Judging speaking fluency:
What do raters do, and what should they do?

Nivja de Jong

Utrecht University

IATEFL TEASIG 2011, September 17th, Innsbruck

PEARSON

Overview

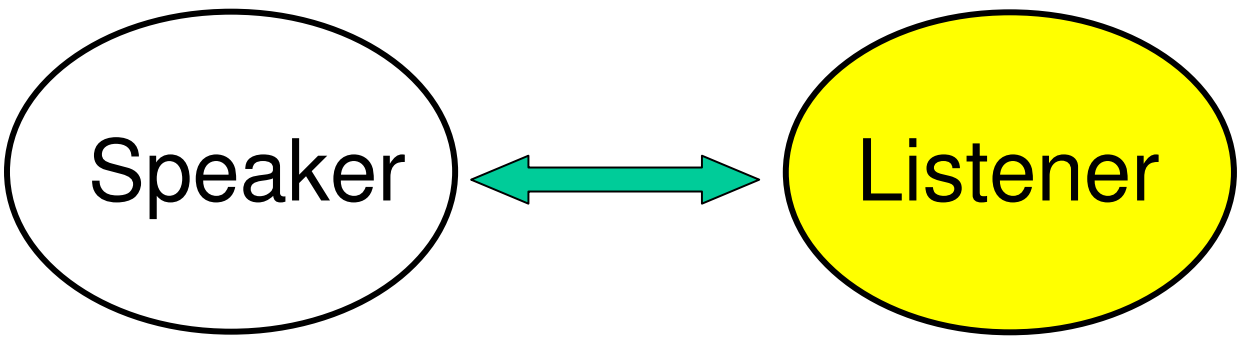
- Definition & viewpoints
- Measures of L1 and L2 fluency
- What should raters take into account when judging speech on fluency?
- What do raters take into account?
- Conclusion & Discussion

Definition of speaking fluency

Lennon (1990, p. 391):

Fluency “is an impression on the listener’s part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently”.

Viewpoints on fluency



What makes L2 speech *sound* more or less fluent?

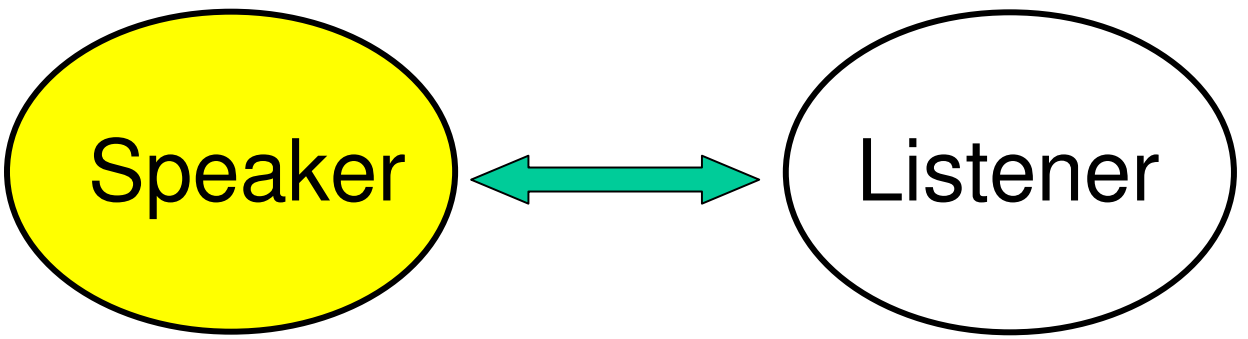
Speech without pauses, filled pauses, repairs

(e.g., Kormos & Dénes, 2004; Riggenbach, 1991, Rossiter, 2009)

Alternative definition of speaking fluency

Fluency is the extent to which the psycholinguistic processes of speech planning and speech production *are* functioning easily and efficiently.

Viewpoints on fluency



What makes L2 speech *be* more or less fluent?

1. Individual characteristics
2. Conceptual planning
3. Formulating & articulating in L2

Aspects of fluency

- **Cognitive fluency**: ability of the L2 speaker to smoothly translate thoughts to L2 speech
- **Utterance fluency**: objective measures of an utterance
- **Perceived fluency**: subjective measure of what listeners perceive

Segalowitz (2010). *Cognitive bases of second language fluency.* New York: Routledge.

