

Chapter 2

Literature Review

This chapter reviews the literature pertaining to the present study. As the whole research can be chronologically broken down into three main phases, covering (1) building an argument for embedding nonverbal delivery into speaking assessment, (2) the formulation and (3) the validation of the rating scale for group discussion in formative assessment, this chapter is accordingly organised into five sections, with the first section reviewing nonverbal delivery relating to the first phase, and the other four sections consecutively addressing the related literature concerning rating scale development and validation.

Specifically, the first section reviews the previous research with regard to nonverbal delivery. Instead of standing still in the arena of language assessment, this section of review will commence with a review on nonverbal delivery in other fields of research; thus, a dearth of the related studies can be felt in the context of language testing. The second section is more concerned with the conceptualisation of *communicative competence*, addressing the issue of what rationale the rating scale development in the case of the present study should be based on. In particular, a link between *nonverbal delivery* and *strategic competence* will be drawn so that a theoretical argument can be tentatively advanced to embed nonverbal delivery into speaking assessment. The third section, appertaining to the categorisations of rating scales in language assessment and the essentials of formative assessment, provides a leeway for determining the basic properties of the rating scale to be designed in this research. In response to the issue of rating scale validation, the fourth and fifth sections, respectively, dwell on the notions of *validity* and *validation*, and quantitative and qualitative approaches to be adopted for validating the rating scale proposed in this study.

2.1 Nonverbal Delivery

In retrospect, the meaning conveyance via nonverbal delivery might be dated back to Greek rhetoric, where Quintilian (AD 35-100), one of the first in recorded history, drew the research attention to the use of gesture. He distinguishes rhetorical delivery into *vox* (voice) and *gestus* (the use of gesture). In a quite similar vein,

Cicero (106-43 BC) particularly expounds on rhetorical skills and conceptualises *sermo corporis* (body language) or *eloquentia corporis* (eloquence of the body). However, the burgeoning of studying nonverbal delivery, such as gesture and eye contact, as a subject in its incubation stage unfortunately enticed limited academic attention afterwards, given the privileging of language in academia. It was not until the Cold War in the twentieth century that nonverbal delivery seemed to be renourished in the research scenario.

Despite a vicissitude of exploring nonverbal delivery as above briefed, its significance in communication has been well documented (Leathers and Eaves 2008). In particular, its communicative functions in specific social and cultural contexts, its impact on intercultural communication as well as its interface with verbal delivery are circumnavigated in a plethora of different disciplines. From a sociological perspective, claims are made that nonverbal delivery exerts great functional significance on society (Leathers and Eaves 2008) and that “the importance of nonverbal behaviour in overall communication effectiveness is obvious, and the difficulties in assessing the skills involved should not blind us to their significance” (Baird 1983, p. 33).

In addition to researching into the significance of nonverbal delivery in communication, how various nonverbal delivery channels convey meaning are championed by scholars such as Ekman and Friesen (1969, 1974), Goldin-Meadow and Singer (2003), Kendon (1981, 1996, 2004), Leathers and Eaves (2008) and McNeill (1979, 1992, 2000, 2005), whose studies will be unfolded and reviewed in the ensuing part on the representative channels of nonverbal delivery. More recently, social semioticians such as Martinec (2000b, 2001, 2004) and Hood (2007, 2011) have also systematised nonverbal delivery, such as gestures, from a systemic functional linguistics perspective. Their works, which will be discussed in-depth in the section concerning MDA approach, are referred to in this research when the rating scale is validated qualitatively.

Given the above, although nonverbal delivery is felt to play a crucial role in communication, there seems an extremely low profile of studies with regard to the employment of nonverbal delivery by EFL learners in their target language communication. In a limited number of such studies, Nambiar and Goon (1993) discover that assessors tend to assign lower grades when only candidates’ voice-recording is provided. By comparison, when assessors rate the same candidates’ performance via video-recording, where both verbal and nonverbal delivery are made accessible, the candidates with satisfactory nonverbal delivery are assigned higher scores because raters need to simultaneously focus on verbal utterance and extra-linguistic cues. Another study pertaining to the employment of nonverbal delivery is an interactional analysis by Neu (1990), who finds that EFL learners might exhibit their communicative competence effectively by synchronised gesturing.

Thus, it can be believed that the inclusion of nonverbal delivery in speaking assessment is intended to not only better discriminate candidates across a range of proficiency levels but also provide more comprehensive feedback for candidates in

relation to what potential progress can be made in their spoken English performance.

In order to render a fuller picture of the previous studies on nonverbal delivery, this section of review will continue with the concrete and representative manifestations of nonverbal channels, viz. eye contact, gesture and head movement (Junghem 1995, 2001). With these nonverbal channels reviewed below, it is anticipated that a stronger theoretical argument for embedding nonverbal delivery into speaking assessment can be built, thus paving the way for a forthcoming empirical argument to be advanced in this study.

2.1.1 Eye Contact

The central role of eye contact in nonverbal delivery has long been acknowledged. A host of researchers are devoted to studying the language of eyes and now arrive at a consensus that there may well be a language of the eyes with its own syntax and grammar (Webbink 1986). Janik et al. (1978) find that attention is focused on the eyes 43.4 % of the communication duration. When eye contact is investigated in a social context, more interest is invited in identifying how eye contact can make meanings in social interactions (Kendon 1967; Street 1993). For example, Bourne and Jewitt (2003) study various purposes of eye contact in young learners' English learning process. Besides, there are also extensive studies aiming at the roles of eye contact in the development of children's language and communication, indicating that eye contact is primal regarding its shared attention by both infants and adults (Tomasello 2003).

Leathers and Eaves (2008) list a total of seven functions that eye contact possibly serves. The first function is attentiveness. Argyle and Cook (1976) emphasise that mutual eye contact "has the special meaning that two people are attending to each other, [which] is usually necessary for social interaction to begin or be sustained" (p. 170). The enlargement of pupils can be an indication that listener's or speaker's attentiveness is accordingly promoted (Hess 1975). Second, persuasive function, with which the persuader wishing to be noticed as trustworthy must maintain eye contact while speaking and being spoken to by the persuadee (Burgoon and Saine 1978; Burgoon et al. 1986; Grootenboer 2006). Third, intimacy, conducive to establishing interpersonal relations, is another function. In interpreting this function, Hornik (1987) and Kleinke (1986) assert that the intensity of eye contact, or the duration of gaze, has a crucial role to play in developing intimacy between persons. Fourth, regulatory function, which refers to alerting the decoder that the encoding process is occurring and continuing by virtue of signalling the encoder whether listening and decoding are occurring, and by indicating when the listener is to speak (Ellsworth and Ludwig 1971; Kalma 1992). Fifth, eye contact can also serve an affective function. Eye contact, along with facial expression, is able to function as a powerful medium of emotional communication (Zebrowitz 1997), or as Schlenker (1980) concisely phrases, "the eyes universally

symbolise affect” (p. 258). Sixth, eye contact has its power function, which largely deals with eyes’ function of exerting authority, or that of performing mesmerisation (Henley 1977; Henley and Harmon 1985). Seventh, impression management function, as its name suggests, means speaker’s efforts in formatting either positive or negative impressions upon the addressees (see Iizuka 1992; Kleinke 1986).

However, it should be noted that the above taxonomy of communicative functions are viewed in such a broad social context that it might not be directly applicable to studying eye contact deployed by EFL learners. For instance, in language assessment context, where candidates perform their oral task, it would be less likely that there are occurrences of eye contact with intimacy or power function as almost no necessity can be felt in this particular setting. In addition, a few communicative functions might be overlapping or serve more than one function as above elaborated, in the case of which judging what function(s) a captured occurrence of eye contact serves might be complicated. The desolation of eye contact from its accompanying verbiage can be another drawback of the above taxonomy. Without synchronised verbal utterance, it would be a practical challenge to fathom what exactly eye contact attempts to convey.

When an occurrence of eye contact is observed and measured, Poggi (2001) proposes a set of measures to analyse eye contact from the perspective of bodily organs, roughly including eyebrows (inner part, medial part and outer part), eyelids (upper or lower), wrinkles and eye (humidity, reddening, pupil dilation, eye position and eye direction). Fine-grained as these measures are, it may be technologically demanding as the observation of various occurrences of eye contact in accordance with the above specified frame might be jeopardised by its complexity and judgment subjectivity. In real practice, when eye contact is measured in this study, in order to provide a leeway for the first phase of this study, where an empirical argument is tentatively built for embedding nonverbal delivery into speaking assessment, the descriptive analysis will be refrained from resorting to the detailed taxonomy of bodily organs. Instead, analyses will be largely based on candidates’ eye contact as is *de facto* presented, mainly from the angles of eye contact directionality and duration because both measures can tentatively help allow an observation of candidates’ frequency and intensity of the various referents they visualise (Cerrato 2005). When the occurrences of eye contact are described and analysed, the taxonomy by Leathers and Eaves (2008) above outlined will be referred to.

Nonetheless, when the rating scale is validated qualitatively, considering more explanatory power and applicability, eye contact will be probed into with an MDA approach, with basically an integrated framework drawn from the studies by Martinec (2000b, 2001, 2004) and Hood (2007, 2011). In such a context, not only the frequency and duration of eye contact as salient measures will be probed into, but also other vehicles carried via eye contact, such as eye contact shift, will also be focused on. The operationalised framework from Martinec’s (2000b, 2001, 2004) and Hood’s (2007, 2011) studies will be further expounded below in detail along with an elaboration on MDA approach in Sect. 2.5 of this chapter.

2.1.2 *Gesture*

Unlike eye contact, whose manifestations mainly refer to such issues as duration, directionality and intensity of pupil fixation, gesture can be instantiated via a plethora of different manifestations. Thus, the question of what constitutes a unit of gesture is contested, with compelling reasons offered for various perspectives. Within the field of nonverbal communication, gesture can be broadly defined as “any distinct bodily action that is regarded by participants as being directly involved in the process of deliberate utterance” (Kendon 1985, p. 215). Kendon (1996) further proposes that a gesture consists of “phases of bodily action that have those characteristics that permit them to be ‘recognised’ as components of willing communicative action” (p. 8). However, this begs the question of recognition by whom. In addition, there can be concerns in the subjectivity involved in identifying unambiguously what is willing communicative gesture. Kendon (2004) explains that a prototypical gesture passes through three phases, namely the preparation, the stroke and retraction, with the stroke phase being the only obligatory element. McNeill (1992) describes the stroke phase as “the phase carried out with the quality of ‘effort’ a gesture in kinetic term” (p. 375). He continues to argue that “[s]emantically, it is the content-bearing part of the gesture” (p. 376). With the above, when gesture is observed in this study, more foci will be placed on the meaning potential it makes though the judgement will basically follow. Kendon’s (2004) proposed prototypical gesture, with the stroke phase as the core.

Following formal instantiation of gestures, quite a few studies decipher what various gestures would supposedly convey in particular settings, viz. their emblematic or iconic meanings. However, they rarely touch upon more than an inventory of providing the respective verbal glosses in various social contexts (e.g. Barakat 1973; Creider 1977; Efron 1941; Green 1968; Saitz and Cervenka 1972; Sparhawk 1978; Wylie 1977), though efforts are also made in response to gestures’ role in generating thinking (Alibali et al. 1997), in enhancing teaching and learning for complex ensembles (Kress et al. 2001) and in coordinating with workplace discourse (Heath and Luff 2007). However, emblematic meaning alone does not constitute all the possible conveyance or function of gestures.

Ekman and Friesen’s (1969) taxonomy of gesture functions encapsulates emblems, illustrators, affect displays, regulators and adaptors. Emblems are gestures with a direct verbal translation consisting of a word or two with a precise meaning known by most of the members of a given culture; thus emblematic gestures are mostly speech independent. For instance, the OK sign made by a fist with the thumb pointing upward is a classic example of an emblem.

Illustrators are used to augment what is being said and to reinforce or de-intensify the perceived strength of emotions experienced by the communicator. Therefore, examples of illustrators can be signals for turn-taking in conversations (pointing at the next turn-holder with an upward palm) or baton (slamming of hand). Given the fact that gestures might be highly associated with the accompanying verbiage when being interpreted, they can be regarded as speech dependent.

The communication of affect displays or emotions is much more closely linked with facial expressions, postures and reflex actions than gestures, such as shivering. Therefore, the function of gesture in this aspect is discarded in this study due to practical and technological constraints.

Regulators, as described by Kendon (2004), are gestures that are habitual and mostly unintentional and that are used by interactants to exercise a mutual influence over the initiation and termination of spoken messages. Therefore, the judgment on gestures falling into this category also requests synchronised speech. In certain cases, such gestures are vital in the sense that interactants can be sensitive to each other's turn-taking prerogatives.

Adaptors, according to Ekman and Friesen (1969), are a source of involuntary information about the psychological states of individuals who exhibit them, which might showcase anxiety, nervousness, etc. Self-adaptors involve the manipulation of the enactor's body such as scratching. Alter adaptors are designed to psychologically or physically protect the enactor from others, which can include folding of arms. Object-focused adaptors involve the unconscious manipulation of objects such as tapping of pens. Therefore, as far as meaning conveyance is concerned, adaptor gestures are usually not recognisably communicative.

The above taxonomy classifies gestures by taking into account formal gestures, their communicativeness and their relation with psychological and physical reactions. To a certain extent, it has to be admitted that this taxonomy provides a comprehensive encapsulation with regard to what functions gesture might serve. However, akin to what is pointed out concerning the weaknesses of categorising eye contact functions above, this taxonomy may also be internally overlapping. For example, a pictogram, such as tracing the movement of signing a cheque when requesting for the bill, can fall into the category of illustrator when there is accompanying verbiage. However, in certain social contexts, such a gesture can convey the intended meaning even without any synchronised verbal utterance. Thus, the taxonomy's isolation from verbal language can be regarded as a main drawback.

In this study, gesture is observed with regard to the movement of hands and arms exclusively. This is because if movements by other bodily parts are also taken into account, it would turn out to be an almost endless inventory encompassing the movements of various bodily parts; what might be even more intriguing is that gesture, if defined too broadly, will be likely to trigger a confusion of hand/arm movement with other synchronised bodily movements as well as a complication for gesture transcription. Against this, unlike the prescribed practice of observing eye contact in light of its directionality and duration in the first phase of the present study, only the gestures with the involvement of hands and arms will be looked into. When the detected gestures are further analysed in relation to their communicative functions, this study will refer to Ekman and Friesen's (1969) taxonomy above reviewed.

However, it should be noted that when the rating scale is validated qualitatively, although the judgment on gesture occurrences still follows the observation of hands or arms, their meaning potentials will be analysed beyond Ekman and Friesen's

(1969) taxonomy in order to maximise the interpretability of various gestures. Therefore, an MDA approach with regard to gestures, particularly Martinec's (2000b, 2001, 2004) and Hood's (2007, 2011) frameworks, will be focused on. An integrated framework, considering whether a gesture is performative or communicative and how it realises metafunctional meanings, will be further clarified in the section of MDA below.

2.1.3 *Head Movement*

Dissimilar to a fervour that solely concentrates on eye contact and gesture, quite few studies, if not none, have been exclusively devoted to a third essential and conspicuous channel of nonverbal delivery, head movement. This nonverbal channel might be slightly akin to eye contact in the sense that the directionality of head movement, in most cases, naturally corresponds to that of eye contact. It is, however, different from gestures in that head movements, with a comparative scarcity in variedness, are overwhelmingly instantiated via head nod or head shake though other vertical or horizontal movements of the head, such as one-way leftward movement from a central position, can also constitute a basic occurrence of head movement under discussion.

In a limited number of studies, a revelation can be made concerning cultural influence on head movement (e.g. Maynard 1987, 1990; Weiner et al. 1972). For instance, head shake can be usually interpreted as negation or disagreement in the Chinese culture, whereas in certain other cultures, such an occurrence can also be understood as agreement (Matsumoto 2006). Let us take head nodding as another example, Jungheim (2001) deems it as a backchannelling signal "giving feedback to indicate the success or failure of communication" (p. 4), especially when interactants intend to (1) show agreement with what is said, (2) pay respect to other speakers, or (3) indicate that they are attentively listening to the speaker in the Japanese culture (see Maynard 1987, 1989, 1990; White 1989).

Considering a dearth of any existing framework concerning the communicative functions of head movement that this study can comfortably rests upon, in building an argument for incorporating head movement as one of the dimensions of nonverbal delivery in speaking assessment, Ekman and Friesen's (1969) aforementioned framework in its general application is tentatively referred to. Since the main purpose of that research phase would just discriminate candidates across the pre-determined proficiency levels, in terms of formal head movement, only head nod (generally interpreted as agreement) and head shake (generally interpreted as disagreement) that are semantically loaded will be investigated. When head movement as one subdimension of the rating scale descriptors for nonverbal delivery is validated, in addition to head nod or shake, more fine-grained head movements, such as vertical or horizontal movements of high frequency in an interval unit, are also taken into account following an integrated framework drawn from Martinec's (2000b, 2001, 2004) and Hood's (2007, 2011) research to be unfolded below.

The above provides a review on nonverbal delivery, with a particular view to the three most representative channels and what approaches this study will adopt in observing and analysing formal nonverbal delivery at different phases of the study. With this section of review addressed, it can be felt that nonverbal delivery, with its proven significance and saliency in communication, should be embedded into speaking assessment, where meaning making is realised not just from verbal language alone. The ensuing section will then review the notion of communicative competence and specifically indicate the role that nonverbal delivery legitimately plays in assessing EFL learners' communicative ability.

2.2 Communicative Competence

In order to more comprehensively and accurately evaluate the multifacets of EFL learners' proficiency, providers of a plethora of worldwide administered language proficiency tests have been aware of the importance of seeking for a sound theoretical rationale to account for what is to be measured (e.g. Grant and Ginther 2000; Schoonen et al. 2002; Wolf-Quintero et al. 1998). In the domain of oral assessment, where communication between the articulator and the addressee determines the evaluation on the candidates' performance, the notion of *communicative competence* that serves as a construct yardstick can never be underestimated. Therefore, the formulation of a rating scale, with a consideration of embedding nonverbal delivery as previously argued, should first of all take that notion into serious consideration, without the delineation of which a rating scale will remain groundless concerning what should be measured.

Chronologically, Chomsky (1965) from the outset conceptualised *competence*, which is regarded as the internal grammar of the speaker and the listener. Chomsky (1965) believes that it is the "ideal" language system that enables speakers to produce and understand an infinite number of sentences and to distinguish grammatical sentences from ungrammatical sentences. Linguistic competence is inclusive of such components as phonetics, phonology, syntax, semantics and morphology. Therefore, linguistic competence, as is explained by Chomsky (1965), can be deemed as an entity or system, far more abstract than the language system per se. In response to Chomsky's notion, Hymes (1972), considering the social dimension of language use, puts forward *communicative competence* with more emphasis on the social nature of language. As Hymes (1972) states, *communicative competence* refers to one's capability and awareness of knowing *when*, *where* and *how* to say *what* with *whom*. Although the notion witnesses great percussion in the applied linguistics community, its operationalisation is challenged (Canale and Swain 1980). In order to offset the weaknesses and improve the bases of how *communicative competence* can be interpreted, Canale and Swain (1980), and Canale (1983) bring forth *communicative competence* model, comprising four domains of competence to be observed. Afterwards, by critiquing and distilling the essence of other researchers' views on what *communicative competence* should be

construed, Bachman (1990), Bachman and Palmer (1996) eclectically put forward the model of *communicative language ability* (CLA), which has been credited as a widely recognised framework with new insights on the language ability. The most recent framework with regard to *communicative competence* is the conceptualisation of *communicative language competence* (Council of Europe 2001) as one of the by-products from *Common European Framework of Reference* (CEFR).

The above brief introduction on the notional evolution of *communicative competence* leads to a necessity that this section of review should outline, critique and compare the above models for reaching the fittest one to bolster the explanation of what domains should be measured in a speaking rating scale and why nonverbal delivery plays a crucial role in light of *communicative competence* assessment.

2.2.1 Hymes' Notion of Communicative Competence

2.2.1.1 Theoretical Groundings

The notion of *communicative competence* is termed by American sociolinguist Dell Hymes, who employs the terminology in a research article entitled *On Communicative Competence*, and defines the notion as “a knowledge of the rules for understanding and producing both the referential and social meaning of language” (1972, p. 270). As a matter of fact, such a notion is long incubated and might be traced back to Hymes' conceptualisation of *communicative competence* in other academic works, such as *The ethnography of speaking* (1962) and *The ethnography of communication* (1964), which include communicative event, constituents of communicative event, the interrelationship among the constituents as well as the expected knowledge and the abilities of a communicator. It can be said, therefore, Hymes' notion of *communicative competence* is progressively enriched, yet it was not until at an international conference on language development that this notion systematically came to the fore.

2.2.1.2 Components of *Communicative Competence*

Hymes (1972) asserts that one's capacity is composed of language knowledge and the ability to use language and that *communicative competence* consists of four parameters that included “communicative form and function in integral relation to each other” (Leung 2005b, p. 119). Concerning communication beyond Chomsky's (1965) demarcation between *competence* and *performance*, he proposes a framework comprising four following questions to explain what *communicative competence* should include.

(To what degree) is something formally possible?

(To what degree) is something feasible in virtue of the means of implementation available?

(To what degree) is something appropriate in relation to a context in which it is used and evaluated?

(To what degree) is something in fact done, actually performed, and what does its doing entail? (Hymes 1972, pp. 270–288).

The first question deals with what is possible considering the language form. Actually, what is possible refers to something acceptable within a formal system that is grammatical, cultural or communicative (Hymes 1972). However, *communicative competence* is not succinctly interpreted when what is possible stands alone; the second question, therefore, touches upon feasibility, such as the memory limitation and perceptual device, or as can be rephrased, concerns what is biologically and psychologically feasible. To illustrate this parameter, Royce (2007) renders an example, where a sentence itself may be grammatically well formed, yet can be so lengthy that it fails to convey what is intended. The third question is more concerned with the appropriateness of language use in particular settings, reflecting the sociological and pragmatic aspects of language use. The last parameter bears upon a communicator's knowledge of probabilities in the sense that whether what is conveyed is actually common determined by whether successful communication can be fulfilled.

Reaffirmed by Hymes' other works (1973, 1974, 1982), his proposition might be interpreted that *communicative competence* includes not only grammatical knowledge but also language user's ability to judge whether what is said is practical, appropriate and probable. That means a language user with expected communicative competence should be aware of the above parameters, and the most salient connotation of performance is "that of imperfect manifestation of underlying system" (Hymes 1972, p. 289).

2.2.1.3 A Critique on Hymes' Notion of *Communicative Competence*

Hymes conceptualises *communicative competence* mainly from an ethnographic perspective, and when further motivated within the community of applied linguistics, this notion exerts great impact on language teaching (e.g. Celce-Murcia et al. 1997; Savignon 1983, 1997). On the surface level, Hymes' (1972) notion of *communicative competence* seems to contradict *linguistic competence* by Chomsky (1965), yet it is not the real case. In fact, Chomsky views linguistic competence as a set of idealised knowledge system in a pure language community (Brumfit and Johnson 1979), while Hymes puts more weight on the contextualised use of language.

On top of that, both perceive language from different angles. As a descriptive linguist, Chomsky focuses more on how language is generated and comprehended and tries to establish *Universal Grammar* (Chomsky 1965) across different languages in human society, while little interest is felt in how language is supposed to be used. By contrast, Hymes, as an ethno- and sociolinguist, ruminates over language

operationalisation in a given context. Therefore, the former's effort is abstracting all languages into a condensed form of representation; whereas the latter analyses languages in their real use. Since their perspectives towards the language vary substantially, why their conceptualisation of language ability differs is understandable.

Although Hymes' notion of *communicative competence* sheds much light on how language ability should be interpreted, particularly attaching significance to language teaching, it is not without limitations. First, in a communicational setting, Hymes' notion seems to lay more stress on the speaker's part, somehow neglecting the interactiveness in communication (Canale and Swain 1980; Johnson and Johnson 1999). According to Hymes' explanation, a communicator should pay attention to such parameters as where and when to say what to whom, all of which depart from an initiator. However, in real communication, it should also be borne in mind that communication goes beyond one-way conveyance; it should be a mutual process as a result of meaning negotiation. What is also under concern in the communication process is the ability to take into consideration the target audience's ability and potential ideas, both at the affective and cognitive levels. Therefore, one limitation of this notion is that the interaction between communicators' internalised ideas as well as their awareness is somehow impaired (Riley 1996).

Second, Hymes' notion overemphasised the determining role that external setting plays in communication fulfilment. As Hymes (1972) explains, the core of language use is the degree to which language is used appropriately. Thus, when language is uttered, it should be not only grammatically acceptable, but also context-specific and suitable on the particular speech occasion (Richards et al. 1992). Although the external speech setting undeniably influences the communicators' selection of speech, its role on language ability might be felt as a determinant if overstressed. Given the above critique, context and speech should not be on an absolute vis-à-vis relation; the selection of speech, though context-specific, does not bear a fixed pattern of expression.

2.2.2 Communicative Competence Model

2.2.2.1 Theoretical Groundings

As divulged above, the notion put forward by Hymes (1972) exerts great impact on language teaching, yet the four parameters are mainly challenged by virtue of their operationalisation. Although plenty of studies with *communicative competence* as a point of departure manage to apply the notion to language teaching, in such domains as syllabus design (Munby 1978) and language classroom teaching (Savignon 1983; Widdowson 1978), such application largely concentrates on a micro-basis. Against this, Canale and Swain (1980) contrive a model with more pertinent foci on the overall reflection of *communicative competence*, comprising *grammatical competence*, *sociolinguistic competence* and *strategic competence*. Later, Canale (1983) adds *discourse competence* for the model expansion.

2.2.2.2 Components of *Communicative Competence Model*

Conjured up with four components, Fig. 2.1 reflects an integration of *communicative competence* model by Canale and Swain (1980) and Canale (1983). Within this model, *grammatical competence* can be understood to include “knowledge of lexical items and rules of morphology, syntax, sentence-grammar semantics and phonology” (Canale and Swain 1980, p. 29), a level mainly pertaining to the comprehension and production of surface meaning as well as the ability to construct well-formed sentences. Therefore, the competence in this respect reflects the knowledge and skills needed to correctly understand and accurately express what is intended to convey (Canale 1983).

As Canale and Swain (1980) explain, *sociolinguistic competence* includes two subcomponents: sociocultural rules of use and rules of discourse, and “knowledge of these rules will be crucial in interpreting utterances for social meaning, particularly when there is a low level of transparency between the literal meaning of an utterance of the speaker’ intention” (p. 30). If these rules are violated, the speaker can be judged as lacking the expected sociolinguistic competence. The significance of *sociolinguistic competence* can be demonstrated by an understanding that *grammatical competence* alone cannot constitute *communicative competence* because successful achievement of communication should go beyond grammatical rules and include contextualised appropriateness.

Strategic competence, or *coping strategy* (Stern 1978), undergoes slightly conceptual extension in the model formulation as the initial *communicative competence* model by Canale and Swain (1980) defines it as consisting of “verbal and nonverbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient competence” (p. 30). However, Canale (1983) further broadens the ends of *strategic competence* as also “[enhancing] the rhetorical effect of utterances” (p. 339). From the extended definition, it might be felt that *strategic competence* does not merely serve a compensatory purpose, but enhances speech production as well.

Discourse competence, emerging as the fourth component in the revised model by Canale (1983), falls into an ability to encompass grammatical forms and semantic meanings to construct a text that surpasses sentential level. Text organisation, both in written and spoken forms, cannot be realised without discourse competence. In addition, discourse competence in conversation can be regarded as a

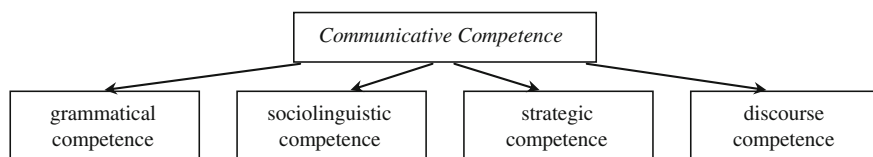


Fig. 2.1 *Communicative Competence Model* (Canale and Swain 1980; Canale 1983)

main point of departure in discourse analysis (e.g. Hatch 1978; Larsen-Freeman 1980; Richards and Schmidt 1983; Sinclair and Coulthard 1975). Particularly, discourse competence can reflect the extent to which EFL learners are competent at initiating, maintaining and terminating a conversation or discussion.

2.2.2.3 A Critique on *Communicative Competence Model*

Admittedly, *communicative competence* model, with an incorporation of Hymes' (1972) notion of *communicative competence*, introduces new components, viz. *discourse competence* and *strategic competence*. It is groundbreaking because it runs counter to Oller's hypothesis that language ability is a unitary construct (see Oller 1983; Oller and Hinofotis 1980). On top of that, this model refutes Chomsky's (1965) notion of *competence* and points out that *competence* fails to account for sociolinguistic appropriateness in a given context above discussed. Shohamy (1996) is also positive about this model as it brings non-linguistic elements into the conceptualisation of communicative competence. In addition, the primacy of *linguistic competence* in this model is also well represented (Jackendoff 1983). Therefore, the influence of *communicative competence* model can be felt in the applied linguistics arena, where abundant studies are somewhat assimilated to this model (e.g. Bachman and Palmer 1982; Kasper and Rose 2002; O'Malley and Chamot 1990; Skehan 1995; Spolsky 1989b; Swain 1985; Tarone and Yule 1989; Verhoeven 1997).

Notwithstanding *communicative competence* model broadens the construct of language ability by absorbing in other components, such as *discourse competence* and *strategic competence*, there is still flaw in the model per se. On condition that the whole model is scrutinised with a top-down approach, its components seem to be powerless to manifest their respective roles, and such behavioural objectives lead to ill-defined domains and problems for testing (Popham 1990). In other words, how the four components interact internally and whether all the components are equally quintessential remain to be specified. If how the components function is not explicitly stated, by default they can only be regarded as independent, dissimilar to a sound model whose components are inter-woven and consummated.

2.2.3 Communicative Language Ability Model

2.2.3.1 Theoretical Groundings

In the early 1990s, Lyle F. Bachman, an American applied linguist, based on the critique on the weaknesses of Lado's (1961) and Carroll's (1961, 1968) interpretations on language ability, develops the prevailing models posited by Halliday (1976), van Dijk (1977), Hymes (1972, 1973, 1982), Savignon (1983), Canale and Swain (1980) and Canale (1983) and conceptualises a new model, where

communicative competence should be constructed as “consisting of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualised communicative language use” (Bachman 1990, p. 84). As Bachman (1990) points out, the underpinnings of the CLA model actually reflect what Candlin (1986) describes *communicative competence* as

[the] ability to create meanings by exploring the potential inherent in any language for continual modification in response to change, negotiating the value of convention rather than conforming to established principle. In sum, a coming together of organised knowledge structures with a set of procedures for adapting the knowledge to solve new problems of communication than do not have ready-made and tailored solutions. (p. 40)

The development of the CLA model is aggregated by a multitude of studies, among which the main sources include *Fundamental considerations in language testing* (Bachman 1990), *Language testing in practice* (Bachman and Palmer 1996) and *What does language testing have to offer* (Bachman 1991). Revolving around the above three sources, the following is devoted to reviewing the model components, in conjunction with a critique.

2.2.3.2 Components of the CLA Model

Bachman (1990) gestates the construct of CLA on the basis of three core components, viz. *language competence*, *strategic competence* and *psychophysiological mechanisms*. Figure 2.2 illustrates the componential breakdown and the internal correlation of the model. As is shown, knowledge structures refer to language users’ social and cultural knowledge and the general knowledge about the material world,

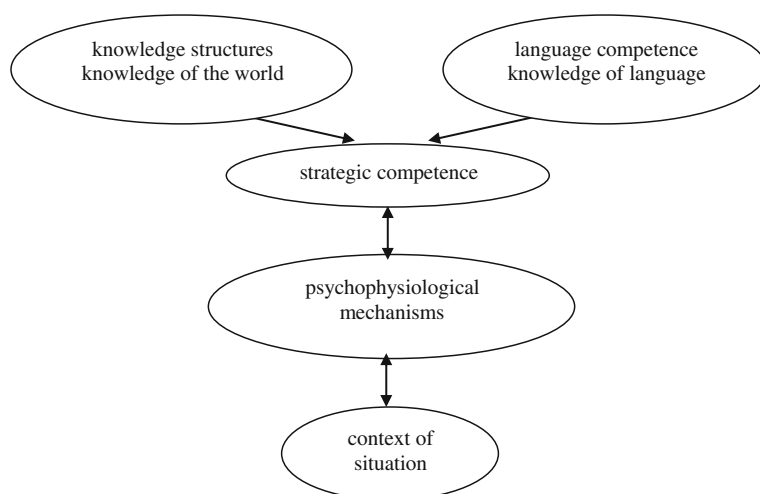


Fig. 2.2 CLA components in communicative language use (Bachman 1990, p. 85)

whereas the context of situation includes the reciprocal sides of the communication, situation, topic and purpose (Bachman 1990). In addition to the knowledge in both regards, the three core parts constituting the CLA model are *language competence*, *strategic competence* and *psychophysiological mechanisms*, all coordinating with the knowledge structures and situation context to depict an overall picture of *communicative competence*.

