ellis, R. (2005). Planning and task-based performance: Theory and research. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 3 - 36). Amsterdam: John Benjamins.
- Ellis, R. & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167 - 192). Amsterdam: John Benjamins.
- Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove: Brooks/Cole.
- Evans, A. N. (2008). *Using basic statistics in the social sciences*. (4th Ed.). Toronto: Pearson Education.

- Faul, F., Erdfelder, E., Buchner, A., Lang, A.G. (2011). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav Res Methods* 41(4), 1149-1460.
- Fiksdal, S. (2000). Fluency as a function of time and rapport. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 128-140). Ann Arbor: University of Michigan Press.
- Fillmore, C. J. (1979). On fluency. In C. Fillmore, D. Kempler, & W.S.Y. Wang (Eds.). *Individual differences in language ability and language behavior* (pp. 85-101). New York: Academic Press.
- Flowerdew, J. & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge Language Education: Cambridge, U.K.
- Forsberg, F. & Fant, L. (2015). Idiomatically speaking – effects of task variation on formulaic language in high proficient users of L2 French and Spanish. In D. Wood (Ed.). *Perspectives on formulaic language in acquisition and communication*. New York: Continuum, p. 47-70.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21, 354–375.
- Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 75 – 94). London/New York: Longman.
- Foster, P. & Skehan, P. (2009). The influence of planning and task type on second language performance. In K. Van den Branden, M. Bygate, & J. M. Norris. (Ed.), *Task-based language teaching: A reader*. (pp. 275-300). John Amsterdam: Netherlands.

- Fox, J. (2004). Test decisions over time: Tracking validity. *Language Testing*, 21(4), 437-465.
- Fox, J. & Hirotnani, M. (2016). Detecting incremental changes in oral proficiency in neuroscience and language testing: Advantages of Interdisciplinary collaboration. In V. Arayadoust & J. Fox, (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 89-120). Cambridge, UK: Cambridge Scholars Press.
- Fox, J., von Randow, J., & Volkov, A. (2016). Identifying students-at-risk through post-entry diagnostic assessment: An Australasian approach takes root in a Canadian university. In V. Arayadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 266–285). Newcastle upon Tyne: Cambridge Scholars Press. On ARES.
- Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6), 709-738.
- Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition*, 29(2), 320-326.
- Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123-48). Amsterdam: John Benjamins.
- Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggensbach (Ed.), *Perspectives on fluency* (pp. 243-265). Ann Arbor: University of Michigan Press.

- Freed, B. F., Segalowitz, N., & Dewey, D. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, 26(1), 275-301.
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13, 208-238.
- Fulcher, G. (2003). *Testing second language speaking*. London: Longman.
- Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.
- Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: the case of the first certificate in English examination. *Language Assessment Quarterly*, 5(2), 89-119.
- Galaczi, E. & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219-236.
- Galante, A. & Thomson, R. I. (2016). The effectiveness of drama as an instructional approach for the development of second language oral fluency, comprehensibility, and accentedness. *TESOL Quarterly*, 50(2), 1-28.
- Gatbonton, E. & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *The Canadian Modern Language Review*, 61(3), 325-353.
- Gilbert, J. B. (2012). *Clear speech: Pronunciation and listening comprehension in North American English*. (3rd Ed.). Cambridge: Cambridge University Press.
- Ginther, A., Dimova, S., and Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379-399.

- Goldman-Eisler, F. (1968). *Psycholinguistics experiments in spontaneous speech*. London: Academic Press.
- Götz, S. (2013). *Fluency in native and non-native English speech*. Amsterdam: John Benjamins.
- Grosjean, F. & Deschamps, A. (1972). Analyse des variables temporelles du français spontané. *Phonetica*, 26, 129-156.
- Grosjean, F. & Deschamps, A. (1975). Analyse contrastive temporelles des l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hesitation. *Phonetica* 31, 144-184.
- Grosjean, F. (1980). Linguistic structures and performance structures: Studies in pause distribution. In H. W. Dechert, D. Mohle, & M. Raupach (Eds.), *Temporal variables in speech* (pp. 91-103). The Hague: Moulton.
- Gullberg, M. (2008). A helping hand? Gestures, L2 learners, and grammar. In S. G. McGafferty & G. Stam (Eds.), *Gesture. Second language acquisition and classroom research* (pp. 185–210). New York: Routledge.
- Guz, E. (2016). Refining the methodology for investigating the relationship between fluency and the use of formulaic language in learner speech. *Research in Language*, 14(2), 95-122.
- Hasselgren, A. (2002). Learner corpora and language testing: Smallwords as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 143-173). Amsterdam: John Benjamins.
- Hinofotis, F. (1983). Cloze as an alternative method of ESL placement and proficiency testing. In J. Oller & K Perkins (Eds.), *Language Testing* (pp. 121- 128). Rowley, MA: Newbury House.

