RISTAL 4 / 2021

Research in Subject-matter Teaching and Learning

Volume 4 – Special Issue edited by Timo Leuders & Katharina Loibl

Citation:
Witzigmann, S., Sachse, S. (2021). Diagnostic competencies of prospective teachers of French as a foreign language: judgement of oral language samples [Special Issue]. *RISTAL, 4*, 71–87.

DOI: https://doi.org/10.23770/

ISSN 2616-7697

Diagnostic competencies of prospective teachers of French as a foreign language: judgement of oral language samples

Stéfanie Witzigmann & Steffi Sachse

Abstract

The diagnostic competence of foreign language teachers includes the ability to judge oral language productions. However, these are difficult to judge because of their complexity. Video-based data can be used to explore the diagnostic processes when judging oral productions. In the present study, prospective teachers evaluate nine video samples twice – directly after first viewing the video and after using a rating scale regarding the linguistic features. This study examines the extent to which prospective primary and secondary level teachers use different information to make a judgement compared with experts. The results show that linguistic features are incorporated to varying degrees into the formation of judgements and vary according to pre-service teacher training. Besides the focus on linguistic features, a high proportion of unexplained variance remains in the judgements of the prospective teachers.

Keywords

Diagnostic competences – speech samples – video vignettes – judgement - holistic and analytic scoring

1 Introduction

Oral language productions are an important objective of (foreign) language teaching, but are also relevant for other subjects like history, etc. where oral competencies are being judged (and graded). In the current era of communicative foreign language teaching, orality is increasingly important, and so requires teachers to judge their students correctly (e.g., Fulcher, 2015). However, oral performance appears to be difficult to judge, as many internal and external factors can influence the judgement of spoken foreign language (e.g., Brown, Iwashita, & McNamara, 2005; Chuang, 2009; Davis, 2016; Winke, Gass, & Myford, 2013).

Perceptible factors are, for example, linguistic features (e.g., pronunciation, extent and complexity of vocabulary, grammatical correctness, but also pragmatic elements), which can be interpreted as cues. It is expected that these cues will contribute to the judgements in varying ways and degrees. Raters can be biased by such factors as their personal characteristics (e.g., knowledge, beliefs, experience), characteristics of the tasks (e.g., type of test, difficulty of task, type of material), characteristics of the learners (e.g., language skills, age, sex/gender, cultural background), or the judgement scales used (e.g., holistic or analytic judgement) (Caban, 2003).

DOI: https://doi.org/10.23770/

Considering possible influences of personal characteristics (e.g., pre-service teacher training, linguistic expertise) and situational characteristics (e.g., linguistic or content features, structure of judgement scale), this study aims to examine prospective teachers' judgement of oral productions for students in comparison with those of experts. Here, the framework DiaCoM (Loibl, Leuders, & Dörfler, 2020) is used to form the overarching theoretical framework and to build a first step in exploring teachers' diagnostic thinking and activities.

2 Theoretical Background

Over the last three decades, the diagnostic skills of teachers have been increasingly investigated. The focus has often been on the accuracy of judgement, i.e., the consistency of a teacher's judgement with an objective judgement of the student's characteristics. Several authors have investigated which factors influence this accuracy of judgement (see e.g., the meta-analysis by Südkamp, Kaiser, & Möller, 2012).

However, the formation of such judgements has been less well researched (Herppich et al., 2018). In the area of the judgement of oral language production, the few relevant studies of teachers provide initial indications of personal and situational characteristics which may influence the judgements (Hochstetter, 2011; Kim, 2009). Little research exists on the judgement processes, which take place in this context. In order to make the formation of judgements researchable, Loibl et al. (2020) developed the framework DiaCoM (Explaining Teachers' **Dia**gnostic Judgements by **Co**gnitive **M**odeling). The DiaCoM framework aims to provide an investigative and interpretive framework for the analysis of cognitive processes in teacher judgement.

The DiaCoM framework conceptualizes diagnostic judgements as inferences of a teacher about students based on the information explicitly or implicitly given in a diagnostic situation. Thus, four components constitute the DiaCoM framework: the characteristics of the diagnostic situation (situational characteristics, SC), the characteristics of the teacher (personal characteristics, PC), the internal information processing during the formation of a diagnostic judgement (diagnostic thinking, DT) and the observable diagnostic verbalization of the teacher (diagnostic behavior, DB). The DiaCoM framework has to be specified for each setting and research question and has already being used in various subjects (e.g., mathematics, biology, elementary social studies and sciences, or economics).

The situation characteristics (SC) specify the context which is presented to the teacher e.g., language productions in video vignettes. This diagnostic situation contains information (i.e., cues) which teachers can (or cannot) perceive. When judging a student's performance, the relevant situation characteristics are observable features (e.g., linguistic or content features). In addition, more general characteristics of the situation can influence the judgement (e.g., the structure of the judgement scale).

Personal characteristics (PC) include the characteristics of persons making judgements (such as age, sex/gender, attitudes, professional background, knowledge, or experience)

DOI: https://doi.org/10.23770/

which can have a certain, sometimes significant influence on the judgement, depending on the diagnostic situation (see, e.g., Eckes, 2015; Shaw, 2007).