Language Competence

Bachman (1990) subcategorises *language competence* into *organisational competence* and *pragmatic competence*, under which each can be further broken down into several subcomponents, as showcased in Fig. 2.3.

1. Organisational competence

The component of *organisational competence* in the CLA model is influenced by Widdowson's (1978) demarcation between *use* and *usage* and Halliday's (1978, 1985), Halliday and Matthiessen's (2004) systemic functional grammar, both governing language users' selection of words on micro- and macro-bases. *Organisational competence*, therefore, determines textual organisation, which involves the abilities to control the linguistic form, to produce and identify grammatically correct sentences (Bachman 1990).

Further divided, *organisation competence* comprises *grammatical competence* and *textual competence* (Bachman 1990). The former refers to an ability to compose discourse or sentence with word as a basic unit. This competence boils down to the capability of mastering grammatical rules, ranging from vocabulary, morphemes, syntax to phonology or graphology. The latter combines the organised discourse or sentences into a text on a larger scale, which enables users to connect a number of

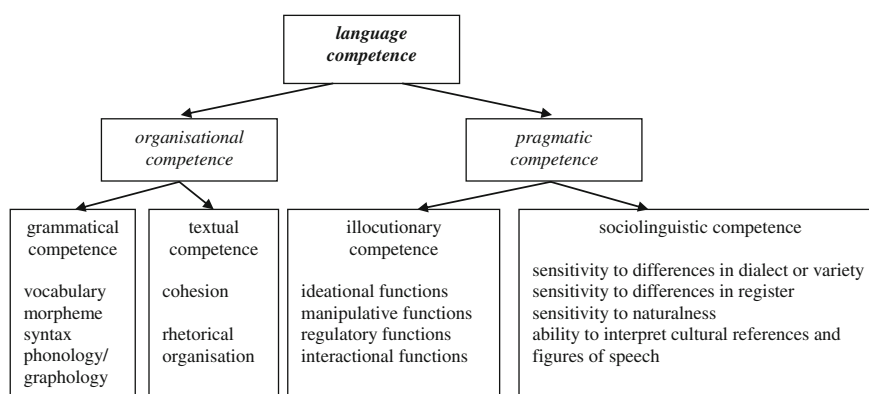


Fig. 2.3 Subcomponents of *language competence* in the CLA model (Bachman 1990, p. 87)

clauses in accordance with the rules stipulating cohesion and rhetorical organisation. Some cohesive devices are salient, such as lexical connection, reference, substitution and omission (Halliday and Hasan 1976); there are also devices with implied functions, regulating the occurring sequence of new and given information in a text. Rhetorical organisation in the CLA model mainly touches upon methods, such as narration, description and classification (McCrimman 1984).

2. *Pragmatic competence*

Pragmatic competence is more concerned about how discourse, clause and intention realise their meanings and functions in a particular context, or as Bachman (1990) pinpoints, this competence deals with “the relationships between (the) signs and their referents on the one hand, and the language users and the context of communication, on the other” (p. 89). Pragmatic competence can be split into two subcomponents: *illocutionary competence* and *sociolinguistic competence*.

Illocutionary competence encompasses “the knowledge of the pragmatic conventions for performing acceptable language functions” (Bachman 1990, p. 90). This concept bears much relevance with Speech Act Theory (Searle 1969), which includes such functions as assertion, warning and imagination. As is shown in Fig. 2.3, illocutionary competence is further classified into four groups: *ideational*, *manipulative*, *heuristic* and *imaginative*. The ideational function is used to “express meaning in terms of our experience of the real world” (Halliday 1973, p. 20), including the language use either to express propositions or to exchange information about such knowledge. The manipulative function is mainly applied to affect the world around us. The abilities falling into this group include the instrumental function used to handle things such as making suggestions, requests, commands and warnings; the regulatory function used to control others’ behaviour by either controlling or formulating the persons or the objects in the environment; and the interactional function, which serves to form, maintain, or change interpersonal relationships. The heuristic function is applied to share with others our knowledge of the world, which frequently occurs in such acts as teaching, learning, problem solving and conscious memorising. The imaginative function enables to create or extend humour or aesthetic values by constructing and communicating fantasies, creating metaphors, attending plays and so forth (Bachman 1990).

Sociolinguistic competence, as another part of *pragmatic competence*, is defined as “the sensitivity to, or control of the conventions of language use that are determined by the features of the use context” (Bachman 1990, p. 94). The sensitivity referred to is linked with the response to which communicators are able to cognate the dialect, language variety, differences in register (Halliday et al. 1964), cultural references and figures of speech as well as the degree to which speakers can appropriately and naturally generate the utterances expected in the target language in a specific language-use context (Pawley and Syder 1983).

Strategic Competence

Akin to one of the components in *communicative competence* model (Canale and Swain 1980; Canale 1983) aforementioned, Bachman (1990) also terms another part that contributes to CLA as *strategic competence*, which deems language use as a dynamic process embedding communicators' judgments, identification, negotiation of relevant information in a particular context. All of the cognitive as well as psychological processes are inter-woven together. Bachman (1990), by distilling the essence of communication strategy previously elaborated, puts forward a much broader concept of strategic competence, comprising of three metacognitive strategies: *assessment strategies*, *planning strategies* and *execution strategies*. Bachman's belief is that in any language-use context, these metacognitive strategies co-occur with all the aspects of language use as an entity, where there is supposed to be no subordinate to another.

1. Assessment component

Bachman (1990) outlines four aspects of assessment strategy, the integration of which are concerned with an ability to decide on the particular language as the desired information conveyance channel, to select what sort of language variety or dialect is able to achieve communicative effectiveness in a particular context, and to identify the interlocutors' knowledge and ability and the degree to which communication is ultimately fulfilled. The particular importance Bachman attaches to is how communicators themselves know the interlocutors' knowledge and ability, as underscored by Corder (1983) that

[t]he strategies adopted by speakers, of course, depend upon their interlocutors...[w]hat we attempt to communicate and how we set about it are determined not only our knowledge of the language but also by our current assessment of our interlocutor's linguistic competence and his knowledge of the topic of discourse. (p. 15)

2. Planning component

The *planning* strategy enables communicators to formulate a plan in realisation of communicative purpose with certain language knowledge selected. If the speakers are interacting in their mother tongue, the knowledge needed derives from the ability in the first language. Nonetheless, if the communication takes place in a bilingual or a second/foreign language setting, what is needed regarding language knowledge can be switched to the abilities either transferred from their first language or those gradually fostered in their interlanguage. The main functions of planning strategy are to select the relevant language knowledge, schemata and mind mapping.

3. Execution component

The strategy of *execution* is a critical stage before communication is realised under the co-functioning of *psychophysiological mechanisms* (see Section "[Psychophysiological Mechanisms](#)"). For instance, in the receptive channel of language input, visual and auditory faculties shall be applied. Bachman (1990)

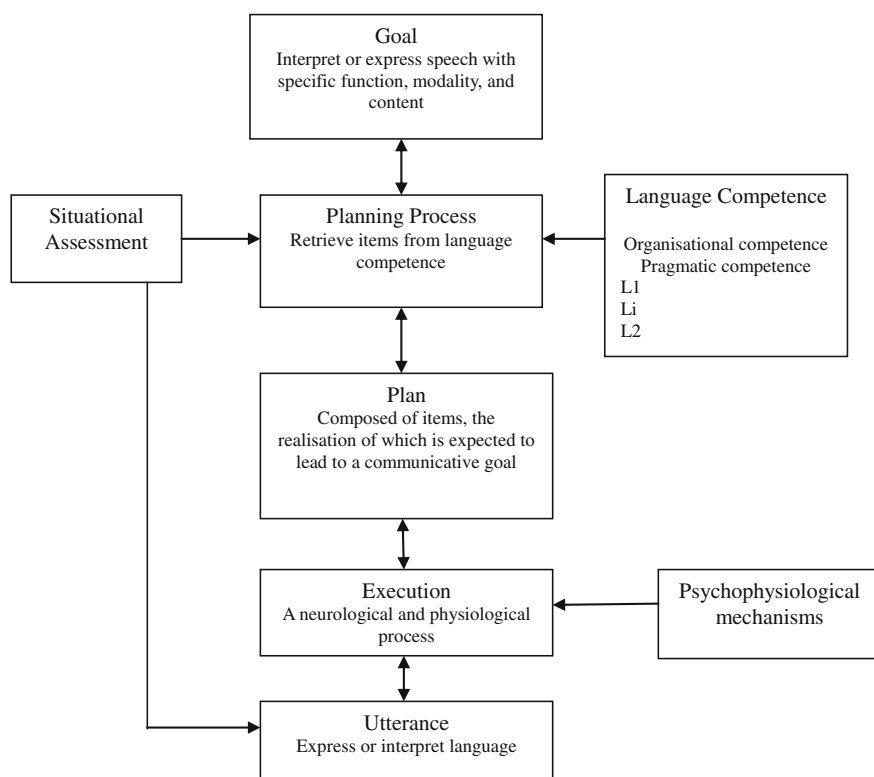


Fig. 2.4 A model of language use (Bachman 1990, p. 103)

holds that the three components of strategic competence, in effect, co-exist in the whole process of communication, interacting with language ability and language-use context. Having integrated the flow chart originally taken from Færch and Kasper's model (1983), Bachman (1990) visualises how the above components and the other parts of the CLA model co-function, as illustrated in Fig. 2.4.

As can be seen, along the central line from *goal* to *utterance*, both language competence and psychophysiological mechanisms exert their respective influences on the planning process and execution stage. The whole process also witnesses the existence of situational assessment impacting planning process and utterance because communicators need to make situation-specific judgments on what communication channels to be adopted to optimise meaning conveyance. Bachman (1991) further contends that language knowledge can only be realised with the involvement of *strategic competence*. Therefore, the strategies concerning *assessment*, *planning* and *execution* are intrinsically interdependent.

Psychophysiological Mechanisms

Psychophysiological mechanisms may be viewed as a third component of the CLA model. Bachman (1990), Bachman and Palmer (1996) associate biological mechanisms with language production and think these mechanisms are “the neurological and physiological processes that are included in the execution phase of language use” (Bachman 1990, p. 107). For instance, when test-takers are required to describe a picture, they not only use linguistic competence to construct sentences, but also employ their visual skill to obtain the non-linguistic information in the picture, their auditory skill to obtain the information in the examiner’s instructions, and their articulatory skill to pronounce the words correctly and to provide appropriate stress and intonation. However, this competence component has been rarely explored either theoretically or empirically in-depth.

2.2.3.3 A Critique on the CLA Model

The above review on the CLA model leads to an embodiment of the interaction of language knowledge within the context of language use, which integrates language knowledge and a series of cognitive strategies. Such a notional presentation is characterised by more explanatory power as the CLA model is epitomised as a leap forward compared with the Canale and Swain’s *communicative competence* model. The CLA model embeds strategic competence and regards it as not just serving a compensatory function, which, to a certain extent, echoes the modified model by Canale (1983). More importantly, the CLA model recognises the roles of cognitive strategies and pragmatic competence, together with their impact on the realisation of *communicative competence*. On the whole, the CLA model has been theoretically sound and empirically verified and has been merited as the state-of-the-art representation (Alderson and Banerjee 2002).

Despite its prevalence, the CLA model is without caveats. McNamara (1990) believes that when performance tests are taken into account, this model seems to be less operationalisable because raters are very likely to assign unbalanced weightings to a particular component of language knowledge. Upshur and Turner (1999), on the same side, believe that a cure-all, construct-only approach to evaluating complex performance may cover the influences that task context and discourse may have on how raters interpret rating scales in the assessment of communicative competence because such a disproportion may beget a biased focus on one component only. In a similar vein, Purpura (2004), when addressing the subcomponent of grammatical competence, contends that since “meaning” plays a central role in the CLA model, the model per se would be more consolidated by how “meaning” should be theoretically defined and how grammatical resources can be employed to express denotative and connotative meanings on the one hand and a variety of pragmatic meanings on the other. Chapelle (1998), from an interactionist perspective towards construct definition, critiques that the CLA model is defined and operated more on the trait basis, and further states that, “[t]rait components can no

longer be defined in context-independent, absolute terms, and contextual features cannot be defined without reference to their impact on underlying characteristics” (p. 43).

In addition to the above, the concrete components of the CLA model seem to be unstable with slightly different wording or naming from the different sources, with which the CLA model is constructed. In Bachman and Palmer’s (1996) *Language testing in practice*, topical knowledge, language knowledge and personal characteristics are interrelated with strategic competence and are all included in the language-use context. This somewhat differs in wording and diverges from the model in its earlier version. Another point is that psychophysiological mechanisms do not have a place in Bachman’s (1990, 1991) description of the CLA model. Likewise, Bachman seems to be subtly uncertain about the categorisation of semantic knowledge as he groups the knowledge of this aspect in the first two versions (Bachman 1990, 1991), whereas the final version of the CLA model (Bachman and Palmer 1996) witnesses the regression of semantic knowledge as part of illocutionary competence and sociolinguistic competence. Despite a few minor weaknesses and possible impracticality of the CLA model above outlined, it has to be admitted that the model per se features a comprehensive, systematic and interrelated reflection of what communicative ability is supposed to be construed. After the review on the last model concerning communicative competence and an integrated review on all the models, more justifications will be rendered as to why this study would refer to the CLA model as the theoretical base when a rating scale with a consideration of incorporating nonverbal delivery as a dimension in assessing speaking is formulated.

2.2.4 Communicative Language Competence Model

2.2.4.1 Theoretical Groundings

With almost the same name yet discrepant academic background with the CLA model, *communicative language competence* (CLC) Model (Council of Europe 2001; North 2010a, b) is a by-product of CEFR (Council of Europe 2001). It is based on the initial considerations of providing a common basis for language syllabi, curriculum guidelines, examinations, textbook and so on, and of relating a Europe credit scheme to fixed points in a framework (van Ek 1975). This framework is inspired by the documents such as *Threshold*, *Vantage*, *Waystage*, *Breakthrough*, *Effective Operational Proficiency* and *Mastery* (Alderson 2010). It is then developed with detailed descriptors for each level of expected behavioural descriptions of language ability in various domains (Little 2006). Therefore, in terms of the theoretical groundings, it is more a political and educational demand than an academic motive though the above documents effectively guide the model formulation and the conceptualisation of *communicative competence* in its own right.

2.2.4.2 Components of the CLC Model

Stipulated by Council of Europe (2001), the CLC Model consists of three domains: “linguistic competences, sociolinguistic competences and pragmatic competences” (p. 108), as is outlined in Fig. 2.5.

As is illustrated, *linguistic competences* are concerned with the “knowledge of and ability to use language resources to form well structured messages” (Council of Europe 2001, p. 109), which can be subcategorised into *lexical competence*, *grammatical competence*, *semantic competence*, *phonological competence* and *orthographic competence*. Judging from the interpretation of these subcomponents in linguistic competences, *linguistic competences* bear much relation with *grammatical competence* of the CLA model, reflecting a mastery of language knowledge in a traditional and narrowed sense.

Sociolinguistic competences refer to the “possession of knowledge and skills for appropriate language use in a social context” (Council of Europe 2001, p. 118). They include linguistic markers of social relations, politeness conventions, expressions of folk wisdom, register difference as well as dialect and accent. *Sociolinguistic competence* in the CLA model is labelled within pragmatic competence; therefore, this subcomponent is somehow elevated as one of the core components in the CLC Model, in the case of which the social realisation of language use is emphasised.

How *pragmatic competences* are defined is largely based on the description of how its subcomponents are made up of. *Pragmatic competences* embed *discourse competences* (abilities to organise, construct and arrange knowledge), *functional competences* (abilities to generate communication-inductive meaning) and *design competences* (abilities to sequence the messages in accordance with schemes and interactiveness) (Council of Europe 2001). Given this, an understanding can be

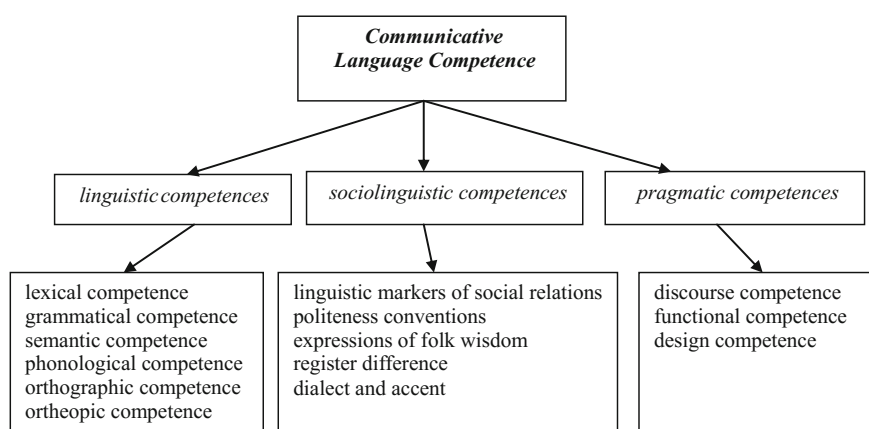


Fig. 2.5 Components of the CLC Model (Council of Europe 2001, pp. 108–129)

reached that *pragmatic competences* in the CLC Model seem to indicate a broader sense of pragmatics, with partial anchoring with *pragmatic competence* in the CLA model.

2.2.4.3 A Critique on the CLC Model

As one of the by-products of CEFR, the CLC Model has provided a Europe-specific reference for language teaching, learning as well as assessment. Council of Europe (2001) claims CEFR to be comprehensive as “it should attempt to specify as full a range of language knowledge, skills and use as possible...and all users should be able to describe their objectives, etc. by reference to it” (p. 7). In that sense, the CLC Model is an important point of reference, but not an instrument of coercion, nor for accountability (Alderson 2010).

Nevertheless, how *communicative competence* is defined in the CLC Model etches much flaw to the model per se. First, the construct of language ability in this model or the descriptors of different levels are basically drawn from teachers’ and learners’ perceptions, with little empirical research or theoretical basis. In addition, the descriptors take “insufficient account of how variations in terms of contextual parameters may affect performances by raising or lowering the actual difficulty level of carrying out the target ‘can-do’ statement” (Weir 2005, p. 281). Although the CLC Model refers to such documents as *Waystage*, *Threshold* and *Vantage*, as previously mentioned, they are barely different from each other (Alderson 2010). While the CEFR claims to cover both aspects of proficiency and development in its six ascending levels of proficiency, it fails to do so consistently (e.g. Alderson et al. 2006; Hulstijn 2011; Norris 2005). A number of researchers (e.g. Cumming 2009; Fulcher 2004; Hulstijn 2007; Spolsky 2008) express concerns regarding the foundation of the CEFR system. Spolsky (2008), for instance, criticises the CEFR as “arbitrary” standard to produce uniformity, whereas Cumming (2009) points out the dilemma of the imprecision of standards such as the CEFR “in view of the complexity of languages and human behaviour” (p. 92).

Second, a comparison between the CLC Model and the previously highlighted models reaches the finding that the CLC Model excludes *strategic competence*, which, though partially included in *pragmatic competences*, is largely abandoned. Therefore, the above-mentioned *pragmatic competences* in a broader sense are no longer how the pragmatic aspect of language use is conventionally conceptualised. As reviewed above, *strategic competences*, playing a quintessential role in language use, should be a subcomponent attached to communicative language ability as a whole. Such abandonment would also cause infeasibility for test development or validation, whose rationale rightly resides in the CLC Model (see Alderson 2002; Morrow 2004).

Third, the naming of *sociolinguistic competences* itself might be problematic. This is because the literal sense of this ability suggests that they seem to be naturally subordinate to *linguistic competences*, which is another core component of the CLC Model that overrides *sociolinguistic competences*.

2.2.5 *An Integrated Review on Communicative Competence*

Endeavouring to seek the fittest model for designing a rating scale with nonverbal delivery included as a dimension, the above review is devoted to an elaboration on *communicative competence*, covering a range of the background of the notion (Hymes 1972) and the subsequent notional evolution (Bachman 1990; Bachman and Palmer 1996; Canale and Swain 1980; Canale 1983; Council of Europe 2001). However, in the process of notional development, there are admittedly also other frameworks relating to *communicative competence*. Celce-Murcia et al. (1997), for instance, extend *communicative competence* model by further dividing sociolinguistic competence into sociocultural competence and actional competence. With regard to the CLA model renovations, Douglas (2000) proposes a model with a particular view to the language use for specific purposes, in the case of which professional or topical knowledge is equally emphasised. Likewise, Purpura (2004) develops an extended model based on the CLA model, where “a model of language knowledge with two interacting components: grammatical knowledge and pragmatic knowledge” (Purpura 2008, p. 60) is proposed.

An in-depth analysis of the above-modified models or frameworks, though excluded here, would be sufficient to constitute an understanding that they are characterised by either domain-specificity or further breakdown dissolving from the CLA model. Therefore, it can be justifiable that the CLA model serves as an umbrella model that covers the notions and models just briefed.

A retrospective review on *communicative competence* model, the CLA model and the CLC Model on a chronological continuum, as illustrated in Fig. 2.6, can provide a better understanding of *communicative competence* and which model can be judged as the fittest. The components with the linkage by arrows as an indication from a developmental point of view mean that they are basically of the same conceptual referents. It can be observed that when the notion is ushered into the CLA model, as the arrows in the figure point, its components are most comprehensive and inclusive, with integrated interactions and mechanisms between different components. Notably, the CLA model substantiates the component of *strategic competence* and incubates *psychophysiological mechanisms*, though the related studies on the latter are unavailable.

When the notion evolves into the CLC Model, *strategic competence* disappears; *design competences* in the CLC Model, judging from the definition that is previously mentioned, have only seemingly partial connection with *psychophysiological mechanisms* in the CLA model, as indicated by a dotted arrow in Fig. 2.6. Therefore, it can be felt that the absence of *strategic competence* in explaining what *communicative competence* is might give rise to the model caveats; thus, it can be naturally argued back that the CLA model should be selected as the fittest model with inclusiveness and explanatory power. All these can enhance a justification that the CLA model is the most appropriate to be the theoretical rationale, based on

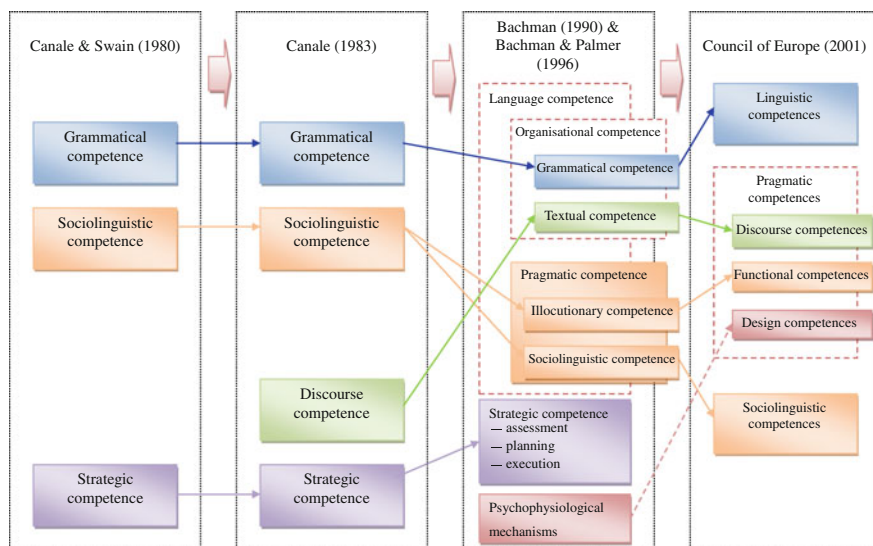


Fig. 2.6 Notional evolution of *communicative competence*

which a rating scale with nonverbal delivery assessment included is to be proposed in the present study.

Gearing the CLA model to the basic properties of spoken language, it can also be felt that this model can be an integrated epitome of *accuracy*, *fluency* and *appropriateness*. Regarding *accuracy*, Skehan (1996) broadly defines it as “the extent to which the language produced conforms to target language norms” (p. 18). In that sense, it not only covers the accurate use of individual word, but also the exactness of phrase, sentence and discourse as a whole. Viewed from the CLA model, this property of spoken language can be reflected from *organisational competence* under *language competence*. In terms of *fluency*, although different researchers vary their perspectives in defining this property (e.g. Brumfit 1984; Færch et al. 1984; Lennon 1990; Sajavaara 1987; Schmidt 1992), this notion is usually associated with three factors: *coherence*, *continuity* and *acceptability*, all of which can also be covered by either *textual competence* or *pragmatic competence* in the CLA model. Though there is no established definition for the property of *appropriateness*, it can be tentatively understood as the extent to which utterances approximate the conventions in a given social context. Therefore, it again falls into the domain of *pragmatic competence* of the CLA model.

In addition to an integrated review on the notional evolution of *communicative competence* as well as an analysis of the common grounds between the CLA model and the basic properties of spoken language, more comments and reflections on the above models as a whole are rendered below. First, from its inception, the notional transmutation actually accompanies a discussion of whether *communicative competence* should be a unitary concept or a multi-componential one; in the case of the

latter, issues arise as to what components can best represent and constitute the construct of the notion. As showcased from the above elaborations, *communicative competence* is multi-componential; thus, when *communicative competence* is assessed, EFL learners are supposed to be assessed in different domains. This also echoes the philosophy of the present study in that a rating scale, particularly in the context of formative assessment, should be designed as analytic instead of holistic. This issue will be re-addressed and further resolved in the next section of this chapter. Second, Connor and Mbaye (2002) pinpoint that a sound model of *communicative competence* offers a convenient framework for categorising components of written and spoken discourse, in which all the possible competences should be reflected in the scoring criteria. A substantial number of test designers also indeed adopted the CLA model to be the basis of rating scale design (e.g. Clarkson and Jensen 1995; Grierson 1995; Hawkey 2001; Hawkey and Barker 2004; McKay 1995; Milanovic et al. 1996). To that end, the selection of the CLA model in the present study can be further justified.

Therefore, following the CLA model, the rating scale to be proposed will comprise two broad dimensions: *language competence* and *strategic competence*. The former is quite self-explanatory within the model with regard to what detailed assessment domains should be looked at; however, strategic competence seems not to be that observable because it is explained in terms of three metacognitive strategies in the model. In that context, enlightened by the definition of strategic competence, which mainly concerns how a speaker resorts to non-linguistic means to sustain communication, and also informed by the review on nonverbal delivery in the previous section, the present study attempts to incorporate nonverbal delivery into the rating scale as one observable dimension to correspond to strategic competence. Although it has to be admitted that nonverbal delivery alone cannot depict a full picture of strategic competence, it can to a large extent provide a detectable and representative profile of candidates' performance in speaking assessment.

With the above, it can be felt that incorporating nonverbal delivery into speaking assessment is well grounded because it is intrinsically rooted in strategic competence in the CLA model. Yet, such a perception largely remains on the theoretical level. If an argument for embedding nonverbal delivery into speaking assessment can be built via an empirical study to verify that the competence in this aspect can indeed discern candidates across a range of proficiency levels, such an argument can be further consolidated. It can also pave the way for the formulation and validation of the rating scale with such a consideration. As aforementioned, this argument will be made in the first phase of this study.

2.3 Rating Scale and Formative Assessment

This section will touch upon the literature concerning rating scale and the context of the rating scale to be proposed in this study, viz. formative assessment. Therefore, this section will review four respects: (1) what is a rating scale in language

assessment? (2) what are different categorisations of rating scales? (3) what is formative assessment and how can it benefit EFL learners? (4) what type of rating scale best accommodates the context of formative assessment? The end of this section will integrate the review to summarise the wide-ranging properties of the rating scale to be proposed.

2.3.1 *Rating Scale*

Scales in language assessment are labelled in various names. Alderson (1991) provides a number of alternatives, such as “band scores, band scales, profile bands, proficiency levels, proficiency scales, [and] proficiency ratings” (p. 71). Similarly, de Jong (1992) also terms rating scales as “guidelines, standards, levels, yardsticks, stages, scales or grades” (p. 43). However, no matter how it is named, considering its function, a rating scale equals to a yardstick against which learners’ performance can be measured in “a hierarchical sequence of performance ranges” (Galloway 1987, p. 27). In describing a rating scale in language assessment, McNamara (2000) suggests that it is a series of ascending descriptions of remarkable features of performance at each language level; on the other hand, Luoma (2004) also states that rating scales are the reflections of test developers’ understanding and expectation of what test construct is. Thus, they “form part of their definition of the construct assessed in the test” (Luoma 2004, p. 59). Davies et al.’s (1999) definition seems more inclusive when they propose that

a rating scale is the description of language proficiency consisting of a series of constructed levels against which a language learner’s performance is judged. Like a test, a proficiency (rating) scale provides an operational definition of a linguistic construct such as proficiency. Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or bands are commonly characterised in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency and cohesion)... Scales are descriptions of groups of typically occurring behaviours; they are not in themselves test instruments and need to be used in conjunction with tests appropriate to the population and test purpose. Raters of judges are normally trained in the use of proficiency scales so as to ensure the measure’s reliability (pp. 153–154).

From the above-integrated definition, it can be thought that, in terms of components, a rating scale includes both the domains to be assessed (construct) and the alignments between examinees’ performance and the predetermined levels of behavioural descriptions. Therefore, as far as language assessment is concerned, any rating scale development should also bear the considerations of the above two components. The present study will consistently follow the definition and confine the concept of rating scale to the context of language assessment only.

The evolving definition of rating scale also orchestrated with the developments of the particular oral rating scales. In the early 1950s, the rating scale of the US Foreign Service Institute (FSI) was first introduced and it has six bands from zero

(*foreignness*) to perfection (*nativeness*). Raters judge the relative amounts of *foreignness* or *nativeness* of each domain: “accent, fluency, comprehension, vocabulary and grammar” (Lowe 1985, p. 19). Afterwards, many other language proficiency test batteries started to apply that rating scale to their scoring, such as the reputed American Council of the Teaching of Foreign Languages (ACTFL) (ACTFL 1986, 1999; North 2000) and other language proficiency tests, oral tests in particular (see Shohamy 1981). The prevailing language assessments in recent years also witness the employment of rating scales, such as IELTS Oral Test, Spoken English of Test of TOEFL and Spoken Test of BEC.

In the context of English language testing in Chinese mainland, similarly, rating scales constitute an integral part in various oral proficiency tests, such as CET-SET, TEM-OT and Oral Test of the Public English Test System (PETS-OT). However, these rating scales for spoken English assessment vary in many aspects. The following section will particularise the prevailing taxonomies of rating scales with a few exemplifications to reflect their respective features.

2.3.2 Taxonomies of Rating Scales

Fulcher (2003), after reviewing different categorisations of rating scale, proposes a framework for describing rating scales from the perspectives of (1) rating scale orientation (Alderson 1991), (2) scoring approach (Hamp-Lyons 1991) and (3) focus (Bachman 1990). In addition, another three ways can categorise rating scales. It is proposed that rating scales can be divided in accordance with how they are designed (Fulcher 2003, 2010; Fulcher and Davidson 2007; Fulcher et al. 2011); whether they are designed based on experts’ intuition, a particular theory, empirical findings or performance decision trees. Alderson and Banerjee (2002) divide rating scales in the angle of task specificity. North (2003) classifies rating scales in terms of band and descriptor layouts. Therefore, in terms of rating scale typology, there can be possibly a total of six taxonomies, as summarised in Table 2.1.

Table 2.1 Taxonomies of rating scales

Orientation • User • Assessor • Constructor	Scoring • Analytic approach • Holistic approach – Holistic scoring – Primary-trait scoring – Multiple-trait scoring	Focus • Real world • Construct
Design • Intuition-based • Theory-based • Empirically driven • Performance decision trees (PDTs)	Task specificity • Generic • Task-specific	Band and descriptor layout • Graphic and numerical scale • Labelled scale • Defined scale

2.3.2.1 User Versus Assessor Versus Constructor

The categorisation of rating scales from the perspective of user-orientation is first proposed by Alderson (1991), whose suggestion leads to a tripartition of user-oriented, assessor-oriented and constructor-oriented scales. This way of classification mainly dwells on the particular informants of a rating scale. User-oriented scales are used to report information about the behaviour of a test-taker at a given level; assessor-oriented scales are designed to provide guidance for rating process, zooming in the quality of the performance expected; constructor-oriented scales are produced to aid test constructors in designing test tasks. As one of the main purposes of this study is to inform teaching practitioners of candidates' performances in formative assessment, primary concern is given to the development of an assessor-oriented rating scale. This concern can also be well supported by North's (2003) argument that scales used to rate second language performance should be mostly assessor-oriented, giving prominence to the aspects of ability as reflected in candidates' performance. However, what needs pondering is that in the context of formative assessment, which will be cast light on later, not only teachers but also peers and learners themselves may play the role of assessors. The forthcoming review on formative assessment will make an argument and justify who should be assessors for the rating scale proposed in this study.

2.3.2.2 Holistic Versus Analytic

The second categorisation is of holistic and analytic scales. This is first brought forth by Shohamy (1981) and has long served as the most salient and best-documented categorisation (e.g. Bachman 1988; Bachman and Savignon 1986; Douglas and Smith 1997; Fulcher 1997; Ingram and Wylie 1993; Underhill 1987; Weir 1990). As this taxonomy is commonly referred to (see Barkaoui 2007; Cooper 1977; Fulcher 2003; Goulden 1992, 1994; Hamp-Lyons 1991; Weigle 2002), more elaborations will be accordingly unfolded in this section of review. Holistic rating scale is also referred to as impressionistic or global scale. It is first defined in the context of writing assessment when Cooper (1977) posits that a holistic rating scale refers to

any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing ... [s]ome holistic procedures may specify a number of particular features and even require that each feature be scored separately, but the reader is never required to stop and count or tally incidents of the feature (p. 4).