- He, A.W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A.W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam, Netherlands: John Benjamins
- Hieke, A. (2005). Linking as a marker of fluent speech. *Language and speech*, 27(4), 343-354.
- Hincks, R. (2010). Speaking rate and information content in English lingua franca oral Presentations. *English for Specific Purposes*, 29(1), 4-18.
- House, J. (1996). Developing pragmatic fluency in English as a foreign language: Routines and metapragmatic awareness. *SSLA*, 18, 225-252.
- House, J. (2013). Developing pragmatic competence in English as a lingua franca: Using discourse markers to express (inter)subjectivity and connectivity. *Journal of Pragmatics*, 59, 57-67.
- Huang, B. H. (2013). The effect of accent familiarity and language teaching experience on raters' judgements of non-native speech. *System*, 41, 770-785.
- Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly*, 13(1), 25-41.
- Huang, B. H., & Jun, S.-A. (2015). Age matters, and so may raters: Rater differences in the assessment of foreign accents. *Studies in Second Language Acquisition*, 37(4), 623-650.
- Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135-159.
- Isaacs, T. (2013). International engineering students' interactional patterns on a paired-speaking test: Interlocutors' perspectives. In K. McDonough & A. Mackey (Eds.). *Second*

- language education in diverse educational contexts*. (pp. 227-246). Amsterdam: John Benjamins.
- Iwashita, N., Brown, A., McNamara, T., & O' Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24-49.
- Jacewicz, E., Fox, R.A., & Wei, L. (2000). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America* 128(2), 839-850.
- Jenkins, J. (2007). *English as a lingua franca: Attitude and identity*. Oxford, UK: Oxford University Press.
- Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal*, 87(1), 90–107.
- Jiang, X., Sanford, R., & Pell, M.D. (2018). Neural architecture underlying person perception from in-group and out-group voices. *Neuroimage*, 18(1), 582-597.
- Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, 64(4), 809-854.
- Kahng, J. (2018). The effect of pause location on perceived fluency, *Applied Psycholinguistics*, 39, 569-591.
- Kang, S. J. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System*, 33(2), 277 - 292.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
- Kennedy, S. & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness in L2

- speech: The role of listener experience and semantic context. *The Canadian Modern Language Review*, 64(3), 459-489.
- Koyama, D., Sun, A., & Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language, Learning & Technology*, 20(1), 148.
- Kachru, B. B. & Smith, L. E. (1988). World Englishes: an integrative and cross-cultural journal of WE-ness. In R. Maxwell (Ed.): *40 years' service to science, technology and education*, 674–8. Oxford: Pergamon Press.
- Kraut, R. & Wulff, S. (2013) Foreign-accented speech perception ratings: a multifactorial case study. *Journal of Multilingual and Multicultural Development*, 34(3), 249-263.
- Koponen, M. & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5-24). Ann Arbor: University of Michigan Press.
- Kormos, J. & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145-164.
- Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Erlbaum.
- Kowal, S., Wiese, R., & O'Connell, D. C. (1983). The use of time in storytelling. *Language and Speech*, 26(4), 377–392.
- Kramersch, C. J. (1986). From language proficiency to interactional competence. *Modern Language Journal*, 70(4), 366-372.
- Lakoff, R. (1975). *Language and Woman's Place*. New York: Harper Colophon.
- Larson-Hall, J. (2013). *A Guide to Doing Statistics in Second Language Research Using SPSS and R*. (2nd Ed.). London: Longman.
- Lam, D. M. K. (2018). What counts as “responding”? Contingency on previous speaker