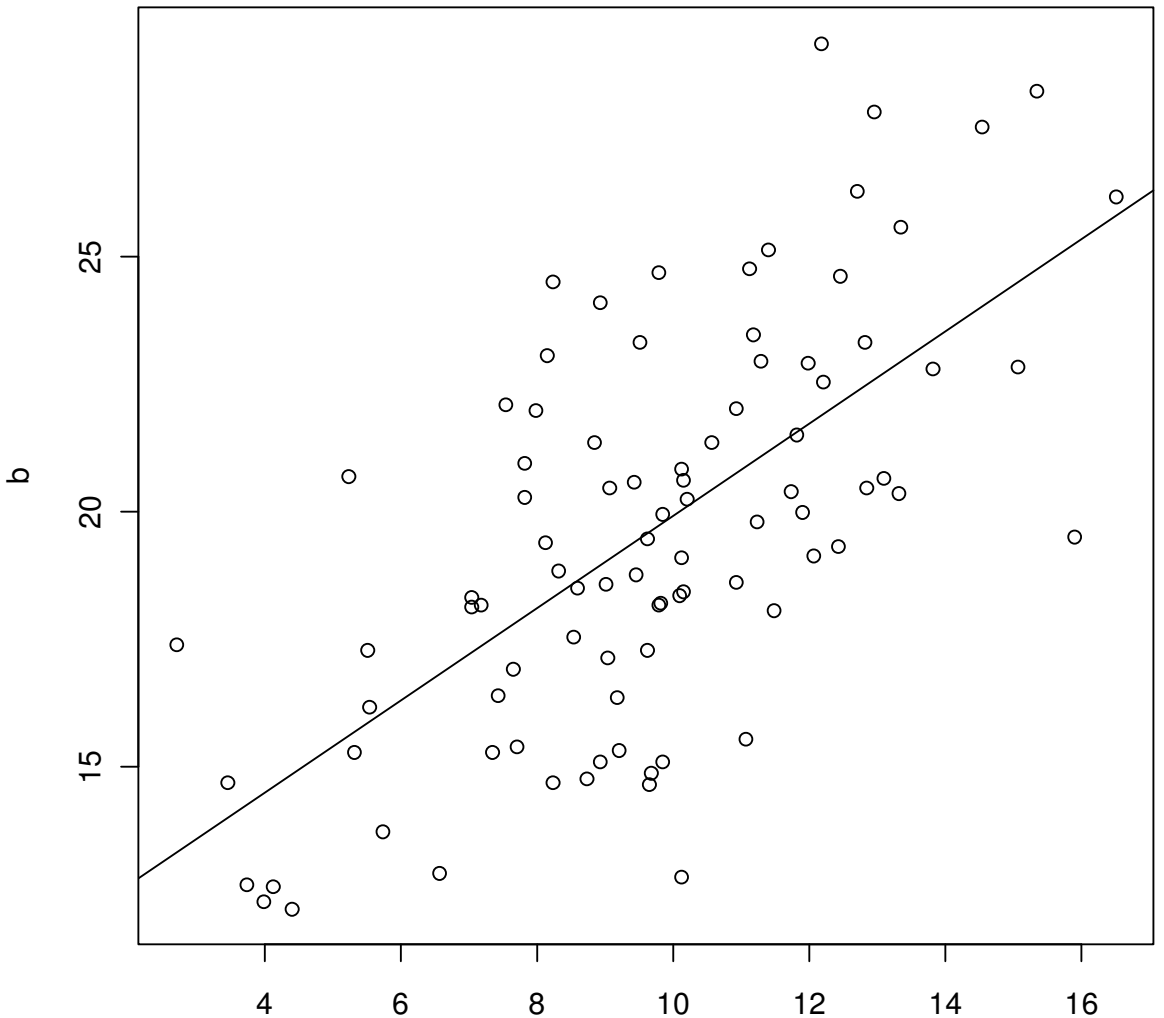
Segalowitz (2010)

- Calculate residuals to measure L2-specific measures of utterance fluency:

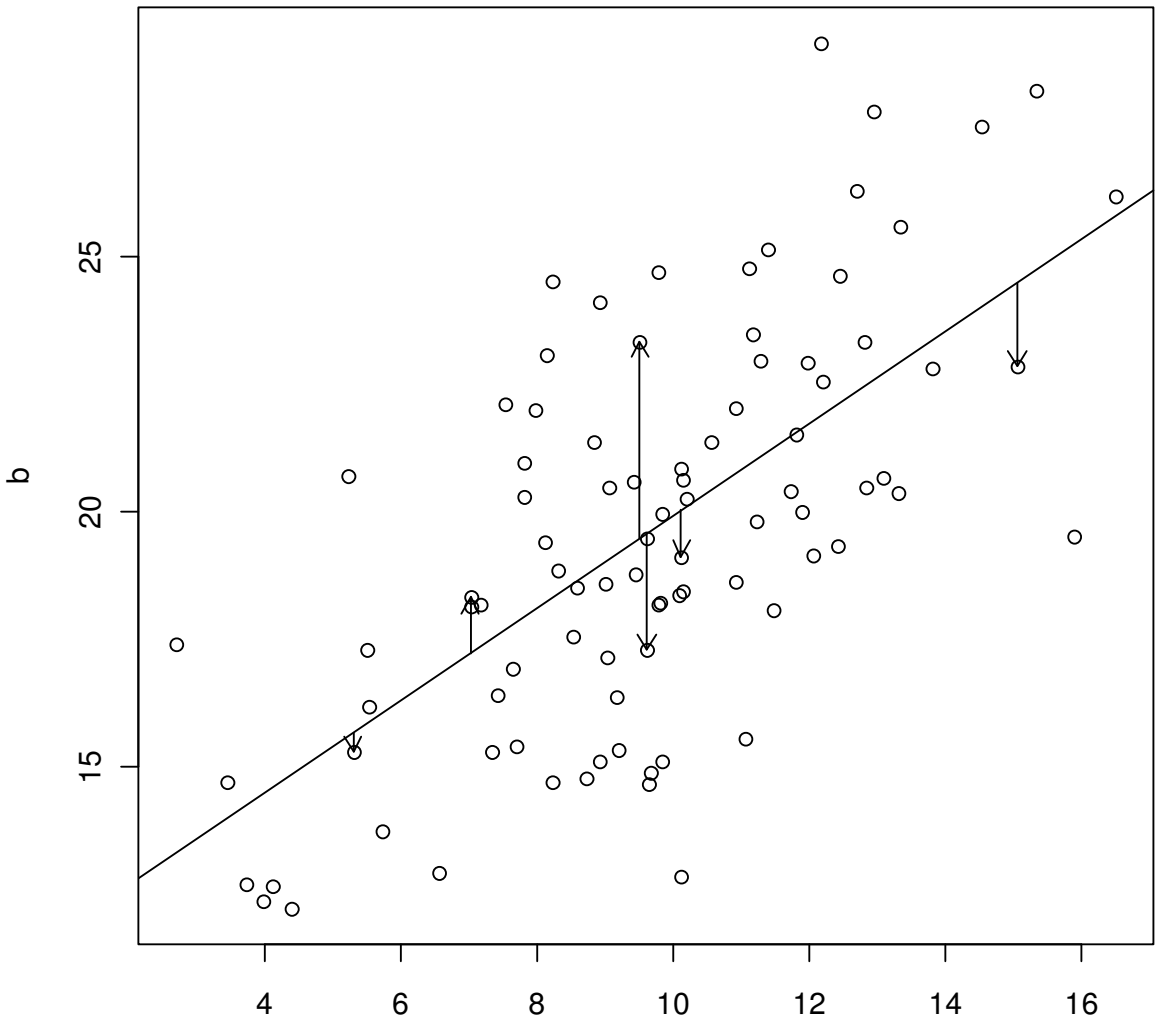
  "partial out sources of variability that are not related specifically to the disfluencies in L2 but that characterize a person's general performance in the given testing conditions" (p. 40)