If further divided in terms of scoring methods, a holistic rating scale can be broken down into holistic, primary-trait and multiple-trait scoring methods (Hamp-Lyons 1991). A *holistic scoring method* requires raters to assign only one score in order to encapsulate the overall performance or features of a candidate in a particular assessment task and its emphasis is on how a candidate excels (White 1985). In most cases, such rating scale descriptors of each proficiency level include

more than one domain of assessment, such as accuracy, vocabulary and fluency. However, all the descriptors of a particular band are grouped together, dissimilar to multi-trait scoring, where different domains of assessment are separately described in detail.

Since only one score is supposed to be given, this scoring method usually triggers controversy in lieu of an incomplete account of targeted construct (Fulcher 2003). It also seems less powerful in explaining the intriguing nature of speaking. Another problem with holistic scoring is that in speaking assessment, raters might overlook one or two aspects, in the case of which candidates might be rated on their strengths instead of being penalised for weaknesses (Bacha 2001; Charney 1984; Cumming 1990; Hamp-Lyons 1990; Knoch 2009). However, holistic scoring is primarily favoured by large-scale language assessments, where the time allocated for rating is of the topmost concern, yet it is spurned by classroom assessment because it provides limited feedback for students and teachers about what might be revealed from assessment per se.

Primary-trait scoring is developed to assess certain expected language functions or rhetorical features elicited by an assessment task (Lloyd-Jones 1977). It was first adopted by the National Assessment of Educational Programme (NAEP) for the purpose of obtaining more information from one single score. As Applebee (2000) explains, regarding writing assessment, “primary trait assessment in its initial formulations focused on the specific approach that a writer might take to be successful on a specific writing task; every task required its own unique scoring guide” (p. 4). Therefore, it can be comprehended that in primary-trait scoring, raters predetermine a main trait for the successful task fulfilment so that scoring criteria are usually reduced to one chief dimension and is therefore context-dependent (Fulcher 2003). Although only one score needs to be assigned in primary-trait scoring, that single score largely depends on the degree to which the candidate addresses the specific requirements of a given oral assessment task (Barkaoui 2007).

This kind of rating scale is advantageous in virtue of its focus on one targeted observable aspect of language performance, and it is a relatively quick way to score speaking performance, especially when rating emphasises one specific aspect of that performance. For example, if candidates are requested to perform a presentation as an assessment task, a rater would rather concentrate on candidates’ *articulation* than *lexical density*. In that case, the primary-trait *articulation* is assessed with a focused weighting. However, just because this way of scoring concentrates on only one primary trait, it would be less fair to argue that the aspect singled out for assessment is primary enough to base a single score on it (Knoch 2009).

Hamp-Lyons (1991) puts forward *multiple-trait scoring*, or *multi-trait scoring* for the rating scale designed to offer feedback to learners and other stakeholders about performance on contextually appropriate and task-specific criteria. As this scoring method per se suggests, it involves evaluating various traits for reaching an overall score. Although this approach is similar to primary-trait scoring in that both methods are holistic in nature, it allows raters to observe more than one dimension. Given that, it can also be regarded as an extended version of holistic scoring method

as the band descriptors of each assessment domain are much more detailed and corporeal.

Since large-scale language assessments usually take rating duration into serious consideration, the rating scales adopted by IELTS Speaking (see Appendix I) and TOEFL *iBT* Independent Speaking Tasks (see Appendix II) are typical of this category. In the former case, a rater is supposed to judge examinees' performance in four aspects, *fluency and coherence*, *lexical resource*, *grammatical range and accuracy*, *pronunciation*, and assigns an overall score according to nine bands (Band 1 to Band 9). What is slightly different in the case of TOEFL is that a number of general descriptions concerning *task fulfilment*, *coherence* or *intelligibility* are also attached in the rating scale in addition to the descriptors of the three individual traits (*delivery*, *language use* and *topic development*). Yet a rater is still expected to accord an overall score to the speech sample within a range of 5 bands (Band 0 to Band 4).

By contrast, Cooper (1977) defines *analytic approach* as requiring the rater "to count or tally incidents of the features" (p. 4). Analytic rating scales are inclusive of separate categories representing different aspects or dimensions of performance. For example, dimensions for oral performance might include fluency, vocabulary and accuracy. Each dimension is scored separately, and then dimension scores are totalled. Analytic rating scales can be extremely similar to multi-trait scoring in the sense that both require raters to assign more than one score to a speech sample. However, their difference consists in the fact that multi-trait scoring is more task-specific, usually focusing on specific features of performance necessary for the success of task fulfilment; the latter is more generalisable to a plethora of assessment tasks with generic dimensions of language production included.

For example, the rating scale for Test of English for Educational Purposes (TEEP) takes this form (see Appendix III). A rater is supposed to tick one number for each of the six assessment domains (*appropriateness*, *adequacy of vocabulary for purpose*, *grammatical accuracy*, *intelligibility*, *fluency and relevance* and *adequacy of content*) and then sum up the subscores. One special example is the rating scale of BEC Oral Test, which combines holistic and analytic rating (see Appendix IV for Level 1). One interlocutor responsible for communicating with candidates marks holistically while another assessor takes charge of analytic marking, with the two scores averaged to a final score subsequently.

However, analytic scoring is criticised inasmuch that various domains apart do not necessarily add up to the whole. In other words, individual subscores for different dimensions might not supply reliable information of what is assessed globally. On the other hand, since scoring is multifaceted, raters might assign correspondingly lower subscores to all the assessment domains if one particular domain is not as satisfactorily performed as expected. Therefore, tendency would be assigning the same low grades across all the domains, known as "halo effect" (Thorndike 1920) or "cross-contamination" (Alderson 1981).

On the positive side, it can be found that if rating is analytically conducted, raters can be refrained from being confused with dimensions as they are supposed to assign subscores to each assessment dimension. Weir (1990) also comments that

analytic rating scales facilitate rater training and scoring calibration, especially for inexperienced raters. In addition, the advantage of adopting analytic over holistic rating scales includes an access to fine-grained information about examinees' language ability (Bachman et al. 1995; Brown and Bailey 1984; Kondo-Brown 2002; Pollitt and Hutchinson 1987) because, from a variety of dimensions, rating analytically may reveal more information about what students are excelled in.

Weigle (2002), in the context of writing assessment, also contends that analytic rating scales are generally accepted to result in higher reliability and construct validity especially for second language writers although they can be time-consuming. This is accorded with Sawaki's (2007) view that in second language assessments, analytic rating scales are often used to assess candidates' language ability within a single modality, viz. speaking in the case of this study. When it comes to the construction of a rating scale for formative assessment, whether analytic or holistic scale is preferred will be further discussed in the follow-up section.

2.3.2.3 Real-World Versus Ability/Interaction

The third categorisation outlined in Table 2.1 is of real-world and ability/interaction rating scales, another demarcation from the perspective of testing situation (see Bachman 1990, pp. 344–348 for details). A *real-world rating scale* stipulates that the assessment tasks are situation-specific, viz. the authentic tasks that are anticipated in real life. Given this, real-world rating scales can be usually applied in performance tests. An *ability/interaction rating scale* relates more to a construct than a task and is designed on the assumption that it is possible to generalise from test scores to real-world situations that may not be modelled in the test tasks. Considering this study concentrates on the development of a rating scale in formative assessment, in which more weighting concerning the general construct of learners' oral performance, including nonverbal delivery, will be given, it should be therefore an **ability/interaction rating scale**.

2.3.2.4 Intuition-Based Versus Theory-Based Versus Empirically Driven Versus PDTs

Regarding the process of rating scale design, North (1996) describes the development of rating scales as condensing the complexity of performance into thin descriptors. The way in which rating scales and rating criteria are constructed and interpreted by raters also act as de facto test constructs (McNamara 2000). Therefore, another categorisation of rating scales takes the perspective of how they are developed: *intuition-based*, *theory-based*, *empirically driven* and *performance decision trees* (PDTs) (Brindley 1991; Fulcher 2003, 2010; Fulcher et al. 2011; North 2003). The first type tends to be a priori measuring instrument or “armchair method of scale development” (Fulcher 2010, p. 209). A priori method usually

refers to constructing the descriptors of the rating scales by an expert, often using his/her own intuitive judgment concerning the nature of language proficiency, along with a consultation with other experts. Therefore, it is believed to be the most prevailing method of generating a rating scale (Knoch 2009). A priori method can be subclassified into more specific development methodologies (North 1994), but they mostly have in common “the lack of any empirical underpinnings, except as post hoc validity studies” (Jarvis 1986, p. 21).

The second type is on the basis of an existing theory or framework. Lantolf and Frawley (1985) expound that the validity of a rating scale can be limited if no linguistic theory or the research in the definition of proficiency is taken into account. As is aforementioned, the advantage of basing a rating scale on a model of *communicative competence* is that “these models are generic and therefore not context-dependent” (Knoch 2009, p. 48), resulting in higher generalisability.

The third type, designed in a post hoc fashion, is likely to be driven by the data elicited from a sample of testees and rating scale developers manage to extract the features that distinguish candidates across various proficiency levels. For example, Fulcher (1987, 1993, 1996a) developed a rating scale of fluency in spoken English assessment based on the distinctive discourse features discernable in candidates’ oral production. Another data-based method of rating scale development is a corpus-based/corpus-driven approach. Hawkey (2001), Hawkey and Barker (2004) manage to design a universal rating scale that covers Cambridge ESOL writing examinations at different proficiency levels.

The latest development of rating scales witnesses the fourth type, which starts with an analysis of the discourse features expected in real-life interaction and then finds its assessment domains in the context of a particular framework as the *trees*. Afterwards, the decision on whether obligatory elements are present in each *tree* is made to determine what should be assessed as reflected in a rating scale (Fulcher 2010). Fulcher et al. (2011) employ a scoring model for service encounters with PDTs and prioritise this method in performance tests within a specific communicative context.

Since it is not quite necessary for the rating scales used in low-stakes speaking assessments to be constructed from data, most of them are developed intuitively. However, when this approach is applied to the formulation of the rating scale for large-scale and high-stakes tests, problems of validation and reliability might arise. For instance, Skehan (1984) and Fulcher (1987, 1993) criticise the English Language Testing Service regarding the intuitively developed rating scale. Likewise, Brindley (1986, 1991), Pienemann and Johnston (1987) find that the rating scale used in Australian Second Language Proficiency Ratings (ASLPR) lacks validity due to its intuitive development. Bachman (1988), Bachman and Savignon (1986), Fulcher (1996b), Lantolf and Frawley (1985, 1988), Matthews (1990) and Spolsky (1993) invalidate the ACTFL scales with either the empirical studies or the reasoning that the scale confuses linguistic with non-linguistic criteria. Therefore, it can be generalised that even though a rating scale is developed intuitively or based on the theoretical underpinnings, it would be better to be validated with or informed by data-driven methods.

Specific to the present study, on the one hand, the development of the rating scale is based on the priori consideration of the CLA model, together with the possible discriminating features informed by the data-driven evidence when an argument for embedding nonverbal delivery into speaking assessment is built. On the other hand, post hoc quantitative and qualitative validation studies will contribute to the finalisation of the rating scale. However, as formative assessment in the context of this study does not fall into professional English tests, almost no need is felt as to apply the PTDs method. Therefore, the rating scale, with nonverbal delivery included as a dimension, is an integration of theory-laden and empirically validated one.

2.3.2.5 Generic Versus Task-Specific

Alderson and Banerjee (2002) divide rating scales in terms of task specificity. One division is *generic scales*, referring to those constructed in advance for almost all sorts of assessment tasks and the other is used to evaluate test-takers' performance on *target specific tasks*. Rating scales and tasks are thus directly linked because the scales describe speaking skills that tasks might elicit (Luoma 2004). However, as different assessment tasks feature discrepant task characteristics, it is questionable as to whether such a generic rating scale can be designable. Since the present study proposes to design a rating scale to be applicable to formative assessment, it would be far beyond a claim of being a generic one because the assessment task in the present study, viz. group discussion to be elaborated below, is prespecified.

2.3.2.6 Graphic and Numerical Versus Labelled Versus Defined

The last categorisation focuses more on the physical layout of rating scales. The most simple type in this categorisation is a *graphic and numerical rating scale*, in which there is a continuum with two points representing both ends of a scale, yet with no descriptors of behaviours expected from candidates (North 2003). Therefore, the subjectivity among various raters becomes the main drawback of such design. The second type is a *labelled rating scale*, viz. a scale with cues attached to various points along the scale. Nonetheless, it can still be regarded as less assessor-friendly as the cues provided might be vague, such as a range from *poor* to *excellent* (Knoch 2009). The third type is a *vertical rating scale with each point elaborately defined* so that succinct space is allowed for longer descriptions. For instance, Shohamy et al.'s (1992) ESL writing scale falls into this type. However, since there is no significant difference in the reliability of different designs (Myford 2002), this study will first aim at a rating scale with sufficient defined behavioural expectations for rater-friendliness, yet subject to revision after the expert judgment in the rating scale formulation phase.

2.3.3 *A Critique on the Existing Rating Scales*

However, classified, rating scales represent “the descriptions of expected outcomes, or impressionistic etchings of what proficiency might look like as one moves through hypothetical points or levels on a development continuum” (Clark 1985, p. 348). This part will continue with a critique on the main existing rating scales for speaking assessment.

North and Schneider (1998) summarise two main weaknesses regarding the rating scales for language proficiency assessment, oral tests included. On the one hand, there is no guarantee that the descriptors of proficiency in a rating scale are accurate or valid; on the other hand, a number of them cannot be regarded as offering criterion-referenced assessment although they generally claim to do so. In particular, the wording in a rating scale is sometimes vague, subjective or hardly measureable (Mickan 2003; Upshur and Turner 1995), such wording as *weak*, *poor* and *better* (Turner and Upshur 2002; Upshur and Turner 1995). Therefore, they result in less consistency and most of them appear in fact to have been produced pragmatically by appeal to intuition and those scales that rating scale developers have access to.

Fulcher (1996b) and North and Schneider (1998) also point out that in the process of rating scale development, it is rare that much account is taken into using a model of *communicative competence* and/or language use and a model of measurement. Less consideration is found when assessment providers use the rank-ordered scale exclusively for one context to another inappropriate context (Spolsky 1986, 1989a). Specifically, there seems arguable possibility of directly borrowing existing rating scales for summative assessment to formative assessment as these two assessment contexts might be mutually distinguishable. Although a good number of rating scales take *strategic competence* into their development, what can be revealed from the examples cited in the review above would follow that few of them systematically observe *competence* in an operationalisable manner. What is even controversial is that nonverbal delivery, as one of the most pronounced components of *strategic competence*, is absent in most, if not all, existing rating scales. Therefore, it should be advisable that a rating scale designed with the CLA model as its underpinnings should be well informed of the above gap.

Over and above, limited literature can be found regarding rating scales exclusively for the context of formative assessment. Since formative assessment can be as important as standard-based or summative assessment, rating scales exclusively in this context also invite concern in light of validity and reliability (Brown and Hudson 1998; Cohen 1994). Similarly, rating scale development with necessary considerations for formative assessment can thus fit into the standardised assessment paradigms to reconceptualise the relationship between formative assessments and standardised summative assessments (Brindley 2002; Lynch 2001; McNamara 2001; Teasdale and Leung 2000). How can the rating scale in this study be designed manageably for formative assessment while also offsetting the weaknesses of the rating scales critiqued above? The following part will interject formative assessment

to expound more on the necessity, feasibility and significance of using an analytic rating scale in formative assessment.

2.3.4 *Formative Assessment*

Formative assessment derives from *formative evaluation* (Scriven 1967), foregrounding the notion on the practice of programme evaluation. Therefore, strictly speaking, formative assessment is not exclusively confined to English learning. Bloom et al. (1971) extend the notion of *formative evaluation* to a much broader sense, stating that

formative evaluation is for us the use of systematic evaluation in the process of curriculum construction, teaching and learning for the purpose of improving any of these three processes...This means that in formative evaluation one must strive to develop the kinds of evidence that will be most useful in the process, seek the most useful method of reporting the evidence, and search for ways of reducing the negative effect associated with evaluation – perhaps by reducing the judgmental aspects of evaluation or, at least, by having the users of the formative evaluation (teachers, students, curriculum makers) make the judgments. (p. 118)

The above definition clarifies that one of the purposes to conduct formative assessment is to diminish, or even remove possible negative backwash of high-stakes tests on language learning (Wang et al. 2006). Against this background, increasing attention is paid to the great potential of formative assessment; conventional summative testing of language learning outcomes also gradually compacts formative modes of assessing language learning as an ongoing process (Davison 2004). However, formative assessment vis-à-vis summative assessment is still under explored (Black and Wiliam 1998; Davies and LeMahieu 2003; Leung 2005a; Leung and Mohan 2004).

2.3.4.1 **Definition**

The notion of formative assessment suggests itself to be opposed to summative assessment. Broadly conceived, formative assessment refers to

the collaborative processes engaged in by educators and students for the *purpose* of understanding the students' learning and conceptual organisation, identification of strength, diagnosis of weaknesses, areas for improvement, and as a *source* of information that teachers can use in instructional planning and students can use in deepening their understanding and improving their achievement. (Cizek 2010, pp. 6–7)

The wording of this broad definition, such as *purpose* and *source*, mainly touches upon the functions of formative assessment. However, because of its broadness, many aspects of formative assessment fail to be specified, such as the referents of educators and the nature of the information source as further guidance in

language learning. As a matter of fact, the most frequently cited definition is brought forth in the seminal article by Black and Wiliam (1998), who define formative assessment as “all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged” (p. 10). This definition specifically narrows down the referents of educators to teachers and students, and the information is stipulated as a kind of feedback, which is regarded as positive impact of formative assessment practices (Allal and Lopez 2005; Brookhart 2004, 2007; Hattie and Timperley 2007; Shute 2008).

Later, Cowie and Bell (1999) further confine the settings of formative assessment to a context, where both assessment and learning take place simultaneously. In a similar vein, from the perspective of formative assessment use, Popham (2008) regards formative assessment as a planned process when the teacher or students use assessment-based evidence to progress learning and instruction. In order to provide a comprehensive definition, Black and Wiliam (2009) propose that assessment is formative:

to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (p. 6)

In this highly inclusive definition, the agents involved in formative assessment are clearer and extended to peers. In addition, formative assessment is no longer just for the sake of evaluation, but also for decision-making. Therefore, the ultimate purpose of formative assessment can enhance teaching and learning. The present study will follow the above definition so that the particular agents for the rating scale can be explicitly stated and desired positive impact of the assessment can be thus achieved.

2.3.4.2 Benefits of Formative Assessment

In fact, the above analyses on the definitions of formative assessment already reveal its functions and purposes, which can also be credited as the benefits in four aspects.

First, as far as the nature of formative assessment is concerned, it provides a wealth of feedback from assessors for learners. Therefore, such feedback is characterised by learner-specificity and full description on an individual basis (Sadler 1989). Herman and Choi (2008) examine the perceptions by teachers and learners to see whether formative assessment is shared with a similar understanding by both sides. The results indicate that the perceptions and attitudes on both sides are consistent, and that it also emphasised the significance of improving learners with the information available from formative assessment. Rea-Dickens (2006, p. 168), however, dissuade that in formative assessment the feedback to learners should be “descriptive” rather than “evaluative” so that it is not negatively perceived.

Second, in lieu of the preference that classroom is usually the primary choice of formative assessment, learners' anxiety can be much lowered due to a familiar environment. Davidson and Lynch (2002), Lynch (2001, 2003) and McNamara (2001) in general agree upon endorsing formative assessment to conventional testing methods as a shift of the locus of control from centralised authority into the hands of classroom teachers and their peers. If assessment environment is familiar to candidates, assumedly they will be at a more advantageous position of giving full play to their potentials.

Third, as formative assessment may include tasks or activities, such as ongoing self-assessment, peer-assessment, projects and portfolios (Cohen 1994), most assessment methods can be task-based. As Ross (2005) point out, one of key appeals formative assessment can provide is the autonomy given to learners. Formative assessment is thus thought to influence learner development through a widened sphere of feedback during their engagement with various learning tasks.

Last, regarding the validity in alignment with traditional standardised assessment, there emerges research on validating formative assessment as a testing method. Huerta-Macias (1995) prioritises the direct face validity of alternatives to conventional achievement tests as sufficient justification for their use. This view is also accorded with the notion of learner and teacher empowerment (Shohamy 2001). Therefore, it can be believed that if a rating scale for formative assessment is also vigorously validated, it can be applied as a valid measure.

With the benefits of formative assessment outlined above, it is necessary to develop, in the formative assessment context, a rating scale with a dimension of nonverbal delivery. In so doing, teachers can assess learners from various dimensions and learners may also have the access to feedback of various aspects for self-enhancement.

2.3.4.3 Group Discussion as a Formative Assessment Task

The above outlines various aspects of benefits that formative assessment might offer. This part of review turns to group discussion as a formative assessment task for assessing EFL learners' spoken English so that why group discussion is particularly chosen as the main assessment task in the present study can be justified.

Prior to unfolding the usefulness of group discussion in formative assessment, the previous studies on group discussion as an assessment task is first reviewed. In the first large-scale study on group discussion concerning the accuracy of test-takers' production, Liski and Puntanen (1983) find that test-takers' performance in group discussions can serve as a fit predictor of their overall academic success. In addition, Fulcher (1996a) also reports that test-takers consider group discussion as a valid form of second language testing and that examinees feel less anxious and more confident when speaking to other discussants instead of examiners or interlocutors (Folland and Robertson 1976; Fulcher 1996a). Fulcher (1996a) also finds that group discussion is an easily organised task compared with picture talk, where an interlocutor or an interview based on speaking prompts will be involved. In

addition, group discussion, similar to paired discussion (Brooks 2009), may elicit richer language functions than oral proficiency interviews (OPI) so that communicative ability can be more comprehensively assessed (Nakatsuhara 2009; van Moere 2007).

Pre-eminently, in the context of formative assessment, group discussion can be assessed not only by instructors but also by learners and their peers on condition that the rating scales and criteria are made transparent and accessible to all the parties concerned (Fulcher 2010; Shepard 2000). However, previous studies also indicate that without substantial experience of applying the scoring criteria to work samples, self-assessments may fluctuate substantially (Ross 1998; Patri 2002). By contrast, peer-assessments are likely to be much more reliable though they can be more lenient than instructor-assessments (Matsuno 2009). Therefore, the present study, instead of including self-assessment as a rating method, resorts to teacher and peer-rating when the proposed rating scale is validated based on the above considerations.

Although the reliability of group discussion as an assessment task in standardised large-scale testing is challenged as raters might not be able to assign reliable scores when candidates are tested in groups (Folland and Robertson 1976; Hilsdon 1995), such unreliability is hardly recorded empirically. In response to that, Nevo and Shohamy (1984) compare 16 language assessment experts' intuition and perception of having group discussion as an assessment task with other forms, such as role-play and OPI, only to find that group discussion ranks the top in terms of task utility standards, but stands in the middle on fairness, probably leading to testing experts' suspicion of the reliability of group discussion. Despite that, scant evidence can be collected to support that group discussion is not reliable.

In terms of task usefulness and task characteristics, group discussion also has a few distinctive features. It is first of all highly interactive and authentic (Kormos 1999; Lazaraton 1996b; van Lier 1989), with all the discussants involved in a meaning-making and negotiating process. It is also characterised by a high degree of feasibility and economy in the sense that formative assessment of this kind can just take place in classrooms and can be time-saving because several students are grouped together to be assessed, thus greatly reducing the time that traditional testing methods would call for (Ockey 2001). Another point inherent in formative assessment that can also credit group discussion is that all candidates are in discussion with familiar faces without interlocutors, which tends to lower their anxiety and eschew more errors arising from the intervention of interlocutors (Ross and Berwick 1992; Johnson and Tylor 1998; Young and He 1998a, Brown 2003). What's more, even though candidates' weaknesses are disclosed in various aspects, they do not feel as ashamed as they would otherwise be in face of generally stern examiners or interlocutors.

To briefly summarise, the above review provides positive evidence that this particular assessment task can be judged as ideal from the perspectives of face validity, reliability, authenticity, interactiveness, impact and practicality, which, incidentally but purposefully, can be accorded with Bachman and Palmers' (1996) framework of test usefulness.

2.3.5 *Properties of the Present Rating Scale*

Integrating the above review on rating scales and formative assessment, procurement can be reached regarding the properties of the rating scale that this study intends to propose. It will be an assessor-oriented analytic rating scale specifically for group discussion in formative assessment. The band and level descriptors aim to be defined and descriptive instead of merely evaluative. The design of the rating scale will be firstly theory-grounded on the construct of the CLA model and preliminary discriminating features identifiable in candidates' nonverbal delivery and then undergo empirical corroboration with data-driven involvement.

2.4 Validity and Validation

As the last phase of the present study sets out to validate a proposed rating scale with nonverbal delivery included as an assessment dimension, it is of importance to review the conceptualisation of validity and the evolution of validation methods. What should be pointed out is that validity is an integral and most basic concept in language assessment because “accepted practices of test validation are critical to decisions about what constitutes a good language test for a particular situation” (Chapelle 1999, p. 254). How validity is defined in reality determines how a test is to be validated.

Historically, test validity is an ever-changing concept and has undergone metamorphoses chronologically (Angoff 1988; Cronbach 1988, 1989; Goodwin 1997, 2002; Goodwin and Leech 2003; Kane 1994, 2001; Messick 1988, 1989a, b; Langenfeld and Crocker 1994; McNamara and Roever 2006; Moss 1992; Shepard 1993; Yang and Weir 1998). Researchers with different perceptions towards validity (e.g. Angoff 1988; Kane 2001; Goodwin and Leech 2003) have various demarcations of its development. Nonetheless, what stands to be certain is that the introduction of *construct* in conceptualising validity is widely regarded as a milestone. Therefore, all demarcations can fall into three phases in terms of how the role of *construct validity* evolves: (1) the preconstruct-validity phase, a period before construct validity was put forward by Cronbach and Meehl (1955); (2) the initial phase of construct validity, a period covering the range from the 1970s to the 1980s, when construct validity was made co-existent with other types of validity in language testing; and (3) the core phase of construct validity, a period when the concept starts to play a quintessential role in test validation.

In recent decades, with the popularity of argument-based validation method, there are also other perspectives of conceptualising validity, among which *Assessment Use Argument* (AUA) is utilised in full swing as “[an] overarching logical structure that provides a basis both for test design and development and for score interpretation and use” (Bachman 2005, p. 24). However, as far as the essence

of AUA is concerned, it still falls into the third phase as this notion calls for evidence collection in support of construct validity.

Therefore, concerning the concept of validity, this part will embark upon a review on a componential notion of validity, followed by a unitary concept of validity, with construct validity as the core. Afterwards, details on the newly established AUA (Bachman 2005; Bachman and Palmer 2010) will also be briefly reviewed; however, the critique on AUA in this section will lead to an argument that caution should be taken in employing AUA as the framework of validating the proposed rating scale. This section of review will wind up with the justification of employing both quantitative and qualitative approaches for the validation of the rating scale to be proposed based on the unitary notion of test validity; in particular, the incorporation of nonverbal delivery calls for a qualitative approach in validating the rating scale.

2.4.1 Validity: A Componential Notion

Prior to a unitary concept, test validity could be viewed as an umbrella term covering several types of validity, yet different researchers hold discrepant yardsticks of test validity taxonomies. For example, Guilford (1946) divides test validity into two components from the perspective of data analysing methods and real use: factorial validity and practical validity. However, Cronbach (1949), again from the angle of data analysis, categorises validity into logical/judgmental validity and analytical/empirical validity. The former is a rather loosely organised, broadly defined set of approaches, including content analyses, and examination of operational issues and test-taking processes, whereas the latter places more emphasis on the use of factor analysis, and especially on correlation(s) between test scores and a criterion measure (Anastasi 1950). Anastasi (1954, 1961, 1976) categorises test validity into four aspects, viz. face validity, content validity, factorial validity and empirical validity. Although the ways of cataloguing test validity vary in the early phase of conceptualisation, they are almost identical in nature: correlating observed test scores with criterion measurement. Thus, all of them except face validity can be grouped into *criterion-related validity*.

2.4.1.1 Criterion-Related Validity

The early phase of test validity stresses test purposes. Guilford (1946) points out that every test is purpose-specific and that one test can be valid for a particular purpose but invalid for another. No matter whether it is test providers or test users, all parties of stakeholders should be responsible for verifying that one test is valid for the particular purpose it serves. In that sense, how validity is defined is closely associated with test purposes, which can be interpreted in the Garrett's (1947) definition that "the validity of a test is the extent to which it measures what it

purports to measure” (p. 394). Similarly, Cureton (1950) also views that test purpose should be the basic issue of test validity and phrased its importance as “how well a test does the job it was employed to do” (p. 621). Against the above viewpoint, test purposes can be twofold: either diagnosing the existing issues or predicting the future performance. Accordingly, *American Psychological Association* (APA), *American Educational Research Association* (AERA) and *National Council on Measurement in Education* (NCME) in their early versions of *Standards for Educational and Psychological Testing* (*Standards*) divides criterion-related validity into *concurrent validity* and *predictive validity* (see APA 1954; APA et al. 1966).

In fact, criterion-related validity is deeply rooted in a realist philosophy of science, which holds that every individual can produce a value on the specific assessment characteristics and the assessment purpose is to estimate or predict that value as accurately as possible. In the context of standardised testing, the “true score”, or the estimates most approximating the “true score”, reflects the extent to which the test has precisely estimated that value (Thorndike 1997). In that sense, the precision of estimation is the degree of test validity.

The above definition reveals that criterion-related validity is concerned with the test per se and that it is a static property attached to test validity (Goodwin and Leech 2003). Therefore, criterion-related validity equals to “the correlation of scores on a test with some other objective measure of that which the test is used to measure” (Angoff 1988, p. 20). A test can be judged as valid or invalid according to the measuring results (Cureton 1950; Gulliksen 1950) and “[i]n a very general sense, a test is valid for anything with which it correlates” (Guilford 1946, p. 429). The key to validating criterion-related validity then lies in how to lay down the criterion measure in order to obtain standardised test scores, without which such validation studies cannot be carried out. Cureton (1950) puts forward the following method.

A more direct method of investigation, which is always to be preferred wherever feasible, is to give the test to a representative sample of the group with whom it is to be used, observe and score performances of the actual task by the members of this sample, and see how well the test performances agree with the task performances. (p. 623)

As revealed above, the first step is sampling the target candidates and observing their performances in the real assessment tasks to assign the corresponding scores. The scores ultimately derived should become the standard scores with reference to the criterion. When other tests are in the process of validation, the newly observed scores will be correlated with the standard scores to see the extent to which the test consistently measures the candidates’ ability. Therefore, “the test is valid in the sense of correlating with other [valid and reliable language] tests” (Oller 1979, pp. 417–418). Ebel (1961), however, would rather think that some language tests can be regarded as valid merely through subjective judgment and that language assessment experts’ judgment on validity can be employed to measure test validity. Once the validity criterion is determined, it is possible to design standard testing for the validation of other tests.

The thorny problem of undertaking criterion-related validation is that there is actually no such standardised test; even though there is one, it needs to be validated itself as well. If a standardised test is developed for the purpose of validating another, the cycle will be reduced to an infinite regression (Kane 2001). Therefore, in order to solve this problem, the content that is covered in a test emerged to the attention of validation studies and became an aid in criterion-related validation, hence content validity.

2.4.1.2 Content Validity

Content validity usually refers to the extent to which the test items or tasks are sufficient enough to represent the domain or universe of the content to be covered in a test. It was explained as “[whether] the behaviours demonstrated in testing constitute a representative sample of behaviours to be exhibited in a desired performance domain” (APA et al. 1974, p. 28).

Angoff (1988), in summarising what content validity in language assessment represents from the aspects of content relevance, content coverage and content significance, posits that a test has content validity when all the test items are representative not only of the domain but also of the number and significance of the domain. Messick (1988), from the interface between content and construct, asserts that “[w]hat is judged to be relevant and representative of the domain is not the surface content of test items or tasks, but the knowledge, skill, or other pertinent attributes measured by the items or tasks” (p. 38).

The main validation method regarding content validity is based on logical judgement, such as expert evaluation and a review of the test content by assessment experts (Angoff 1988). Since much subjectivity is involved, this validation method is usually controversial (Guion 1977; Kane 2001). Given this, there used to be a call for empirical validation on expert evaluation (Bachman et al. 1995). However, in direct performance tests, there are indeed advantages for expert evaluation (Cronbach 1971), which is still being utilised in many assessment settings (Kane 2001). Cronbach (1971) also puts forward equivalent tests for content validation, in which two sets of scores obtained from two different tests with the same content coverage are correlated. A low-correlation coefficient can indicate that at least one of them does not have high content validity, yet it is challenging to determine which particular test it is. On the other hand, if the correlation coefficient is high, it can be generally thought that both tests have content validity.

