

**'Matched guise technique' vs. 'Open guise technique' in the elicitation of language attitudes: Insights from a comparative study**

This is a sociolinguistic methods conference, so I believe we are all here pretty familiar with the matched-guise technique in speaker evaluation, as introduced in Lambert et al. (1960). A quick recap: essentially, you play voices to informants and have them provide some evaluation of these voices. But here's now the crucial aspect of this technique that my paper will be dealing with today, in Lambert et al.'s own words:

“[Informants] were not told that they were going to hear some of the voices twice, but rather that they would hear 10 recorded male voices, all reading the same passage, five in French and five in English. [...] There was no indication that any [informant] became aware of the fact that bilingual speakers were used.” Lambert et al. (1960: 44)

So this is what you do in the 'classic' format of the matched-guise technique: you present speaker recordings to informants, and you act as if these were all produced by different people, even though some of them really are produced by the *same* person using different linguistic varieties.

Now, for the past couple of years I have been thinking a lot about this - what I call the 'disguising' aspect of the matched guise technique. This is because much of my research has been dealing with the rhetorical effects of style-shifting, in the context of Austrian German, and I have been using speaker evaluation studies to inform this research, to elicit the social meanings people associate with standard Austrian German on the one hand, and dialectal Austrian German on the other. Now, when you look at rhetorical style-shifting in interaction, what happens a lot there is actually that one and the same person uses different linguistic varieties within one flow of talk. And this is, of course, obvious to all participants present. And in order to recreate this situation with a speaker evaluation experiment, to find out how informants will judge different linguistic varieties in juxtaposition when they *know* the speaker is the same, as they do in interaction, I have used what I call the 'open guise' technique. Under the open-guise, I tell my informants very openly that they will hear the same speakers in different versions of talk and I ask them to give me their impressions accordingly. And this does work: even if I *tell* informants that it's really the *same* person they will hear in two different versions of presentation, they will still produce some rating differences. This is fascinating in and of itself, and I will come back to this in more detail in a little bit, but first I

want to establish my main focus today, which is sort of at one level up from this. Because, what hasn't been *empirically* explored yet, to my knowledge, is what role exactly the disguising ploy in classic matched guise studies might actually play. In other words, does it make any difference, and what type of difference, in the ratings in speaker evaluation experiments whether I tell my informants that they will hear the same person twice, or whether I *pretend* that they hear all *different* speakers.

This is precisely what I have tested for my present paper. So what I'll be doing is I'll be comparing the results from a classic matched-guise technique speaker evaluation experiment with one that uses the open guise.

I'll actually start by describing the set-up of the open guise study, which I did first, about a year ago.

For this study, my informants were 49 Austrian university students, 90% females, because the experiment was done in a humanities course. The informants' age range was between 18 and 30 years; they are all Austrian-born and raised, with at least one Austrian parent; all indicated Austrian German as their native language.

My speakers, whose recordings I presented to the informants, were two females and one male from the Middle Bavarian-Austrian dialect region. For each, I used two different versions of the same text on genetically engineered food – one in Standard Austrian German, one in Middle Bavarian-Austrian dialect. (I had some good reasons for using this text, which I describe in Soukup 2009, in case you are interested.)

So I gave my informants a questionnaire with five-point bipolar semantic differential scales using 22 fairly classic personality items, and I asked them to rate each recording of a speaker on these scales.

I made it very clear that the informants would hear each speaker twice, although I did not explicitly mention the difference in *language variety* used. I told the informants to consider: "How do the speakers come across to a public audience in each of the two ways of presenting the text?"

Now for the matched guise experiment. I conducted it the next semester, last fall, in the same lecture course with the new cohort of students, so the student population was essentially the same. Here, my informants were 37 Austrian university students, 81% female, same age, same sociolinguistic background as before.

My speakers were seven different females from the Middle-Bavarian-Austrian dialect region, speaking either in dialect or standard. Six of them were actually my filler voices, and one was my test speaker. As per the classic matched guise design, I used this speaker twice,

once in each variety; and I actually used the exact same recordings as in the open guise earlier.

The speakers presented the same text as before, and I used the exact same semantic differential scales, this time setting up the task with “How do the speakers come across to a public audience in their individual way of presenting the text?”