- Gather both L1 and L2 data
- Regress L2 measure on L1 measure
- Save residuals

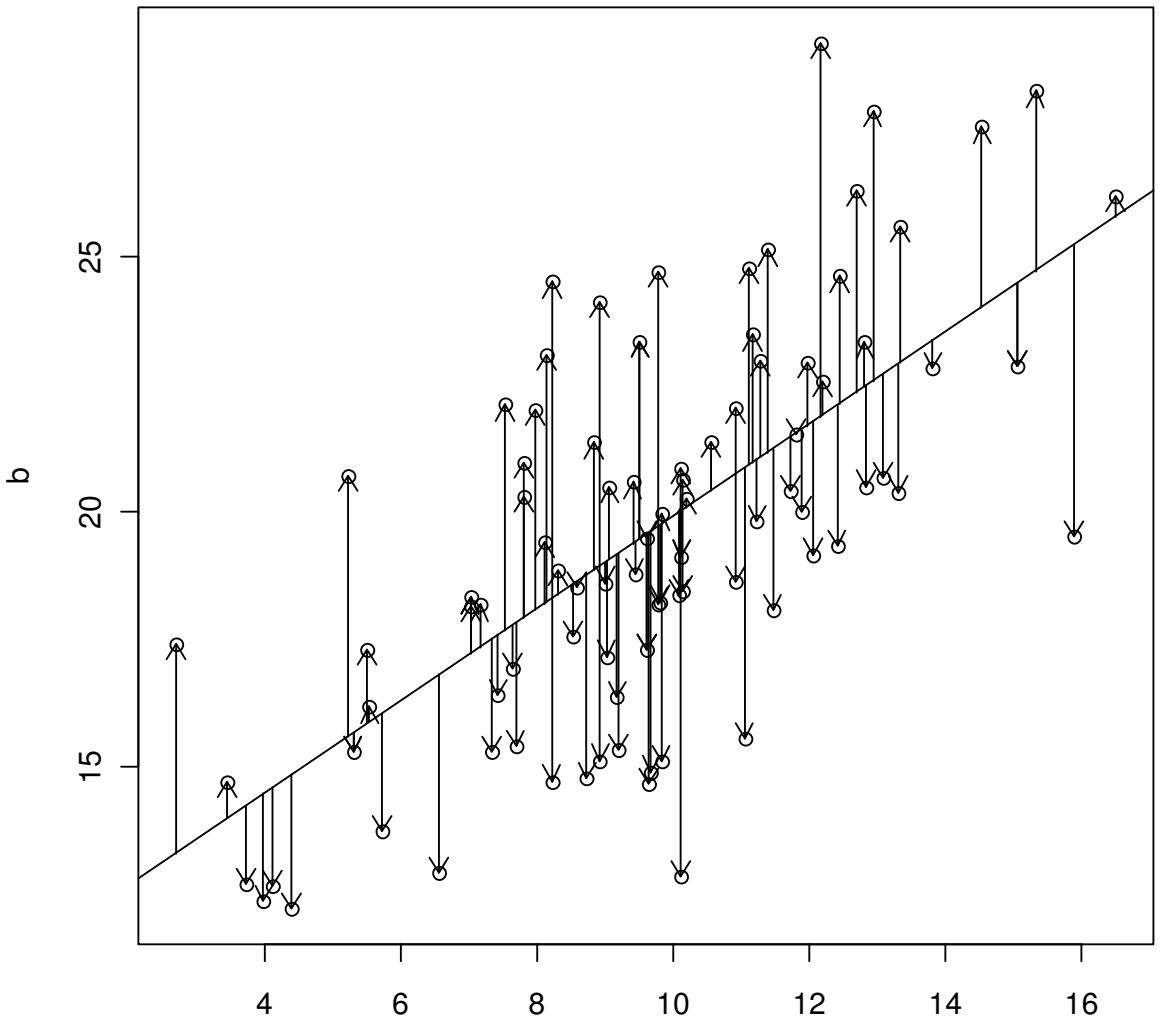
Segalowitz' proposal



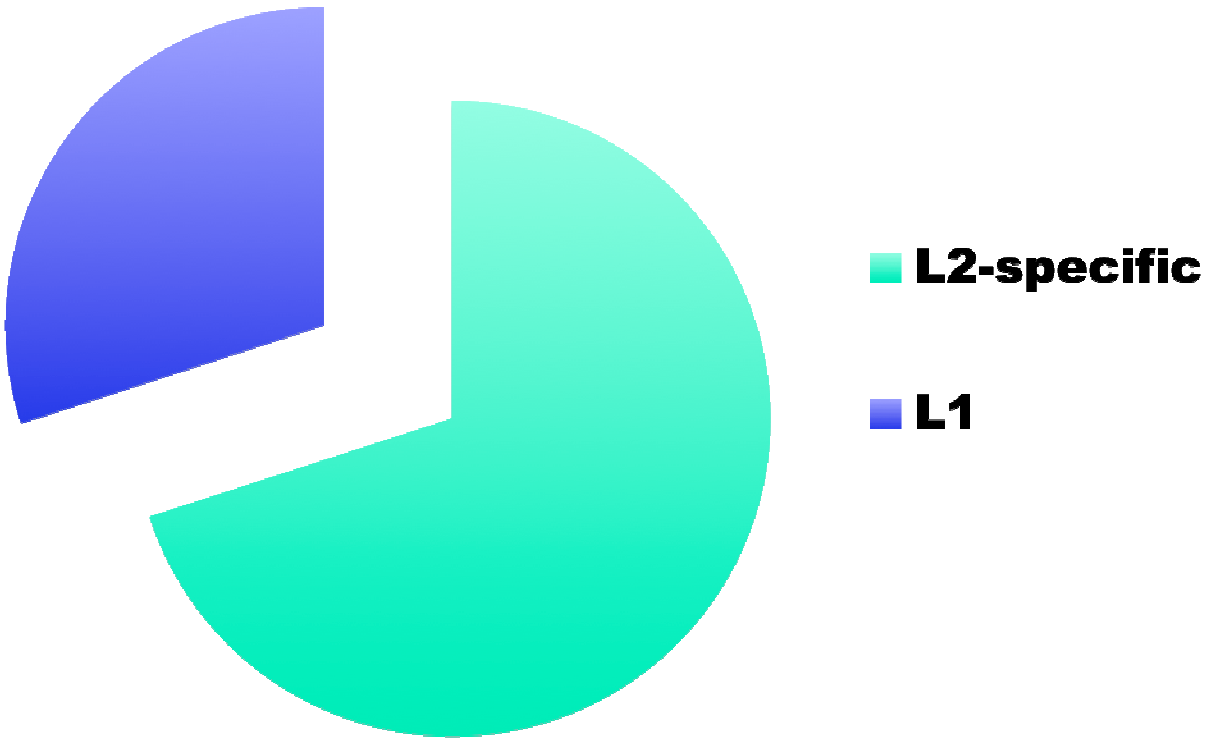
Segalowitz' proposal



Segalowitz' proposal



L2 measure of fluency



Measures of L2 utterance fluency

Fluency has a multifaceted nature (Tavakoli & Skehan, 2005):

- Breakdown fluency (e.g., time filled with speech, no. of pauses, filled pauses)
- Speed fluency (e.g., speech rate measured as words per minute, syllables per minute)
- Repair fluency (e.g. false starts, repetitions)

Research questions

- Are residualized scores better predictors of L2 proficiency than original L2 measures?

  ⇨ What should raters take into account?

- Are residualized scores better predictors of ratings on fluency than original L2 measures?

  ⇨ What do raters take into account?

Method RQ 1

- English and Turkish native speakers with Dutch as L2, intermediate – advanced (n = 53)
- 8 speaking tasks in L2 (Dutch)
- 8 similar speaking tasks in L1 (English/Turkish)
- To measure overall L2-proficiency: productive vocabulary knowledge task

  In previous research (with N = 181): *r* = .79 with overall speaking proficiency.

  (De Jong et al., to appear in SSLA)

An example of a speaking task

Presentation Window

File

Hieronder ziet u wat u een maand geleden op straat hebt gezien.