Unlike criterion-related validation, whose problem lies in the availability of a real standardised test, content validation is challenged in that the representativeness of test content can be barely guaranteed. On the one hand, the domain or universe of a test cannot be easily operationalised because what is assessed can be either language knowledge and language skills, or complicated performances or processes. On the other hand, the number of test items, coverage of test materials and method of sampling all impact the representativeness of a test content as well as its facility and discriminating power (Angoff 1988). Latent variables mentioned above

may give rise to the under-representativeness of a test. Messick (1989b, 1992, 1996) detects two points that might jeopardise the content validity of a test, viz. construct under-representation and construct irrelevance variance. The former might lead to negative washback in lieu of an over-emphasis on partial learning content; the latter may increase or decrease the difficulty of a test because what is covered is somewhat irrelevant to what is supposed to be assessed, thus incurring test unfairness.

Another two points concerning content validity are also worth mentioning. One is confirmationist/conformist bias mentioned by Cronbach (1988) and Kane (2001). Such bias refers to the practice that researchers or test developers are liable to adopt a confirmationist approach when validating the test content from the perspectives of relevance, coverage and significance. By contrast, a falsificationist approach is rarely used. In so doing, content validity, in all likelihood, can be exaggeratingly verified. The other point is that the consideration of test content exerts influence on scores, yet the content should be embedded in the test rather than in the test response (Messick 1975, 1980, 1988, 1989a, b). Therefore, “in a fundamental sense, content-related evidence does not qualify as validity evidence” (Messick 1988, p. 38). In addition, Messick (1988) also cautions researchers that when scores are interpreted, the related skills on the part of high-achievers can be generalised while it does not necessarily lead to the same interpretation that low-achievers do not possess the expected skills. It is because low-achievers might not perform well in a particular testing environment. This point is also regarded as one of the constraints in collecting content-related evidence in test validation (Messick 1975).

2.4.1.3 Construct Validity

Construct validity is first conceptualised by Paul Meehl and Robert Challman upon their draft of the *Standards* (1954), and further nourished by Cronbach and Meehl (1955). The introduction of construct validity, together with criterion-related validity and content validity, signifies the beginning of a “trinity view” of test validity (see APA et al. 1966). Therefore, construct validity has been regarded as a hallmark in the evolution of test validity. However, when this notion is first conceptualised, it is treated as a mere supplement to criterion-related validity (Cronbach and Meehl 1955). This is because when the criterion measure is not available, researchers would turn to an indirect validation method, which highlights the trait or quality underlying the test instead of test behaviour or scores on the criteria. Then “the trait or quality underlying the test” is just what *construct* is. The *Standards* (APA et al. 1974) put a psychological construct as

[a]n idea developed or “constructed” as a work of informed, scientific imagination; that is, it is a theoretical idea developed to explain and to organise some aspects of existing knowledge. Terms such as “anxiety”, “clerical aptitude”, or “reading readiness” refer to such constructs, but the construct is much more than the label; it is a dimension understood or inferred from its network of interrelationships. (p. 29)

Construct in this case is only a theoretical idea, representing the abstraction of constructed terms or labels, equal to the understanding of or inference from the relationships between theories. Ebel and Frisbi (1991) think that *construct* refers to those human behaviours or mental processes that can hardly be measured directly. Therefore, it can be a hypothesised abstraction, trait or variable. Bachman (1990) explains the notion in a much simpler way. He regards construct as those to be measured. Since abstraction cannot be directly measured, observed data cannot be directly used for reasoning or inference. Therefore, *construct validity* refers to the extent to which a theory or trait can be reflected by the observed data.

In construct validation, first of all a construct theory needs to be found that embodies human behaviours. After that, hypotheses are put forward and tests are administered to obtain data. Then, whether the theory-deducted hypotheses can be verified by means of all statistical methods, such as correlation analysis, is another issue. In such a process, the construct theory is the prerequisite as well as the crux. Cronbach and Meehl (1955) adopt a hypothetic-deductive model of theories (HD model) as the framework of constructing theories. HD model (Suppe 1977) treats theories as axiomatic systems and regards the core of theory as a combination of a series of axioms, as reflected as empirical laws. The implicit concept in theories is connected with axioms and correlated with the explicit observable variables. If the observation results are consistent with the theoretical hypotheses, it can serve as a proof that the observation (test) is valid (Hempel 1965; Kane 2001). In that sense, construct validation involves mutual verification of measure and constructed theories. On the one hand, constructed theories guide the collection, analysis and interpretation of the data; on the other hand, the data can serve to testify, modify and even nullify the constructed theories (Angoff 1988).

The above elaboration indicates that construct validity is a fairly complex process. It cannot be simply reflected by one correlation coefficient; rather, it involves evidence collection and reasoning and must be inferred from observation and data analyses. Considering the intriguing nature of construct validity, Campbell and Fiske (1959) put forward an MTMM approach, which includes both theoretical explanation and empirical verification. Theoretically, if the method and trait are the same, they should be highly correlated; empirically, the correlation coefficient between different methods and same traits is also known as convergent validity, which should be significantly higher than discriminant validity, a coefficient between same methods and different traits. In order to claim that the measures of a rating scale have construct validity, it is required that both convergence and discrimination should be demonstrated. This approach will be further detailed in the next section of the literature review.

2.4.1.4 Face Validity

Face validity, as its name suggests, usually refers to the degree to which all those at the surface level, such as the language and instructions used in a test, whether the layout or the printing quality of a test paper can be acceptable to candidates and the

public (Hughes 2003). Whether test validity in this regard should also be treated as a component of validity has long been debatable. Because face validity is only confined to the acceptability of the test paper at the surface level without any involvement of psychological measurement, it cannot truly reflect the validity of a test in the strictest sense, nor can it be a yardstick of measuring the degree of validity for a test. Mosier (1947), when criticising the ambiguity of face validity, thinks that “any serious consideration of face validity should be abandoned” (cited from Angoff 1988, p. 23). Angoff (1988) also mentions that “superficial judgments of the validity of a test made solely on the basis of its appearance can easily be very wrong” (p. 24).

Although face validity is challenged as incompetent to be one of the components of validity, quite a number of researchers have noted its importance. Anastasi (1982) believes that “the language and contexts of test items can be expressed in ways that would look valid and be acceptable to the test-taker and the public generally” (p. 136). Likewise, Nevo (1985) also acknowledges the usefulness of face validity and thinks that face validity should also be reported in test validation.

In brief summary, test validity at the first evolution phase was perceived as a componential entity, with criterion-related validity, content validity and construct validity as its tenets. However, in the case of this study, when a rating scale with a consideration of embedding nonverbal delivery into speaking assessment is validated, it seems quite impractical to accumulate evidence from all the above three aspects of validity. After the following part, where light will be shed on validity as a unitary notion and as playing a core role in all sources of validity, this review can justify that in validating the proposed rating scale, construct validity will be mainly scrutinised.

2.4.2 *Validity: A Unitary Notion*

Although Cronbach and Meehl (1955) augment the significance of construct validity as a determinant responsible for test performance in almost all tests, the *Standards* (APA et al. 1966, 1974) still categorise test validity into three or four components (*predictive validity* and *concurrent validity* can be folded into *criterion-related validity*) and views it as a supplement co-existing with criterion-related and content validity.

It was not until in the early 1980s that measurement researchers, such as Cronbach (1971, 1980, 1988, 1989) and Messick (1975, 1980, 1988, 1989a, b), started to emphasise the inferences and decisions made from test scores. By then, the overarching validity with construct validity as the core turned to be gradually and pervasively accepted. The unitary concept of test validity is reflected by the wholeness of validity and the supplementary nature of validity evidence. In theory, it holds that validity is a multifaceted entirety; in practice, construct validity can be verified from all the sources possible.

2.4.2.1 Definition and Significance

The two latest versions of *Standards* (AERA et al. 1985, 1999) define *construct validity* from a unitary perspective. However, the *Standards* (AERA et al. 1999) are added with test use and consequence, reflecting a further extension of test validity. The *Standards* (AERA et al. 1985) regard validity as

the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. **Validity, however, is a unitary concept.** Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from test scores. (p. 9)

However, the *Standards* (AERA et al. 1999) explain validity as

the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. ...The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. (p. 9)

...these sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. **Validity is a unitary concept.** It is the degree to which all of the accumulated evidence supports the intended interpretation of test scores for the intended purposes. (p. 11)

The above definitions of validity can be comprehended as the extent to which all sorts of evidence can support the score interpretation and use. Hence, validity is a unitary concept with construct validity as the core. Compared with the componential notion, there are three distinguishable features for this view.

First, dissimilar to a preference to classifying validity into several components (see Angoff 1988; Langenfeld and Crocker 1994; Messick 1988, 1989b, 1995; Shepard 1993), this view no longer treats validity as divisible; rather, it is a unifying force (Goodwin and Leech 2003; Messick 1988). The previous criterion-related validity and content validity are also embedded into the evidence collection concerning “content relevance”, “content coverage”, “predictive utility” and “diagnostic utility” (Messick 1980, p. 1015). The validation process includes collecting evidence from various sources, interpreting and using the evidence for verification. In light of construct validity, Cronbach (1988) puts forward two kinds of validation programmes as follows.

The weak programme is sheer exploratory empiricism; any correlation of the test score with another variable is welcomed. ...The strong programme, spelled out in 1955 (Cronbach & Meehl) and restated in 1982, by Meehl and Golden, calls for making one’s theoretical ideas as explicit as possible, then devising deliberate challenges. (pp. 12–13)

It can be seen that the weak programme focuses on the correlation between test scores and other variables, while the strong one tends to seek theory-based ideas. The former holds that evidence should be gathered from a variety of sources so that its advantage consists in its diversity and complementariness. However, just as

Kane (2001) points out, the weakness of this programme is its opportunistic strategy; in other words, it seeks “readily available data rather than more relevant but less accessible evidence” (p. 326). The strong programme follows an approach of validation-through-falsification, viz. “an explanation gains credibility chiefly from falsification attempts that fail” (Cronbach 1988, p. 13). Yet it also has its weakness in that this approach is limited in its utility given an absence of a well-grounded theory to test (Kane 2001).

The unitary notion of validity lays more emphasis on complementariness, instead of alternativeness, of evidence. This view is widely accepted and reinforced since the 1980s. Bachman (1990) notes that “it is important to recognise that none of these by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores” (p. 237). In a similar vein, Weir (2005) also emphasises that

[v]alidity is multifaceted and different types of evidence are needed to support any claims for the validity of scores on a test. These are not alternatives but complementary aspects of an evidential basis for test interpretation...No single validity can be considered superior to another. Deficit in any one raises questions as to the well-foundedness of any interpretation of test scores. (p. 13)

Second, the unitary concept of validity has transferred its focus from test per se to the interpretation of test scores, or more precisely, to the extent to which the score interpretation can be supported by the evidence. In 1986, English Testing Service (ETS) sponsored a symposium themed *Test validity for the 1990s and beyond* and most of the keynote speeches are compiled in the proceedings by Wainer and Braun (1988). On the first page of the prelude, there is a footnote to the effect that a test itself cannot be claimed to be valid; rather, the inferences made from the test scores should be used as the sources of validation.

In fact, Cronbach (1971) shares the above view when stating that “one validates not a test, but an interpretation of data arising from a specified procedure” (p. 447) and “one does not validate a test, but only a principle for making inferences” (Cronbach and Meehl 1955, p. 297). Based on this, McNamara and Roever (2006) even elevate Cronbach’s view to such a height that there is no such thing as a truly valid test, but only defensible interpretations to a certain degree. Therefore, the unitary concept of validity shows that test validity is manifested in score interpretation rather than test per se.

Third, after the unitary concept of test validity is put forward, the test use and its consequence also invite great concern. Although they are not new in validity studies, the *Standards* (1985) include neither of them into the definition of validity. With the maturing of the unitary concept, there has been an increasing awareness of and concern over the intended and unintended purposes, potential and actual consequences (Cronbach 1988; Linn 1994; Messick 1989b, 1994; Shepard 1993). Fitting into that trend, the new version *Standards* (1999) officially include the test use and consequence into the definition of validity.

However, there are also researchers (e.g. Dwyer 2000; Popham 1997) who prefer to confine validity to the boundary of score interpretation and traditional

psychological measurement rather than extend it to the language policy domain. The discussion (see Linn 1997; Mehrens 1997; Popham 1997; Shepard 1997) addresses both positive and negative sides of including test use and consequence into validation, yet winds up with no consensus. The dispute over validity development in recent decades also concentrates on the inclusion or exclusion of test use and its consequences. Proponents tend to include them into the scope of test validation, and they focus on differential item function, backwash effects and social consequence of tests (e.g. Bachman and Palmer 1996; Cheng 2005; Green 2007; Hamp-Lyons 1997; Hughes 2003; Shohamy 2001). However, there are also a good number of researchers on the opposing side. Kunnan (2000, 2004, 2005, 2008, 2010) maintain that the study of test fairness should be placed in a larger scope, rather than treated as a subordinate element of test validity. Bachman (2005), Bachman and Palmer (2010) put forward AUA, splitting test use from test validity. McNamara and Roever (2006) also think that test validity should not be extended to cover political and social dimensions because score use and social consequence would fail to reflect the role of language testing in a social dimension.

2.4.2.2 Multidimensionality of the Unitary Concept

Although the second evolution stage of test validity deems the notion as a unitary one, it is still etched with many dimensions. Messick is among the first proponents of a unitary concept of test validity and his works (1975, 1980, 1988, 1989a, b, 1992, 1994, 1995, 1996) exert far-reaching significance. Messick (1995) defines *validity* as

nothing less than an evaluative summary of both the evidence for and the actual – as well as the potential – consequences of score interpretation and use. This comprehensive view of validity integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility. (p. 742)

As can be interpreted from the above definition, the unitary concept is reflected in *an evaluative summary* and *a comprehensive view*. In other words, this concept encompasses the test content, test criterion and test consequence with hypotheses and empirical verification. Therefore, this concept is characterised by its multidimensionality, where score interpretation, test use, evidential basis and consequential basis are interacting with each other for a comprehensive evaluation, as illustrated in Table 2.2 (Messick 1988, p. 42).

Table 2.2 Facets of validity (Messick 1988, p. 42)

	Test interpretation	Test use
Evidential basis	(1) Construct validity	(3) Construct validity + relevance/utility
Consequential basis	(2) Value implications	(4) Social consequences

Messick (1988) explains the four pairs of interaction as follows.

(1) an inductive summary of convergent and discriminant evidence that the test scores have a plausible meaning or construct interpretation; (2) an appraisal of the value implications of the test interpretation; (3) a rationale and evidence for the relevance of the construct and the utility of the scores in particular applications; (4) an appraisal of the potential social consequences of the proposed use and of the actual consequences when used. (p. 42)

Talking about test validation, Messick (1989b, 1995, 1996) also suggests that evidence from six distinguishable aspects should be collected in order to verify the overall validity. These six aspects are “content, substantive, structural, generalisability, external and consequential aspects of construct validity” (p. 248).

When evidence is collected for further verification, one way is to structure all the evidence in the form of arguments (Cronbach 1980, 1988, House 1980) because validity arguments provide a comprehensive evaluation of the intended interpretation and uses of test scores (Cronbach 1988). Following Cronbach and Messick, Kane (1990, 1992, 2001, 2002, 2004, 2006, 2010), Kane et al. (1999) develop an interpretive framework of arguments to provide guidance for justifying interpretations of test scores and use. Later, Mislevy (2003), Mislevy et al. (2002, 2003) propose an evidence-centred design (ECD), at the heart of which is what is referred to as an evidentiary argument. In the recent development, argument-based approach remains trendy in test validation (e.g. Bachman and Palmer 2010; Chapelle et al. 2008, 2010; Xi 2010). Therefore, the following part will review and critique the most representative framework of argument-based approach AUA and justify why the present study will still employ the unitary notion of test validity instead of resorting to this newly established framework.

2.4.3 *Argument-Based Validation and AUA*

Bachman (2005) first puts forward AUA, and Bachman and Palmer (2010) later revises and enriches the framework with a number of tests in real-life settings. Thus, as aforementioned, this framework is inviting an increasing number of test validation studies, so a review on its essence clearly becomes necessary in the present study. Then what should be the essence of AUA? In fact, any argument-based framework lays its foundation on the base argument whose structure makes explicit the reasoning logic employed to justify the plausibility of the conclusion or claim. AUA is of no exception. Therefore, the structure of the base argument is of crucial importance; a minor modification may divert the general direction of reasoning, thus resulting in utterly different outcomes. Since AUA resides its base argument structure in the Toulmin model, it is necessary to obtain a full understanding of the Toulmin argument structure and its reasoning logic before a critique on the framework can be made.

2.4.3.1 The Toulmin model

Toulmin does not explicitly put forward the notion of “the Toulmin model” himself, but rather regards it as “one of the unforeseen by-products of the uses of argument” (Toulmin 2003, p. viii). The aim of Toulmin’s writing the book is strictly philosophical, to criticise the syllogism or demonstrative deductions in general. His major viewpoint is that the form of syllogism is simplistic and ambiguous with no practical use in daily arguments. To do justice to the situation, Toulmin builds up a pattern of argument analysis. This pattern can be illustrated with the typical example of the Toulmin model (see Fig. 2.7): by appealing to the datum (D)—“Harry was born in Bermuda”, one can make a claim (C) about Harry’s nationality—“So, presumably, Harry is a British subject”. The step from the datum to the claim is guaranteed by the implicit warrant—“A man born in Bermuda will generally be a British subject”, which is an inference drawn on the British Nationality Acts, and whose authority relies on its backing which makes an account of the British statutes and other legal provisions. Considering the potential exceptional conditions, such as “Both Harry’s parents may be aliens” and “Harry might have changed his nationality since birth”, a qualifier—“presumably” is included to indicate a tentative modality in the claim.

This is clearly a judgmental reasoning process. According to Toulmin (2003), the rationality of a logical or practical argument is guaranteed by “Data such as D entitle one to draw conclusions, or make claims, such as C’, or alternatively ‘Given data D, one may take it that C’” (p. 91). In other words, the data “on which the claim is based” (p. 90) should reveal sufficient bearing of warrants, which are “general, hypothetical statements, which can act as bridges, and authorise the sort of step to which our particular argument commits us” (p. 91). Meanwhile, the warrants should be further supported by the backing, which Toulmin defines as “straight-forward matters-of-fact” (p. 96) to provide other assurances for the reasoning

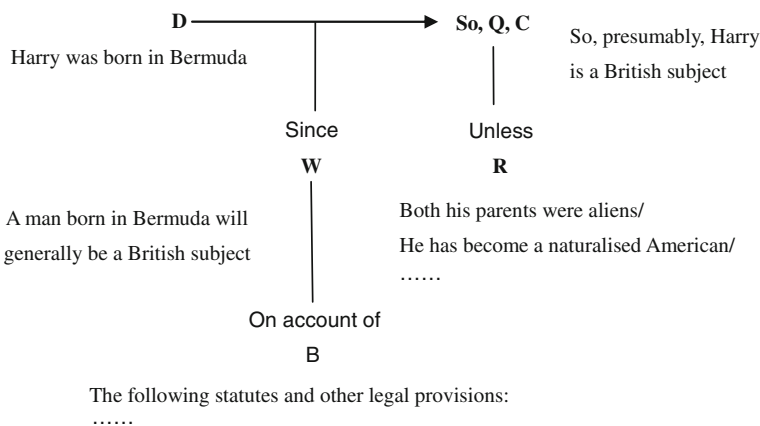


Fig. 2.7 A Toulmin model example (Toulmin 2003, p. 97)

process. Contrary to the syllogistic argument, a Toulmin argument is not considered universally true, so Toulmin includes two important additional elements in his model: a rebuttal to represent “the exceptional conditions which might be capable of defeating or rebutting the warranted conclusion” (p. 94) and a qualifier to indicate “the degree of force which our data confer on our claim in virtue of our warrant” (p. 93).

In making a claim, there is no denying of the possibility of potential rebuttals, but there is no such exclusion either; otherwise, no conclusion can ever be made, since it is just impossible to exclude all rebuttals. What to do is to include a properly worded qualifier to indicate the strength of the claim. Apart from a mechanism of handling exceptional situations, the Toulmin model is superior to the syllogism in still another aspect: differentiating a substantial argument from an analytic one. By breaking down the ambiguous major premise into warrants and backing, syllogistic arguments can be separated into two different types: either in the form of “D; W; so C” or “D; B; so C”. While the former is guaranteed by warrants (hypothetical statements), the claim is arguable and the argument is a substantial one; the latter is guaranteed by backing, namely matters-of-fact or truths, in which the claim is but a tautology of the fact or truth contained in the premises and there is no real argument involved.

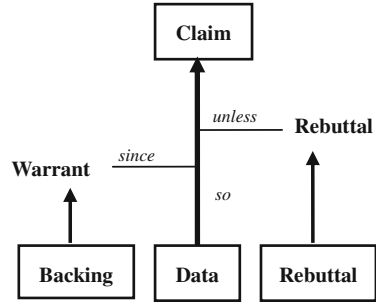
As far as the reasoning mechanism is concerned, the basic principle of reasoning that undergirds the Toulmin model is the law of large probability. This is just the opposite of hypothesis testing whose principle of reasoning is based on the law of small probability. Nevertheless, there exist remarkable resemblances between the two: while the warrant of the Toulmin model is just like the confidence level in a hypothesis testing; the rebuttal corresponds to the significance level (α). The larger probability the warrants entail, the smaller probability the rebuttals. To ensure that the claim is plausible, the rebuttals must be rare and exceptional so as to guarantee that the warrants are highly probable and the step from the data to the claim is secured. Thus, before a claim is made, it should be ensured that the warrants lend to a certain level of confidence; on the other hand, not all warrants of whatever probability can be rejected; otherwise, rational reasoning could be almost impossible.

2.4.3.2 The Base Argument of AUA

When applying the Toulmin model to build up the framework, Bachman (2005) makes a few changes to the basic structure of the model: (1) the *Q* element has been removed; (2) the *rebuttal* remains at its original position, but a new component, *rebuttal data*, has been added to justify the *rebuttal*—to “support, weaken or reject the alternative explanation” (Bachman 2005, p. 10); and (3) Bachman and Palmer (2010) change *rebuttal data* into *rebuttal backing*.

As can be seen in Fig. 2.8, all changes are targeted at the elements that Toulmin employs to attack the syllogism: the qualifier is gone, while the rebuttal is reinforced. This is somehow against Toulmin’s intention. The *qualifier* is what makes a

Fig. 2.8 AUA base argument
(Bachman 2005, p. 9)



Toulmin argument, without which the Toulmin claim is reduced back to a syllogistic one: being either yes or no, all or none; without which there is no need to consider the *rebuttals* in the first place. On the other hand, the *rebuttals*, generally exceptional and rare though negative to the claim, need to be considered for the claim to be plausible, but have to be ignored if any claim is to be made at all. However, in the modified versions, the *qualifier* is nowhere to be found, whereas the *rebuttal* is not let go.

2.4.3.3 Reasoning Logic

With the above, AUA is not entirely consistent with Toulmin’s argument model, especially in terms of its base argument. Then, what is its reasoning logic? An analysis of the roles of *rebuttal* and *rebuttal backing* will help to reach some insights. As is mentioned in the earlier discussion about substantial and analytical arguments, the backing includes straightforward matters-of-fact or truths and when factual backing is used to guarantee a claim, or a hypothetical statement for that matter, no reasoning is involved and no argument is necessary.

However, Bachman and Palmer (2010) change *Rebuttal Data* (the *Rebuttal* within a frame) in Fig. 2.8 into *Rebuttal Backing*. Therefore, as long as the rebuttal is to be verified within the reasoning process from the data to the claim, the reasoning process is undermined. As long as the rebuttal cannot be ignored, the claim is hardly convincing, or even predictable. In that case, the whole logic reasoning process falls into a never-ending regression.

The example in Fig. 2.9 is to illustrate how the rebuttal is supported by the rebuttal backing and thus the claim is rejected (Bachman and Palmer 2010). Based on the data *Jim is going to the hospital*, the claim, *Jim is sick*, is to be made (no claim yet); although the warrant, *People often go to the hospital when they are sick*, should provide enough guarantee to make the claim, we must check whether the rebuttal *Jim could be visiting someone who is in the hospital*, is true or not; it is true that *Jim is visiting his partner in the hospital*, so *Jim is not sick*.

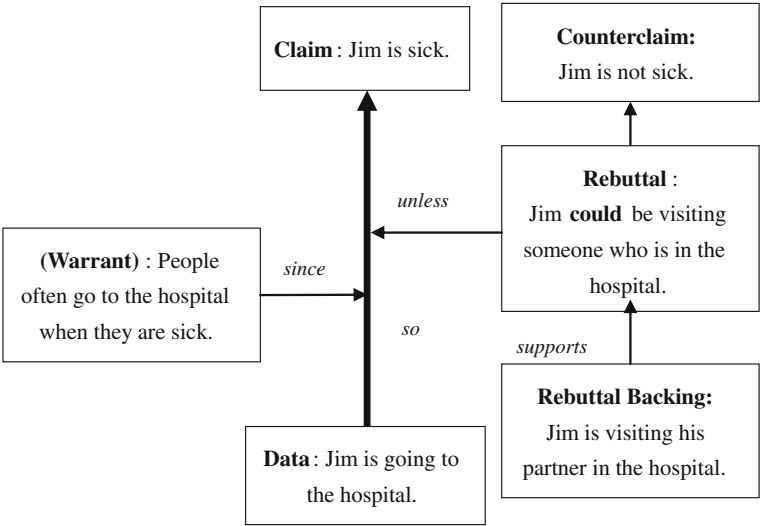


Fig. 2.9 Structure of example practical argument (Bachman and Palmer 2010, p. 97)

However, the above reasoning turns to: *Jim is going to the hospital*, so *Jim is not sick*. This does not seem to be the result of the Toulmin reasoning. If assembled in the form of a Toulmin argument, the reasoning should be as follows.

- A: Jim is going to the hospital (*SINCE* people often go to the hospital when they are sick, *UNLESS* they are going to the hospital for some other reasons,) *SO PRESUMABLY* Jim is sick.
- B: Jim is visiting his partner in the hospital (*SINCE* we can take it that people are not sick themselves when they are visiting someone in the hospital, *UNLESS* they are indeed sick themselves,) *SO PROBABLY* Jim is not sick.

This is how arguments are supposed to be settled. As can be seen, each side has its own claim; each claim is justified with a separate reasoning process; each process is guaranteed with its own warrant. Most importantly, both sides take into consideration the rebuttal, but neither is trying to verify the rebuttal in the same reasoning process, instead a proper qualifier is included.

If there is a need to reason by the logic of AUA, the rebuttal has to be verified as well. As can be seen in Fig. 2.9, even if Jim is visiting his partner, he may still be sick himself. If this rebuttal needs to be verified, chances are that the validation will fall into an endless paradoxical cycle. In other words, before any claim is made, the rebuttals must be verified first. As a consequence, another verification process is embedded in the current one so that in terms of model construction the model always contains “a self” within the model itself.

In brief summary, although argument-based validation can be viewed as a step forward in comparison with the unitary concept of test validity, caution might be taken in applying it to validate the proposed rating scale in question. In particular, it needs further exploration as to how to embed all sorts of validity arguments into a

coherent and significantly sufficient argument with construct validity as the core. Therefore, in terms of validation for the rating scale with nonverbal delivery embedded, the present study will still refer to a unitary notion of validity.

2.5 Rating Scale Evaluation and Validation

The previous section reviews the evolution of *validity* in language testing and justifies the application of a unitary concept as the theoretical base of validation for the rating scale to be proposed in the present study. Then, when it comes to the validation of rating scales, it is still felt necessary to review how rating scales can be validated.

With regard to the facets of rating scale validity, Knoch (2009) tailors Bachman and Palmer's (1996) framework of test usefulness and excludes the facet of inter-activeness because that is not an integral tenet that should necessarily be applied to rating scale validation. In addition, Knoch's (2009) revises the framework emphasises the role of construct validity of a rating scale and puts forward three criteria in validity evaluation as follows.

The scale provides the intended assessment outcome appropriate to purpose and context and the raters perceive the scale as representing the construct adequately...The trait scales successfully discriminate between test takers and the raters report that the scale is functioning adequately...The rating scale descriptors reflect current applied linguistics theory as well as research. (p. 65)

To briefly interpret the above criteria, in validating a rating scale, three aspects should be taken into account: (1) the extent to which a rating scale reflects the construct; (2) the extent to which a rating scale discriminates candidates across various proficiency levels; and (3) the extent to which a rating scale manifests a selected theory. Therefore, at the phase of rating scale validation, these three criteria serve as the guidelines in constructing the phase-specific research questions.

In terms of rating scale validation methods, both quantitative and qualitative methods are well documented. A majority of previous studies employ quantitative methods to validate a rating scale. Because a rating scale with explicitly defined categories facilitates consistent rating, a few studies examine whether differences between score categories are clear using multifaceted Rasch measurement (Bonk and Ockey 2003; McNamara 1996) or other factors impacting scoring results (Lumley and O'Sullivan 2005; O'Loughlin 2002). Besides, multidimensional scaling has also been applied to the scale development for different tests and rater groups (Chalhoub-Deville 1995; Kim 2009). More robust statistical methods, such as an MTMM approach and differential item functioning analysis, have been used for the validation of classroom assessment (Llosa 2007), of speaking tests (Kim 2001) or of a rating scale (Yamashiro 2002). It might be found that with an ever-growing involvement of statistical tools into the language assessment community, an increasing number of sophisticated statistical methods have been applied into and enriched the study of rating and rating scales.

On the other hand, qualitative methods are also increasingly employed in test validation studies (Lazaraton 2008), including speaking assessment validation (e.g. Lazaraton 1992, 2002, 2008). Commonly adopted methods can be rater verbal protocols and analysis of test discourse (e.g. Brown et al. 2005; Cumming et al. 2006). By aligning the rater verbal protocol with the descriptors stipulated in the rating scale, researchers are able to validate a rating scale supposedly reflective of the underlying construct a particular test intends to elicit. More elaborations will be made on qualitative approaches to test validation in the last section of this chapter.

In order to obtain more sources for the validation of the rating scale, both quantitative and qualitative methodologies will be employed in the present study. On the quantitative side, as the rating scale to be proposed touches upon formative assessment with a consideration of embedding nonverbal delivery as an assessment dimension, different traits from candidates' performances as reflected in their group discussions can be measured via different methods, such as teacher-rating and peer-rating; therefore, an MTMM approach will be adopted, which is rather suitable and powerful in addressing the extent to which different measures or methods that assess one given construct are substantially correlated among themselves. As for the qualitative side, since the main argument for validating the proposed rating scale is to validate the dimension of nonverbal delivery, an MDA approach will be used. Further justifications will be made after the related qualitative approaches to assessment validation are shed light on.

2.5.1 *Quantitative Validation Methods*

MTMM is first introduced by Campbell and Fiske (1959), who direct the attention of construct validity research typically to the extent to which data exhibit evidence in three areas, or meet three requirements. One is the concern of *convergent validity* (CV), referring to the extent to which different assessment methods concur in their measurement of the same trait. These values are supposed to be moderately high if the construct validity is probed into. The second concern is *discriminant validity* (DV), indicating the extent to which independent assessment methods diverge in their assessment of different traits. Contrary to the requirement for CV, the values for DV should demonstrate minimal convergence. The last consideration is method effects (MEs), deemed as an extension of DV. MEs represent bias that could possibly derive from using the same method in the assessment of different traits; correlations among these traits would be typically higher than those measured by different methods.

The original MTMM design (Campbell and Fiske 1959) receives criticism because more external, multiple and quantifiable criteria are expected to be incorporated into model perception (e.g. Marsh 1988, 1989; Schmitt and Stults 1986). Widaman (1985) also adds to the effect that the original MTMM design somehow fails to explicitly state the requirement of uncorrelated methods. Contingent upon these criticisms, Widaman (1985) proposes an approach of nested-model comparisons, where a baseline model is first perceived to be compared with other

alternative models which might be trait-correlated or method-correlated. This way of model formulation also signifies that MTMM per se derives from structural equation modelling (SEM).

Jöreskog (1993) categorises three types of model formulation modes: (1) strictly confirmatory, in which a single model is formulated and this model will be tested with empirical data and is either accepted or rejected based on interpretable parameter estimates; (2) model comparison, in which several alternative models are specified and tested with empirical data; (3) model generating, in which a tentative model is specified and keeps being testified based on an SEM analysis and substantive theory until a satisfactory model emerges. Widaman's (1985) framework of alternative model comparison mentioned above squarely falls into the second mode.

In the arena of language testing studies, since MTMM was first applied by Bachman and Palmer (1981) to examine the construct validity of the FSI oral interview, it has been extensively used in understanding the factor structure of test performance and language ability (e.g. Bachman and Palmer 1989; Hale et al. 1989; Turner 1989), in testing hypothesised relationships among test-taker characteristics and test performance (e.g. Kunnan 1995; Purpura 1999; Sasaki 1993), in multi-faceted approaches in construct validation (e.g. Bachman and Palmer 1981, 1982), in multi-sample analyses based on the salient personal attributes (e.g. Bae and Bachman 1998) and in validating classroom assessment (Llosa 2007). It can be felt that, however, this approach keeps a comparatively low profile when applied to rating scale validation, especially a validation study with a special view to observing whether different rating methods might lead to a similar measurement of the same construct. Considering the fact that this study investigates whether, if so, how different scoring methods, viz. teacher-rating and peer-rating, measure the given construct, Widaman's (1985) framework of alternative MTMM model comparison is adopted to investigate the relative effects of different scoring methods on the targeted construct of communicative language ability as reflected in rating scale with a dimension of nonverbal delivery included. In that case, the fittest and most interpretable MTMM model can be found.