So here is how the two experiments were matched up very closely: same type of informants, same overall design, two recordings were also the exact same, and these are the ones I will use in my comparative analysis later. But I did not only use the same speaker recordings in both experiments; I also tried to ensure that, at least for the first time the informants heard my ‘test speaker’ in the matched guise, her standard recording was preceded by the exact same female dialect recording (by another speaker) as in the open guise. So I was hoping to control also for the preceding stimulus environment as far as possible. For the same reason, when the second, *dialectal* recording of my test speaker came round in the matched guise, it was preceded by a *standard* recording of a filler voice. For the other filler voices used in the matched-guise, I used a pseudo-randomized order that mixed up standard and dialect as much as possible to throw off the informants’ expectations at least a little bit.

As you can see, I had four filler voices between the recordings of the same test speaker in the matched-guise study. This seems fairly standard or even actually a lot – many experiments such as Lamberts’ original seem to have deemed it sufficient to have only three or even two filler voices in between the same speakers.

Also for the matched guise, I actually told the informants beforehand that they would hear *nine* speakers total, and provided nine rating pages in the questionnaire. The last speaker did not really exist, but I didn’t want to have my test speaker mentally come ‘last’ for my informants, also similar to the open guise, where I’d played a male speaker at the end, following my test speaker.

So instead of playing a ninth speaker, however, I actually showed my informants in the matched guise study the following text on a Powerpoint slide:

“Two of the 'speakers' you just heard were really one and the same person. Does this surprise you now? Please indicate : YES // NO”

And I also asked

“Which 'speakers' do you believe were really the same person?”

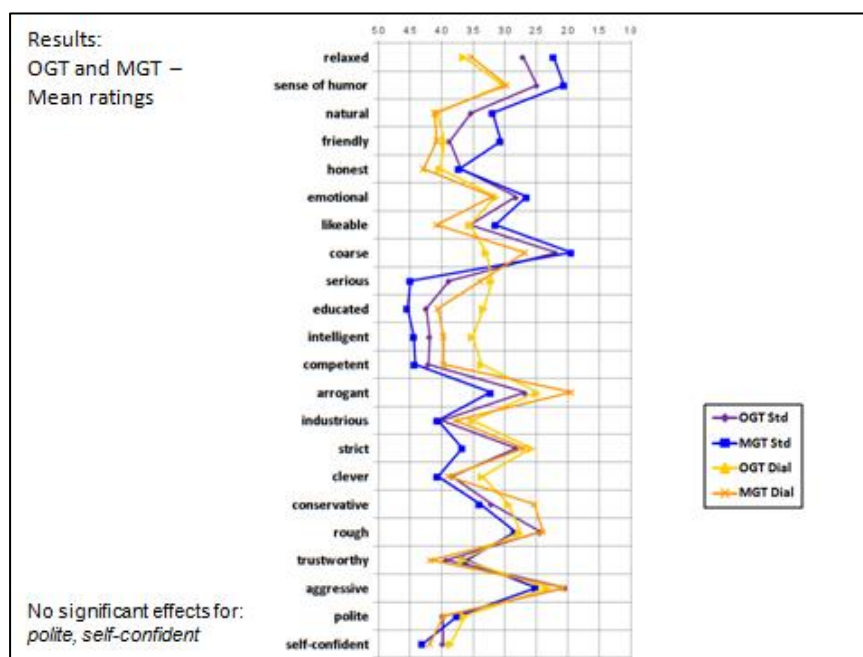
This is how I tried to check whether the matched-guise had actually worked or no – at least as far as the students would admit or could recall. Of course, such recall is a difficult task if sprung on you after eight recordings, but I was hoping to get at least some ballpark idea of

what was going on. As it turned out, 65% of the informants answered that they were indeed surprised to find that they had heard one person twice. And, only six informants indicated the correct pairing of recordings – interestingly, five of those who guessed correctly had actually just before indicated that they were surprised by the fact.

Okay but in general, the disguising ploy in the matched-guise technique seems to have worked for most informants. But one lesson to take away from this is that we cannot *ever* expect it to work 100% - especially, I presume, if fewer filler voices are used than the four I used here.

But at any rate, this now takes us back to the question of what role it might play in an experimental outcome to *know* that the speaker in the two guises *is* one and the same. So, what are the differences in results between my open guise and my matched-guise experiment? Again, I will of course only look at the results for my test speaker, the one I used in both experiments with the same recordings in Austrian Standard and dialect.