De rechter vraagt u om precies te beschrijven wat u heeft gezien.

- Begin uw woorden met "Geachte rechter..."
- Vertel dat u het ongeluk heeft gezien
- Beschrijf in detail wat er gebeurd is

Bereid nu voor wat u wilt gaan zeggen

1

2

3

4

spreekenglish

These pictures show the scene you just saw.
The officer asks for your objective account of what happened.

- Address the police officer as 'officer'

- Explain that you saw what happened

- Give the police officer your account of the incident


1


2


3

Please prepare what you are going to say.


4


5

Measures in L1 and in L2

- Breakdown fluency:
  - Mean length of pause between utterances (> 250ms)
  - Mean length of pause within utterances (> 250ms)
  - Number of silent pauses (> 250ms; per second)
  - Number of filled pauses (per second)
- Speed fluency:
  - Syllable duration (i.e. inversed articulation rate)
- Repair fluency:
  - Number of corrections (per second)
  - Number of repetitions (per second)

Vocabulary knowledge

- Fill in the gap format (116 items):
  - 90 content-words, 9 from each 1000-frequency band (Corpus of Spoken Dutch)
  - 26 multi-word units (prepositional phrases and verb-noun collocations)

Analyses

1. Descriptives (L1 / L2; Turkish / English)
2. Predict L2 fluency measures from L1 measures, save residuals
3. Predict L2 vocabulary knowledge from
    - L2 original fluency measures
    - L2 residualized scores