When a decision is made on whether the data fits the model, the related goodness-of-fit statistics are referred to. Following the well-documented literature, this study would judge comparative fit index (CFI) and non-normed fit index (NNFI), whose values, if greater than 0.95, indicate acceptable model fit (Hu and Bentler 1999; Raykov and Marcoulides 2006). Adjusted goodness-of-fit index (AGFI), Tucker-Lewis index (TLI) and standardised root mean square residual (SRMR) in each model comparison will also be looked into. As reported in Hu and Bentler (1999), good model fit is indicated by AFGI and TLI values greater than 0.95, RMSEA values less than 0.06 and SRMR values less than 0.08. In addition, the root mean square error of approximation (RMSEA) is also calculated. Small residuals less than 0.05 indicate a small discrepancy between the observed correlation matrix and the correlation matrix estimated from the model (Hu and Bentler 1999). Therefore, when the proposed rating scale is validated, the above indices will be referred to for reaching the fittest model for data interpretation.

2.5.2 *Qualitative Validation Methods*

The previous section of review elaborates on the quantitative validation method to be adopted for validating a rating scale. This part of review will turn to the qualitative method for the rating scale validation in this study.

One important issue needs to be addressed before a presentation of the details of the qualitative validation method, that is, the necessity of applying a qualitative approach to assessment validation. It might be noticed that if a study only concentrates on the statistical methods for validating language tests or rating scales, the limitation would be that such validation can only be conducted after test administration because no score can be accessed beforehand. In other words, there can be no *priori* validation in its own right. In that context, there has been a growing awareness that language testers should consider more innovative approaches to test validation, “approaches that promise to illuminate the assessment process itself, rather than just assessment outcomes” (Lazaraton 2002, p. xi). Jacobs (1988) also emphasises the significance of qualitative approaches to test validation and views them as a must-do instead of might-do undertaking when he asserts that

[q]ualitative methods have been sufficiently successful that at this point the task is not to decide whether or not to admit them into the methodological arsenal of practising researchers; the task is to articulate their rationale so that they can be used in an informed and self-conscious fashion. (p. 248)

Against the above, it can be argued that without probing into the *de facto* assessment processes, especially if candidates’ performance is not investigated analytically with a qualitative approach, a full picture of what is tested conforms to what is intended to test will never be depicted. Therefore, for triangulation it is quintessential to apply a qualitative approach to validating the rating scale to be proposed.

As far as rating scale validation is concerned, there are mainly two prevailing qualitative methods: verbal protocol analysis (VPA) and discourse-based approach, particularly conversation analysis (CA). They are either singly adopted for rating or test validations, or orchestrated with other quantitative methods to triangulate research findings. The ensuing part is devoted to a review on both methods, followed by the details of MDA so that further justifications in addition to what is previously argued in the section of nonverbal delivery can be made for adopting MDA as the qualitative validation method in this study.

2.5.2.1 **Verbal Protocol Analysis**

When a need of microscopically looking at the process of rating arises, researchers might resort to VPA, based on which raters’ mental processing of what is being assessed and how their judgments are made can be verbally recorded and reflected. Green (1998) points out that VPA “is a methodology which is based on the assertion that an individual’s verbalisations may be seen to be an accurate record of

information that is (or has been) attended to as a particular task is (or has been) carried out” (pp. 1–2).

If the terrain is surveyed where VPA is empirically adopted, an overwhelming popularity can be found among those who bent on writing assessment rating. For instance, Cumming (1990) uses VPA to compare experienced and novice raters in their judgments on the criterion range of analytic assessment; Cumming et al. (2001, 2002) also examine the criteria extracted from the VPA data to come up with the general categories for essay evaluation. Similarly, this method is also employed by other studies in either describing rating process or comparing raters with various extraneous variables or characteristics (e.g. Connor and Carrel 1993; Erdosy 2004; Lumley 2002, 2005; Milanovic et al. 1996; Smith 2000; Vaughan 1991; Weigle 1994, 1999; Wolfe 1997; Wolfe et al. 1998).

However, applying VPA to speaking assessment rating seems to be underexplored. One of the few studies in spoken language assessment using that method is conducted by Brown et al. (2005). In their study, they use VPA to investigate rater orientation in the context of academic English assessment. The study finds that expert EAP teachers generally assess test-takers’ vocabulary skills and frequently comment on the adequacy of their vocabulary for a particular purpose. Ducasse and Brown (2009), also using VPA, finds teacher-raters can identify three interaction parameters in assessing paired oral communication, which yields implications for a fuller understanding of the construct of effective interaction.

It has to be admitted that using VPA enables researchers to validate a rating scale in terms of the extent to which raters score the candidates’ products in line with what is stipulated in the descriptors of a rating scale. In other words, it can mainly enhance the degree of scoring validity. However, when it comes to the construct validation of a rating scale, this method seems to be less powerful because it is very likely that the data elicited from rater verbal protocol does not necessarily cover whole thinking processes. Thus, VPA may record an incomplete reoccurrence of rater’s mind (Barkaoui 2011). When evaluating this method, Green (1998), Lumley and Brown (2005) also point out a few drawbacks of VPA. Besides its conspicuous disadvantage of time consumption, this method might also result in individual differences in the sense that respondents might either produce long or short reports of their mental processing. If not enough due attention is paid to the wording of verbal report elicitation, respondents’ verbal reports might also be disrupted as they could be somewhat coerced to “keep talking” (Ericsson and Simon 1993).

Coupled with the above drawbacks, most studies using this method outlined above also justify themselves in choosing VPA because most previous studies on the rating in writing assessment also heavily rely on this method. Considering the applicability of VPA in the present study, whose focus differs significantly from writing assessment, and given the practicality issue that VPA in the context of oral assessment might also consume even more time than that of writing, this method has to be discarded.

2.5.2.2 Conversation Analysis

Another main qualitative approach favoured by test validation researchers, as is aforementioned, is a discourse-based approach, which can be mostly represented by CA with its origin in sociology (Goodwin and Heritage 1990). As a matter of fact, CA covers an extensive scope of research applicability, ranging from validating language tests and rating scales to broadening the boundary of investigating the organisational structure of conversation. The latter is usually achieved by identifying the reoccurring patterns of naturally occurring conversation produced by speakers with various demographic variables. In other words, researchers manage to generalise the generic stages of conversation and try to model them on a turn-by-turn basis.¹ Turning back to how CA can be employed in language assessment, investigators analyse candidates' performances based on the transcription conventions of CA and further look into how their performances are aligned with a test construct or a rating scale. More specifically, CA can be viewed as an instrument to observe whether elicited performance by candidates can be correlated with test construct, as reflected by a rating scale. If qualitative descriptions of elicited data can be proven to provide positive correlation with the defined construct to a certain degree, it can then be reckoned that a test or a rating scale features construct validity.

Lazaraton (2002) summarises a few salient features of CA. One is that CA often deals with single cases, which is largely based on descriptive rather than statistical analyses. Given this, analysis results are usually "situationally invoked standards that are part of the activity they seek to explain" (Pomerantz and Fehr 1997, p. 67). Another is that CA "rejects the use of investigator-stipulated theoretical and conceptual definitions of research questions" (Pomerantz and Fehr 1997, p. 66). Therefore, it is usually not the practice of CA to formulate a hypothesis of what conversation patterns are before the data is analysed and generalised into "talk rules".

Unlike VPA that is much embraced by written assessment and validation, CA has its place in the empirical studies of speaking assessment given its nature that resides in spoken language. Young and He (1998b) conduct a number of studies on the assessment of spoken English with a discourse-based approach, particularly with CA. In that edited manuscript, researchers compare oral proficiency interview with natural conversation by looking at the turn, sequence and repair (e.g. Egbert 1998; He 1998). In addition, as reviewed in the section of rating scale, Fulcher (1993, 1996a) analyse candidates' responses and native speakers' talks qualitatively to operationalise the notion of *fluency*. Although he does not explicitly state his adoption of CA, the whole research procedure follows a CA approach in that he analytically extracts the (dis)fluency features differentiating learners across different proficiency levels on a turn-by-turn basis.

Similarly, from the lens of second language acquisition, Young (1995) analyses the rating scales of the ACTFL OPI Guidelines (ACTFL 1986) and Cambridge Assessment of Spoken English with a discourse-based approach. It is discovered

¹For detailed descriptions of *turn*, refer to Sacks (1992), Sacks et al. (1974), Oreström (1983).

that both rating scales share the weakness that there is a dearth of continuous development of language acquisition that is supposed to be reflected in rating scales. Another large-scale application of CA is mainly conducted by Lazaraton (1991, 1992, 1995, 1996a, b) on a series of Cambridge EFL examinations in both interview conversation structure of spoken language assessment and interlocutor/candidate behaviours. In these studies, she not only aligns candidates' responses with possible communicative functions to see whether the tests really elicit the intended construct, but also profiles the role of interlocutor in certain assessment settings.

Therefore, CA clearly serves as a necessary and reasonable complement to the validation of language tests. Psathas (1995) evaluates CA as "an approach and a method for studying social interaction, utilisable for a wide, unspecified phenomena... it is a method that can be taught and learned, that can be demonstrated and that has achieved reproducible results" (p. 67). However, largely due to its constraint of being applied to small-scale data, one of the criticisms against CA is that the analytic methodology itself and its descriptive categories adopted might be too vaguely defined to be usable and replicable to the studies of a similar nature (Brown and Yule 1983; Cortazzi 1993; Eggins and Slade 1997; Wolfson 1989). On the other hand, since CA is a method that involves much training and practice, most researchers have to consume more time to familiarise themselves with the transcribing conventions than to transcribe data (Hopper et al. 1986). On top of that, Schiffrin (1994) and Levinson (1983) also notice that CA seems less capable of bridging the gap between language form and language function.

Having critiqued the above, this part might call for an awareness that although CA is conducive to tracking speakers' utterances on a turn-by-turn basis, it is not equally powerful and explanatory to synchronise what happens non-linguistically with what is uttered verbally. The section reviewing rating scales already reiterates that a majority of prevailing rating scales do not assess candidates' nonverbal delivery. If all meaning-making resources need to be probed into, CA seems to be a dispreferred option. This is because although it might be argued that nonverbal delivery could still be transcribed using a "second-line" (Lazaraton 2002, p. 71), this method can neither align verbal delivery with nonverbal channels on a large scale, nor could it analyse interactions among different nonverbal channels, such as eye contact, gesture and head movement, as previously reviewed. Therefore, CA seems beyond its strength to be applied to the present study. In order to find a method that is able to scrutinise more meaning-generation resources, this study turns to an emerging discourse-based approach: MDA.

2.5.2.3 Multimodal Discourse Analysis

Having outlined the advantages of qualitative methods to assessment validation along with their complementarity to quantitative methods, this part then continues with a review on MDA, the qualitative validation method to be adopted in this study. Previous studies in speaking assessment are heavily dependent on the

transcription of verbal language, generally known as a single semiotic system. Nevertheless, the call for extending a semiotic system to a multifaceted one has long been overdue.

Halliday (1978, 1985), Chafe (1994), Halliday and Matthiessen (2004) contend that gesture, facial expression and so forth that accompany the discourse should also be regarded as semiotic modes capable of generating meanings. Likewise, there is already an explicit acknowledgement that communication is inherently multimodal, that literacy is not confined to language (Kress and van Leeuwen 2001; Levine and Scollon 2004) and that “all texts are multimodal” (Stein 2008, p. 25). Norris (2004) also shares the view that “all interactions are multimodal” and multimodality “steps away from the notion that language always plays the central role in interaction, without denying that it often does” (p. 3). Matthiessen (2007) regards multimodality as “an inherent feature of all aspects of our lives ... throughout human evolution” (p. 1). Zhu (2007) even points out certain possible danger if discourse is analysed monomodally. Therefore, it can be felt that an investigation into discourse should not strictly follow the propriety of verbal language exclusively. Instead, the weakness of the qualitative approaches outlined above should be overcome and the possibility of perceiving all the possible meaning-making resources, or other modalities and inter-semiotics, should be accordingly explored.

Definition and Research Scope

Anterior to unfolding what MDA can offer, two key concepts need to be clarified in foregrounding the notion. Since there are different approaches to MDA to be shed light on later, the definitions of key concepts could also be slightly different. If MDA is defined in a stratum-by-stratum manner, Stöckl (2004) views multimodal as “communicative artefacts and processes which combine various sign systems (modes) and whose production and reception calls upon the communicators to semantically and formally interrelate all sign repertoires present” (p. 9). Then, what is the point of mediating multimodal with discourse analysis? The main reason is that quite a portion of meaning is conveyed through nonverbal channels. In that case, communication should not be understood as a process realised only by one particular sensory organ. Therefore, the discourse elicited in such settings is multimodal discourse (Zhang 2009).

The stratum of multimodal discourse naturally extends to the method with which multimodal discourse is examined, viz. MDA. Jewitt (2006) thinks that MDA is a perspective from which discourse is analysed when all the communicative modes are deemed as meaning-making resources and that it depicts an approach that “understand[s] communication and representation to be more than language, and which attend to the full range of communicational forms people use—image, gesture, gaze, posture, and so on—and the relationships between them” (Jewitt 2009, p. 14). O’Halloran (2011), in a similar vein, defines MDA as “[extending] the study of language per se to the study of language in combination with other

resources, such as image, scientific symbolism, gesture, action, music and sound” (p. 120).

Having noted that MDA also looks at meaning-making resources other than verbal language alone, this section maps out the terrains that this particular method can cover. Simpson (2003) points out six domains that MDA mainly focuses on: (1) multimodality and new media; (2) application of multimodality in the academic and educational context; (3) multimodality and literacy; (4) construction of multimodal corpora; (5) multimodality and typology; and (6) MDA and its rationale. Baldry and Thibault (2006), however, posit six slightly different topics for MDA research: (1) what is multimodal text; (2) how to transcribe and analyse such text; (3) what technologies are needed to analyse multimodal texts and construct multimodal corpora; (4) how meaning potential can be exponentially increased when meaning-making resources from multimedia are applied to hypertext; (5) how to relate language studies to multimodality and multimedia; and (6) to what extent MDA can bring changes to linguistics. It can be felt that two things might be shared even though the above research domains vary slightly from each other. One is that the ultimate purpose of MDA is to perceive all the meaning-making resources, particularly those beyond the boundary of verbal language. The other is a trend that MDA can be applied to large-scale research by means of corpus construction. Bateman et al. (2004) and Bateman (2008) also believe that one of the multimodal study foci is to formulate an analytical framework for dealing with multimodal data in corpora. In fact, this domain is also foregrounded by the fact that previous discourse-based analysis methods usually fail to quantitatively account for and generalise research findings.

Nonetheless, even though MDA sets explicit directions for research and further development, there are still different approaches to or streams of MDA, as is foreshadowed. In order to select a suitable approach for this study and be consistent in a line of analysis, the following part introduces these approaches and reviews how they are applied to the studies related to Chinese EFL learners, and then justifications will be made to account for selecting MDA in this study.

Approaches to Multimodality

Broadly divided, there could be two approaches to MDA with different theoretical underpinnings. One of the approaches lay its foundation on Halliday’s (1978, 1985) social semiotics to language studies, in which all potential meanings are structured and construed in the sets of interrelated systems. Therefore, this stream is usually known as systemic functional multimodal discourse analysis (SF-MDA), whose bases are established by the works of Kress and van Leeuwen (1996, 1998, 2001, 2002, 2006; Kress et al. 2001, 2005; van Leeuwen 1999, 2001), O’Toole (1994, 2010), Baldrey and O’Halloran (2005, 2008a, 2011) and so forth. The other stream of MDA, whose rationale can be traced back to activity theory (Engeström 1987; Daniels 2001) (AT-MDA), draws upon interactional sociolinguistics and intercultural communication. That stream includes mediated discourse theory

(MDT) (Norris 2002, 2004; Norris and Jones 2005; Scollon 2001; Scollon and Scollon 2004) and situated discourse analysis (SDA) (Gu 2006a, b, 2007, 2009).

SF-MDA

One of the main reasons why SF-MDA emerges and develops exponentially is that its underpinnings can be directly loaned from systemic functional linguistics (SFL). Specifically, SF-MDA absorbs the notion of language as social semiotic and meaning potential and extends the boundary of meaning-making resources. In addition, with reference to metafunctional meanings, SF-MDA also believes that multimodal discourse is also multifunctional in that discourse is embedded with ideational, interpersonal and textual meanings. SF-MDA also develops the theory concerning register and associates the interpretation of discourse with the particular context of the discourse. All these features provide SF-MDA with a fit platform on which all the SFL-related theories, without any further modification, can immediately serve as its strong support.

Within the scenario of SF-MDA, most studies concentrate on the analyses and interpretations of pictorial system, especially within a framework of analysing visual text and its communicative meaning (Kress and van Leeuwen 1996, 2006). Congruent with ideational, interpersonal and textual meanings in the SFL studies, this framework describes meanings as not only *representational* (the representation of entities, physical or semiotic), but also *interactive* (images constructing the nature of relations among viewers and what is viewed) and *compositional* (the distribution of information value or the relative emphasis among elements of the image). Therefore, how images convey meanings also conforms to certain grammatical rules, which are beyond the conventional sense of grammar in linguistics. In their follow-up work (Kress and van Leeuwen 2001), having noted that the drawback of their framework lies in the isolated grammar for each individual modality, Kress and van Leeuwen (2001) draw the attention of perceiving all the modalities in a coherent context. The broad framework is supposed to identify the four strata of meaning making in any communicative practice, including *discourse*, *design*, *production* and *distribution*.

Other representative researchers also mainly take the lens of MDA on images. For instance, O'Toole (1994, 2010) applies the visual arts grammar to the analyses of paintings and architecture and reaches similar terms regarding meaning making: *representational meaning*, *modal meaning* and *compositional meaning*. Likewise, SF-MDA is also tailored to study other semiotic resources, including visual images (Kress and van Leeuwen 2006; O'Halloran 2008b); mathematical symbols (O'Halloran 2005), movement and gesture (Martinec 2000b, 2001, 2004), video texts and Internet sites (Djonov 2006; Iedema 2001; Lemke 2002; O'Halloran 2004) and three-dimensional sites (Ravelli 2000) as well.

The above research on SF-MDA frames indicates that this stream does have much to offer, especially when meaning-making resources other than verbal language are probed into. However, when briefing SF-MDA, Jewitt (2009) thinks this stream is not without flaws. It might dawn upon this review that most of the

analyses on images, symbols and among others, if not all, are rather impressionistic. In other words, if perceived by different researchers with varied cultural or educational background, the interpretations might diverge to a certain extent. The reason might be that SF-MDA is already linked the signifier with the signified to a great extent, yet the way their relevancy is interpreted is still based on subjective perceptions.

Another limitation pointed out by Jewitt (2009) is that “MDA is a kind of ‘linguistic imperialism’ that imports and imposes linguistic terms on everything” (p. 26). However, this limitation can be justified as most SF-MDA studies are undertaken within linguistics field. If MDA is intended to interpret a language system, there should be no “linguistic imperialism” to speak of. It might also be controversial that SF-MDA is only concerned with static discourse, such as image, architecture, as how they convey meanings through different channels. Nevertheless, this flaw can again be defended by the fact that even though most SF-MDA studies focus on those static discourses, it does not necessarily follow that it would be powerless in dealing with dynamic discourses, such as situated discourses embodying human actions. This can be supported by Hood’s (2007, 2010, 2011) studies, in which an SF-MDA approach is adopted to present a multimodal analysis of a poet’s performance and the role of body language in face-to-face teaching. Therefore, although this flaw exists, it might be caused by a lower profile of SF-MDA on dynamic discourse instead of the powerlessness of the approach per se.

AT-MDA

In addition to applying SFL to studying various modalities, a host of researchers are also interested in basing their MDA studies on activity theory.

By integrating sociolinguistics, ethnolinguistics, intercultural communication, Scollon (2001 and Scollon and Scollon 2003) proposes MDT that integrates social activity with discourse. This is a step forward in that previous discourse analysis studies usually neglect the significance of activity, whereas sociology theories, in most cases, do not take discourse into account either. Unlike a conventional sense of discourse analysis, which treats a text or a genre as the unit of analysis, MDT mainly looks at mediated action and “social actors as they are acting because these are the moments in social life when the discourses in which we are interested are instantiated in the social world as social action, not simply as material objects” (Scollon 2001, p. 3). According to Scollon (2001), any social actor conducts a mediated action by means of material objects (including the actor’s own dress, body and so forth) in the material world. Based on Scollon’s framework of AT-MDA, Norris (2002, 2004) devises a MDA framework, where mediated action is still taken as the unit of analysis. Her framework substantiates AT-MDA in the sense that she further distinguishes different sorts of mediated actions into low-level action (a simple gesture) and high-level action (a series of concrete actions) and that the framework quantifies the degree of complexity for high-level actions by ushering in the notion of *mode density* (Norris and Jones 2005).



Fig. 2.10 Content and medium layers in agent-oriented modelling (Gu 2006a)

Informed by activity theory, Gu (2006a, b) also establishes another AT-MDA framework, setting a perspective on studying multimodal texts from content unit and medium unit. Figure 2.10 is an example to illustrate a distinction of these two units (Gu 2006a). The top-screen shots are a series of contiguous actions by an attendant providing in-flight service; thus, this can be viewed as a content unit, symbolising concrete acts of service. A medium unit is realised by the duration of and the time frame of this act. Thus, a multimodal text is composed when these two units are combined. Based on these considerations, Gu (2006b, 2009) proposes agent-oriented modelling (AOM)² to frame situated discourse by social actors, in the case of which total saturated experience can be distinguished from total saturated signification. The former refers to “face-to-face interaction with naked senses and embodied messages”, while the latter is more concerned with “the total of meaning constructed out of the total saturated experience by the acting co-present individuals” (Gu 2009, p. 436).

Compared with SF-MDA, it can be informed that AT-MDA usually takes a stance that does not hurriedly establish a link between the signifier and the signified. In other words, in dealing with multimodal texts, AT-MDA usually faithfully presents what can be observed objectively. However, as the issue of interpreting the observation remains to be resolved, this stream tends to advocate more objective methods, such as layman validation of what a particular gesture is signified (Gu, personal communication, 5 December 2010). Objective though it appears, it can still be foreseeable that layman validation will result in even more inconsistency in interpretation because layman involvement in great numbers might end up with diversified interpretations.

In addition, AT-MDA lays a comparatively higher demand on technology literacy, the area of which most researchers in applied linguistics might find challenging, especially with regard to transcription, markup and modelling languages. Although Gu (2009) signals a tripartite division of labour in corpus-based MDA studies to facilitate research, logistic issues of how different parties are pooled

²Gu (2006b) uses the term agent-oriented modelling language (AML), yet he later changes the term to agent-oriented modelling (AOM) because AOM perceives the modelling as a methodology, while AOML emphasises its relation with UML as the modelling metalanguage (Gu 2009).

together still remain to be solved instantly. What distinguishes AT-MDA from SF-MDA can also be that most AT-MDA studies, if not all, deal with situated or mediated discourses; therefore, this approach can satisfactorily explain dynamic discourse, viz. discourse that usually embeds humans' contiguous actions.

MDA in the Chinese EFL Context

When MDA is applied as a domain of enquiry in discourse-based academia, studies tend to explore a diversified range of meaning-making resources, such as films (Baldry and Thibault 2006; Iedema 2001; Martinec 2000a; O'Halloran 2004; Tseng and Bateman 2010), animation (O'Toole 2011) and colour (van Leeuwen 2011). In addition, a number of school subjects are also investigated multimodally, such as mathematics (O'Halloran 2000, 2005, 2009), science (Guo 2004; Jewitt 2002; Kress et al. 2001; Kress 2000), English (Daly and Unsworth 2011; Jewitt 2002, 2011; Kress et al. 2005; Macken-Horarik et al. 2011; Unsworth and Chan 2009) and history (Derewianka and Coffin 2008). A variety of media are also the research foci of MDA, such as picture books (Guijarro and Sanz 2009; Martin 2008; Painter 2007, 2008; Painter et al. 2013), comic books (Kaindl 2005), newspapers (Bateman et al. 2006; Caple 2008; Knox 2008; Macken-Horarik 2004), advertisements (Feng 2011; O'Halloran and Lim 2009), documents (Baldry and Thibault 2006; Bateman 2008), television advertisements (Thibault 2000; Baldry and Thibault 2006; Tan 2009), websites (Lemke 2002; Kok 2004; Djonov 2006, 2008; Tan 2010) and online virtual world (Maiorani 2009).

All the above studies to a great extent consolidate the theoretical base of MDA and further inform the directions to which MDA can be applied in the Chinese EFL context, where this method per se is just emerging. Studies that apply MDA as the rationale in the Chinese EFL context would mostly focus on how teaching and learning can be facilitated by multimodal inputs and how meanings can be instantiated in an unconventional fashion. A review of the previous MDA studies can help us to better understand the *status quo* of its application in the Chinese EFL context and further inform us of how this approach can be further lent itself to the arena of language assessment, rating scale validation in particular.

The MDA studies previously reviewed can be basically categorised into three aspects. First, a number of studies link multimodality with multi-literacy, stressing how EFL learners' multi-literacy can be fostered by an input with a combination of verbal language and other meaning-generative means, such as visual and auditory channels (e.g. Chen 2008; Zhu 2008). Other studies in this stream encourage an interrelation between verbal language and other visual input by analysing EFL learners' PowerPoint slides in order to highlight how other channels can enhance meaning making (e.g. Hu and Dong 2006; Wei 2009; Zhang 2010). Second, MDA studies in China also integrate multimodality with how specific micro-skill language teaching can be administered (Zhang 2010), such as listening (Long and Zhao 2009; Wang 2009) and speaking (Zhang and Wang 2010). The third category that applies MDA to the studies pertinent to the Chinese EFL context is pertaining

to English textbook evaluation. With meaning-making resources as a point of departure, researchers critique the layout, illustrations or colours of language textbooks in relation to what is conveyed verbally. For instance, by referring to the framework of ideational meaning, Chen and Wang (2008) assess the image-text relations and their differences across a range of scaffolding stages. Similarly, Chen and Huang (2009) adopt the framework of interpersonal meaning to further examine potential problems in language textbook compilation.

Common grounds of the above studies can be twofold. On the one hand, most of these MDA studies are based on exploring the possibility of improving language learning and teaching with a repertoire of modalities. In a sense, their advantage can be seen via a perception that multimodal input can stimulate sensory organs of EFL learners. On the other hand, most MDA studies above follow an SF-MDA approach, whereas the paucity in applying AT-MDA can be evidently found. Part of the reason can be that AT-MDA might not be that suitable to account for meaning-making channels as far as printed texts are concerned. This is because that AT-MDA mostly deals with mediated or situated discourse, or dynamic discourse as a whole.

Another point worth attention is that although there are prolific studies addressing the issue of how different modalities may interact for an enhancement of effective language learning, a similar issue of how learners' employment of different modalities in their output is yet to be resolved. Additionally, MDA is still an untouched approach in language assessment. Therefore, if a rating scale particularly with nonverbal delivery incorporated also takes the above points into account and is further validated with this approach, this study can enrich qualitative validation methods in language testing.

An Integrated Evaluation

Having reviewed both approaches to MDA and how MDA is employed in the Chinese EFL context, this part comes to an integrated evaluation and justifies this approach in the present study. What needs to be addressed first is that there is no absolute distinction as to which approach is right or wrong. Gu (2006b) expresses his concern over foreseeable collaboration between SF-MDA and AT-MDA though both approaches have solid foundations in their own right. Indeed, considering the ultimate research purpose and explanatory power, both approaches are not contradictory; their divergence only lies in different perspectives of looking at multimodal discourses and meaning-making resources. SF-MDA treats multimodal texts on the basis of social semiotics in its fullest sense. By comparison, as AT-MDA focuses more on how discourse is realised in a social activity context, it can be fully operated in dynamic discourses.

This study adopts SF-MDA based on the following considerations. Regarding the nature and aims of this study, which intend to design and validate a rating scale with a consideration of embedding nonverbal delivery in speaking assessment, it should be noticed that nonverbal delivery will be looked into to a great extent. As is

critiqued above, AT-MDA seems less explored in dealing with static discourse, while SF-MDA can be applied to both static and dynamic discourses though previous studies have not rendered great concern for dynamic discourse. In that case, if full use is made of SF-MDA to probe into a static discourse and more potentials of SF-MDA are tapped to analyse a dynamic discourse, this study can not only qualitatively analyse how candidates perform, but also benefits SF-MDA in terms of its extended scope of applicability. It may be argued that both SF-MDA and AT-MDA can be applied to the present study in an interwoven manner as both have their strengths in approaching different types of multimodal texts. However, adopting SF-MDA does not necessarily mean that both approaches are not reconcilable; rather, the decision on adopting SF-MDA follows the principle of consistently referring to the same framework and applying it to qualitatively validate the rating scale to be proposed.

By static discourse, it mainly refers to the transcription of candidates' verbal language, while dynamic discourse takes a closer look at candidates' nonverbal delivery. To be more specific, at the rating scale validation stage, when candidates' performances are investigated to be aligned with their analytic scores and the descriptors regarding verbal utterances, all possible meaning-making resources will be analysed with SF-MDA as the theoretical framework. On the other hand, when how candidates perform and synchronise their verbal language with nonverbal delivery, then SF-MDA will also be referred to.

Apart from a consideration of discourse nature, another concern is that since MDA will only be adopted in the qualitative stage of rating scale validation, the randomly selected samples will not be that large in scale compared with those when MTMM, a quantitative approach, is utilised. Therefore, the previously mentioned weakness of SF-MDA's reliability in directly bridging what is signifier and what is signified can be offset to the minimum degree. Otherwise, if all the samples are to be analysed with an SF-MDA approach, it is felt that analyses will wind up with an almost endless inventory, giving rise to other logistic issues jeopardising the practicality or implementation of this study. Furthermore, as is also aforementioned, AT-MDA demands higher level of technology literacy, which might constrain this study.

It could also be argued that since nonverbal delivery can be probed into within the paradigm of nonverbal communication studies, why this study will adopt SF-MDA as the validation method for the rating scale to be proposed. Scollon and Scollon (2009) also note the similarities between the current interests in multimodality with the research in the field of nonverbal communication, as best represented by the works by Pike (1967), Ruesch and Kees (1956) and Hall (1959). However, while acknowledging that the work in nonverbal communication can inform multimodal studies, they highlight that "it is not simply a return" as the crucial difference is that "[n]o longer is language taken to be the model by which these other phenomena are studied, but, rather, language itself is taken to be equally grounded in human action with material means in specific earth-grounded sites of engagement" (Scollon and Scollon 2009, p. 177).

Based on all the above considerations, this study employs an SF-MDA approach in the qualitative validation of the rating scale to be proposed and the fine-grained reference to MDA henceforth is SF-MDA. At this stage, however, what still leaves blank is how to apply the framework of MDA to operationalise the rating scale validation. The next part will sketch out an operationalised framework informed by MDA and provides a revised one drawn from Martinec's (2000b, 2001, 2004) and Hood' (2007, 2011) studies.

Applying MDA to Rating Scale Validation

In line with an MDA approach, three strata of meaning-making resources are focused on in the present study. The first stratum is a *semiotic system*. As is illustrated before, all semiotic resources available for meaning generation can be regarded as modes. Jewitt (2006) adopts *mode* as a foregrounding stratum as "concentrating on the semiotic resources of individual modes as they feature in a text is one way to 'prise open' a text" (p. 40). Within this stratum, this study can find how candidates deploy a range of resources and further assign meanings with them.

However, the first stratum only deals with whether or not these semiotic systems are utilised, rather than how they are put into use in relation to meaning making. If attention is only placed on one semiotic system, the texts will then be fragmented to realise only part of meaning potential. In that case, no interaction between different semiotic systems can be instantiated. This naturally leads to the second stratum of the framework, namely the *metafunctions* of meaning. Halliday (1978, 1985) classifies all social functions into three metafunctions. Concurrently each social semiotic is conveyed with the construal of the world around us and inside us (*ideational meaning*), meaning relating to interaction between speaker and addressees (*interpersonal meaning*) and how it is structured and created (*textual meaning*). Likewise, MDA also applies metafunctions to all modes so as to see how different modes interact and how their juxtaposition and relations realise meanings.

When discourse metafunctionality is associated with the present study, how different semiotic systems, especially those instantiated via nonverbal channels, are interrelated can be discerned. For example, candidates' nonverbal delivery is supposed to construct ideational, interpersonal or textual meaning in group discussion in formative assessment. Let us take gestures as an example, candidates' gestures can instantiate *ideational meaning* (the social ensemble of a particular gesture), *interpersonal meaning* (how gesture is made to influence the interpersonal relation and intangible distance) as well as *textual meaning* (how a gesture is frequently made to achieve transition in expression).