- contribution as a feature of interactional competence. *Language Testing*, 35(3), 377-401.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387-417.
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Ann Arbor: University of Michigan Press.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levesque, C., Zuehlke, N. A., Stanek, L.R., Deci, R.M. (2004). Autonomy and competence in German and American university students: A comparative study based on self-determination theory. *Journal of Educational Psychology*, 96(1), 68-84.
- Lin, P. (2018). *The prosody of formulaic sequences: A corpus and discourse approach*. Bloomsbury Academic: London.
- Lintunen, P., Peltonen, P., & Webb, J. (2015). Tone units as indicators of L2 fluency development: evidence from native and learner English. In J. A. Mompean & J. Fouz-González (Eds.). *Investigating English Pronunciation: Trends and Directions*. (pp. 196-218). Basingstoke: Palgrave Macmillan.
- Logan, G. (1998). Toward an instance theory of automatization. *Psychological Review*, 95(1), 492-527.
- Lounsbury, F. G. (1954). Transitional probability, linguistic structure, and systems of habit-family hierarchies. In C.E. Osgood & T.A. Sebok (Eds.), *Psycholinguistics: a survey of theory and research problems* (pp. 93-101). Bloomington: Indiana University Press.
- Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.
- MacIntyre, P. D. (1994). Variables underlying willingness to communicate: A causal analysis. *Communication Research Reports*, 11(2), 135-142.

- MacIntyre, P. D., Dörnyei, Z., Clément, R., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, 82(4), 545-562.
- Magne, V., Suzuki, S., Suzukida, Y., Ilkan, M., Tran, M., & Saito, K. (2019). Exploring the dynamic nature of second language listeners' perceived fluency: A Mixed-Methods approach. *TESOL Quarterly*, 53(4), 1139-1150.
- May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, 8(2), 127-145.
- McCafferty, S. G., & Stam, G. (2008). *Gesture: Second language acquisition and classroom research*. New York: Routledge.
- McCarthy, M. (2006). Fluency and confluence: What fluent speakers do. In M. McCarthy (Ed.) *Explorations of Corpus Linguistics*. Cambridge: Cambridge University Press.
- McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, 1(1), 1-15.
- McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, 11(2), 113-128.
- Mehnert U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20. 52-83.
- Michel, M., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL*, 45(3), 241-259.
- Mohle, D. (2005). A comparison of the second language speech production of different native speakers. In H.W. Dechert, D. Mohle, & M. Raupach (Eds.), *Second language productions* (pp. 26-49). Tubingen, Germany: Gunter Narr Verlag.
- Morales-López, E. (2000). Fluency levels and the organization of conversation in non-native

- Spanish' speech. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 266-286). Ann Arbor: University of Michigan Press.
- Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97.
- Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, 23(1), 451-468.
- Nakatsuhara, F. (2006). The impact of proficiency-level on conversational styles in paired Speaking tests. *Research Notes 25*, University of Cambridge ESOL Examinations, 15–20. Retrieved from www.cambridgeenglish.org/images/23144-research-notes-25.pdf
- Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, 28(4), 483–508.
- Nation, I.S.P. (1989). Improving speaking fluency. *System*, 17(3), 377-384.
- Nation, I.S.P. (2013). *Learning vocabulary in another language*. (2nd Ed.). Cambridge: Cambridge Applied Linguistics.
- Nattinger, J. R. & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
- Nematizadeh, S. & Wood, D. (2019). Willingness to communicate and second language speech fluency: An investigation of affective and cognitive dynamics. *CMLR*, 75(3), 197-215.
- Nematizadeh, S. (2019). *Willingness to communicate and second language speech fluency: A complex dynamic systems perspective*. (Unpublished doctoral dissertation). Carleton University, Ottawa, Canada.