The diagnostic situation combined with personal characteristics (e.g., language skills or teaching experience) are those which induce a diagnostic process in the person (i.e., diagnostic thinking, DT). When teachers judge oral foreign language skills, several cognitive processes are initiated (e.g., the perception of content and linguistic features, and the integration of these individual features into an overall judgement). These processes can in turn translate into observable diagnostic behavior (DB) (e.g., verbal statements or written answers), which can be used as indicators for judgement processes.

2.1 Influence of the Situation in the Judgement Process

Regarding the evaluation of oral language productions, the structure of the evaluation scale can make a decisive contribution to the formation of judgement processes. In general, a distinction can be drawn between holistic or analytic scoring (see Bachman & Palmer, 2010; Chuang, 2009; Harsch & Martin, 2013).

In the case of holistic scoring, an overall judgement is made which does not explicitly judge individual sub-dimensions. Speech production as a whole is usually judged relatively quickly after viewing the speech product and based on the impression gained (Hinger & Stadler, 2018; Knoch, 2011).

In analytic scoring, the language construct is differentiated into individual sub-dimensions. These individual features (linguistic or content-related aspects) are evaluated separately (see e.g., Brookhart, 2013).

In our article, we use the term global judgement to distinguish the formation of an overall judgement after an analytic evaluation from the first intuitive holistic overall judgement. Empirical studies show that the linguistic construct can be reliably assessed in both modes of judgement (holistic and analytic) (see e.g., Chuang, 2009; Metruk, 2018). Holistic judgements also have the advantage of being able to show the implicit use of individual linguistic features.

Accurate judgement of oral speech production should be based on construct-relevant characteristics wherever possible. In the field of oral production, these characteristics can be identified in the areas of communicative language competences: linguistic competence, pragmatic competence, and sociolinguistic competence (Council of Europe, 2020). Six characteristics at the linguistic and pragmatic competence level (phonology, vocabulary, grammar, fluency, comprehensibility, and communication strategies) are selected in the present study. When judging fluency, our study will be based exclusively on the speaking rate (Fulcher, 2015). In our study, the judgement of comprehensibility should include linguistic as well as paralinguistic elements - needed when conveying a message. Furthermore, communication strategies are classified into compensation strategies (paralinguistic strategies such as gesture/mimicry, onomatopoeia, code-switching when used as compensation) and communication-enhancing strategies

DOI: https://doi.org/10.23770/

(gesture/mimicry, onomatopoeia, prosodic means such as volume and pitch) to reinforce a statement.

2.2 Influence of Personal Characteristics on the Judgement Process

Personal characteristics include those traits (such as age, sex/gender, attitudes, professional background, knowledge, and experience) which may influence judgements (see e.g., Eckes, 2015; Shaw, 2007). This is also referred to as a rater-mediated judgement. When judging oral language production, only the characteristics of the learners' language abilities should be accurately identified and judged. However, judgements are often influenced by a number of factors, such as judgement biases (e.g., tendency towards strict/mild, central tendency, or halo effects), the difficulty of tasks, rating scale structures, personal attitudes, and motivational factors (Caban, 2003; Lenske, 2016).

In the field of foreign language research, particularly in oral judgements, previous studies have been able to show raters' influence on test results (including Bachman, Lynch, & Mason, 1995; Lumley & McNamara, 1995). Recent studies have also attempted to investigate the relationship between rater background and judgements of speaking skills, though with a clear focus on raters' leniency (Carey, Mannell, & Dunn, 2011; Sundqvist, Wikström, Sandlund, & Nyroos, 2018; Yan & Ginther, 2018; Zhang & Elder, 2011). However, these studies are not really comparable, because of the many factors which can influence the judgement of oral language samples (e.g., (teaching) experience of evaluators, difficulty of tasks, language level of speakers). For Duijm, Schoonen, and Hulstijn (2018), two essential background characteristics of raters can nevertheless be identified: linguistic expertise and exposure to L2 speech. Although they focused only on the characteristics of speech accuracy and fluency, they found that high exposure to speech increases the sensitivity of the raters to perceive improvements in these characteristics. These results may indicate that high linguistic expertise and the implicit knowledge of essential speech features in oral samples are particularly important for identifying and judging them.

In Germany, the system of teacher education shows that future teachers undergo different types of formal qualification depending on their pre-service training, i.e., if they aim at becoming primary, lower, or upper secondary school teachers (Cortina Kai S. & Thames, 2013). Training for primary school teachers is in general shorter than for secondary school teachers. While the training for primary school teachers includes - in addition to the chosen foreign language - an extensive pedagogical study as well as the subjects' mathematics or German, the training for upper secondary school teachers is based on a strong scientific focus on two subjects. In Baden-Württemberg, teacher training for the primary level takes place at the universities of Education, while universities provide teacher training for the upper secondary school teachers. In the context of foreign language teaching, studies have shown that the knowledge differed depending on which school type teachers are trained for. There is a significant proportion of primary school teachers teaching a foreign language who did not undergo a subject-specific university education (Kolb, 2011; Porsch & Wilden, 2017). Kolb (2011) was able to show that secondary school teachers of English in Germany use different methodological

DOI: https://doi.org/10.23770/

forms of work than primary school teachers. The focus at the primary level is predominantly on oral forms of work (including oral judgements), while these are less important at the secondary level (at upper secondary school oral judgements were used by only 9% of the teachers surveyed (n=270)). At the secondary level, formal language aspects (such as grammar or vocabulary work) are predominant. Kolb (2011) also found in her study that primary school teachers pay less attention to formal features of language than do secondary school teachers.