Descriptives: L2 Vocabulary knowledge

Turkish	English
55.1(24.7)	56.7 (22.0)

Descriptives: breakdown fluency

Measure	L1 English speakers	L2 (Dutch) English speakers	L1 Turkish speakers	L2 (Dutch) Turkish speakers
Pause within (ms)	552 (110)	711 (205)	635 (132)	739 (158)
Pause between (ms)	650 (162)	820 (276)	677 (168)	893 (238)
# silent pauses per second	0.30 (0.06)	0.34 (0.05)	0.24 (0.06)	0.34 (0.06)
# filled pauses per second	0.15 (0.09)	0.20 (0.13)	0.19 (0.09)	0.24 (0.12)

Descriptives: breakdown fluency

Measure	L1 English speakers	L2 (Dutch) English speakers	L1 Turkish speakers	L2 (Dutch) Turkish speakers
Pause within (ms)	552 (110)	711 (205)	635 (132)	739 (158)
Pause between (ms)	650 (162)	820 (276)	677 (168)	893 (238)
# silent pauses per second	0.30 (0.06)	0.34 (0.05)	0.24 (0.06)	0.34 (0.06)
# filled pauses per second	0.15 (0.09)	0.20 (0.13)	0.19 (0.09)	0.24 (0.12)

Descriptives: speed fluency

Measure	L1 English speakers	L2 (Dutch) English speakers	L1 Turkish speakers	L2 (Dutch) Turkish speakers
Syllable duration (ms)	215 (25)	286 (68)	189 (26)	294 (49)

Descriptives: repair fluency

Measure	L1 English speakers	L2 (Dutch) English speakers	L1 Turkish speakers	L2 (Dutch) Turkish speakers
# repetitions per second	0.05 (0.04)	0.04 (0.04)	0.01 (0.01)	0.04 (0.03)
# repairs per second	0.03 (0.02)	0.04 (0.02)	0.04 (0.01)	0.05 (0.02)