- North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.
- O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, 29(1), 557-582.
- O'Brien, M.G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, 64(4), 715-748.
- Ockey (2011). Self-consciousness and assertiveness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning*, 61(3), 968-989.
- Ockey, G. J. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. *English for Specific Purposes*, 35, 17-29. doi:10.1016/j.esp.2013.11.003
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277-295.
- Pallant, J. (2007). *SPSS survival manual—A step by step guide to data analysis using SPSS for windows* (3rd ed.). Maidenhead: Open University Press.
- Paragon Testing Enterprises (2015). *Oral Language Sample Test*. Retrieved from: https://www.cael.ca/wcontent/uploads/2015/10/OLT_practice_test_October_2002.pdf
- Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York: Longman.
- Pawley, A. & Syder, F. H. (2000). The one clause-at-a-time hypothesis. In H. Riggensbach (Ed.),

- Perspectives on fluency* (pp. 163-199). Ann Arbor: University of Michigan Press.
- Pawlikowska-Smith, G. (2002). *Canadian language benchmarks 2000: English as a second language – for adults*. Minister of Public Works and Government Services Canada.
- Pellegrino, F., Coup'è, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 87, 539–558.
- Pickering, L. (2006). Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics*, 26, 219-233.
- Pike, K. L (1972). General characteristics of intonation. In D. Bolinger (Ed.) *Intonation*. London: Penguin Books.
- Pizziconi, B. (2009). Stereotypes of communicative styles: Japanese indirectness, ambiguity, and vagueness. In R. Gómez Morón, M. Padilla Cruz, Lucía Fernández Amaya, & M. O. Hernández López (Eds). *Pragmatics applied to language teaching and learning*. Cambridge: Cambridge Scholars Publishing.
- Plough, I. Banerjee, J. & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, 35(3), 427-445.
- Poyatos, F. (2005). Linguistic fluency and non-verbal cultural fluency. In A. Wolfgang (Ed.) *Nonverbal behavior: Perspectives, applications, intercultural insights* (pp. 431-460). Liangton, NY: C.J. Hogrefe.
- Poyatos, F. (1987). *Cross-cultural perspectives in nonverbal communication*. Toronto: Hogrefe.
- Peltonen, P. & Lintunen, P. (2016). Integrating quantitative and qualitative approaches in L2 fluency analysis: A study of Finnish-speaking and Swedish-speaking learners of English at two school levels. *European Journal of Applied Linguistics*, 4, 209-238.
- Peltonen, P. (2017a) Temporal fluency and problem-solving in interaction: An exploratory study

- of fluency resources in L2 dialogue. *System*, 70, 1-13.
- Peltonen, P. (2017b). L2 fluency in spoken interaction: A case study on the use of other-repetitions and collaborative completions. In M. Kuronen, P. Lintunen, & T. Nieminen (Eds.) *Insights into Second Language Speech*. (pp. 118-138). Jyväskylä: The Finnish Association of Applied Linguistics.
- Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *The Modern Language Journal*, 102(4), 676-692.
- Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *The Canadian Modern Language Review*, 69(3), 324-348.
- Préfontaine, Y. & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *Modern Language Journal*, 99(1), 96-112.
- Préfontaine, Y. Kormos, J. & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, 33(1), 53-73.
- Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of Acoustic Society of America*, 123(1), 1104-1113.
- Raupach, M. (1980). Temporal variables in first and second language speech production. In H. W. Dechert, D. Mohle, & M. Raupach (Eds.), *Temporal variables in speech* (pp. 263-270). The Hague: Moulton.
- Raupach, M. (2005). Formulae in second language speech production. In H.W. Dechert, D. Mohle, & M. Raupach (Eds.), *Second language productions* (pp. 114-137). Tübingen: Gunter Narr Verlag.