3 Research Questions

As we are interested in which personal or situational characteristics affect judgement accuracy, we conducted this study with prospective teachers to see if personal characteristics such as low experience in judgement or type of pre-service teacher training influence judgements. Those cues (characteristics of language productions) which are actually contained in presented video vignettes were empirically tested by utilization of experts' knowledge (see Witzigmann & Sachse, 2020). Thus, this study examines the judgement of students' oral productions (in French) by prospective primary and secondary level teachers compared with those of experts. It examines which cues are integrated into an overall judgement. Two different diagnostic judgements had to be made within the study: a first, holistic overall judgement followed by rating scale supported analytic judgements; and a second, overall judgement. The aim is to investigate the relationship between the diagnostic situation (in our case the specific linguistic characteristics of the language samples) and personal characteristics during the judgement by answering the following questions:

1. Do prospective primary and secondary level teachers judge oral language samples of foreign language learners differently than experts?
2. Which different (linguistic) information do prospective primary and secondary level teachers use to make their judgements compared to experts? Do the groups integrate linguistic features in different ways?

In line with the need for research pointed out in section 2, we analyze what kind of information (linguistic features in video vignettes) in the diagnostic situation (two different diagnostic judgements) is perceived by the prospective teachers (primary versus secondary level) and how this information is processed. Measuring teachers' traits, the specification of the diagnostic situation (with holistic and analytic scoring), and expert knowledge will help us to better understand the information processing of (future) teachers of French.

DOI: https://doi.org/10.23770/

4 Empirical Design

4.1 Sample

Participants in the study were prospective French teachers for primary (n=38) and secondary (n=32) school levels (18.7% male, 81.43% female). All prospective teachers come from the states of Baden-Württemberg and Hesse, but from different universities and locations. Teacher training varies considerably in terms of structure and content between the prospective teachers. In Baden-Württemberg, teacher training for the primary level takes place at universities of education, while teacher training for the secondary level (advanced level, called “Gymnasium”) takes place at universities. However, at both types of university, judgements of language productions in foreign language teaching is dealt with in only a few sessions during their teacher training. Thus, prospective teachers often have no or only minimal experience with this matter (possibly acquired during school internships M = 9.07 (SD = 9.09) weeks).

A group of subject-matter education researchers (N=13) of the target language French formed the group of experts. They are characterized by their area-specific knowledge in the domain of "French Education" (professors and lecturers at universities). In the context of our study, their judgement can be seen as an aggregated expert judgement. We already know from this group of experts that their judgements are relatively homogeneous due to their high Interclass Correlation Coefficient (ICC) close to 1 (ICC = .954, F = 32.95; $p < .001$), particularly with regard to the integration and composition of their judgements by linguistic features (R^2 see table 5) (Witzigmann & Sachse, 2020).

4.2 Design of the Study

Because oral language productions are only subject to direct observation, the diagnostic situation was realized based on videotaped speech samples (stimulus material) and integrated into a computer survey tool. In contrast to audio recordings, video vignettes allow the perception of paralinguistic features such as gestures and facial expressions (Hsieh & Davis, 2019). Four school classes at different schools (2 primary school classes (4th grade) and 2 secondary school classes (6th grade)) were selected. The students were videotaped individually, and all were confronted with the same tasks. In the first part - stimulated by four questions from a native speaker - they were asked to introduce themselves, their family, their hobbies, and their favorite animal. In the second part, they were asked to describe a comic strip (eight pictures with a short incident between a dog (the main protagonist) and other dogs). A total of nine video vignettes were selected (five boys and four girls). The speech samples are real speech situations: students were not specifically prepared for the questions and did not know the comic strip.

In the first round, the video vignettes were viewed, and immediately afterward a holistic overall judgement (Diagnostic Behavior 1 - DB1) on a scale of 0-100% (100% = what could be considered for each rater as a perfect performance for the given group of students) had to be given. In the second round, the prospective teachers and experts received a rating scale (for the characteristics see section 2.1) to re-evaluate the same video

DOI: https://doi.org/10.23770/

vignettes they had just seen. Here, each linguistic feature had to be judged on a numerical 10-level scale (very bad to very good, fig. 1). An overall global judgement (Diagnostic Behavior 2 - DB2) was requested again after each video vignette and the analytic judgement.

	very bad									very good
pronunciation	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
	very poor									very differentiated
vocabulary	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩

Fig. 1: Linguistic features, example items.

In the present study, the individual elements of the framework (see 2.) are specified and examined in the following way: prospective teachers perceive the nine video vignettes of students and generate an overall judgement. We assume that they identify individual features of the language productions (e.g., pronunciation, comprehensibility, or communication strategies) by viewing the video vignette and then integrate the judgement of the different features into an overall judgement. They do this in the first round implicitly (holistically) (DB1), and in the second round more explicitly after the analytic evaluation of the linguistic features (DB2). The analysis of personal characteristics such as type of education degree is thought to be related to the judgement (diagnostic behavior, DB).

4.3 Data Analysis

The data was analyzed using SPSS for Windows (version 27.0) and the open-source statistics program R (R Core-Team, 2018) with the desktop platform R-Studio (RStudio, Inc., 2018).