What should be noted is that analysing these metafunctions alone would still be incomplete. Against this, this study will extend to a third stratum: intersemiotic relations. In other words, when different modes are utilised and what metafunctions they instantiate are interpreted, how they interact with each other will also be scrutinised. It is possible that different nonverbal semiotic systems can be mutually

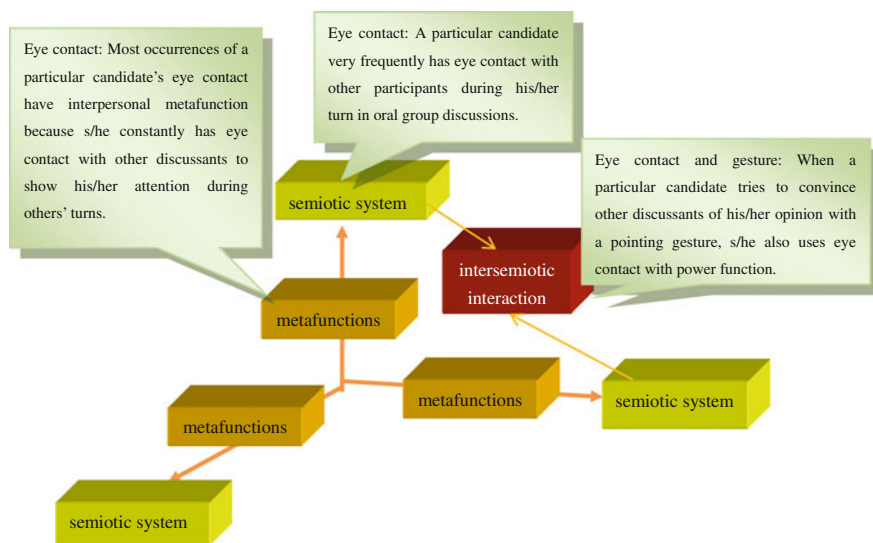


Fig. 2.11 Three-stratum MDA framework: an example

enhanced, which can be judged as the inter-semiotic complementariness, while the reversed way is also possible given the fact that one nonverbal semiotic system is not in full conformity with another synchronised nonverbal channel or accompanying verbiage.

Figure 2.11 depicts an example illustrating the mechanism of a three-stratum MDA framework. Although this figure is three-dimensional in appearance, it does not necessarily follow that only three strata are assigned to the observed semiotic system. As a matter of fact, the number of dimensions is determined by the number of modes observed. Therefore, if a multimodal text is analysed, the discourse per se can be actually n -dimensional. Alongside each semiotic system, three metafunctions are concurrently embedded and the relation between two modes or among more than two semiotic systems, if any, can lead to intersemiotic interaction. Therefore, it does not necessarily mean that only two semiotic systems will interact; three or even more can also interact beyond a mere depiction in Fig. 2.11.

The example illustrated in Fig. 2.11 is a semiotic system of *eye contact*. On the stratum of *semiotic system*, it might be observed that a particular candidate very frequently has eye contact with other participants during his/her own turn in the group discussion. Then, this meaning-making resource is analysed from three *metafunctions*; it can be interpreted that most occurrences of his/her eye contact have an underlying interpersonal metafunction because s/he constantly has eye contact with other discussants during others' turns to show his/her attentiveness. Elevated to a higher stratum of *inter-semiotic relations* in the framework, this semiotic system can be found to interact with other semiotic systems, such as gestures. The semiotic system of eye contact co-ordinates well when s/he gazes

with a power function to convince other discussants of his/her own opinion with an upward pointing index finger. In that manner, the three strata are comprehensively probed into and candidates' performance can be qualitatively aligned with the rating scale descriptors and the subscores assigned by teacher and peer raters.

The above general framework provides a sketch of how candidates' nonverbal delivery can be analysed from an MDA perspective. In order to particularise a repertoire of nonverbal delivery channels and re-address the analysis framework that is held back in the previous section of review on nonverbal delivery, this study will mainly refer to Martinec's (2000b, 2001, 2004) and Hood's (2007, 2011) studies in qualitatively validating the rating scale to be proposed.

Nonverbal Delivery: Communicative Versus Performative

This study, in the phase of rating scale validation, divides nonverbal delivery into *communicative* channel and *performative* channel, which is also in alignment with Kendon's (1981, 2004) and Cienki's (2008) studies in describing nonverbal delivery, particularly with regard to gestures. In terms of its relationship with verbal language, communicative channel is further classified into *language correspondent* channel, *language independent* channel and *language dependent* channel. Language correspondent channels refer to those that co-occur with accompanying verbiage, but their meanings can be accessed and interpreted without relying on speech. Language independent channels occur in the absence of language and generate meaning on their own. *Language correspondent* channels can be distinguished from *language independent* channels mainly by the criteria of whether there is accompanying verbiage in the occurrence of nonverbal delivery. *Language dependent* channels also co-occur with language but request the accompanying verbiage for a full access to and interpretation of their meanings.

Performative channels are mainly nonverbal delivery practically performed for the execution of a task. It may not be semantically loaded or wilfully performed to convey meaning. An example of performative eye contact can be a sudden downward eye contact directionality shift when the discussant is questioned. Another example of performative gesture can be scratching one's neck to ease an itch. While the primary intent of performative channels is not to communicate, they may, at times, be construed to convey meaning, thus serving as communicative channels as aforementioned. For instance, an act of scratching one's head can be a performative gesture as a reflex to an itch. However, it can also be interpreted as a communicative gesture to suggest uncertainty. As observed, the boundary between the classification of Communicative and performative channels might be nebulous. Nonetheless, the intended meanings are usually disambiguated when a particular occurrence of nonverbal delivery is interpreted in a co-contextualised manner. Hence, it is arguably useful not to disregard performative channels in this study, despite them being not primarily communicative in nature.

It can be felt that instead of describing the communicative functions of nonverbal delivery channels as reviewed in the first section of this chapter, the above taxonomy is one step forward in that it also considers the role of accompanying

verbiage and how it interacts with what happens non-linguistically. In addition to the above demarcation of nonverbal channels in relation to verbal language, more fine-grained frameworks (Hood 2007, 2011; Martinec 2000b, 2001, 2004) are reviewed below for an integrated framework for validating the rating scale.

Martinec's Taxonomy on Actions

Martinec (2000b) proposes that actions can be classified into Presenting Action, Representing Action and indexical action. Martinec (2000b, p. 243) defines Presenting Action as “most often used for some practical purpose” and “communicates non-representational meanings”. They are classified as performative channels in this study. Representing Actions “function as a means of representation” and are semantically loaded. They are classified as communicative channels in this study. In terms of its relationship with language, Representing Action can also be described as Language correspondent channel or language independent channel in this study. Indexical action usually only co-occurs with accompanying verbiage and “in order to retrieve its full meaning, one has to have access to the second-order context which is represented simultaneously in Indexical action and concurrent speech” (Martinec 2000b, p. 244). Indexical Action is therefore classified as communicative channel and is described as language dependent channel in this study. A synthesis of the above review can reach an integrated framework, as is outlined in Fig. 2.12, where the taxonomy of communicative and performative channels and Martinec's (2000b, 2001, 2004) taxonomy of action types, along with their relationship with verbal language, are hierarchically connected.

According to Martinec (2000b, p. 247), *Presenting Action* can be “seen as part of our experience of reality, formed in our interaction with it by means of our perceptions and motor actions”. As such, Martinec (2000b) adapts the Hallidayan

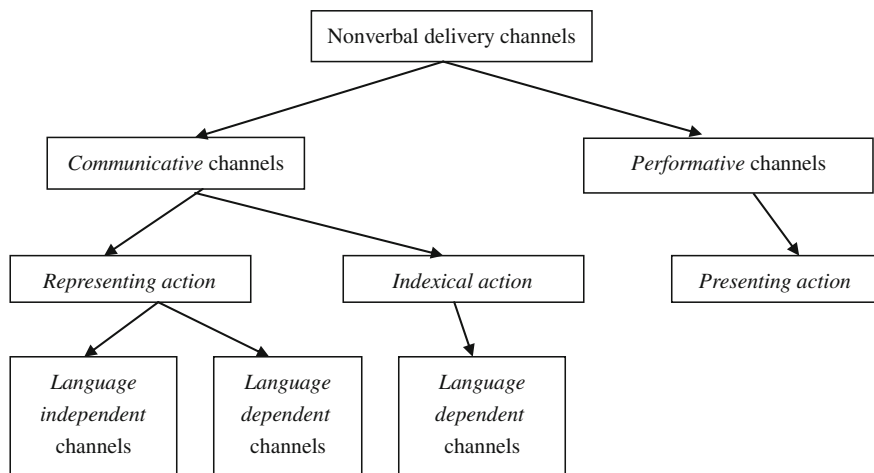


Fig. 2.12 An integrated taxonomy of nonverbal delivery channels

processes of transitivity (Halliday 1978, 1985) to *Presenting Action*. The different types of *Presenting Actions* are distinguished according to the processes of transitivity in systemic functional theory. They are *Material* process, *Behavioural* process, *Mental* process, *Verbal* process and *State* process.

The classification for Material processes, defined by an obvious expansion of effort, such as moving a chair forward, can be straightforward. Martinec (2000b, p. 247) claims that “behavioural processes are similar to Material processes in that they involve an expenditure of energy but they differ in that the main participant, called Behavior, must be conscious”. This distinction can arguably be blurred as almost all occurrences of nonverbal delivery must necessarily be enacted by a conscious individual though not all of them are intentional. Martinec (2000b) further describes an act of *kicking a ball* as Material processes and an act of grooming such as combing as Behavioural processes. It may be controversial that in both situations, there must necessarily be an enactor of the action that is conscious. Perhaps a more distinct classification is whether an action is directed to self, described as a Behavioural process, or directed to others or to objects, described as a Material process. Examples of Behavioural process might include laughing, and physiological processes like coughing.

Martinec (2000b, p. 249) also proposes the category of State processes to describe processes without salient movement, or those without obvious consumption of energy, such as sitting and standing. Verbal processes have two realisations: visual and auditory. Martinec (2000b, p. 248) asserts that “the visual realization is the lip movement which articulates sound in the way that is done for speech” and “the auditory realization is speech sounds”. As neither facial expression nor lip movement is within the scope of nonverbal delivery channels to be investigated in this study, Verbal processes in Presenting Action are discarded accordingly.

Martinec (2000b) believes that there are no Mental processes in action unlike in language as “they are processes of cognition and take place in the mind, which is not directly observable” (p. 250). However, these “processes of cognition” might be expressed in language as mental processes through such mental verb as *think* and *consider*. In a similar vein, it is arguably possible to identify the realisation of mental processes. From the analysis of candidates’ performance in nonverbal delivery, it might be found that indicators of cognition may be suggested by an act of a finger pursing at the chin.

Representing Action can be certain nonverbal delivery with a signifying function in a given sociocultural context (Martinec 2000b). They are either universally recognisable or within a semiotic community. The ideational meanings instantiated by Representing Action are classified as *Participants*, *Processes* and *Circumstances*, and they are usually realised in the case of gestures. *Participants* can be the physical entities that gestures refer to, such as a Representing gesture with the reference to an object, such as “village”. Martinec (2000b, p. 253) suggests only two kinds of *Processes* for Representing Actions: *static* and *dynamic*. For example, certain ongoing actions can fall into this category, such as “scuba-diving”. *Circumstances* can be those indicating concrete directions or locations, such as a gesture accompanying the verbiage of “outdoors”.

Table 2.3 Ideational meaning of nonverbal delivery channels

	Ideational meaning		
	Presenting action	Representing action	Indexical action
① Eye contact	✓ Processes	✓ Entity	✓ Importance ②③
② Gesture	Material ①②	Participants ②	✓ Receptivity ②③
③ Head movement	Behavioural ②	Process ②	✓ Relation ②
	State ①②	Circumstances ②	✓ Agreement ①
	Mental ①②③		✓ Uncertainty ①
			✓ Defensiveness ②

The third category of action delineated by Martinec (2000b) is indexical action. They are communicative channels and language dependent channels because they necessarily accompany language for an accurate interpretation. From the data of candidates' group discussion, certain indexical actions can be interpreted via an understanding of the accompanying verbiage.

However, it should be noted that the above framework needs revisiting when applied to describing and validating nonverbal delivery occurrences by candidates in the Chinese EFL context. In addition, three main nonverbal delivery channels, viz. eye contact, gesture and head movement, will instantiate intended meanings by various means. Not all the realisation of meaning potentials reviewed above can be generated by or come into effect through three main nonverbal channels. For example, in presenting actions, only gesture is able to realise a behavioural process because neither eye contact nor head movement would be embedded with something behavioural or a bodily motion.

Integrating the above review, therefore, this part synthesises an analytical framework for a repertoire of actions with regard to their possible ideational meaning with an MDA approach, as is outlined in Table 2.3. What is supposed to be observed in each type of action for a specific nonverbal channel is indicated by a figure following the observation point. For Presenting Action, which does not virtually serve a signifying function or embody semantic meanings, its ideational meaning is usually realised via Processes, which might incorporate Material, Behavioural, State and Mental processes, as previously reviewed. Independent of language, Representing Actions realise their ideational meaning through entities, which include Participants, Process and Circumstances. As for indexical actions, usually requesting a co-contextualisation for interpretation, realise their ideational meanings by possibly indicating their importance, receptivity, relation or other context-specific meanings in certain semiotic contexts.

Hood's Taxonomy on Nonverbal Delivery Metafunctions

The above analytical framework deals with ideational meaning that would be possibly instantiated by nonverbal channels. This part of review will continue with Hood's (2007, 2011) studies on nonverbal delivery metafunctions, particularly in relation to interpersonal and textual meanings so that a complete analytical framework for qualitatively validating the rating scale can be constructed.

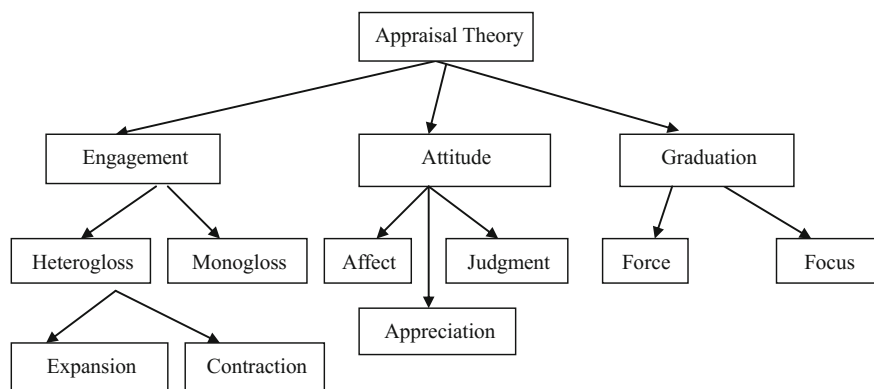


Fig. 2.13 The structure of Appraisal Theory (Martin and White 2005, p. 38)

Building on the work by McNeill (1992, 1998, 2000) and Enfield (2009) in cognitive studies as well as Kendon's (1980, Kendon 2004) research in psychology, Hood (2007, 2011) takes an SFL perspective to investigate nonverbal delivery, with a special to gestures. In terms of interpersonal meanings, Hood (2011), informed by Appraisal Theory (Martin 1995, 2000; Martin and White 2005), identifies gesture that embodies attitude, engagement and graduation, as illustrated in Fig. 2.13. Hood (2011) further argues that nonverbal channels, such as gestures, can express feelings and values in attitude can grade meaning along various dimensions in graduation and can expand or contract space for others during interaction in engagement.

In Appraisal Theory, attitudes can instantiate a variety of interpersonal meanings. However, considering the three main nonverbal channels in the present study, a polemic set of values that broadly classify attitudes as *Positive* and *Negative* are proposed. This is because, unlike facial expression, eye contact, gesture and head movement generally signify either positive or negative attitude instead of affect, appreciation and judgment, as outlined in Fig. 2.13. For instance, positive attitude can be embodied in an occurrence of head nod, while negative attitude can be instantiated by the gesture of crossing both hands before the chest when a candidate intends to interrupt other speakers.

Graduation in interpersonal meaning is also elaborated by Hood (2004, 2006). She is concerned, however, that "by grading an objective (ideational) meaning the speaker gives a subjective slant to the meaning, signalling for the meaning to be interpreted evaluatively" (Hood 2011, p. 43). In line with Appraisal Theory, Hood (2011) extends *graduation as force* to the meanings of intensity, size, quantity, scope and *graduation as focus* to specificity. Instead of addressing all the aspects, this study will mainly look at the pace of different nonverbal delivery occurrences, such as the frequency of head nod in an interval unit.

The third aspect of Appraisal Theory is *engagement*. Specific to gestures, engagement is realised via the positioning of the hands to expand or contract negotiation space for other addressees. In describing interpersonal meanings instantiated by teachers' gestures, Hood (2011) suggests an open palm or palms-up

position as “[embodying] an elicitation move on the part of the teacher, enacting an expansion of heteroglossic space, inviting student voices into the discourse” (p. 46). By contrast, a palms-down gesture contracts space for negotiation. However, in addition to expansion and contraction, there can also be neutral engagement, which takes up most of the time in candidates’ group discussion, and possibility, which, for example, can be instantiated by an occurrence of placing the left hand against the tip of the nose with the index finger and the thumb gently touching the face.

Although it has to be admitted that the above taxonomy is intentionally applied in investigating gestures, with moderations, it can be applied to eye contact and head movement as well. In fact, as this taxonomy covers almost all the possible interpersonal meanings instantiated by gestures, which supposedly convey more meanings than eye contact or head movement, the application of this taxonomy can be justified to analyse interpersonal meaning of candidates’ eye contact and head movement in this study. It is a similar case when the taxonomy of textual meanings is applied to eye contact and head movement below.

Therefore, in this study, interpersonal metafunction generally covers representing and indexical actions, as listed in Table 2.4. In line with Hood’s (2007, 2011) work on interpersonal meaning of nonverbal delivery, interpersonal meaning can be realised via attitude, engagement and graduation. Irrespective of any nonverbal channel, attitude is categorised into positive and negative. The judgment can be facilitated and triangulated with reference to synchronised verbal utterances. Engagement is broken down into expansion, contraction, neutral and possibility, and graduation is realised by the pace of nonverbal channels (fast, medium or slow).

For textual meanings, Hood (2011) describes the identification, waves of interaction, salience and cohesion in gesture. Mainly the wave of gestures can be realised via an occurrence of repeated action, for example, constant or rhythmic beat at a certain object. Each wavelength presents a peak where prominence is given to the meaning conveyed (Martinec 2004). This can be especially true in indexical gestures, where beats are supposed to offer an enhancement of importance intended in the ideational meanings.

Following Hood’s (2011) line of analysis, another aspect of textual meaning can be realised through pointing. Hood (2011) proposes not only the dimension of specificity but also the dimension of directionality accorded by Martinec’s (2004) study. Hood (2011, p. 38) also argues that variation in bodily resources can be interpreted “as varying along a cline of specificity”. In that sense, textual meanings can be interpreted differently when pointing is realised by different fingers or a combination of more than one finger, or by a palm.

Table 2.4 Interpersonal meaning of nonverbal delivery

	Interpersonal meaning	
	Representing action	Indexical action
Eye contact	✓ Attitude: Positive, Negative	
Gesture	✓ Engagement: Neutral, Expansion, Contraction, Possibility	
Head movement	✓ Graduation: Fast, Medium, Slow	

Table 2.5 Textual meaning of nonverbal delivery

	Textual meaning	
	Representing action	Indexical action
Eye contact	★ Gaze target ✓ Directionality: various objects or no direction ✓ Specificity: duration of gaze	
Gesture	★ Pointing ✓ Directionality: various objects ✓ Specificity: Hand, index finger, thumb, thumb and index finger ★ Wavelength ✓ Rhythm: Once or consecutively many times	
Head movement	★ Wavelength ✓ Rhythm: Occurrences of head nod/shake in a unit interval ★ Amplitude ✓ Specificity: Angle of head movement	

Slightly different from the application of the taxonomy of interpersonal meaning to eye contact and head movement, Hood’s (2007, 2011) framework with regard to textual meaning is somewhat extended and revised for an analysis of eye contact and head movement. Table 2.5 presents the analytical framework for textual meaning of nonverbal delivery channels. For eye contact, the target which gaze is aimed at can achieve its textual meaning mainly from the perspectives of directionality and specificity. Various objects or no concrete object (direction) from candidates’ gaze, as well as how long a gaze fixes on an object can interpret textual meaning. In a quite similar vein, gesture realises its textual meaning by pointing. Nonetheless, the specificity of pointing is different from that of eye contact in that it is more concerned with how different fingers, or a combination of fingers, specify intended textual meaning. Apart from that, gestures can also achieve textual meanings via wavelength, which might be observed in terms of gesturing rhythm. The textual meanings instantiated by head movement are also two-faceted. Apart from the wavelength in the form of frequency in a unit interval, the amplification of head movement, namely the angle of movement is also one of the observation foci.

2.6 Summary

Revolving around three key phases of the present study, viz. (1) building an argument for embedding nonverbal delivery into speaking assessment, and the issues of (2) how to design and (3) how to validate a rating scale with such a consideration informed by the argument, this chapter reviews the related literature. The first section reviews the topical issue of this study: nonverbal delivery. The next two sections address the issue of the rating scale design, while the last two

sections pave the way for the concrete procedures of how to validate a rating scale, especially the notion of validity and validation methods.

Specifically, the first section mainly pinpoints the significance of nonverbal delivery in communication and in a repertoire of research fields and also outlines the previous studies on three most representative channels of nonverbal delivery. In that sense, a theoretical argument for incorporating nonverbal delivery into speaking assessment can be felt to call for a corresponding empirical argument.

In the second section, by comparing and contrasting the evolution of *communicative competence* related models, the section outlines their components and respective strengths and weaknesses, justifies the employment of the CLA model as the theoretical framework for the rating scale design and points out the quintessential role of nonverbal delivery in the CLA model. The third section also responds to the issue of rating scale design. With a review on the prevailing taxonomies of rating scales in language assessment and the exemplifications of a few existing rating scales used by main language testing batteries, this section explicitly informs the formulation of the rating scale with nonverbal delivery embedded as an assessment dimension. Moreover, by highlighting the context where a rating scale is to be applied, the properties that the rating scale supposedly processes are also accorded.

The fourth section is devoted to conceptualising *validity* and *validation*. An overview is provided regarding three evolution phases of validity in language assessment scenario, based on which this study justifies itself in adopting a unitary notion of validity with construct validity as the core. In terms of validation methods, the last section argues the necessity of using both quantitative and qualitative methodologies in rating scale validation. MTMM is reviewed so that a glimpse is rendered of how this quantitative method will be adopted to verify the construct validity of the rating scale with teacher-rating and peer-rating as different scoring methods and different subdimensions on the scale as traits. The last section introduces MDA in detail, ranging from its theoretical origin, different streams of research and its application both worldwide and in the Chinese EFL context. The end of the last section provides fine-grained frameworks informed by an MDA approach so that the proposed rating scale can be validated quantitatively by an alignment of candidates' nonverbal delivery performance with the corresponding rating scale descriptors and the subscores they are assigned.

References

- ACTFL. 1986. *ACTFL proficiency guidelines*. Hasting-on-Hudson: American Council on the Teaching of Foreign Languages.
- ACTFL. 1999. *Revised ACTFL proficiency guidelines—Speaking*. Yonkers: American Council on the Teaching of Foreign Languages.
- AERA, APA, and NCME. 1985. *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

- AERA, APA, and NCME. 1999. *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.
- Alderson, J.C. 1981. Report of the discussion on general language proficiency. In *Issues in language testing*, ed. J.C. Alderson, and A. Hughes, 87–92. London: The British Council.
- Alderson, J.C. 1991. Bands and scores. In *Language testing in the 1990s*, ed. J.C. Alderson, and B. North, 71–86. London: Modern English Publications and the British Council.
- Alderson, J.C. (ed.). 2002. *Common European Framework of Reference for Languages: learning, teaching, assessment: case studies*. Strasbourg: Council of Europe.
- Alderson, J.C. 2010. *The Common European Framework of Reference for Language*. Invited seminar at Shanghai Jiao Tong University, Shanghai, China, Oct 2010.
- Alderson, J.C., and J. Banerjee. 2002. Language testing and assessment (Part 2). *Language Teaching* 35(2): 79–113.
- Alderson, J.C., N. Figueras, H. Kuiper, and G. Nold. 2006. Analyzing tests of reading and listening in relation to the Common European Framework of Reference: the experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly* 3(1): 3–30.
- Alibali, M.W., L. Flevares, and S. Goldin-Meadow. 1997. Assessing knowledge conveyed in gesture: do teachers have the upper hand? *Journal of Educational Psychology* 89: 183–193.
- Allal, L., and L.M. Lopez. 2005. Formative assessment of learning: a review of publication in French. In *Formative assessment: improving learning in secondary classrooms*, ed. J. Looney, 241–264. Paris: Organisation for Economic Cooperation and Development.
- Anastasi, A. 1950. *Some implications of cultural factors for test construction*. New York: Educational Testing Service.
- Anastasi, A. 1954. *Psychological testing*. New York: Macmillan.
- Anastasi, A. 1961. *Psychological testing*, 2nd ed. New York: Macmillan.
- Anastasi, A. 1976. *Psychological testing*, 4th ed. New York: Macmillan.
- Anastasi, A. 1982. “What do intelligence tests measure?” In *On educational testing: Intelligence, performance standards, test anxiety, and latent traits*, eds. S.B. Anderson, and J.S. Hemlick, 5–28. San Francisco, CA: Jossey-Bass, Inc.
- Angoff, W. 1988. Validity: an evolving concept. In *Test validity*, ed. H. Wainer, and H.I. Braun, 19–32. Hillsdale: Lawrence Erlbaum Associates.
- APA. 1954. Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement* 51(2): 1–38.
- APA, AERA, and NCME. 1966. *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.
- APA, AERA, and NCME. 1974. *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.
- Applebee, A.N. 2000. Alternative models of writing development. In *Perspectives on writing: research, theory, practice*, ed. R. Indrisano, and J.R. Squire, 90–111. Newark: International Reading Association.
- Argyle, M., and M. Cook. 1976. *Gaze and mutual gaze*. Cambridge: Cambridge University Press.
- Bacha, N. 2001. Writing evaluation: what can analytic versus holistic essay scoring tell us? *System* 29: 371–383.
- Bachman, L.F. 1988. Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition* 10(2): 149–164.
- Bachman, L.F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L.F. 1991. What does language testing have to offer? *TESOL Quarterly* 25(4): 671–704.
- Bachman, L.F. 2005. Building and supporting a case for test use. *Language Assessment Quarterly* 2(1): 1–34.
- Bachman, L.F., and A.S. Palmer. 1981. The construct validation of the FSI oral interview. *Language Learning* 31: 67–86.
- Bachman, L.F., and A.S. Palmer. 1982. The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16(4): 449–465.

- Bachman, L.F., and A.S. Palmer. 1989. The construct validation of self-ratings of communicative language ability. *Language Testing* 6(4): 449–465.
- Bachman, L.F., and A.S. Palmer. 1996. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.
- Bachman, L.F., and A.S. Palmer. 2010. *Language assessment in practice: developing language tests and justifying their use the real world*. Oxford: Oxford University Press.
- Bachman, L.F., and S.J. Savignon. 1986. The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal* 70(3): 380–390.
- Bachman, L.F., B.M. Lynch, and M. Mason. 1995. Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing* 12(2): 238–257.
- Bae, J., and L.F. Bachman. 1998. A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing* 15(3): 380–414.
- Baird, L.L. 1983. *The search for communication skills*. Educational Testing Service Research Report, No. 83-14. Princeton: Educational Testing Service.
- Baldry, A., and P. Thibault. 2006. *Multimodal transcription and text analysis*. London: Equinox.
- Barakat, R.A. 1973. Arabic gestures. *Journal of Popular Culture* 6(4): 749–787.
- Barkaoui, K. 2007. Rating scale impact on EFL essay marking: a mixed-method study. *Assessing Writing* 12(2): 86–107.
- Barkaoui, K. 2011. Think-aloud protocols in research on essay rating: an empirical study of their veridicality and reactivity. *Language Testing* 28(1): 51–75.
- Bateman, J.A. 2008. *Multimodality and genre: a foundation for the systematic analysis of multimodal documents*. London: Palgrave Macmillan.
- Bateman, J., J. Delin, and R. Henschel. 2004. Multimodality and empiricism: preparing for a corpus-based approach to the study of multimodal meaning-making. In *Perspectives on multimodality*, ed. E. Ventola, C. Cassily, and M. Kaltenbacher, 65–88. Philadelphia: John Benjamins.
- Bateman, J.A., J. Delin, and R. Henschel. 2006. Mapping the multimodal genres of traditional and electronic newspapers. In *New directions in the analysis of multimodal discourse*, ed. T.D. Royce, and W.L. Bowcher, 147–172. Mahwah: Lawrence Erlbaum Associates.
- Black, P., and D. Wiliam. 1998. Assessment and classroom learning. *Assessment in Education* 5 (1): 7–74.
- Black, P., and D. Wiliam. 2009. Developing the theory of formative assessment. *Educational Measurement, Evaluation and Accountability* 21(1): 5–31.
- Bloom, B.S., J.T. Hasting, and G.F. Madaus (eds.). 1971. *Handbook of formative and summative evaluation of student learning*. New York: McGraw-Hill.
- Bonk, W.J., and G.J. Ockey. 2003. A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing* 20(1): 89–110.
- Bourne, J., and C. Jewitt. 2003. Orchestrating debate: a multimodal approach to the study of the teaching of higher order literacy skills. *Reading: Literacy and Language*, UKRA, July, 64–72.
- Brindley, G. 1986. *The assessment of second language proficiency: issues and approaches*. Adelaide: National Curriculum Resource Centre.
- Brindley, G. 1991. Defining language ability: the criteria for criteria. In *Current developments in language testing*, ed. S. Anivan, 139–164. Singapore: Regional Language Centre.
- Brindley, G. 2002. Issues in language assessment. In *The Oxford handbook of applied linguistics*, ed. R.B. Kaplan, 459–470. Oxford: Oxford University Press.
- Brookhart, S.M. 2004. Classroom assessment: tensions and intersection in theory and practice. *Teachers College Record* 106(3): 429–458.
- Brookhart, S.M. 2007. Expanding views about formative classroom assessment: a review of the literature. In *Formative classroom assessment: theory into practice*, ed. J.H. McMillan, 43–62. New York: Teachers College Press.
- Brooks, L. 2009. Interacting in pairs in a test of oral proficiency: co-constructing a better performance. *Language Testing* 26(3): 341–366.

- Brown, A. 2003. Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20(1): 1–25.
- Brown, A., N. Iwashita, and T. McNamara. 2005. *An examination of rater orientations and test taker performance on English for academic purposes speaking tasks*. TOEFL Monograph Series, No. TOEFL-MS-29. Princeton: Educational Testing Service.
- Brown, J.D., and K.M. Bailey. 1984. A categorical instrument for scoring second writing skills. *Language Learning* 34(1): 21–42.
- Brown, J.D., and T. Hudson. 1998. The alternatives in language assessment. *TESOL Quarterly* 32(4): 653–675.
- Brown, G., and G. Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.
- Brumfit, C.J. 1984. *Communicative methodology in language teaching: the roles of fluency and accuracy*. Cambridge: Cambridge University Press.
- Brumfit, C.J., and K. Johnson. 1979. *The communicative approach to language teaching*. Oxford: Oxford University Press.
- Burgoon, J.K., and T. Saine. 1978. *The unspoken dialogue: an introduction to nonverbal communication*. Boston: Houghton Mifflin Company.
- Burgoon, J.K., D.A. Coker, and R.A. Coker. 1986. Communicative effects of gaze behavior: a test of two contrasting explanations. *Human Communication Research* 12: 495–524.
- Campbell, D.T., and D.W. Fiske. 1959. Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin* 56: 81–105.
- Canale, M. 1983. From communicative competence to communicative language pedagogy. In *Language and communication*, ed. J.C. Richards, and R.W. Schmidt, 2–27. London: Longman.
- Canale, M., and M. Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1): 1–47.
- Candlin, C.N. 1986. Explaining communicative competence limits of testability? In *Toward communicative competence testing: proceedings of the second TOEFL invitational conference*, ed. C.W. Stansfield, 38–57. Princeton: Educational Testing Service.
- Caple, H. 2008. Intermodal relations in image nuclear news stories. In *Multimodal semiotics: functional analysis in contexts of education*, ed. L. Unsworth, 125–138. London: Continuum.
- Carroll, J.B. 1961. The nature of data, or how to choose a correlation coefficient. *Psychometrika* 35(4): 347–372.
- Carroll, J.B. 1968. The psychology of language testing. In *Language testing symposium: a psycholinguistic perspective*, ed. A. Davies, 46–69. London: Oxford University Press.
- Celce-Murcia, M., Z. Dörnyei, and S. Thurrell. 1997. Direct approaches in L2 instruction: a turning point in communicative language teaching? *TESOL Quarterly* 31(1): 141–152.
- Cerrato, L. 2005. Linguistic functions of head nods. In *Gothenburg papers in theoretical linguistics 92: proceedings from 2nd Nordic conference on multi-modal communication*, ed. J. Allwood, and B. Dorriots, 137–152. Sweden: Gothenburg University.
- Chafe, W. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.
- Chalhoub-Deville, M. 1995. Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12(1): 16–33.
- Chapelle, C.A. 1998. Field independence: a source of language test variance? *Language Testing* 15(1): 62–82.
- Chapelle, C.A. 1999. Validity in language assessment. *Annual Review of Applied Linguistics* 19: 254–272.
- Chapelle, C.A., M.K. Enright, and J. Jamieson (eds.). 2008. *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
- Chapelle, C.A., M.K. Enright, and J. Jamieson. 2010. Does an argument-based approach to validity make a difference? *Educational Measurements: Issues and Practice* 29(1): 3–13.
- Charney, D. 1984. The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English* 18(1): 65–81.
- Chen, R. 2008. Some words on writing a multimodal lesson ware for English teaching. *Journal of Fujian Education Institute* 1: 75–77.