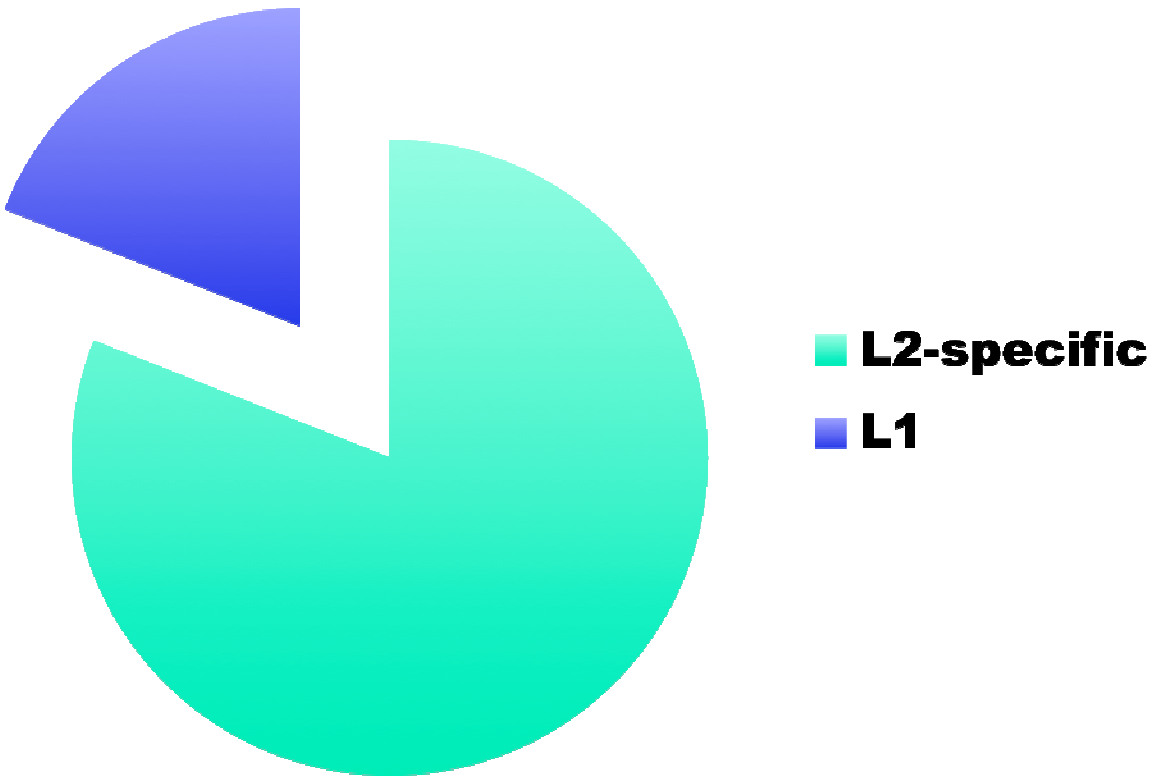
Predicting L2 measures from L1 measures

	Intercept	Slope	Adjusted intercept for Turkish	Total R^2 (%)
Syllable duration	-0.30 (0.19)	0.48 (0.15)	0.64 (0.29)	18.8
Pause within	0.00 (0.11)	0.65 (0.11)		42.2
Pause between	0.00 (0.09)	0.76 (0.09)		57.3
# silent pauses	0.00 (0.12)	0.58 (0.15)		33.1
# filled pauses	0.00 (0.11)	0.73 (0.10)		52.2
# repetitions	-0.30 (0.16)	0.79 (0.14)	0.63 (0.26)	43.4
# corrections	0.00 (0.11)	0.65 (0.11)		42.8

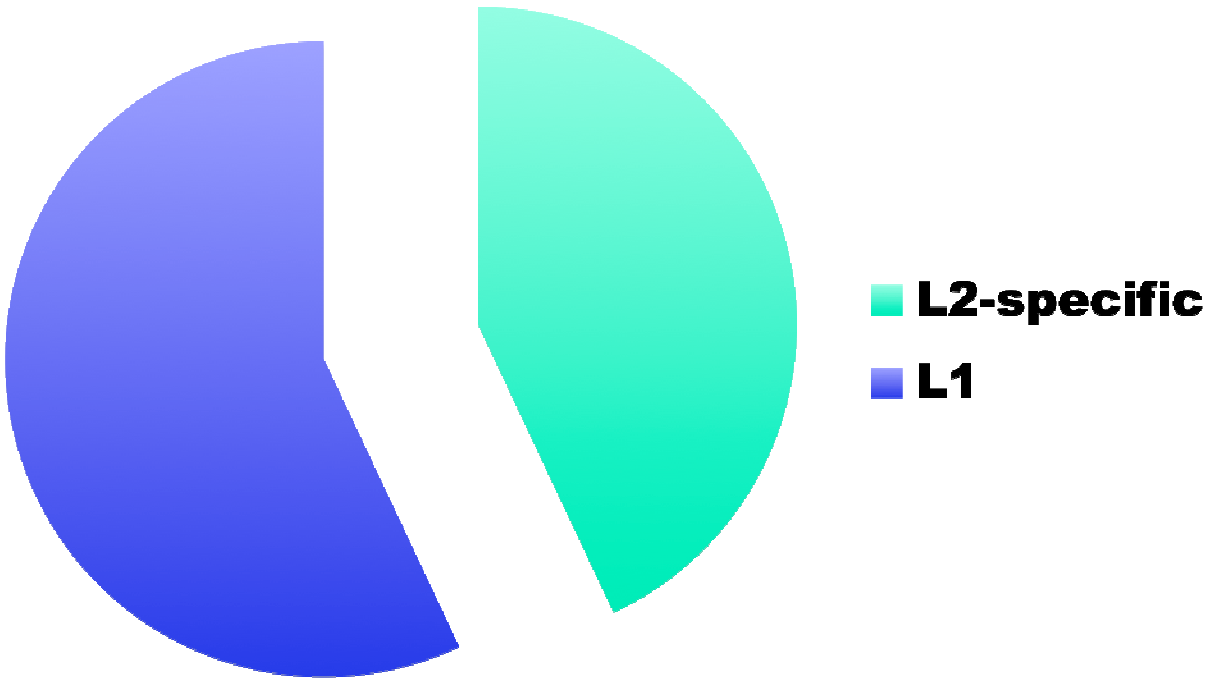
Predicting L2 measures from L1 measures

	Intercept	Slope	Adjusted intercept for Turkish	Total R^2 (%)
Syllable duration	-0.30 (0.19)	0.48 (0.15)	0.64 (0.29)	18.8
Pause within	0.00 (0.11)	0.65 (0.11)		42.2
Pause between	0.00 (0.09)	0.76 (0.09)		57.3
# silent pauses	0.00 (0.12)	0.58 (0.15)		33.1
# filled pauses	0.00 (0.11)	0.73 (0.10)		52.2
# repetitions	-0.30 (0.16)	0.79 (0.14)	0.63 (0.26)	43.4
# corrections	0.00 (0.11)	0.65 (0.11)		42.8