Nine video vignettes were evaluated from each rater. The data structure contains ratings where the stimuli/language samples are clustered. Here we have a repeated-measurement design with n = 70 and a within factor with nine conditions. Accordingly, all analyses were performed separately for each rater over all stimuli and then averaged. To find out which of the rated features of the rating grid influence the first and second judgement and in which way, dominance analyses were carried out. This method is used to compare the predictors in all possible constellations to determine which predictor is dominant. The results of multiple regression are used to determine the dominance of one predictor over another. The elucidations of variance for a focal variable are averaged over the different constellations with all possible preceding combinations of variables. Dominance analysis is particularly suitable when predictors are intercorrelated (Azen & Budescu, 2006).

5 Results

DOI: https://doi.org/10.23770/

Question 1: Do prospective primary and secondary level teachers judge oral language samples (SID) of foreign language learners differently than experts?

Tab. 1: Mean scores, standard deviation, and range by n=13 experts for both diagnostic behaviors

SID		DB1-Experts				DB2-Experts			
	sex	***M***	*SD*	*min*	*max*	***M***	*SD*	*min*	*max*
SID-1	m	**52.08**	17.47	*20*	*80*	**51.69**	18.04	*20*	*80*
SID-2	w	**70.69**	18.19	*25*	*98*	**57.15**	15.50	*35*	*80*
SID-3	m	**20.00**	10.02	*5*	*35*	**24.53**	10.45	*5*	*40*
SID-4	w	**63.07**	18.65	*40*	*95*	**58.31**	16.82	*30*	*85*
SID-5	m	**29.92**	15.38	*5*	*60*	**32.77**	14.05	*15*	*58*
SID-6	m	**48.69**	17.02	*20*	*72*	**44.77**	16.32	*20*	*75*
SID-7	w	**75.69**	13.17	*50*	*100*	**74.62**	10.21	*62*	*95*
SID-8	w	**62.23**	18.66	*25*	*90*	**53.31**	13.60	*35*	*77*
SID-9	m	**82.00**	13.71	*40*	*95*	**77.69**	15.24	*35*	*92*

Note: SID = Students' oral language productions; DB1 = Diagnostic behavior (holistic); DB2 = Diagnostic behavior (global).

Corresponding to our assumption that differences in pre-service training influence oral judgement, we subdivided the participants into two groups: prospective teachers at primary level and secondary level (see section 2.2). For both diagnostic behaviors – first holistic judgement (DB1) and second global judgement after the analytic rating grid (DB2) – mean scores, standard deviation, and range were calculated for experts and prospective teachers (see tables 1 and 2).

Tab. 2: Mean scores, standard deviation, and range by n=38 prospective primary level teachers and n=32 prospective secondary level teachers for both diagnostic behaviors

SID		*DB1-PT-PRIM*			*DB2-PT-PRIM*			*DB1-PT-SEC*			*DB2-PT-SEC*		
	S	***M*** *(SD)*	*min*	*max*	*M (SD)*	*min*	*max*	***M*** *(SD)*	*min*	*max*	***M*** *(SD)*	*min*	*max*
SID-1	m	**53.55** (18.6)	*25*	*85*	**54.21** (16.2)	*30*	*90*	**44.22** (20.6)	*10*	*80*	**51.56** (17.2)	*15*	*80*
SID-2	w	**63.55** (18.9)	*25*	*95*	**60.13** (19.5)	*20*	*95*	**55.00** (20.7)	*15*	*85*	**51.88** (19.9)	*10*	*90*
SID-3	m	**33.16** (17.3)	*5*	*85*	**33.29** (13.7)	*5*	*70*	**24.84** (15.6)	*5*	*60*	**27.19** (14.2)	*5*	*55*
SID-4	w	**71.32** (14.1)	*40*	*95*	**67.50** (15.9)	*30*	*95*	**67.50** (17.9)	*35*	*95*	**65.78** (18.2)	*20*	*95*
SID-5	m	**46.32** (17.9)	*10*	*90*	**44.61** (13.6)	*25*	*75*	**34.53** (17.1)	*10*	*70*	**35.78** (17.4)	*5*	*75*
SID-6	m	**58.16** (17.8)	*15*	*90*	**55.13** (14.5)	*30*	*85*	**52.81** (17.1)	*20*	*80*	**54.38** (17.6)	*15*	*80*
SID-7	w	**68.82** (17.8)	*30*	*95*	**63.55** (15.1)	*35*	*90*	**70.16** (18.6)	*25*	*95*	**66.88** (16.3)	*30*	*95*
SID-8	w	**64.21** (17.3)	*25*	*95*	**63.95** (14.2)	*35*	*95*	**61.09** (17.1)	*20*	*85*	**62.97** (18.6)	*30*	*95*
SID-9	m	**71.18** (17.1)	*35*	*100*	**71.18** (16.2)	*40*	*95*	**74.84** (18.1)	*30*	*95*	**73.75** (22.2)	*20*	*100*

Note: SID = Students' oral language productions; S = Sex; PT-PRIM = Prospective primary level teachers (n=38); PT-SEC = Prospective secondary level teachers (n=32).

DOI: https://doi.org/10.23770/

For the majority of prospective teachers, the ranges of the judgements on the 0-100% scale are larger than within the experts: the language productions are sometimes rated higher and sometimes lower by the prospective teachers than by the experts. Correlations between the first and second judgement do not show large deviations on average among the experts (r= .752, p= .009) and prospective teachers (r= .734, p< .001). For all groups, there is a high degree of consistency between the first judgement (DB1) and the second judgement (DB2). The ranking order of the prospective teachers is only partially consistent with that of the experts (see table 3).