The statistical tests I used in my analysis are 2x2 mixed ANOVAs with post hoc *t* tests. My variables are, on the one hand, the type of language variety used (standard vs. dialect), and on the other hand, the type of experiment – matched guise vs. open guise. The cut-off level I assumed for statistical significance is .05. I paid particular attention to effect sizes in my analysis, using eta squared as effect size measure for the ANOVA results and Cohen's *d* for the post-hoc *t* tests; and in the following I will mainly draw on results for which there is at least a medium effect.

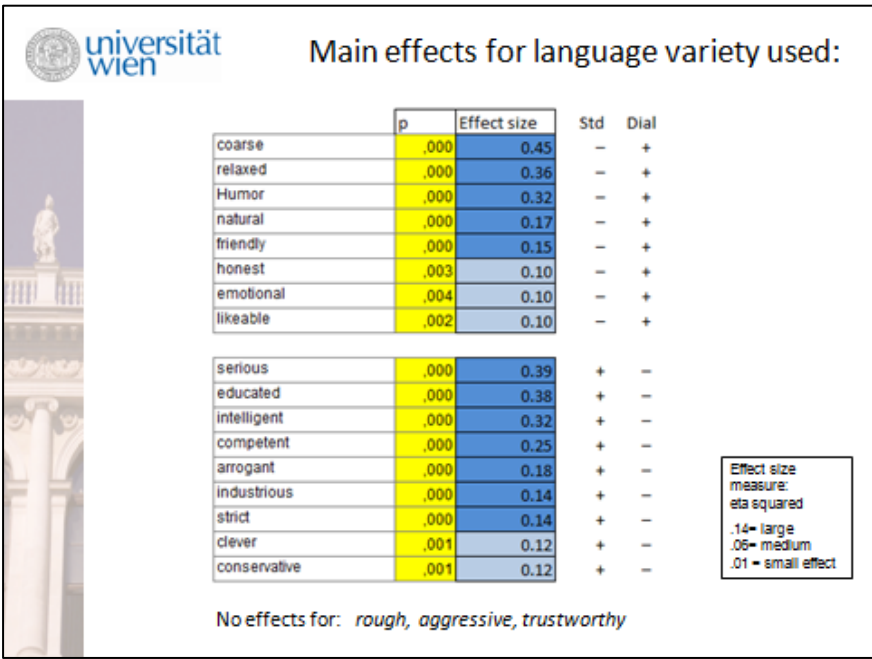


Okay so in this overall graph, which is based on the mean ratings, we can already see some patterns where the two outcomes run in clear parallel at least for some items. In this graph, the

means for the standard recordings are in dark colors and the dialect ones in lighter colors. For a lot of items, the general order is the same for both experiments, with the dialect recordings being rated either lower or higher than the standard. But there are also some crossovers and overlaps that suggest some differences between the types of experiment.

So what's the picture in terms of statistically significant results from the ANOVAs. First off, I'll actually exclude from discussion two items on which there were no statistical effects on any count, namely *polite* and *self-confident*. So these didn't turn out to be meaningful in my experiments.

For all the other items, let's first look at any main effects for the variable 'language variety used'.



**Main effects for language variety used:**

	p	Effect size	Std	Dial
coarse	.000	0.45	-	+
relaxed	.000	0.36	-	+
Humor	.000	0.32	-	+
natural	.000	0.17	-	+
friendly	.000	0.15	-	+
honest	.003	0.10	-	+
emotional	.004	0.10	-	+
likeable	.002	0.10	-	+
serious	.000	0.39	+	-
educated	.000	0.38	+	-
intelligent	.000	0.32	+	-
competent	.000	0.25	+	-
arrogant	.000	0.18	+	-
industrious	.000	0.14	+	-
strict	.000	0.14	+	-
clever	.001	0.12	+	-
conservative	.001	0.12	+	-


Effect size measure: eta squared  
 .14 = large  
 .06 = medium  
 .01 = small effect

No effects for: *rough, aggressive, trustworthy*

Here, you can see that overall, dialect is rated as sounding significantly more *coarse*, but also more *relaxed*, *humorous*, *natural*, *friendly*, *honest*, *emotional* and *likeable* than the standard, in order of effect size. (And we are getting some pretty large main effects here in fact.) At the same time, the standard was held to sound significantly more *serious*, *educated*, *intelligent*, *competent*, *industrious*, and *clever*, but also more *arrogant*, *strict* and *conservative*. Interestingly, there were no main effects for *rough* and *aggressive* nor for *trustworthy*.