- Chen, Y., and G. Huang. 2009. Multimodal construal of heteroglossia: evidence from language textbooks. *Computer Assisted Foreign Language Education* 6: 35–41.
- Chen, Y., and H. Wang. 2008. Ideational meaning of image and text-image relations. *Journal of Ningbo University (Education Edition)* 1: 124–129.
- Cheng, L. 2005. *Changing language teaching through language testing: a washback study*. Cambridge: Cambridge University Press.
- Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.
- Cienki, A. 2008. Why study metaphor and gesture? In *Metaphor and Gesture*, eds. A. Cienki and C. Müller, 5–26. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Cizek, G.J. 2010. An introduction to formative assessment: history, characteristics and challenges. In *Handbook of formative assessment*, ed. H.L. Andrade, and G.J. Cizek, 3–17. New York: Routledge.
- Clark, J.L. 1985. Curriculum renewal in second language learning: an overview. *Canadian Modern Language Review* 42(3): 342–360.
- Clarkson, R., & M.T. Jensen. 1995. Assessing achievement in English for professional employment programmes. In *Language assessment in action*, ed. G. Brindley, pp. 165–194. Sydney, Macquarie University: National Centre for English Language Teaching and Research.
- Cohen, A. 1994. *Assessing language ability in the classroom*, 2nd ed. Boston: Heinle and Heinle Publishers.
- Connor, U., and P.L. Carrel. 1993. The interpretation of the tasks by writers and readers in holistically rated directed assessment of writing. In *Reading in the composition classroom: second language perspectives*, ed. J.G. Carson, and I. Leki, 141–160. Boston: Heinle & Heinle.
- Connor, U., and A. Mbaye. 2002. Discourse approaches to writing assessment. *Annual Review of Applied Linguistics* 22: 263–278.
- Cooper, C.R. 1977. Holistic evaluation of writing. In *Evaluating writing: describing, measuring, judging*, ed. C.R. Cooper, and L. Odell, 3–31. Urbana: NCTE.
- Corder, S.P. 1983. Strategies of communication. In *Strategies in interlanguage communication*, ed. C. Færch, and G. Kasper, 15–19. London: Longman.
- Cortazzi, M. 1993. *Narrative analysis*. London: Falmer Press.
- Council of Europe. 2001. *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Cowie, B., and B. Bell. 1999. A model of formative assessment in science education. *Assessment in Education* 6(1): 102–116.
- Creider, C. 1977. Towards a description of East African gestures. *Sign Language Studies* 14: 1–20.
- Cronbach, L.J. 1949. *Essentials of psychological testing*. New York: Harper & Row.
- Cronbach, L.J. 1971. Test validation. In *Educational measurement*, 2nd ed, ed. R.L. Thorndike, 443–507. Washington, DC: American Council on Education.
- Cronbach, L.J. 1980. Validity on parole: how can we go straight? New directions for testing and assessment: Measuring achievement over a decade. *Proceedings of the 1979 ETS invitational conference*, pp. 99–108. San Francisco: Jossey-Bass.
- Cronbach, L.J. 1988. Five perspectives on validity argument. In *Test validity*, ed. H. Wainer, and H.I. Braun, 3–17. Hillsdale: Lawrence Erlbaum Associates.
- Cronbach, L.J. 1989. Construct validation after thirty years. In *Intelligence: measurement, theory, and public policy*, ed. R. Linn, 147–167. Urbana: University of Chicago.
- Cronbach, L.J., and P.C. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52(4): 281–302.
- Cumming, A. 1990. Expertise in evaluating second language composition. *Language Testing* 7(1): 31–51.
- Cumming, A., R. Kantor, and D.E. Powers. 2001. *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: an investigation into raters' decision making and development of a preliminary analytic framework*. TOEFL Monograph Series, No. TOEFL-MS-22. Princeton: Educational Testing Service.
- Cumming, A. 2009. Language assessment in education: tests, curricula and teaching. *Annual Review of Applied Linguistics* 29: 90–100.

- Cumming, A., R. Kantor, and D.E. Powers. 2002. Decision making while rating ESL/EFL writing tasks: a descriptive framework. *Modern Language Journal* 86: 67–96.
- Cumming, A., R. Kantor, K. Baba, U. Erdosy, K. Eouanzoui, and M. James. 2006. *Analysis of discourse features and verification of scoring levels for independent and integrated tasks for the new TOEFL*. Princeton: Educational Testing Service.
- Cureton, E.E. 1950. Validity. In *Educational measurement*, ed. E.F. Lingquist, 621–694. Washington, DC: American Council on Education.
- Daly, A., and L. Unsworth. 2011. Analysis and comprehension of multimodal texts. *Australian Journal of Language and Literacy* 34(1): 61–80.
- Daniels, H. 2001. *Vygotsky and pedagogy*. London: Routledge.
- Davidson, F., and B. Lynch. 2002. *Testcraft: a teacher's guide to writing and using language test specifications*. New Haven: Yale.
- Davies, A., and P. LeMahieu. 2003. Assessment for learning: reconsidering portfolio and research evidence. In *Optimising new modes of assessment: in search of qualities and standards*, ed. M. Sergers, F. Dochy, and E. Cascallar, 141–169. Dordrecht: Kluwer Academic Publishers.
- Davies, A., A. Brown, C. Elder, K. Hill, T. Lumley, and T. McNamara. 1999. *Dictionary of language testing*. Cambridge: Cambridge University Press.
- Davison, C. 2004. The contradictory culture of teacher-based assessment: ESL assessment practices in Australian and Hong Kong secondary schools. *Language Testing* 21(3): 305–334.
- de Jong, J.H.A.L. 1992. Assessment of language proficiency in the perspective of the 21st century. *AILA Review* 9: 39–45.
- Derewianka, B., and C. Coffin. 2008. Visual representations of time in history textbooks. In *Multimodal semiotics*, ed. L. Unsworth, 187–200. London: Continuum.
- Djonov, E.N. 2006. *Analysing the organisation of information in websites: from hypermedia design to systemic functional hypermedia discourse analysis*. Unpublished Ph.D. thesis, University of New South Wales, Australia.
- Douglas, D., and J. Smith. 1997. *Theoretical underpinnings of the Test of Spoken English revision project*. TOEFL Monograph Series, No. TOEFL-MS-9. Princeton: Educational Testing Service.
- Douglas, D. 2000. *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- Ducasse, A.M., and A. Brown. 2009. Assessing paired orals: raters' orientation to interaction. *Language Testing* 26(3): 423–443.
- Dwyer, C.A. 2000. Excerpt from validity: theory into practice. *The Score* 22(4): 6–7.
- Ebel, R.L. 1961. Must all tests be valid? *American Psychologist* 16(10): 640–647.
- Ebel, R. L., and D. A. Frisbie. 1991. *Essentials of educational measurement*, 5th ed. Englewood Cliffs, NJ: Prentice—Hall.
- Efron, D. 1941. *Gesture, race and culture*. The Hague: Mouton.
- Egbert, M.M. 1998. Miscommunication in language proficiency interviews of first-year German students: a comparison with natural conversation. In *Talking and testing: discourse approaches to the assessment of oral proficiency*, ed. R. Young, and W. He, 147–172. Philadelphia: John Benjamins.
- Eggs, S., and D. Slade. 1997. *Analysing casual conversation*. London: Cassell.
- Ekman, P., and W.V. Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry* 32: 88–106.
- Ekman, P., and W.V. Friesen. 1974. Detecting deception from body or face. *Journal of Personality and Social Psychology* 29: 288–298.
- Ellsworth, P.C., and L.M. Ludwig. 1971. Visual behaviour in social interaction. *Journal of Communication* 21(4): 375–403.
- Enfield, N.J. 2009. *The anatomy of meaning: Speech, gesture, and composite utterances*. Cambridge: Cambridge University Press.
- Engstrom, Y. 1987. *Learning by expanding: an activity theoretical approach to developmental research*. Helsinki: Orienta-Konsultit Oy.

- Erdosy, M.U. 2004. *Exploring variability in judging writing ability in a second language: a study of four experienced raters of ESL compositions*. TOEFL Research Report, No. RR-03-17. Princeton: Educational Testing Service.
- Ericsson, K.A., and H. Simon. 1993. *Protocol analysis*. Cambridge: MIT Press.
- Færch, C., and G. Kasper (eds.). 1983. *Strategies in interlanguage communication*. London: Longman.
- Færch, C., et al. 1984. *Learner language and language learning*. Philadelphia: Multilingual Matters Ltd.
- Feng, D. 2011. Visual space and ideology: a critical cognitive analysis of spatial orientations in advertising. In *Multimodal studies: exploring issues and domains*, ed. K.L. O'Halloran, and B.A. Smith, 55–75. London: Routledge.
- Folland, D., and D. Robertson. 1976. Towards objective in group oral testing. *ELT Journal* 30(2): 156–167.
- Fulcher, G. 1987. Tests of oral performance: the need for data-based criteria. *ELT Journal* 41(4): 287–291.
- Fulcher, G. 1993. *The construction and validation of rating scales for oral tests in English as a foreign language*. Unpublished Ph.D. thesis. University of Lancaster, UK.
- Fulcher, G. 1996a. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13(2): 208–238.
- Fulcher, G. 1996b. Invalidating validity claims for the ACTFL oral rating scale. *System* 24(2): 163–172.
- Fulcher, G. 1997. The testing of speaking in a second language. In *Encyclopaedia of language and education*, vol. 7, ed. C. Clapham, and D. Corson, 75–85., Language testing and assessment New York: Springer.
- Fulcher, G. 2003. *Testing second language speaking*. London: Longman/Pearson Education.
- Fulcher, G. 2004. Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly* 1(4): 253–266.
- Fulcher, G. 2010. *Practical language testing*. London: Hodder Education.
- Fulcher, G., and F. Davidson. 2007. *Language testing and assessment: an advanced resource book*. London: Routledge.
- Fulcher, G., F. Davidson, and J. Kemp. 2011. Effective rating scale development for speaking tests: performance decision trees. *Language Testing* 27(1): 1–25.
- Galloway, V.B. 1987. From defining to developing proficiency: a look at the decisions. In *Defining and developing proficiency: guidelines, implementations, and concepts*, ed. H. Byrnes, and M. Canale, 25–73. Lincolnwood: National Textbook Company.
- Garrett, H.E. 1947. *Statistics in psychology and education*, 3rd ed. New York: Longmans, Green & Company.
- Goldin-Meadow, S., and M.A. Singer. 2003. From children's hands to adults' ears: Gesture's role in teaching and learning. *Developmental Psychology* 39: 509–520.
- Goodwin, L.D. 1997. Changing conceptions of measurement validity. *Journal of Nursing Education* 36: 102–107.
- Goodwin, L.D. 2002. Changing conceptions of measurement validity: an updated on the new standards. *Journal of Nursing Education* 41: 100–106.
- Goodwin, C., and J.C. Heritage. 1990. Conversation analysis. *Annual Review of Anthropology* 19: 283–307.
- Goodwin, L.D., and N.L. Leech. 2003. The meaning of validity in the new standards for educational and psychological testing: implications for measurement courses. *Measurement and Evaluation in Counseling and Development* 36(3): 181–191.
- Goulden, N.R. 1992. Theory and vocabulary for communication assessments. *Communication Education* 41(3): 258–269.
- Goulden, N.R. 1994. Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Development in Education* 27: 73–82.
- Grant, L., and L. Ginther. 2000. Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing* 9: 123–145.

- Green, J.R. 1968. *A gesture inventory for the teaching of Spanish*. Philadelphia: Chilton Books.
- Green, A. 1998. *Verbal protocol analysis in language testing research: a handbook*. Cambridge: Cambridge University Press.
- Green, A. 2007. Washback to learning outcomes: a comparative study of IELTS preparation and university pre-session language courses. *Assessment in Education* 14(1): 75–97.
- Grierson, J. 1995. Classroom-based assessment in intensive English centres. In *Language assessment in action*, ed. G. Brindley, 239–270. Sydney: National Centre for English Language Teaching and Research.
- Grootenboer, H. 2006. Treasuring the gaze: eye miniature portraits and the intimacy of vision. *Art Bulletin* 88(3): 496–507.
- Gu, Y. 2006a. Multimodal text analysis: a corpus linguistic approach to situated discourse. *Text & Talk* 26(2): 127–167.
- Gu, Y. 2006b. Agent-oriented modelling language, Part 1: modelling dynamic behaviour. *Proceedings of the 20th international CODATA conference, Beijing*, pp. 21–47. Beijing: Information Centre, Chinese Academy of Social Sciences.
- Gu, Y. 2007. Learning by multimedia and multimodality. In *E-learning in China: Sino-UK initiatives into policy, pedagogy and culture*, ed. H. Spencer-Oatey, 37–56. Hong Kong: The Hong Kong University Press.
- Gu, Y. 2009. From real life situated discourse to video-stream data-mining: an argument for agent-oriented modelling for multimodal corpus compilation. *International Journal of Corpus Linguistics* 14(4): 433–466.
- Guijarro, A.J.M., and M.J.P. Sanz. 2009. On interaction of image and verbal text in a picture book: a multimodal and systemic functional study. In *The world told and the world shown: multisemiotic issues*, ed. E. Ventola, and A.J.M. Guijarro, 107–123. Hampshire: Palgrave Macmillan.
- Guilford, J.P. 1946. New standards for test evaluation. *Educational and Psychological Measurement* 6(3): 427–438.
- Guion, R.M. 1977. Content validity: the source of my discontent. *Applied Psychological Measurement* 1(1): 1–10.
- Gulliksen, H. 1950. *Theory of mental tests*. Hillsdale: Lawrence Erlbaum Associates.
- Guo, L. 2004. Multimodality in biology textbooks. In *Multimodal discourse analysis: systemic-functional perspectives*, ed. K.L. O'Halloran, 196–219. London: Continuum.
- Hale, G.A., D.A. Rock, and T. Jirele. 1989. *Confirmatory factor analysis of the TOEFL*. TOEFL Research Report, No. RR-32. Princeton NJ: Educational Testing Service.
- Hall, E.T. 1959. *The silent language*. New York: Doubleday.
- Halliday, M.A.K. 1973. *Explorations in the functions of language*. London: Edward Arnold.
- Halliday, M.A.K. 1976. The form of a functional grammar. In *Halliday: system and function in language*, ed. G. Kress, 101–135. Oxford: Oxford University Press.
- Halliday, M.A.K. 1978. *Language as social semiotic: the social interpretation of language and meaning*. London: Edward Arnold.
- Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Arnold.
- Halliday, M.A.K., and R. Hasan. 1976. *Cohesion in English*. London: Longman.
- Halliday, M.A.K., and C.M.I.M. Matthiessen. 2004. *An introduction to functional grammar*, 3rd ed. London: Edward Arnold.
- Halliday, M.A.K., A. McIntosh, and P. Stevens. 1964. *The linguistic sciences and language teaching*. Bloomington: Indiana University Press.
- Hamp-Lyons, L. 1990. Second language writing: assessment issues. In *Second language writing: research insights for the classroom*, ed. B. Kroll, 69–87. New York: Cambridge University Press.
- Hamp-Lyons, L. 1991. Scoring procedures for ESL contexts. In *Assessing second language writing in academic contexts*, ed. L. Hamp-Lyons, 241–276. Norwood: Ablex.
- Hamp-Lyons, L. 1997. Washback, impact and validity: ethical concerns. *Language Testing* 14(3): 295–303.

- Hatch, E. 1978. Discourse analysis and second language acquisition. In *Second language acquisition: a book of readings*, ed. E. Hatch, 401–435. Rowley: Newbury House.
- Hattie, J., and H. Timperley. 2007. The power of feedback. *Review of Educational Research* 77(1): 81–112.
- Hawkey, R. 2001. Towards a common scale to describe L2 writing performance. *Cambridge Research Notes* 5: 9–13.
- Hawkey, R., and F. Barker. 2004. Developing a common scale for the assessment of writing. *Assessing Writing* 9(2): 122–159.
- He, W. 1998. Answering questions in LPIs: a case study. In *Talking and testing: discourse approaches to the assessment of oral proficiency*, ed. R. Young, and W. He, 101–116. Philadelphia: John Benjamins.
- Heath, C.C., and P. Luff. 2007. Gesture and institutional interaction: figuring bids in auctions of fine art and antiques. *Gesture* 7(2): 215–240.
- Hempel, C.G. 1965. *Aspects of scientific explanation and other essays in the philosophy of science*. Glencoe: Free Press.
- Henley, N.M. 1977. *Body politics: power, sex, and nonverbal communication*. Englewood Cliffs: Prentice-Hall.
- Henley, N.M., and S. Harmon. 1985. The nonverbal semantics of power and gender: a perceptual study. In *Power, dominance, and nonverbal behavior*, ed. S.L. Ellyson, and J.F. Dovidio, 151–164. New York: Springer.
- Herman, J.L., and K. Choi. 2008. *Formative assessment and the improvement of middle school science learning: The role of teacher accuracy*. CRESST Report 740. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Hess, E.H. 1975. *The tell-tale eye: how your eyes reveal hidden thoughts and emotions*. New York: van Nostrand Reinhold.
- Hilsdon, J. 1995. The group oral exam: advantages and limitations. In *Language testing in the 1990s: the communicative legacy*, ed. C. Alderson, and B. North, 189–197. Hertfordshire: Prentice Hall International.
- Hood, S. 2004. Managing attitude in undergraduate academic writing: A focus on the introductions to research reports. In *Analysing academic writing: Contextualized frameworks*, eds. L.J. Ravelli, and R.A. Ellis, 24–44. London: Continuum.
- Hood, S. 2006. The persuasive power of prosodies: Radiating values in academic writing. *Journal of English for Academic Purposes*, 5(1):37–49.
- Hood, S.E. 2007. *Gesture and meaning making in face-to-face teaching*. Paper presented at the Semiotic Margins Conference, University of Sydney.
- Hood, S.E. 2010. *Mimicking and mocking identities: the roles of language and body language in Taylor Mali's "Speak with conviction"*. Invited seminar at the Hong Kong Polytechnic University, 4 November 2010.
- Hood, S.E. 2011. Body language in face-to-face teaching: a focus on textual and interpersonal meaning. In *Semiotic margins: meanings in multimodalities*, ed. S. Dreyfus, S. Hood, and S. Stenglin, 31–52. London: Continuum.
- Hopper, R., S. Koch, and J. Mandelbaum. 1986. Conversation analysis methods. In *Contemporary issues in language and discourse processes*, ed. D.G. Ellis, and W.A. Donohue, 169–186. Hilldale: Lawrence Erlbaum Associates.
- Hornik, J. 1987. The effect of touch and gaze upon compliance and interest of interviewees. *The Journal of Social Psychology* 127: 681–683.
- House, E.T. 1980. *Evaluating with validity*. Beverly Hills: Sage Publications.
- Hu, L.T., and P.M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modelling: A Multidisciplinary Journal* 6: 1–55.
- Hu, Z., and J. Dong. 2006. How meaning is construed multimodally: a case study of a PowerPoint presentation contest. *Computer Assisted Foreign Language Education* 3: 3–12.
- Huerta-Macias, A. 1995. Alternative assessment: responses to commonly asked questions. *TESOL Journal* 5(1): 8–11.

- Hughes, A. 2003. *Testing for language teachers*, 2nd ed. Cambridge: Cambridge University Press.
- Hulstijn, J.H. 2007. The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91(4): 663–667.
- Hulstijn, J.H. 2011. Language proficiency in native and nonnative speakers: an agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly* 8(3): 229–249.
- Hymes, D.H. 1962. The ethnography of speaking. In *Anthropology and human behaviour*, ed. T. Gladwin, and W.C. Sturtevant, 13–53. Washington: The Anthropology Society of Washington.
- Hymes, D.H. 1964. Introduction: toward ethnographies of communication. *American Anthropologist* 6(6): 1–34.
- Hymes, D.H. 1972. On communicative competence. In *Sociolinguistics*, ed. J. Pride, and J. Holmes, 269–293. Harmondsworth: Penguin.
- Hymes, D.H. 1973. Toward linguistic competence. *Texas working papers in sociolinguistics (working paper No. 16)*. Austin, Tx: Centre for Intercultural Studies in Communication, and Department of Anthropology, University of Texas.
- Hymes, D.H. 1974. *Foundations in sociolinguistics: an ethnographic approach*. Philadelphia: University of Pennsylvania Press.
- Hymes, D.H. 1982. *Toward linguistic competence*. Philadelphia: Graduate School of Education, University of Pennsylvania.
- Iedema, R. 2001. Analysing film and television: a social semiotic account of hospital: an unhealthy business. In *Handbook of visual analysis*, ed. T. van Leeuwen, and C. Jewitt, 183–204. London: Sage.
- Iizuka, Y. 1992. Extraversion, introversion and visual interaction. *Perceptual and Motor Skills* 74: 43–59.
- Ingram, D., and E. Wylie. 1993. Assessing speaking proficiency in the international English language testing system. In *A new decade of language testing research: selected papers from the 1990s language testing research colloquium*, ed. D. Douglas, and C. Chapelle, 220–234. Alexandria: TESOL Inc.
- Jacobs, E. 1988. Clarifying qualitative research: A focus on traditions. *Educational Researcher*, 17 (1):16–24.
- Jackendoff, R. 1983. *Semantics and cognition*. Cambridge: MIT Press.
- Janik, S.W., A.R. Wellens, M.L. Goldberg, and L.F. Dell’Osso. 1978. Eyes as the centre of focus in the visual examination of human faces. *Perceptual and Motor Skills* 47: 857–858.
- Jarvis, G.A. 1986. Proficiency testing: a matter of false hopes? *ADFL Bulletin* 18: 20–21.
- Jewitt, C. 2002. The move from page to screen: the multimodal reshaping of school English. *Journal of Visual Communication* 1(2): 171–196.
- Jewitt, C. 2006. *Technology, literacy and learning: a multimodal approach*. London: Routledge.
- Jewitt, C. 2009. An introduction to multimodality. In *The Routledge handbook of multimodal analysis*, ed. C. Jewitt, 14–27. London: Routledge.
- Jewitt, C. 2011. The changing pedagogic landscape of subject English in UK classrooms. In *Multimodal studies: exploring issues and domains*, ed. K.L. O’Halloran, and B.A. Smith, 184–201. London: Routledge.
- Johnson, K., and H. Johnson. 1999. *Encyclopaedic dictionary of applied linguistics: a handbook for language teaching*. Malden: Blackwell Publishers Inc.
- Johnson, M., and A. Tylor. 1998. Re-analysing the OPI: how much does it look like natural conversation? In *Talking and testing: discourse approaches to the assessment of oral proficiency*, ed. R. Young, and W. He, 27–51. Philadelphia: John Benjamins.
- Jöreskog, K.G. 1993. Testing structural equation models. In *Testing structural equation models*, ed. D. Bollen, and J.S. Long, 294–316. Newbury Park: Sage Publications.
- Jungheim, N.O. 1995. Assessing the unsaid: the development of tests of nonverbal ability. In *Language testing in Japan*, ed. J.D. Brown, and S.O. Yamashita, 149–165. Tokyo: JALT.
- Jungheim, N.O. 2001. The unspoken element of communicative competence: evaluating language learners’ nonverbal behaviour. In *A focus on language test development: expanding the*

- language proficiency construct across a variety of tests*, ed. T. Hudson, and J.D. Brown, 1–34. Honolulu: University of Hawaii, Second Language Teaching and Curriculum Centre.
- Kaindl, L. 2005. Multimodality in the translation of humour in comics. In *Perspectives on multimodality*, ed. E. Ventola, C. Charles, and M. Kaltenbacher, 173–192. Amsterdam: John Benjamins.
- Kalma, A. 1992. Gazing in triads: a powerful signal in floor apportionment. *British Journal of Social Psychology* 31: 21–39.
- Kane, T. M. 1990. *An argument-based approach to validation*. Iowa: The American College Testing Program.
- Kane, M.T. 1992. An argument-based approach to validity. *Psychological Bulletin* 112(3): 527–535.
- Kane, M.T. 1994. Validating interpretative arguments for licensure and certification examinations. *Evaluation and the Health Professions* 17(2): 133–159.
- Kane, M.T. 2001. Current concerns in validity theory. *Journal of Educational Measurement* 38(4): 319–342.
- Kane, M.T. 2002. Validating high-stakes testing programs. *Educational Measurement: Issues and Practice* 21(1): 31–41.
- Kane, M.T. 2004. Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135–170.
- Kane, M.T. 2006. Validation. In *Educational measurement*, 4th ed, ed. R. Brennan, 17–64. Westport: American Council on Education and Praeger.
- Kane, M.T. 2010. Validity and fairness. *Language Testing* 27(2): 177–182.
- Kane, M.T., T. Crooks, and A. Cohen. 1999. Validating measures of performance. *Educational Measurement: Issues and Practice* 18(2): 5–17.
- Kasper, G., and K.R. Rose. 2002. *Pragmatic development in a second language*. Oxford: Blackwell.
- Kendon, A. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26: 22–63.
- Kendon, A. 1980. Gesticulation and speech: Two aspects of the process of utterance. In *The relationship of verbal and nonverbal communication*, ed. M.R. Key, 207–227. The Hague: Mouton and Co.
- Kendon, A. 1981. The organization of behavior in face-to-face interaction: observations on the development of a methodology. In *Handbook of research methods in nonverbal behavior*, ed. P. Ekman, and K. Scherer, 440–505. Cambridge: Cambridge University Press.
- Kendon, A. 1985. Some uses of gesture. In *Perspectives on silence*, ed. D. Tannen, and M. Saville-Troike, 215–234. Norwood: Ablex.
- Kendon, A. 1996. Gesture in language acquisition. *Multilingual* 15: 201–214.
- Kendon, A. 2004. *Gesture: visible action as utterance*. Cambridge: Cambridge University Press.
- Kim, M. 2001. Detecting DIF across the different language groups in a speaking test. *Language Testing* 18(1): 89–114.
- Kim, Y. 2009. An investigation into native and non-native teachers' judgments of oral English performance: a mixed methods approach. *Language Testing* 26(2): 187–217.
- Klinke, C.L. 1986. Gaze and eye contact: a research review. *Psychological Bulletin* 100(1): 78–100.
- Knoch, U. 2009. *Diagnostic writing assessment: the development and validation of a rating scale*. Frankfurt: Peter Lang.
- Knox, J.S. 2008. Online newspapers and TESOL classrooms: a multimodal perspective. In *Multimodal semiotics: functional analysis in contexts of education*, ed. L. Unsworth, 139–158. London: Continuum.
- Kok, A.K.C. 2004. Multisemiotic mediation in hypertext. In *Multimodal discourse analysis: systemic-functional perspectives*, ed. K.L. O'Halloran, 131–159. London: Continuum.
- Kondo-Brown, K. 2002. A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing* 19(1): 3–31.

- Kormos, J. 1999. Simulating conversations in oral-proficiency assessments: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing* 16(2): 163–188.
- Kress, G. 2000. Design and transformation: new theories of meaning. In *Multiliteracies: literacy learning and the design of social futures*, ed. B. Cope, and M. Kalantzis, 153–161. South Yarra: Macmillan Publishers Australia Pte Ltd.
- Kress, G., et al. 2001. *Multimodal teaching and learning: the rhetorics of the science classroom*. London: Continuum.
- Kress, G., and T. van Leeuwen. 1996. *Reading images: the grammar of visual design*. London: Routledge.
- Kress, G., and T. van Leeuwen. 1998. The (critical) analysis of newspaper layout. In *Approaches to media discourse*, ed. A. Bell, and P. Garrett, 186–219. Oxford: Blackwell.
- Kress, G., and T. van Leeuwen. 2001. *Multimodal discourse: the modes and media of contemporary communication*. London: Edward Arnold.
- Kress, G., and T. van Leeuwen. 2002. Colour as a semiotic mode: notes for a grammar of colour. *Visual Communication* 3: 343–368.
- Kress, G., and T. van Leeuwen. 2006. *Reading images: the grammar of visual design*, 2nd ed. London: Routledge.
- Kress, G., et al. 2005. *English in urban classrooms: a multimodal perspective on teaching and learning*. London: Routledge.
- Kunnan, A.J. 1995. *Test taker characteristics and test performance: a structural modelling approach*. Cambridge: Cambridge University Press.
- Kunnan, A.J. (ed.). 2000. *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.
- Kunnan, A.J. 2004. Test fairness. In *European language testing in a global context*, ed. M. Milanovic, and C.J. Weir, 27–48. Cambridge: Cambridge University Press.
- Kunnan, A.J. 2005. Language assessment from a wider context. In *Handbook of research in second language learning*, ed. E. Hinkel, 779–794. Mahwah: Lawrence Erlbaum Associates.
- Kunnan, A.J. 2008. Towards a model of test evaluation: using the test fairness and wider context frameworks. In *Multilingualism and assessment: achieving transparency, assuring quality, sustaining diversity. Papers from the ALTE Conference in Berlin, Germany*, ed. L. Taylor, and C.J. Weir, 229–251. Cambridge: Cambridge University Press.
- Kunnan, A.J. 2010. Fairness matters and Toulmin's argument structures. *Language Testing* 24(2): 183–189.
- Lado, R. 1961. *Language testing*. New York: McGraw-Hill.
- Langenfeld, T.E., and L.M. Crocker. 1994. The evolution of validity theory: publish school testing, the courts, and incompatible interpretations. *Educational Assessment* 2(2): 149–165.
- Lantolf, J., and W. Frawley. 1985. Oral proficiency testing: a critical analysis. *The Modern Language Journal* 69(3): 337–345.
- Lantolf, J., and W. Frawley. 1988. Proficiency, understanding the construct. *Studies in Second Language Acquisition* 10(2): 181–196.
- Larsen-Freeman, D. (ed.). 1980. *Discourse analysis in second language research*. Rowley: Newbury House.
- Lazaraton, A. 1991. *A conversation analysis of structure and interaction in the language interview*. Unpublished Ph.D. thesis, University of California at Los Angeles, USA.
- Lazaraton, A. 1992. The structural organisation of a language interview: a conversational analytic perspective. *System* 20(3): 373–386.
- Lazaraton, A. 1995. Qualitative research in TESOL: a progress report. *TESOL Quarterly* 29: 455–472.
- Lazaraton, A. 1996a. Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13(2): 151–172.
- Lazaraton, A. 1996b. A qualitative approach to monitoring examiner conduct in CASE. In *Studies in language testing 3: performance testing, cognition, and assessment: selected papers from*

- the 15th Language Testing Research Colloquium, Cambridge and Arnhem, ed. M. Milanovic, and N. Saville, 18–33. Cambridge: Cambridge University Press.
- Lazaraton, A. 2002. *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.
- Lazaraton, A. 2008. Utilising qualitative methods for assessment. In *Encyclopaedia of language and education*, 2nd edn. Vol. 7: *Language Testing and Assessment*, pp. 197–209. New York: Springer.
- Leathers, D.G., and H.M. Eaves. 2008. *Successful nonverbal communication: principles and applications*, 4th ed. New York: Pearson Education Inc.
- Lemke, J.L. 2002. Travels in hypermodality. *Visual Communication* 1(3): 299–325.
- Lennon, P. 1990. Investigating fluency in EFL: a quantitative approach. *Language Learning* 40(3): 387–417.
- Leung, C. 2005a. Convivial communication: recontextualising communicative competence. *International Journal of Applied Linguistics* 15(2): 119–143.
- Leung, C. 2005b. Classroom teacher assessment of second language development: construct as practice. In *Handbook of research in second language teaching and learning*, ed. E. Hinkel, 869–888. Mahwah: Lawrence Erlbaum Associates.
- Leung, C., and B. Mohan. 2004. Teacher formative assessment and talk in classroom contexts: assessment as discourse and assessment of discourse. *Language Testing* 21(3): 335–359.
- Levine, P., and R. Scollon (eds.). 2004. *Discourse and technology: multimodal discourse analysis*. Washington: Georgetown University Press.
- Levinson, S.C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
- Linn, R.L. 1994. Performance assessment: policy promises and technical measurement standards. *Educational Researcher* 23(9): 4–14.
- Linn, R.L. 1997. Evaluating the validity of assessments: the consequences of use. *Educational Measurement: Issues and Practice* 16(2): 14–16.
- Liski, E., and S. Puntanen. 1983. A study of the statistical foundations of group conversation tests in spoken English. *Language Learning* 33(2): 225–246.
- Little, D. 2006. The Common European Framework of Reference for Languages: content, purpose, origin, reception and impact. *Language Teaching* 39(3): 167–190.
- Llosa, L. 2007. Validating a standards-based classroom assessment of English proficiency: a multi-trait multi-method approach. *Language Testing* 24(4): 489–515.
- Lloyd-Jones, R. 1977. Primary trait scoring. In *Evaluating writing: describing, measuring, judging*, ed. C.R. Cooper, and L. Odell, 33–66. Urbana: National Council of Teachers of English.
- Long, Y., and P. Zhao. 2009. The interaction study between multimodality and metacognitive strategy in college English listening comprehension teaching. *Computer Assisted Foreign Language Education* 4: 58–74.
- Lowe, P. 1985. The ILR proficiency scale as a synthesising research principle: the view from the mountain. In *Foreign language proficiency in the classroom and beyond*, ed. C.J. James, 9–54. Lincolnwood: National Textbook Company.
- Lumley, T. 2002. Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing* 19: 246–276.
- Lumley, T. 2005. *Assessing second language writing: the rater's perspective*. New York: Peter Lang.
- Lumley, T., and A. Brown. 2005. Research methods in language testing. In *Handbook of research in second language teaching and learning*, ed. E. Hinkel, 855–933. Mahwah: Lawrence Erlbaum Associates.
- Lumley, T., and B. O'Sullivan. 2005. The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing* 22(4): 415–437.
- Luoma, S. 2004. *Assessing speaking*. Cambridge: Cambridge University Press.
- Lynch, B. 2001. Rethinking assessment from a critical perspective. *Language Testing* 18(4): 333–349.
- Lynch, B. 2003. *Language assessment and programme evaluation*. New Haven: Yale.