Tab. 3: Mean scores and ranking by experts and prospective teachers for both diagnostic behaviors

SID		DB1-EXP		DB2-EXP		DB1-PT-PRIM		DB2-PT-PRIM		DB1-PT-SEC		DB2-PT-SEC	
	sex	*M*	rank	*M*	rank	*M*	rank	*M*	rank	*M*	rank	*M*	rank
SID-1	m	52.08	6	51.69	6	53.55	7	54.21	7	44.22	7	51.56	7
SID-2	w	70.69	3	57.15	4	63.55	5	60.13	5	55.00	5	51.88	6
SID-3	m	20.00	9	24.53	9	33.16	9	33.29	9	24.84	9	27.19	9
SID-4	w	63.07	4	58.31	3	71.32	1	67.50	2	67.50	3	65.78	3
SID-5	m	29.92	8	32.77	8	46.32	8	44.61	8	34.53	8	35.78	8
SID-6	m	48.69	7	44.77	7	58.16	6	55.13	6	52.81	6	54.38	5
SID-7	w	75.69	2	74.62	2	68.82	3	63.55	4	70.16	2	66.88	2
SID-8	w	62.23	5	53.31	5	64.21	4	63.95	3	61.09	4	62.97	4
SID-9	m	82.00	1	77.69	1	71.18	2	71.18	1	74.84	1	73.75	1

Note: SID = Students' oral language productions; EXP = Experts (n=13); PT-PRIM = Prospective primary level teachers (n=38); PT-SEC = Prospective secondary level teachers (n=32).

Among prospective primary level teachers, the ranking for language productions does not concord with the experts except for the two lowest judgements at the first judgement. In the second judgement (following the analytic rating grid), however, the best language production is identified in accordance with the experts. In the case of prospective secondary school teachers, the two lowest and two best students are recognized in the first judgement. In the second judgement, prospective secondary level teachers rank five of nine language productions in the order aggregated by the experts (the three lowest and the two best productions). At the extremes, however, all raters agree on the lowest and best language productions. For the first holistic judgement (DB1), t-tests show statistically significant difference between judgements of prospective primary and secondary school teachers for the two lowest productions (SID-3: $t(68)= 2.09$, $p= .040$; SID-5: $t(68)= 2.08$, $p= .007$). For the second judgement, a statistically significant difference was found in the fifth language production (second-lowest performing) (SID-5: $t(68)= 2.38$, $p= .020$). On average, prospective teachers at the primary level judge students more positively than those at the secondary level.

Question 2: Which different (linguistic) information do prospective primary and secondary level teachers use to make their judgements compared to experts? Do the groups integrate linguistic features in different ways?

DOI: https://doi.org/10.23770/

The correlative analyses and the tolerance values (*TOL* (1-R^2): Pronunciation= .140; Lexicology= .049; Grammar= .114; Fluency= .097; Comprehensibility= .095; Communication-enhancing strategies= .197; Compensation strategies= .159) indicate multicollinearity. Due to this multicollinearity and to determine the integration of the individual linguistic features (i.e., cues), dominance analyses were performed (see table 4).

Tab. 4: Variance explained (R^2) by single cues and total as a result of dominance analysis with overall judgement one and two, respectively, as criterion and averaged over n=13 experts and n=70 prospective teachers

Experts			Prospective Teachers		
	DB1	DB2		DB1	DB2
Lexicology	.217	.230	Lexicology	.108	.161
Comprehensibility	.161	.189	Fluency	.092	.117
Grammar	.146	.173	Comprehensibility	.088	.131
Pronunciation	.128	.110	Grammar	.083	.111
Communication str.	.123	.079	Compensation str.	.072	.082
Fluency	.114	.139	Pronunciation	.064	.091
Code-Switching	.074	.064	Com.-enhancing str.	.045	.048
R^2	**.964**	**.984**		**.552**	**.742**

Note: In the dominance analyses, 7 linguistic features are used as predictors and the overall judgement (i.e., DB1, DB2) as a criterion. The 9 language samples are the data basis of each dominance analysis. Communication str. = Communication strategies; Compensation str. = Compensation strategies; Com.-enhancing str. = Communication-enhancing strategies.

Following feedback from the experts, the rating grid was changed to make a distinction between communication-enhancing-strategies (e.g., paraphrasing, use of silence, intonation) and compensation strategies (e.g., code-switching, gestures/mimicry, onomatopoeia) instead of only communication strategies and code-switching (see Witzigmann & Sachse, 2020). Vocabulary is the predominant predictor in both groups. For the experts, comprehensibility (DB1= 16.1%; DB2= 18.9%) and grammar (DB1= 14.6%; DB2= 17.3%) are also strong predictors. For prospective teachers, fluency (DB1= 9.2%) and comprehensibility (DB2= 13.1%) form the second strongest predictors. After using the rating grid, their linguistic predictors contribute more strongly to the total variance explanation (R^2= .742) but are still far behind the experts (R^2= .984). In particular, significantly fewer linguistic features are integrated by prospective teachers in the first judgement than by experts.