So for these ratings, the experimental findings seem to line up overall. And these main effects yield a rather 'classic' and perhaps expected picture of evaluation of a standard vs. a non-standard variety. The dialect loses out mainly on what Zahn & Hopper (1985) in their meta-study have termed items of 'superiority' (like *competence* and *intelligence*), but the dialect wins on items of social attractiveness like sounding *natural* and *friendly*.

However, when we now look at the results in terms of main effects of the variable ‘type of experiment’, and for interaction effects of the two variables ‘type of language variety’ and ‘type of experiment’, then for some of these items the results also show some clear differences in patterning. Here are again the ANOVA results:



OGT vs. MGT

Main effects for type of experiment			Interaction effects		
	p	Effect size		p	Effect size
educated	.000	0.17	rough	.001	0.12
coarse	.001	0.13	trustworthy	.002	0.11
serious	.002	0.11	aggressive	.002	0.11
clever	.004	0.10	arrogant	.001	0.10
competent	.008	0.08	likeable	.002	0.10
strict	.010	0.08	friendly	.001	0.10
intelligent	.016	0.07	strict	.027	0.05
friendly	.036	0.05	educated	.024	0.04

Effect size measure: eta squared  
.14= large // .06= medium // .01 = small effect

As you can see, there was a main effect for type of experiment for the items *educated*, *coarse*, *serious*, *clever*, *competent*, *strict*, *intelligent*, and *friendly*. And there was an interaction effect of the two variables ‘type of experiment’ and ‘type of language variety used’ for *rough*, *trustworthy*, *aggressive*, *arrogant*, *likeable*, *friendly*, *strict*, and *educated*.

This doesn’t as yet tell us any reasons for the discrepancies, so I conducted a series of post-hoc *t* tests to find out where the differences in the outcome between the two kinds of experiments might lie. Here is now a summary and overview of the insights generated through this post hoc analysis.

In total, we have here seven items for which I found a difference in ratings in the matched-guise study, but *not* in the open guise. These are:

*friendly*, *likeable*, *trustworthy*, *strict*, *arrogant*, *rough*, and *aggressive*.

In reverse, there was a difference in the *open guise* but not in the *matched-guise* for the item *clever*.

At the same time, there are items on which *both* types of experiment found a *main effect* of speaker’s language variety used, but the effect was *stronger* in one of the experiments. So, in both experiments the standard came out as sounding less *relaxed*, showing a lower *sense of humor*, and sounding more *serious*, but the effect was stronger in the matched-guise.

Similarly, in both experiments the dialect was rated as sounding more *coarse* and less *competent, intelligent* and *educated* than the standard, but this effect was stronger in the open guise.

Finally, for the items *natural, industrious, conservative, honest, emotional*, there were no significant interaction- or between-study effects.

Now, what is really fascinating to me, and I hope to you also, is the fact that once we rearrange a bit those items that showed an effect according to the outcome I just mentioned, a clear pattern seems to emerge. This pattern becomes clearer when we apply to my list of items Zahn & Hopper's meta-classification of language attitudinal scales.

What Zahn & Hopper (1985) did is to subject thirty semantic differential items as typically used in speaker evaluations to factor analysis, and they were able to provide evidence for a three-factor model solution for these items, with the following factors:

Superiority [Highest loading items: 'literate – illiterate', 'educated – uneducated', 'upper class – lower class'] – attractiveness [Highest loading items: 'sweet – sour', 'nice – awful', 'good natured – hostile'] – dynamism [Highest loading items: 'active – passive', 'talkative – shy', 'aggressive – unaggressive']

If we now apply this model to my own results, here's the picture that emerges.