- Révész, A., Ekiert, M., & Torgersent, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 37(6), 828-848.
- Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, 14(4), 423-441.
- Riggenbach, H. (1998). Evaluating Learner Interactional Skills: Conversation at the Macro Level. In R. Young & A. Weiyun He (eds.): *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency. Studies in Bilingualism, vol. 14*. Amsterdam: John Benjamins Publishing.
- Riggenbach, H. (2001). Sample analysis: Hesitation phenomena in second language fluency. In A. Wennerstrom (Ed). *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.
- Roach, P. (1998). *English phonetics and phonology: A practical course*. Cambridge, U.K. Cambridge University Press.
- Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: a triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction*. (pp. 287-318). Cambridge: Cambridge University Press.
- Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, 65(3), 395-412.
- Rossiter, M. J., Derwing, T. M., Manimtim, L. G. Thomson, R. I. (2010). Oral fluency: The neglected component in the communicative language classroom. *The Canadian Modern Language Review*, 66(4), 583-606.
- Rowe, B. M. & Levine, D. P. (2015). *A course introduction to linguistics*. Routledge: London.

- Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics*, 39(3), 593-617.
- Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38(4), 439-462
- Sajavaara, K. (1988). Cross-linguistic and cross-cultural intelligibility. In P.H. Lowenberg (Ed.), *Language spread and language policy: Issues, implications, and case studies*, (pp. 250-265). Washington: Georgetown University Press.
- Saldaña, J. (2009). *The coding manual for qualitative researchers*. Los Angeles: CA: Sage.
- Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System*, 45,(2) 79-91.
- Sato, T. (2013). The influential features on linguistic laypersons' evaluative judgments of second language oral proficiency. *JLTA Journal*, 16, 1-14.
- Schelgoff, E. A. (2007). *Sequence organization in interaction: A primer in conversational analysis..* Cambridge: Cambridge University Press.
- Schmidt, R. (1992). Psycholinguistic mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, 14(1), 357-385.
- Schneider, W. & Schrifin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review*, 84(1), 1-66.
- Segalowitz, N. (2000). Automaticity and attention skill in fluent performance. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 200-219). Ann Arbor: University of Michigan Press.

- Segalowitz, N. & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition, 26*(2), 173-199.
- Segalowitz, N. (2007). Access fluidity, attention control, and the acquisition of fluency in a second language. *TESOL Quarterly, 41*(1), 181-186.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. London/New York: Routledge.
- Shea, C. & Leonard, K. (2019). Evaluating measures of pausing for second language fluency research. *CMLR, 75*(3), 216-235.
- Shimada, K., Hirotsu, M., Yokokawa, H., Yoshida, H., Makita, K., Yamazaki-Murase, M., Tanabe, H.C. & Sadato, N. (2015). Fluency-dependent cortical activation association with speech production and comprehension in second language learners. *Neuroscience, 300*, 474-492.
- Shintani, N., Saito, K., & Koizumi, R. (2018). The relationship between multilingual raters' language background and their perceptions of accentedness and comprehensibility of second language speech. *International Journal of Bilingual Education and Bilingualism, 1* – 18.
- Skehan, P. & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193 - 218). Amsterdam: John Benjamins.

- Skehan, P. (2009). A framework for the implementation of task-based instruction. In K. Van den Branden, M. Bygate, & J. M. Norris. (Ed.), *Task-based language teaching: A reader*. (pp. 83-108). John Amsterdam: Netherlands.
- Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 54(2), 97-111.
- Spolsky, B. (1990). Oral examinations: an historical note. *Language Testing*, 7(2), 158-173.
- Suzuki, S., & Kormos, J. (2019). linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, , 1-25.
- Tannen, D. (2005). *Conversational style: Analyzing talk among friends*. Oxford: Oxford University Press.
- Tannen, D. (1986). *That's not what I meant! How conversational style makes or breaks relationships*. New York: HarperCollins.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). Amsterdam: John Benjamins.
- Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, 65, 71–79.
- Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 54(2), 133–150.

- Tavakoli, P., Campbell, C. & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, 50(2), 447-471.
- Tavakoli, P. & Hunter, A.M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, 22(3), 330-349.
- Taylor, G. & Wigglesworth, A. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing*, 26(3), 325-329.
- Teddlie C. & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage.
- Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics*, 4(2), 91-112.
- Thomson, H. (2017). Building speaking fluency with multiword expressions. *TESL Canada Journal*, 34 SI(3), 26.
- Thomson, R. I. (2015). Fluency. In M. Reed & J. M. Lewis (Eds.) *The Handbook of English Pronunciation* (pp. 209-226). Hoboken, NJ: Wiley.
- Thomson, R. I. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang & A. Ginther (Eds). *Assessment in second language pronunciation*. Routledge: London.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84-119.
- Upshur, C. E. & Turner, J.A. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49(1), 3-12.