The results of the total variance analysis are particularly interesting regarding prospective teachers. While there is a high explanation of variance by the linguistic features among the experts, almost half of the total variance remains unexplained (R^2= .552) for prospective teachers and at least 25% for the second judgement (R^2= .742). In order to examine whether the different groups integrate the individual features differently in their overall judgement, a distinction was again made according to the type of teacher

DOI: https://doi.org/10.23770/

training (see table 5). This is the average dominance for the group of prospective teachers.

Tab. 5: Variance explained (R^2) by single cues and total as a result of dominance analysis with overall judgement one and two, respectively, as criterion and averaged over n=38 prospective primary level teachers and n=32 prospective secondary level teachers

	Prospective Teachers Primary level		Prospective Teachers Secondary level	
	DB1	DB2	DB1	DB2
Lexicology	.109	.163	.118	.166
Fluency	.092	.112	.101	.129
Comprehensibility	.090	.131	.094	.136
Grammar	.076	.103	.102	.125
Pronunciation	.065	.093	.069	.094
Compensation str.	.060	.071	.096	.101
Com.-enhancing str.	.044	.048	.050	.051
R^2	**.536**	**.721**	**.631**	**.802**

In the holistic first judgement, vocabulary and fluency are the two major indicators of the variance explanation by prospective primary level teachers. At the secondary level, grammar and compensation strategies are strong predictors of variance explanation in addition to vocabulary and fluency. Furthermore, grammar and compensation strategies are more important for the explanation of variance at the secondary level (Grammar= 10.2%; Compensation strategies= 9.4%) than at the primary level.

6 Discussion and Limitations

To better understand the judgement of oral language productions, this article aimed to investigate the judgement of oral language production using video samples. We compared an immediate judgement delivered after the presentation of a diagnostic situation (nine video samples to judge in two different ways) by giving two judgements – the first, holistic, one after the first viewing, and a second one after using a rating grid regarding the linguistic features of the language productions. We were interested in the factors underlying the judgements of prospective teachers of different school types and experts (university teachers).

Our results indicate less accurate judgements in prospective teachers compared to experts. The ranking order of the prospective teachers is only partially consistent with that of the experts. In principle, a deficiency in diagnostic competencies may be due to a limited experience with judgements during university training. These results are in line with Jacob (2012) and Sundqvist et al. (2018), who showed that experience can influence judgements. Interestingly, the judgements of prospective primary level teachers are further from the experts' judgements than those of prospective secondary level teachers. This could be explained by the longer teacher training and the stronger focus on subject-

DOI: https://doi.org/10.23770/

specific characteristics among prospective secondary level teachers. Examining the differences in the ranking order, prospective primary level teachers' judgements are milder than among experts - except for the two or three best learners, where the judgements are stricter. The results of the prospective secondary level teachers are sometimes milder and sometimes stricter compared to the experts. The found tendency to overestimate student performance can also have positive effects on student learning development, especially among weaker learners. However, with the stricter overall judgements, there is a risk that existing language features will not be adequately recognized, and that performance will be underestimated. This tendency may be due to a deficiency in the ability of perception and interpretation of (linguistic) features in oral language productions. In practice, this needs to be integrated and trained to a higher extent in the education of foreign languages, but also for judgements of oral productions in other subjects.

The higher total variance explanation for the second judgement shows that the prospective teachers pay more attention to the cues and integrate them into their judgement. As indicated by Hochstetter (2011), the evidence we found points to the use of a rating grid if raters are not familiar with judgements. Also noticeable is that prospective primary level teachers pay less attention to the given linguistic features. Here, other moderators seem to be integrated into the overall judgement. A possible explanation could be their training, which includes extensive pedagogical studies in addition to their subject-specific studies. In further research, additional moderator variables should be included to investigate which personal characteristics are responsible for the resulting judgements. The question remains as to whether other groups of participants, such as in-service teachers at the primary level, would come to similar findings. This should be pursued in future studies.

The most surprising result is that approximately 25% of the total variance by prospective teachers remains unexplained, despite the use of a rating grid. From a methodological point of view, it is important to note that no additional qualitative data in the form of think-aloud protocols were collected within this study. Further research in the field of teachers' diagnostic judgements should integrate think-aloud protocols or other process data (e.g., eye movements) to get more insights into their judgements.

Further, our results indicate that both groups show high consistencies between the first judgement and the second judgement. We would have expected that, by focusing externally on linguistic features through the rating grid, prospective teachers would have been more likely to adjust their judgements on the second judgement – i.e., focusing on the criteria would have triggered a change in their judgements. Hence, it would be interesting to investigate this in an experimental design where analytic and holistic rating scales are systematically varied (for writing judgements see Keller, Jansen, & Vögelin, 2019). However, our design cannot exclude the raters' possible need for consistency between judgements or that the overall holistic initial judgement may influence the judgement of the individual linguistic features.

DOI: https://doi.org/10.23770/

Also, we would like to discuss our results in terms of how the different groups judge the oral language samples or which (linguistic) features they use to make their judgements. Regarding the integration of individual (linguistic) features with prospective primary and secondary level teachers compared to experts, our results show that vocabulary and comprehensibility remain the main predictors for all groups and these contribute strongly to variance explanation. This indicates that today great importance is attached to a large vocabulary and comprehensible message transmission, which are considered important indicators of communicative ability. Although grammatical correctness still remains a relatively important factor for experts, prospective teachers place more weight on fluency. Fluency could also be easier for prospective teachers to identify than grammatical correctness.