OGT		MGT		Post-hoc analysis	Zahn & Hopper (1985)
relaxed	Std low	Std lower	0.5	both: St- (MGT effect stronger within; between: MGT Std lower)	Attractiveness (Dynamism)
humor	Std low	Std lower	0.6	both: St- (MGT effect stronger within; between: MGT Std lower)	Attractiveness
friendly		Std lower	0.9	MGT St-, OGT no diff; between: MGT Std much lower	Attractiveness
likeable		Dial higher	-0.5	MGT D+, OGT no diff; between: MGT Dial higher	Attractiveness
trustworthy		Dial higher	-0.5	MGT D+, OGT no diff; between: MGT Dial higher	Attractiveness
serious	Std high	Std higher	-0.8	both: St+ (MGT effect stronger within); between: MGT Std much higher	Attractiveness
strict		Std higher	-0.8	MGT St+, OGT no diff; between: MGT Std much higher	Attractiveness
arrogant		Std higher	-0.5	MGT St+, OGT no diff; between: MGT Std higher;	Attractiveness
		Dial lower	0.5	MGT Dial lower	
rough		Std higher	-0.5	MGT St+, OGT no diff; between: MGT Std higher	Attractiveness
aggressive		Std higher	-0.5	MGT St+, OGT no diff; between: MGT Std higher	Attractiveness (Dynamism)
sophisticated	Dial lower	Dial low		both: D- (OGT effect stronger); between: OGT Dial much lower	Superiority
competent	Dial lower	Dial low		both: D- (OGT effect stronger); between: OGT Dial lower	Superiority
clever	Dial lower	Dial low		OGT D-, MGT no diff; between: OGT Dial lower	Superiority
intelligent	Dial lower	Dial low		both: D- (OGT effect stronger); between: OGT Dial lower	Superiority
educated	Std lower	Dial low		both: D- (OGT effect stronger); between: OGT Std lower, OGT Dial much lower	Superiority
	Dial lower				

Effect size measure: Cohen's d  
.02 = small / .05 = medium / .08 = large effect size

I have here lined up items for which the effect is stronger in the matched guise or for which the open guise showed no significant effect at all, and those for which the reverse is the case.

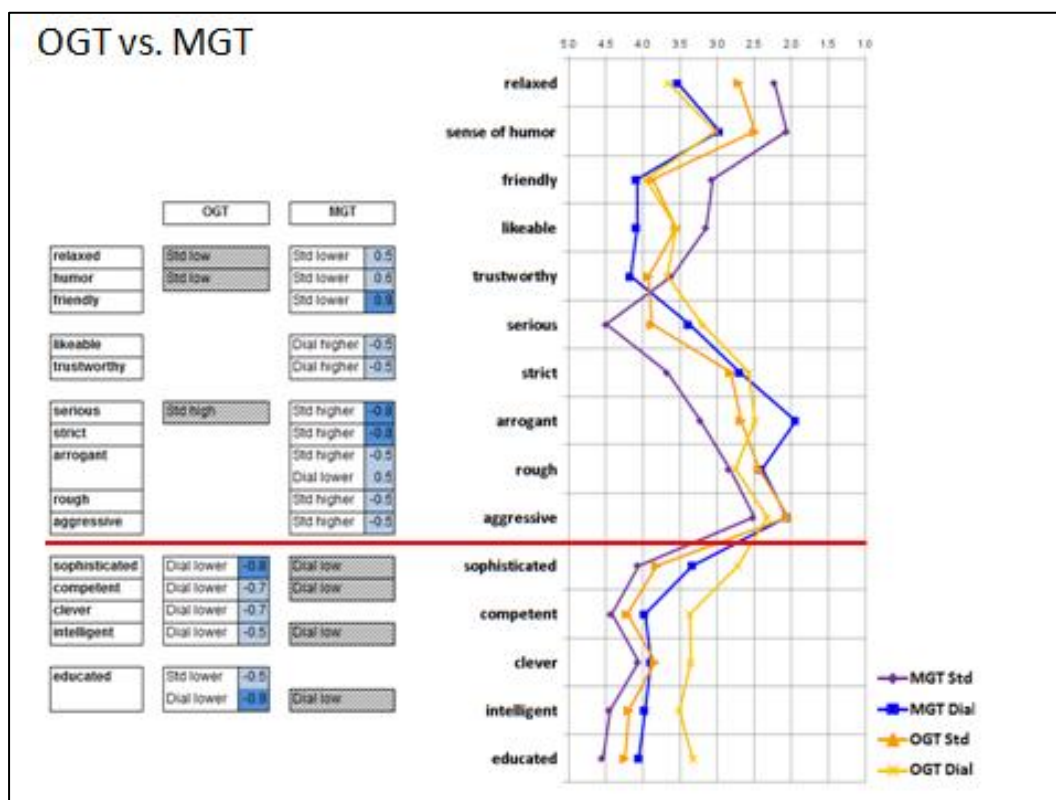


The effect sizes shown are the Cohen's  $d$ 's from post-hoc  $t$  tests using experiment-type as independent variable.