- Van Dessel, P., Mertens, G., Smith, C. T., De Houwer, J., (2017). *Experimental Psychology*, 64(1), 299-314.
- Vogt, P. W. (2000). *Quantitative research methods for professionals*. Boston: Pearson Education.
- Wainer, H. & Thissen, D. (1994). On examinee choice in educational testing. *Language Testing*, 64(1), 159-195.
- Waniek-Klimczak, E. (2014). Selected observations on the effect of rhythm on proficiency, accuracy, and fluency in non-native English speech. In W. Szubko-Stiraek, L. Salski, & P. Stalmaszyk. (Eds). *Language Learning, Discourse and Communication. Second Language Learning and Teaching*, (pp. 167-181). Springer: Cham.
- Watanabe, M., Keikichi H., Yasuharu D., Nobuaki, M. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, 50(1), 81–94.
- Weir, C. J. (2005). *Language Testing and Validation: An evidence-based approach*. Houndgrave, UK: Palgrave-Macmillan.
- Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102-127). Ann Arbor: University of Michigan Press.
- Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press: Oxford.
- Wennerstrom, A. & Siegel, A.F. (2003) Keeping the floor in multiparty conversations: intonation, syntax, and pause. *Discourse Processes*, 36(2), 77-107.
- Wigglesworth, G. (2000). Issues in the development of oral tasks for competency-based assessments of second language performance. In G. Brindley (ed.), *Studies in Immigrant*

- English Language Assessment* (Vol. 1) [Research Series 11]. National Centre for English Language Teaching and Research Macquarie University: Sydney.
- Williams, K. (2018). A mixed-methods study exploring perceptions of speech fluency. *CONTACT*, 44(2) 35-40.
- Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252.
- Wolf, J. P. (2008). The effects of backchannels on fluency in L2 oral task production. *System*, 36(1), 24-39.
- Wood, D. (2006). Uses and functions of formulaic sequences in second-language speech: An exploration of the foundations of fluency. *CMLR*, 63(1), 13-33.
- Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *The Canadian Journal of Applied Linguistics*, 12(1), 39-57.
- Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence, and classroom applications*. London/New York: Continuum.
- Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. London: Bloomsbury.
- Wood, D. (2016). Willingness to communicate and second language speech fluency: An idiodynamic investigation. *System* 60(2), 11-28.
- Wray, A. & Perkins. M. (2000). The functions of formulaic language: an integrated model. *Language & Communication*, 20, 1 – 28.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.
- Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test.

- Language Learning*, 61(4), 1222–1255.
- Yeh, C. J. & Inose, M. (2003). International students' reported English fluency, social support satisfaction, and social connectedness as predictors of acculturative stress. *Counselling Psychology Quarterly*, 16(1), 15-28.
- Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14,403-424.
- Young, R. (1995). Conversational styles in language proficiency interviews, *Language Learning*, 45(1), 3–42.
- Young, R. & Hallebeck, G. B. (1998). 'Let them eat cake!' or how to avoid losing your head in cross-cultural conversations. In R. Young, & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 355-382). Amsterdam: John Benjamins.
- Young, R. & He, A. (1998). Language proficiency interviews: a discourse approach. In R. Young, & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins.
- Yuan, F. & Ellis, R. (2003). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167-192). Amsterdam: John Benjamins.
- Yule. G. (1997). *Pragmatics*. Oxford: Oxford University Press.
- Zhang, B., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher ratings: Competing or complementary constructs? *Language Testing*, 28(1), 31–50.
- Ziegler, N., Seals, C., Ammons, S., Lake, J., Hamrick, P., & Rebuschat, P. (2013). Interaction in

conversation groups: The development of L2 conversational styles. In K. McDonough & A. Mackey (Eds.). *Second language education in diverse educational contexts*. (pp. 269-292). Amsterdam: John Benjamins.