The result of the different weighting of the cues between the first overall judgement (holistic) and the second one is not entirely surprising for the prospective teachers. Compared to the experts - who placed almost twice as much emphasis on communication strategies in the first than in the second judgement - there are hardly any noticeable changes here. This could be explained by the low level of experience of prospective teachers with judgements. In contrast to the experts, they do not have the features implicitly in their minds (Witzigmann & Sachse, 2020) during the first judgement, which is also indicated by the low values of the dominance analysis. As previous studies have shown, the desire for a high degree of consistency seems to influence the judgement. Here, further studies seem necessary which integrate process data such as think-aloud protocols in order to gain more insights into the judgement processes.

When comparing the two types of teacher training, it is noticeable that both groups pay particular attention to vocabulary and fluency. This finding validates previous results by Iwashita, Brown, McNamara, and O'Hagan (2008). Further, prospective secondary school teachers integrate the individual features of grammar and compensation strategies more in their judgement than prospective primary school teachers. This could be due to teacher training. The universities which train teachers for upper secondary education offer several courses exclusively in grammar and translation. At the universities of education, for example, courses with translation have not been taught for a long time.

Both the low level of agreement among prospective teachers of ranking the language productions compared with the experts and the total variance explanation by linguistic features show that there is a need to focus more on the acquisition of diagnostic competences in the field of oral language production in practice and research. Overall, the present study on judgement processes among (prospective) teachers could only provide a first insight into this complex. We should note that this paper focused on a few influencing factors and was conducted only with prospective teachers. Furthermore, our study did not specify what 100% perfect performance means. Further studies should, for example, work with benchmarks. This could be done with a first video showing the “perfect” performance for the given learning level. Our results indicate that the communicative oral language ability is very complex and therefore it is not easy to understand the process of diagnostic thinking. In addition, other studies should include process data

DOI: https://doi.org/10.23770/

(e.g., think-aloud protocols, eye tracking) to gain a deeper insight into the judgements processes (e.g., in mathematics, Becker, Spinath, Ditzen, & Dörfler, 2020).

These first results allow to better understand the process of cognitive modelling of diagnostic judgements even if not all aspects of teachers' diagnostic activities can be considered. By limiting the setting by immediate judgements delivered after the presentation of a diagnostic situation (whether materials or students), taking into account influencing factors such as personal characteristics of raters (e.g. knowledge, experience or beliefs) and specifying it to the particular research goal (and subject), the DiaCoM framework seems to be a useful theoretical basis for a better understanding of judgement processes in different educational contexts (Loibl et al., 2020). In other subject-specific fields (see, e.g., Becker et al., 2020; Rieu, Loibl, & Leuders, 2020 for mathematics or Hoppe, Renkl, Seidel, Rettig, & Rieß, 2020 for biology), first experimental manipulation of (prospective) teachers' traits (e.g., stress) or framing of the diagnostic situation (e.g., task difficulty) allow first theoretical assumptions about the use of knowledge during the judgement process. We believe that our results can be a further component of the processes underlying diagnostic judgements in educational settings and we will explore in further studies more influencing variables (e.g., beliefs, L2 exposure, or conducted with in-service teachers).

Funding:

This work was supported by the Ministry of Science, Research, and the Arts Baden-Württemberg, Germany as part of the Research Training Group "Diagnostic Competences of Teachers" (DiaKom).

References

Azen, R., & Budescu, D. v. (2006). Comparing Predictors in Multivariate Regression Models: An Extension of Dominance Analysis. *Journal of Educational and Behavioral Statistics*, *31*, 157–180. https://doi.org/10.3102/10769986031002157

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*, 239–257.

Bachman, L. F., & Palmer, A. S. (2010). *Language testing in practice: Designing and developing useful language tests*. *Oxford applied Linguistics*. Oxford: Oxford Univ. Press.

Becker, S., Spinath, B., Ditzen, B., & Dörfler, T. (2020). Der Einfluss von Stress auf Prozesse beim diagnostischen Urteilen – eine Eye Tracking-Studie mit mathematischen Textaufgaben. [The influence of stress on processes of diagnostic judgement-an eye tracking study based on mathematical word problems]. *Unterrichtswissenschaft*, *48*, 531–550. https://doi.org/10.1007/s42010-020-00078-4

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, Virginia: ASCD.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks*. Princeton, NJ: Educational Testing Service.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, *21*, 1–44.

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. https://doi.org/10.1177/0265532210393704

Chuang, Y.-Y. (2009). Foreign language speaking assessment: Taiwanese College English teachers' scoring performance in the holistic and analytic rating methods. *Asian EFL Journal*, *11*, 150–173.

Cortina Kai S., & Thames, M. H. (2013). Teacher Education in Germany. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 49–62). Boston: Springer-Verlag.

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe Publishing. Retrieved from http://www.coe.int/lang-cefr

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*, 117–135.

Duijm, 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*, 501–527. https://doi.org/10.1177/0265532217712553

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd, Revised ed.). *Language Testing and Evaluation: Vol. 22*. Frankfurt a.M.: Peter Lang.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Abingdon, London, New York: Routledge.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, *20*, 281–307.

Herppich, S., Praetorius, K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, *76*, 181–193.