L2 Syllable duration



L2 Mean length of pauses between utterances



Predicting vocabulary scores from L2 original measures and from residuals

	L2 original R^2 (%)	L2 residuals R^2 (%)
Syllable duration	28.2	39.4
Pause within	1.3	3.3
Pause between	1.6	0.0
# silent pauses	17.9	14.9
# filled pauses	5.8	7.1
# repetitions	10.8	9.4
# corrections	7.1	7.3

Predicting vocabulary scores from L2 original measures and from residuals

	L2 original R^2 (%)		L2 residuals R^2 (%)
Syllable duration	28.2	⟺	39.4
Pause within	1.3		3.3
Pause between	1.6		0.0
# silent pauses	17.9		14.9
# filled pauses	5.8		7.1
# repetitions	10.8		9.4
# corrections	7.1		7.3

Conclusions RQ1

- L2 original scores are in most cases just as well related to a measure of L2 proficiency as L2 residuals
- For the speed measure syllable duration, residuals are better predictors of L2 proficiency than the original scores
- For all other measures: L2-specific fluency explains L2 proficiency; adding information about L1 behavior does not lead to worse predictions about L2 proficiency

Conclusions RQ1 (2)

- Pause durations (original score and residualized score) are hardly related to a measure of overall L2 proficiency
- *Replication of earlier finding* with N = 179, different L1's, and more measures related to L2 proficiency

  (De Jong et al., *to appear in Applied Psycholinguistics*)
- What should raters do?
  - Judge L2-specific syllable duration
  - **Not** judge duration of pauses

RQ2: Are residualized scores better predictors of ratings on fluency than original L2 measures?

Method RQ2

- Speech materials:

  90 L2 items: 3 L2 Dutch task performances of 15 English and 15 Turkish native speakers (20 sec excerpts)

  24 L1 Dutch items: 3 L1 task performances of 8 Dutch native speakers (20 sec excerpts)

- Rating task:

  20 native Dutch listeners rating on fluency

  (9-point scale)

Analyses RQ2

Re-calculate residuals for these 20-second excerpts.

Predict L2 ratings on fluency from

- L2 original fluency measures
- L2 residualized scores

Predicting ratings from L2 original measures and from residuals

Predicting ratings by:	L2 original R^2 (%)
Syllable duration	50.6
Pause duration within + between	21.8
# silent pauses	36.9
# filled pauses	10.6
# repetitions	15.4
# corrections	12.8
All measures	77.7

Predicting ratings from L2 original measures and from residuals

Predicting ratings by:	L2 original R^2 (%)	L2 residuals R^2 (%)
Syllable duration	50.6	49.5
Pause duration within + between	21.8	10.6
# silent pauses	36.9	37.3
# filled pauses	10.6	6.1
# repetitions	15.4	16.1
# corrections	12.8	8.4
All measures	77.7	66.0

Conclusion RQ2

- L2-specific fluency explains most of the variance of L2 ratings on fluency; adding information about *L1 pause duration*, leads to better predictions about L2 ratings
- What do raters do?
  - Mostly judge L2-specific syllable duration √
  - Judge duration of silent pauses X

Overall conclusion

- Ratings on fluency are related to pause duration, whereas overall proficiency is not related to pause duration

- Ratings on fluency are indeed related to temporal measures of fluency: 78% variance explained (just as they were instructed)

Discussion

Should we instruct raters

- To pay attention to speed of speech mostly
- To ignore duration of silent pauses

Should we use measures of utterance fluency, instead of or in addition to subjective judgments?

Should we sample L1-speech, to measure L2-specific utterance fluency?

Do we want to measure perceived fluency, or do we want to measure (L2-specific) cognitive fluency?

Acknowledgements

Utrecht Fluency team:
Hans Rutger Bosker
Anne-France Pinget
Hugo Quené

WiSP team (UvA):
Jan Hulstijn
Rob Schoonen
Margarita Steinel
Arjen Florijn

Research assistants:
Rachel Hanson
Veysel Yüce
Iske Bakker
Cem Keskin
Erica Bouma

Funding:
NWO (Dutch Research Council)
Pearson Language Tests

Questions?

n.dejong@uu.nl