Appendix A: Carleton University REB Clearance

CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

The Carleton University Research Ethics Board-A (CUREB-A) has granted ethics clearance for the research project described below and research may now proceed. CUREB-A is constituted and operates in compliance with the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2).

Ethics Protocol Clearance ID: Project # 107627

Project Team Members: Kent Williams (Primary Investigator)

David Wood (Research Supervisor)

Project Title: Developing and piloting a rating scale for assessing English as a second language (ESL) speech fluency on a paired conversational task [Kent Williams]

Funding Source (If applicable):

Effective: **October 04, 2017**

Expires: **October 31, 2018.**

Restrictions:

This certification is subject to the following conditions:

1. Clearance is granted only for the research and purposes described in the application.
2. Any modification to the approved research must be submitted to CUREB-A via a Change to Protocol Form. All changes must be cleared prior to the continuance of the research.
3. An Annual Status Report for the renewal of ethics clearance must be submitted and cleared by the renewal date listed above. Failure to submit the Annual Status Report will result in the closure of the file. If funding is associated, funds will be frozen.
4. A closure request must be sent to CUREB-A when the research is complete or terminated.
5. Should any participant suffer adversely from their participation in the project you are required to report the matter to CUREB-A.

Failure to conduct the research in accordance with the principles of the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans 2nd edition* and the *Carleton University Policies and Procedures for the Ethical Conduct of Research* may result in the suspension or termination of the research project.

Upon reasonable request, it is the policy of CUREB, for cleared protocols, to release the name of the PI, the title of the project, and the date of clearance and any renewal(s).

Please contact the Research Compliance Coordinators, at ethics@carleton.ca, if you have any questions or require a clearance certificate with a signature.

CLEARED BY:

Date: October 04, 2017

Andy Adler, PhD, Chair, CUREB-A

Bernadette Campbell, PhD, Vice-Chair, CUREB-A

Appendix B: Renison University College at University of Waterloo REB Clearance

OFFICE OF RESEARCH ETHICS

Notification of Ethics Clearance of Application to Conduct Research with Human Participants

Faculty Supervisor: David Wood **Department:** Carleton University; Applied Linguistics and Discourse Studies
Student Investigator: Kent Williams **Department:** Renison College
Collaborator: Julia Williams **Department:** Renison College

ORE File #: 22499

Project Title: Developing and piloting a rating scale for assessing English as a second language (ESL) speech fluency on a paired-interactive task _copy

Human Research Ethics Committee (HREC) Clinical Research Ethics Committee (CREC) is pleased to inform you the above named study has been reviewed and given ethics clearance.

Approval to start this research is effective on the ethics clearance date which is: 11/13/2017 (m/d/y)

University of Waterloo Research Ethics Committees are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS, 2nd edition), International Conference on Harmonization: Good Clinical Practice (ICH-GCP), the Ontario Personal Health Information Protection Act (PHIPA), the applicable laws and regulations of the province of Ontario. Both Committees are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREC) and IRB00007409 (CREC).

The above named study is to be conducted in accordance with the submitted application (Form 101/101A) and the most recent approved versions of all supporting materials.

Ethics clearance for this study is valid until: 11/13/2018 (m/d/y). Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified by the Research Ethics Committee (Form 105). Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports by the expiry date will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review:

- Delegated review
 Full committee review meeting date: _____ (m/d/y)

Signed on behalf of: HREC Chair HREC Vice-Chair CREC Chair CREC Vice-Chair

- Julie Joza, Acting Chief Ethics Officer, jajoza@uwaterloo.ca, ext. 38535
 Heather Root, Senior Manager, heather.root@uwaterloo.ca, ext. 30469
 Karen Pieters, Manager, kpieters@uwaterloo.ca, ext. 30495
 Joanna Eidse, Research Ethics Advisor, jeidse@uwaterloo.ca, ext. 37163
 Laura Strathdee, Research Ethics Advisor, lstrathd@uwaterloo.ca, ext. 30321
 Erin Van Der Meulen, Research Ethics Advisor, ervandermeulen@uwaterloo.ca, ext. 37046

This is an official document. Retain for your files.

You are responsible for obtaining any additional institutional approvals that might be required to complete this study.