Hinger, B., & Stadler, W. (2018). *Testen und Bewerten fremdsprachlicher Kompetenzen. [Testing and evaluation of foreign language skills]*. *Narr Studienbücher*. Tübingen: Narr Francke Attempto.

Hochstetter, J. (2011). *Diagnostische Kompetenz im Englischunterricht der Grundschule: Eine empirische Studie zum Einsatz von Beobachtungsbögen. [Diagnostic competence in primary school English teaching: An empirical study on the use of observation sheets]*. *Giessener Beiträge zur Fremdsprachendidaktik*. Tübingen: Narr.

Hoppe, T., Renkl, A., Seidel, T., Rettig, S., & Rieß, W. (2020). Exploring How Teachers Diagnose Student Conceptions about the Cycle of Matter. *Sustainability*, *12*, 4184. https://doi.org/10.3390/su12104184

Hsieh, C.-N., & Davis, L. (2019). The effect of audiovisual input on academic listen-speak task performance. In S. Papageorgiou & K. M. Bailey (Eds.), *Global research on teaching and learning English: Vol. 6. Global perspectives on language assessment: Research, theory, and practice* (pp. 96–107). London: Routledge.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics*, *29*, 24–49.

DOI: https://doi.org/10.23770/

Jacob, A. (2012). Examining the relationship between student achievement and observable teacher characteristics: Implications for school leaders. *International Journal of Educational Leadership Preparation*, 1–13.

Keller, S. D., Jansen, T., & Vögelin, C. (2019). Can an instructional video increase the quality of English teachers' assessment of learner essays? *RISTAL*, *2*(1), 140–161 https:/doi.org/10.23770/rt1829

Kim, Y.-H. (2009). Exploring rater and task variability in second language oral performance assessment. In A. Brown & K. Hill (Eds.), *Tasks and Criteria in Performance Assessment* (91-110). Peter Lang.

Knoch, U. (2011). *Diagnostic writing assessment: The Development and Validation of a Rating Scale*. *Language Testing and Evaluation*. Frankfurt a.M.: Peter Lang GmbH Internationaler Verlag der Wissenschaften.

Kolb, A. (2011). Kontinuität und Brüche: Der Übergang von der Primar- zur Sekundarstufe im Englischunterricht aus der Perspektive der Lehrkräfte. [Continuity and breaks: The transition from primary to secondary English teaching from the teachers' perspective]. *Zeitschrift Für Fremdsprachenforschung*, *22*, 145–175.

Lenske, G. (2016). *Schülerfeedback in der Grundschule.: Untersuchung zur Validität. [Pupil feedback in the primary school: study on validity]*. Münster, New York: Waxmann.

Loibl, K., Leuders, T., & Dörfler, T. (2020). A Framework for Explaining Teachers' Diagnostic Judgements by Cognitive Modeling (DiaCoM). *Teaching and Teacher Education*, *91*, 103059. https://doi.org/10.1016/j.tate.2020.103059

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–71. https://doi.org/10.1177/026553229501200104

Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *Journal of Teaching English for Specific and Academic Purposes*, *6*, 179–189. https://doi.org/10.22190/JTESAP1801179M

Porsch, R., & Wilden, E. (2017). The Introduction of EFL in Primary Education. Challenges for EFL Teachers in Germany. In E. Wilden & R. Porsch (Eds.), *The professional development of primary EFL teachers: National and international research* (pp. 59–75). Münster: Waxmann Verlag GmbH.

Rieu, A., Loibl, K., & Leuders, T. (2020). Förderung diagnostischer Kompetenz von Lehrkräften bei Aufgaben der Bruchrechnung. [Promotion of diagnostic competence of teachers in tasks dealing with fractions]. *Herausforderung Lehrer*innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion*, *3*, 492–509. https://doi.org/10.4119/HLZ-3167

Shaw, S. D. (2007). Modelling facets of the assessment of writing within an ESM environment. *Research Notes*, *27*, 14–19. Retrieved from https://www.cambridgeenglish.org/images/23146-research-notes-27.pdf

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *3*, 743–762.

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2018). The teacher as examiner of L2 oral tests: A challenge to standardization. *Language Testing*, *35*, 217–238.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252. https://doi.org/10.1177/0265532212456968

Witzigmann, S., & Sachse, S. (2020). Verarbeitung von Hinweisreizen beim Beurteilen von mündlichen Sprachproben von Schülerinnen und Schülern durch Hochschullehrende im Fach Französisch. *Unterrichtswissenschaft*, *48*, 551-571. https://doi.org/10.1007/s42010-020-00076-6

Yan, X., & Ginther, A. (2018). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 67–88). Milton Park, Abingdon, Oxon, New York, NY: Routledge.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*, 31–50. https://doi.org/10.1177/0265532209360671

Stéfanie Witzigmann

worked many years as a secondary school teacher for the subjects French, Arts and Home Economics. 2001 she has been assigned to teacher training at different Universities of Education. Her research focuses on Content and Language Integrated Learning (CLIL), Video-based classroom research and empirical foreign language research. Since 2017 she is working on her habilitation in the project DiaCoM (Diagnostic Judgements by Cognitive Modeling) at the University of Education in Heidelberg.

Steffi Sachse

is a Professor of Developmental Psychology with a focus on language development. She teaches prospective special education teachers and her research focuses on dual language learning, identification of language delays or disorders and language interventions within the context of early education contexts.

DOI: https://doi.org/10.23770/