Here in the final column is Zahn and Hopper's classification. I would actually argue that in the context of my study, both *relaxed* and *aggressive* are rather expressive of social attractiveness than of 'dynamism', as Zahn and Hopper would have it, but we can argue about that. And note also that I have actually recoded *coarse* into its opposite *sophisticated* here, because this shows the patterning more nicely.

Okay so what we see here is that the matched guise clearly shows stronger effects for items related to *social attractiveness*. The open guise shows stronger effects for items related to *superiority*. There are some overlaps, where in both experiments differences between dialect and standard come out equally in the ratings. But where there are differences, this is the general trend.

You can see this also in this next visualization:



There are stronger effects for the matched guise, especially for the standard variant, for the *attractiveness* items, but stronger effects in the open guise on *superiority*, especially for the dialect variant.

Recall again that whereas in the matched-guise, the informants did *not* know that two recordings featured one and the same speaker, in the open guise, the informants knew this full well. And so it seems that this knowledge does indeed play a role in speaker evaluation. The



informants in the open guise apparently did not think that changing speaking style affected the speaker's projected personality in terms of social attractiveness. Humor was affected, but that's pretty much it there. But changing speech style *did* affect how the informants perceived the speaker's projected intellectual competence, particularly for the dialect guise.

On the other hand, the informants in the matched guise, who did *not* know they were listening to one speaker multiple times in two different guises, additionally and more strongly assumed differences in personality that were to do with attractiveness. And they did not appreciate the standard guise much in this context.

So here we now seem to finally have an empirically generated hint at the effect and role of the disguising ploy in the matched guise technique.

Honestly, I am not yet quite 100% sure what to make of all this, and I'll be really happy for your suggestions – especially regarding the literature in social psychology that I should perhaps be linking up with here. Also, of course, I am only dealing with one test speaker here.

But I do have two final thoughts.

First, maybe there's something about the directionality of the switch that plays a role in the open guise: if you didn't start out sounding friendly and likeable, maybe you are unable to catch up. But on the other hand, one can apparently always use language shifting to dumb it down.


And this brings me to my second thought: I said at the beginning that I'm using these types of experiments to inform my research on rhetorical style-shifting in Austrian German. What I've found there is that people do use shifts from standard into dialect to ridicule opponents, for example in TV discussions, which is the data set I am mostly working with. This is a salient pattern in my data – when producing off-hand, derogatory comments, like supposedly 'stupid answers' to stupid questions, people are more likely than in other contexts to shift from Austrian standard into dialect. Humor of course plays a role there too. This very much corresponds with my open guise findings: dialect is used to sound funny and stupid and ignorant. I pretty much couldn't find any examples where someone would switch into the dialect for a positive effect that would correspond with increasing their likeability. Again, maybe it's hard to switch to sounding friendlier if you didn't start out that way.

All in all, there's much food for thought here, I believe. What I certainly hope has been established, however, is that we need to choose our experimental design for speaker evaluation studies very carefully and with the role of the disguising ploy in mind. Here's a quote from William Labov, master of sociolinguistic methods, to close us out:

“An essential feature of this methodology [the matched-guise technique] is that the subjects not be aware that the same speaker is presented in different guises. An alternation of three or four intervening subjects is sufficient to ensure that this be so. Every once in a while someone carries out a matched guise experiment with a single speaker, not realizing the force of this condition, and the results are very pale by comparison.” Labov (2006 [1966]:271)

I heartily disagree – the results are not pale at all. But they are noticeably and significantly *different*. It is this difference between open guise and matched guise that we need to reckon with in the future.

THANK YOU!



**References:**

Labov, William. 2006. *The social stratification of English in New York City*. 2<sup>nd</sup> ed. Cambridge University Press.

Lai, Mee Ling. 2007. Exploring language stereotypes in post-colonial Hong Kong through the matched-guise test. *Journal of Asian Pacific Communication* 17/2:225-244.

Lambert, Wallace E., Richard Hodgson, Robert C. Gardner, and Samuel Fillenbaum. 1960. Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology* 60/1:44-51.

Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana: University of Illinois Press.

Soukup, Barbara. 2009. *Dialect use as interaction strategy*. Vienna: Braumüller.

Zahn, Christopher J., and Robert Hopper. 1985. Measuring language attitudes: The Speech Evaluation Instrument. *Journal of Language and Social Psychology* 4/2:113-123.