- Macken-Horarik, M. 2004. Interacting with the multimodal text: reflections on image and verbiage in ArtExpress. *Visual Communication* 3(1): 5–26.
- Macken-Horarik, M., L. Love, and L. Unsworth. 2011. A grammatics ‘good enough’ for school English in the 21st century: four challenges in realising the potential. *Australian Journal of Language and Literacy* 34(1): 9–23.
- Maiorani, A. 2009. *The Matrix phenomenon. A linguistic and multimodal analysis*. Saarbrücken: VDM Verlag.
- Marsh, H.W. 1988. Multi-trait multi-method analyses. In *Educational research methodology, and evaluation: an international handbook*, ed. J.P. Keeves, 570–578. Oxford: Pergamon.
- Marsh, H.W. 1989. Confirmatory factor analysis of multi-trait multi-method data: many problems and a few solutions. *Applied Psychological Measurement* 15: 47–70.
- Martin, J.R. 1995. Interpersonal meaning, persuasion and public discourse: Packing semiotic punch. *Australian Journal of Linguistics*, 15(1):33–67.
- Martin, J.R. 2000. Beyond exchange: Appraisal systems in English. In *Evaluation in text: Authorial stance and the construction of discourse*, eds. S. Hunston, and G. Thompson 142–175. Oxford: Oxford University Press.
- Martin, J.R. 2008. Intermodal reconciliation: mates in arms. In *New literacies and the English curriculum*, ed. L. Unsworth, 112–148. London: Continuum.
- Martin, J.R. and P.R.R., White. 2005. *The language of evaluation: Appraisal in English*. London: Palgrave.
- Martinec, R. 2000a. Types of processes in action. *Semiotica* 130(3): 243–268.
- Martinec, R. 2000b. Construction of identity in Michael Jackson’s “Jam”. *Social Semiotics* 10(3): 313–329.
- Martinec, R. 2001. Interpersonal resources in action. *Semiotica* 135(1): 117–145.
- Martinec, R. 2004. Gestures that co-occur with speech as a systematic resource: the realisation of experiential meanings in indexes. *Social Semiotics* 14(2): 193–213.
- Matsumoto, D. 2006. Culture and cultural worldviews: Do verbal descriptions about culture reflect anything other than verbal descriptions of culture? *Culture and Psychology*, 12(1):33–62.
- Matsuno, S. 2009. Self-, peer- and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing* 26(1): 75–100.
- Matthews, M. 1990. The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations. *English Language Teaching Journal* 44(2): 117–121.
- Matthiessen, C.M.I.M. 2007. The multimodal page: a systemic functional exploration. In *New directions in the analysis of multimodal discourse*, ed. T.D. Royce, and W.L. Bowcher, 1–62. Mahwah: Lawrence Erlbaum Associates.
- Maynard, S.K. 1987. Interactional functions of a nonverbal sign: head movement in Japanese dyadic casual conversation. *Journal of Pragmatics* 11: 589–606.
- Maynard, S.K. 1989. *Japanese conversation: self-contextualisation through structure and interactional management*. Norwood: Albex.
- Maynard, S.K. 1990. Understanding interactive competence in L1/L2 contrastive context: a case of backchannel behaviour in Japanese and English. In *Language proficiency: defining, teaching, and testing*, ed. L.A. Arena, 41–52. New York: Plenum Press.
- McCrimman, J.M. 1984. *Writing with a purpose*, 8th ed. Boston: Houghton Mifflin.
- McKay, P. 1995. Developing ESL proficiency descriptions for the school context: the NLLIA ESL band scales. In *Language assessment in action*, ed. G. Brindley, 3–34. Sydney: National Centre for English Language Teaching and Research.
- McNamara, T. 1990. Item response theory and the validation of an ESP test for health professionals. *Language Testing* 7(1): 52–76.
- McNamara, T. 1996. *Measuring second language performance*. London: Longman.
- McNamara, T. 2000. *Language testing*. Oxford: Oxford University Press.
- McNamara, T. 2001. Language assessment as social practice: challenges for research. *Language Testing* 18(4): 333–349.
- McNamara, T., and C. Roever. 2006. *Language testing: the social dimension*. Oxford: Blackwell Publishing.

- McNeill, D. 1979. *The conceptual basis of language*. Hilldale: Lawrence Erlbaum Associates.
- McNeill, D. 1992. *Hand and mind: what gestures reveal about thought*. Chicago: The University of Chicago Press.
- McNeill, D. 1998. Speech and gesture integration. In *The nature and functions of gesture in children's communication. New directions for child development*, eds. J.M. Iverson, and S. Goldin-Meadow, 11–27. San Francisco: Jossey-Bass Inc, Publishers.
- McNeill, D. (ed.). 2000. *Language and gesture*. Cambridge: Cambridge University Press.
- McNeill, D. 2005. *Gesture and thought*. Chicago: The University of Chicago Press.
- Mehrens, W.A. 1997. The consequences of consequential validity. *Educational Measurement: Issues and Practice* 16(2): 16–18.
- Messick, S. 1975. The standard problem: meaning and values in measurement and evaluation. *American Psychologist* 30(10): 955–966.
- Messick, S. 1980. Test validity and the ethics of assessment. *American Psychologist* 35(11): 1012–1027.
- Messick, S. 1988. The once and future issues of validity: assessing the meaning and consequences of measurement. In *Test validity*, eds. H. Wainer, and H.I. Braun, 33–45. Hillsdale: Lawrence Erlbaum Associates.
- Messick, S. 1989a. Meaning and value in test validation: the science and ethics of assessment. *Educational Researcher* 18(2): 5–11.
- Messick, S. 1989b. Validity. In *Educational measurement*, 3rd ed, ed. R.L. Linn, 13–103. New York: American Council on Education & Macmillan Publishing Company.
- Messick, S. 1992. Validity of test interpretation and use. In *Encyclopaedia of educational research*, 6th ed, ed. M.C. Alkin, 1487–1495. New York: Macmillan.
- Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessment. *Educational Research* 2(2): 13–23.
- Messick, S. 1995. Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice* 14(4): 5–8.
- Messick, S. 1996. Validity and washback in language testing. *Language Testing* 13(3): 241–256.
- Mickan, P. 2003. *What's your score? An investigation into language descriptors for rating written performance*. Canberra: IELTS Australia.
- Milanovic, M., N. Saville, A. Pollitt, and A. Cook. 1996. Developing and validating rating scales for CASE: theoretical concerns and analyses. In *Validation in language testing*, ed. A. Cumming, and R. Berwick, 15–38. Philadelphia: Multilingual Matters Ltd.
- Mislevy, R.J. 2003. Substance and structure in assessment arguments. *Law, Probability, and Risk* 2(4): 237–258.
- Mislevy, R.J., L.S. Steinberg, and R.G. Almond. 2003. On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives* 1(1):3–67.
- Mislevy, R.J., R.G. Almond, and L.S. Steinberg. 2002. On the roles of task model variables in assessment design. In *Generating items for cognitive tests: theory and practice*, ed. S. Irvine, and P. Kyllonen, 97–128. Hillsdale: Lawrence Erlbaum Associates.
- Morrow, K. (ed.). 2004. *Insights from the Common European Framework*. Oxford: Oxford University Press.
- Mosier, C.I. 1947. A critical examination of the concepts of face validity. *Educational and Psychological Measurement* 7(2): 191–205.
- Moss, P.A. 1992. Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research* 62(3): 229–258.
- Munby, J. 1978. *Communicative syllabus design*. Cambridge: Cambridge University Press.
- Myford, C.M. 2002. Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education* 15(2): 187–215.
- Nakatsuhara, F. 2009. *Conversational styles in group oral tests: how is the conversation co-constructed?* Unpublished Ph.D. thesis, The University of Essex, UK.
- Namibar, M.K., and C. Goon. 1993. Assessment of oral skills: a comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal* 24 (1): 15–31.

- Neu, J. 1990. Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In *Developing communicative competence in a second language: series on issues in second language research*, ed. C.R. Scarcella, S.E. Andersen, and D.S. Krashen, 121–138. New York: Newbury House Publishers.
- Nevo, D., and E. Shohamy. 1984. *Applying the joint committee's evaluation standards for the assessment of alternative testing methods*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Nevo, B. 1985. Face validity revisited. *Journal of Educational Measurement* 22(4): 287–293.
- Norris, S. 2002. *Theoretical framework for multimodal discourse analysis presented via the analysis of identity construction of two women living in Germany*. Unpublished Ph.D. thesis, Georgetown University, USA.
- Norris, S. 2004. *Analysing multimodal interaction: a methodological framework*. London: Routledge.
- Norris, J.M. 2005. Book review: common European Framework of Reference for Languages: learning, teaching, assessment. *Language Testing* 22(3): 399–405.
- Norris, S., and R.H. Jones (eds.). 2005. *Discourse in action: introducing mediated discourse analysis*. London: Routledge.
- North, B. 1994. *Scales of language proficiency: a survey of some existing systems*. Washington, DC: Georgetown University Press.
- North, B. 1996. *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement*. Unpublished Ph.D. thesis, Thames Valley University, UK.
- North, B. 2000. *The development of a common framework scale of language proficiency*. New York: Peter Lang Publishing Inc.
- North, B. 2003. *Scales for rating language performance: descriptive models, formulation styles, and presentation formats*. TOEFL Monograph, No. TOEFL-MS-24. Princeton: Educational Testing Service.
- North, B. 2010a. Levels and goals: central frameworks and local strategies. In *The handbook of educational linguistics*, ed. B. Spolsky, and F.M. Hult, 220–230. Malden: Wiley-Blackwell.
- North, B. 2010b. *Assessment, certification and the CEFR: an overview*. Plenary speech at IATEFL TEA SIG & EALTA conference, Barcelona, Spain.
- North, B., and G. Schneider. 1998. Scaling descriptors for language proficiency scales. *Language Testing* 15(2): 217–262.
- O'Halloran, K.L. 2000. Classroom discourse in mathematics: a multisemiotic analysis. *Linguistics and Education* 10(3): 359–388.
- O'Halloran, K.L. 2004. Visual semiosis in film. In *Multimodal discourse analysis: systemic-functional perspectives*, ed. K.L. O'Halloran, 109–130. London: Continuum.
- O'Halloran, K.L. 2005. *Mathematical discourse: language, symbolism and visual images*. London: Continuum.
- O'Halloran, K.L. 2008a. Inter-semiotic expansion of experiential meaning: hierarchical scales and metaphor in mathematics discourse. In *New developments in the study of ideational meaning: from language to multimodality*, ed. C. Jones, and E. Ventola, 231–254. London: Equinox.
- O'Halloran, K.L. 2008b. Systemic functional-multimodal discourse analysis (SF-MDA): constructing ideational meaning using language and visual imagery. *Visual Communication* 7(4): 443–475.
- O'Halloran, K. 2009. Historical changes in the Semiotic landscape: From calculation to computation. In *The routledge handbook of multimodal analysis*, ed. C. Jewitt, 98–113. UK: Routledge.
- O'Halloran, K.L. 2011. Multimodal discourse analysis. In *Continuum companion to discourse analysis*, ed. K. Hyland, and B. Paltridge, 120–137. London: Continuum.
- O'Halloran, K.L., and F.V. Lim. 2009. Sequential visual discourse frames. In *The world told and the world shown: multisemiotic issues*, ed. E. Ventola, and A.J.M. Guijarro, 139–156. Hampshire: Palgrave Macmillan.

- O'Loughlin, K.K. 2002. The impact of gender in oral proficiency testing. *Language Testing* 19(2): 169–192.
- O'Malley, J.M., and A.U. Chamot. 1990. *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.
- O'Toole, M. 1994. *The language of displayed art*. London: Leicester University Press.
- O'Toole, M. 2010. *The language of displayed art*, 2nd ed. London: Routledge.
- O'Toole, M. 2011. Art vs. computer animation: integrity and technology in “South Park”. In *Multimodal studies: exploring issues and domains*, ed. K.L. O'Halloran, and B.A. Smith, 239–252. London: Routledge.
- Ockey, G.J. 2001. Is the oral interview superior to the group oral? *Working paper on language acquisition and education*, International University of Japan, vol. 11, pp. 22–41.
- Oller, J.W. 1979. *Language tests at school*. London: Longman.
- Oller, J.W. 1983. Evidence for a general language proficiency factor: an expectancy grammar. In *Issues in language testing research*, ed. J.W. Oller, 3–10. Rowley: Newbury House.
- Oller, J.W., and F.B. Hinoftotis. 1980. Two mutually exclusive hypotheses about second language ability: indivisible or partially divisible competence. In *Research in language testing*, ed. J.W. Oller, and K. Perkins, 13–23. Rowley: Newbury House.
- Oreström, B. 1983. Turn-taking in English conversation. *Lund Studies in English* 66, CWK Gleerup.
- Painter, C. 2007. Children's picture book narratives: reading sequences of images. In *Advances in language and education*, ed. A. McCabe, M. O'Donnell, and R. Whittaker, 40–59. London: Continuum.
- Painter, C. 2008. The role of colour in children's picture books. In *New literacies and the English curriculum*, ed. L. Unsworth, 89–111. London: Continuum.
- Painter, C., J.R. Martin, and L. Unsworth. 2013. *Reading visual narratives: Image analysis of children's picture books*. Bristol: Equinox Publishing.
- Patri, M. 2002. The influence of peer feedback on self- and peer-assessment. *Language Testing* 19 (2): 109–132.
- Pawley, A., and F.H. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In *Language and communication*, ed. J.C. Richards, and R.W. Schmidt, 191–225. London: Longman.
- Pienemann, M., and M. Johnston. 1987. Factors influencing the development of language proficiency. In *Applying second language acquisition research*, ed. D. Nunan, 89–94. Adelaide: National Curriculum Resource Centre.
- Pike, K.L. 1967. *Language in relation to a unified theory of the structure of human behaviour*, 2nd ed. The Hague: Mouton & Co.
- Poggi, I. 2001. The lexicon of the conductor's face. In *Language, vision and music*, ed. P. McKeivitt, S. Nuallsin, and C. Mulvihill, 271–284. Amsterdam: John Benjamins.
- Pollitt, A., and C. Hutchinson. 1987. Calibrating graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing* 4(1): 72–92.
- Pomerantz, A., and B.J. Fehr. 1997. Conversation analysis: An approach to the study of social action as sense making practices. In *Discourse as social action, discourse studies: a multidisciplinary introduction*, vol. 2, ed. T.A. van Dijk, 64–91. London: Sage Publications.
- Popham, W.J. 1990. *Modern educational measurement: a practitioner's perspective*. New York: Prentice Hall.
- Popham, W.J. 1997. Consequential validity: right concern—wrong concept. *Educational Measurement: Issues and Practice* 16(2): 9–13.
- Popham, W.J. 2008. *Transformative assessment*. Alexandria: Association for Supervision and Curriculum Development.
- Psathas, G. 1995. *Conversation analysis: the study of talk-in-interaction*. Thousand Oaks: Sage.
- Purpura, J. 1999. *Learner strategy use and performance on language tests: a structural equation modelling approach*. Cambridge: Cambridge University Press.
- Purpura, J. 2004. *Assessing grammar*. Cambridge: Cambridge University Press.

- Purpura, J. 2008. Assessing communicative language ability. In *Encyclopaedia of language and education*, eds. E. Shohamy, and N.H. Hornberger, 2nd edn. Vol. 7: *language testing and assessment*, pp. 53–68. New York: Springer.
- Ravelli, L.J. 2000. Beyond shopping: constructing the Sydney Olympics in three-dimensional text. *Text* 20(4): 489–515.
- Raykov, T., and G.A. Marcoulides. 2006. *A first course in structural equation modeling*, 2nd ed. Mahwah: Lawrence Erlbaum Associates, Inc.
- Rea-Dickens, P. 2006. Currents and eddies in the discourse of assessment: a learning-focused interpretation. *International Journal of Applied Linguistics* 16(2): 163–188.
- Richards, J.C., and R.W. Schmidt. 1983. Conversation analysis. In *Language and communication*, ed. J.C. Richards, and R.W. Schmidt, 117–153. London: Longman.
- Richards, J.C., et al. 1992. *Longman dictionary of language teaching and applied linguistics*. London: Longman.
- Riley, P. 1996. Developmental sociolinguistics and the competence/performance distinction. In *Performance and competence in second language acquisition*, ed. G. Brown, K. Malinkjaer, and J. Williams, 114–135. Cambridge: Cambridge University Press.
- Ross, S.J. 1998. Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing* 15(1): 1–20.
- Ross, S.J. 2005. The impact of assessment method on foreign language proficiency growth. *Applied Linguistics* 26(3): 317–342.
- Ross, S.J., and R. Berwick. 1992. The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14(2): 159–176.
- Royce, T. 2007. Multimodal communicative competence in second language contexts. In *New directions in the analysis of multimodal discourse*, ed. T. Royce, and W. Bowcher, 361–390. New York: Routledge.
- Ruesch, J., and W. Kees. 1956. *Nonverbal communication: notes on the visual perception of human relations*. Berkeley: University of California Press.
- Sacks, H. 1992. *Lectures on conversation*, vol. 1&2. Cambridge: Blackwell.
- Sacks, H., E.A. Schegloff, and G. Jefferson. 1974. A simplest systematic for the organisation of turn-taking in conversation. *Language* 50: 696–735.
- Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18(2): 119–144.
- Saitz, R., and E.J. Cervenka. 1972. *Handbook of gestures*. Mouton: The Hague.
- Sajavaara, K. 1987. Second language speech production: factors affecting fluency. In *Psycholinguistic models of production*, ed. H.D. Dechert, and M. Raupach, 45–65. Norwood: Ablex.
- Sasaki, M. 1993. Relationships among second language proficiency, foreign language aptitude and intelligence: a structural equation modelling approach. *Language Learning* 43: 313–344.
- Savignon, S.J. 1983. *Communicative competence: theory and classroom practice; texts and contexts in second language learning*. Reading: Addison-Wesley.
- Savignon, S.J. 1997. *Communicative competence: theories and classroom practice*. New York: McGraw-Hill.
- Sawaki, Y. 2007. Construct validation of analytic rating scales in a speaking assessment: reporting a score profile and a composite. *Language Testing* 24(3): 355–390.
- Schiffrin, D. 1994. *Approaches to discourse*. Oxford: Basil Blackwell.
- Schlenker, B.R. 1980. *Impression management: the self-concept, social identity, and interpersonal relations*. Monterey: Brooks/Cole.
- Schmidt, R. 1992. Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition* 3: 357–385.
- Schmitt, N., and D.M. Stults. 1986. Methodology review: analysis of multi-trait multi-method matrices. *Applied Psychological Measurement* 10: 1–22.
- Schoonen, R., A. Van Gelderen, K. De Glopper, J. Hulstijn, P. Snellings, A. Simis, and M. Stevenson. 2002. Linguistic knowledge, metacognitive knowledge, and retrieval speed in L1,

- L2 and EFL writing: a structural equation modelling approach. In *New directions for research in L2 writing*, ed. S. Ransdell, and M.L. Barbier, 101–122. Dordrecht: Kluwer Academic.
- Scollon, R. 2001. *Mediated discourse: the nexus of practice*. London: Routledge.
- Scollon, R., and S.W. Scollon. 2003. *Discourses in place: language in the material world*. London: Routledge.
- Scollon, R., and W.B.K. Scollon. 2004. *Nexus analysis: Discourse and the emerging internet*. London: Routledge.
- Scollon, R., and S.W. Scollon. 2009. Multimodality and language: a retrospective and prospective view. In *The Routledge handbook of multimodal analysis*, ed. C. Jewitt, 170–180. London: Routledge.
- Scriven, M. 1967. The methodology of evaluation. In *Perspectives on curriculum evaluation*, ed. R.W. Tylor, R.M. Gagne, and M. Scriven, 39–83. Chicago: Rand McNally.
- Searle, J.R. 1969. *Speech act: an essay in the philosophy of language*. Cambridge: Cambridge University Press.
- Shepard, L.A. 1993. Evaluating test validity. In *Review of research in education*, vol. 19, ed. L. Darling-Hammond, 405–450. Washington DC: American Educational Research Association.
- Shepard, L.A. 1997. The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–8, 13, 24.
- Shepard, L.A. 2000. The role of assessment in a learning culture. *Educational Researcher* 29(7): 4–14.
- Shohamy, E. 1981. Inter-rater and intra-rater reliability of the oral interview and concurrent validity with cloze procedure. In *The construct validation of tests of communicative competence*, ed. A.S. Palmer, J.M. Groot, and G.A. Trosper, 94–105. Washington, DC: TESOL.
- Shohamy, E. 1996. Competence and performance in language testing. In *Performance and competence in second language acquisition*, ed. G. Brown, K. Malmkjaer, and J. William, 138–151. Cambridge: Cambridge University Press.
- Shohamy, E. 2001. *The power of tests: a critical perspective of the uses of language tests*. London: Longman.
- Shohamy, E., C.M. Gordon, and R. Kraemer. 1992. The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal* 76: 27–33.
- Shute, V.J. 2008. Focus on formative feedback. *Review of Educational Research* 78(1): 153–189.
- Simpson, J. 2003. *Report on BAAL/CUP seminar on multimodality and applied linguistics*. Reading, UK.
- Sinclair, J.M., and M. Coulthard. 1975. *Towards an analysis of discourse*. Oxford: Oxford University Press.
- Skehan, P. 1984. Issues in the testing of English for specific purposes. *Language Testing* 1(2): 202–220.
- Skehan, P. 1995. Analysability, accessibility and ability for use. In *Principles and practice in applied linguistics*, ed. G. Cook, and B. Seidlhofer, 91–106. Oxford: Oxford University Press.
- Skehan, P. 1996. Second language acquisition research and task-based instruction. In *Challenge and change in language teaching*, ed. J. Willis, and D. Willis, 17–30. Oxford: Heinemann.
- Smith, D. 2000. Rater judgments in the direct assessment of competency-based second language writing ability. In *Studies in immigrant English language assessment*, vol. 1, ed. G. Brindley, 159–189. Sydney: Macquarie University.
- Sparhawk, C.M. 1978. Contrastive identificational features of Persian gesture. *Semiotica* 24: 49–86.
- Spolsky, B. 1986. A multiple choice for language testers. *Language Testing* 3(2): 147–158.
- Spolsky, B. 1989a. Communicative competence, language proficiency and beyond. *Applied Linguistics* 10(2): 138–156.
- Spolsky, B. 1989b. *Conditions for second language learning: introduction to a general theory*. Oxford: Oxford University Press.

- Spolsky, B. 1993. Testing and examinations in a national foreign language policy. In *National foreign language policies: practice and prospects*, ed. K. Sajavaara, S. Takala, D. Lambert, and C. Morfit, 124–153. Jyväskylä: Institute for Education Research, University of Jyväskylä.
- Spolsky, B. 2008. Introduction: language testing at 25: maturity and responsibility? *Language Testing* 25(3): 297–305.
- Stein, P. 2008. *Multimodal pedagogies in diverse classrooms: representation, rights and resources*. London: Routledge.
- Stern, H.H. 1978. *The formal-functional distinction in language pedagogy: a conceptual clarification*. Paper presented at the 5th AILA congress, Montreal, Canada.
- Stöckl, H. 2004. In between modes: language and image in printed media. In *Perspectives on multimodality*, ed. E. Ventola, C. Charles, and M. Kaltenbacher, 9–30. Amsterdam: John Benjamins.
- Street, B.V. (ed.). 1993. *Cross-cultural approaches to literacy*. Cambridge: Cambridge University Press.
- Suppe, F. 1977. *The structure of scientific theories*, 2nd ed. Urbana: University of Illinois Press.
- Swain, M. 1985. Communicative competence: some roles of comprehensible input and comprehensible output in its development. In *Input in second language acquisition*, ed. S. Gass, and C. Madden, 235–256. New York: Newbury House.
- Tan, S. 2009. A systemic functional framework for the analysis of corporate television advertisements. In *The world told and the world shown: multisemiotic issues*, ed. E. Ventola, and A.J.M. Guijarro, 157–182. Hampshire: Palgrave Macmillan.
- Tan, S. 2010. Modelling engagement in a web-based advertising campaign. *Visual Communication* 9(1): 91–115.
- Tarone, E.E., and G. Yule. 1989. *Focus on the language learner: approaches to identifying and meeting the needs of second language learners*. Oxford: Oxford University Press.
- Teasdale, A., and C. Leung. 2000. Teacher assessment and psychometric theory: a case of paradigm crossing? *Language Testing* 17(2): 163–184.
- Thibault, P.J. 2000. The multimodal transcription of a television advertisement. In *Multimodality and multimodality in the distance learning age*, ed. A. Baldry, 311–385. Campobasso, Italy: Palladino.
- Thorndike, E.L. 1920. A constant error in psychological ratings. *Journal of Applied Psychology* 4: 469–477.
- Thorndike, R.M. 1997. *Measurement and evaluation in psychology and education*. Upper Saddle River: Merrill.
- Tomasello, M. 2003. *Constructing a language: a usage-based theory of language acquisition*. London: Harvard University Press.
- Toulmin, S.E. 2003. *The uses of argument*. Cambridge: Cambridge University Press.
- Tseng, C., and J. Bateman. 2010. Chain and choice in filmic narrative: an analysis of multimodal narrative construction in *The Fountain*. In *Narrative revisited*, ed. C.R. Hoffmann, 213–244. Amsterdam: John Benjamins.
- Turner, C.E. 1989. The underlying factor structure of L2 cloze test performance in Francophone, University-level students: Causal modelling as an approach to construct validation. *Language Testing*, 6(2):172–197.
- Turner, C.E., and J.A. Upshur. 2002. Rating scales derived from student samples: effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly* 36 (1): 49–70.
- Underhill, N. 1987. *Testing spoken English*. Cambridge: Cambridge University Press.
- Unsworth, L., and E. Chan. 2009. Bridging multimodal literacies and national assessment programs in literacy. *Australian Journal of Language and Literacy* 32(3): 245–257.
- Upshur, J.A., and C.E. Turner. 1995. Constructing rating scales for second language tests. *ELT Journal* 49(1): 3–12.
- Upshur, J.A., and C.E. Turner. 1999. Systematic effects in the rating of second language speaking ability: test method and learner discourse. *Language Testing* 16(1): 82–111.

- van Dijk, T.A. 1977. *Text and context: exploration in the semantics and pragmatics of discourse*. London: Longman.
- van Ek, J.A. 1975. *The threshold level in a European unit/credit system for modern language learning by adults*. Strasbourg: Council of Europe.
- van Leeuwen, T. 1999. *Speech, sound and music*. London: Macmillan.
- van Leeuwen, T. 2001. Visual racism. In *The semiotics of racism*, ed. R. Wodak, and M. Reisigl, 333–350. Vienna: Passagen Verlag.
- van Leeuwen, T. 2011. *The language of colour: an introduction*. London: Routledge.
- van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly* 23(3): 489–508.
- van Moere, A. 2007. *Group oral test: how does task affect candidate performance and test score?* Unpublished Ph.D. thesis, The University of Lancaster, UK.
- Vaughan, C. 1991. Holistic assessment: what goes on in the rater's mind? In *Assessing second language writing in academic contexts*, ed. L. Hamp-Lyons, 111–125. Norwood: Ablex.
- Verhoeven, L. 1997. Sociolinguistics and education. In *The handbook of sociolinguistics*, ed. F. Coulmas, 389–404. Oxford: Blackwell.
- Wainer, H., and H.I. Braun (eds.). 1988. *Test validity*. Hilldale: Lawrence Erlbaum Associates.
- Wang, Y. 2009. The design of multimodal listening autonomous learning and its effect. *Computer Assisted Foreign Language Education* 6: 62–65.
- Wang, L., G. Beckett, and L. Brown. 2006. Controversies of standardised assessment in school accountability reform: a critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education* 19(4): 305–328.
- Webbink, P. 1986. *The power of the eyes*. New York: Springer.
- Wei, Q. 2009. A study on multimodality and college students' multiliteracies. *Computer Assisted Foreign Language Education* 2: 28–32.
- Weigle, S.C. 1994. Effects of training on raters of ESL compositions. *Language Testing* 11(2): 197–223.
- Weigle, S.C. 1999. Investigating rater/prompt interactions in writing assessment: quantitative and qualitative approaches. *Assessing Writing* 6(2): 145–178.
- Weigle, S.C. 2002. *Assessing writing*. Cambridge: Cambridge University Press.
- Weiner, M., et al. 1972. Nonverbal behaviour and nonverbal communication. *Psychological Review* 79: 185–214.
- Weir, C.J. 1990. *Communicative language testing*. Englewood Cliffs: Prentice Hall Regents.
- Weir, C.J. 2005. Limitations of the Common European Framework of Reference for Languages (CEFR) for developing comparable examinations and tests. *Language Testing* 22(3): 281–300.
- White, E.M. 1985. *Teaching and assessing writing*. San Francisco: Jossey-Bass Inc.
- White, S. 1989. Backchannels across cultures: a study of Americans and Japanese. *Language in Society* 18: 59–76.
- Widaman, K.F. 1985. Hierarchically tested covariance structure models for multi-trait multi-method data. *Applied Psychological Measurement* 9: 1–26.
- Widdowson, H.G. 1978. *Teaching language as communication*. Oxford: Oxford University Press.
- Wolfe, E.W. 1997. The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing* 4(1): 83–106.
- Wolfe, E.W., C. Kao, and M. Ranney. 1998. Cognitive differences in proficient and non-proficient essay scorers. *Written Communication* 15: 465–492.
- Wolfe-Quintero, K., S. Inagaki, and H.-Y. Kim. 1998. *Second language development in writing: measures of fluency, accuracy and complexity*. Honolulu: University of Hawaii at Manoa.
- Wolfson, N. 1989. *Perspectives: sociolinguistics and TESOL*. New York: Newbury House.
- Wylie, L. 1977. *Beaux gestes: a guide to French body talk*. New York: E. P. Dutton.
- Xi, X. 2010. How do we go about investigating test fairness? *Language Testing* 27(2): 147–170.
- Yamashiro, A.D. 2002. *Using structural equation modelling for construct validation of an English as a foreign language public speaking rating scale*. Unpublished Ph.D. thesis, Temple University, USA.

- Yang, H., and C.J. Weir. 1998. *Validation study of the national College English Test*. Shanghai: Shanghai Foreign Language Education Press.
- Young, R. 1995. Discontinuous language development and its implications for oral proficiency rating scales. *Applied Language Learning* 6: 13–26.
- Young, R., and W. He. 1998a. Language proficiency interviews: a discourse approach. In *Talking and testing: discourse approaches to the assessment of oral proficiency*, ed. R. Young, and W. He, 1–24. Philadelphia: John Benjamins.
- Young, R., and W. He (eds.). 1998b. *Talking and testing: discourse approaches to the assessment of oral proficiency*. Philadelphia: John Benjamins.
- Zebrowitz, L.A. 1997. *Reading faces: window to the soul?*. Boulder: Westview Press.
- Zhang, D. 2009. On a synthetic theoretical framework for multimodal discourse analysis. *Foreign Languages in China* 1: 24–30.
- Zhang, Z. 2010. A co-relational study of multimodal PPT presentation and students' learning achievements. *Foreign Languages in China* 3: 54–58.
- Zhang, D., and L. Wang. 2010. The synergy of different modes in multimodal discourse and their realisation in foreign language teaching. *Foreign Language Research* 2: 97–102.
- Zhu, Y. 2007. Theory and methodology of multimodal discourse analysis. *Foreign Language Research* 5: 82–86.
- Zhu, Y. 2008. *Studies on multiliteracy ability and reflections on their effects on teaching*.

Nonverbal Delivery in Speaking Assessment
From An Argument to A Rating Scale Formulation and
Validation

Fan, J.

2016, XXV, 305 p. 85 illus., 31 illus. in color., Hardcover

ISBN: 978-981-10-0169-7