In G. Fulcher & F. Davidson (Eds.), The Routledge handbook of language testing. New York, 2012

12

Social dimensions of language testing

Richard F. Young

> Toward the end of the academic year, a man with a clipboard turned up at one of the capital city's best high schools. He sauntered from classroom to classroom, ignoring the students and instead engaged in seemingly trivial chitchat with the teachers, twenty minutes at a time.
>
> Tell me, what subjects are your specialties? How long have you worked here? Can you explain to me a little about how you prepare your lessons? The inspector didn't seem to be particularly interested in what the teachers said. He only cared about how they said it.
>
> Olga Muravyova teaches biology and geography. She laughed nervously as she recalled her meeting with the inspector. 'He wrote a report saying that I understood all the questions, that I answered all the questions, but that I made some errors. That is actually what he claimed,' Ms. Muravyova said. 'Of course that is hard to hear.'
>
> After the inspector told her that she had failed the test, he told her to attend Estonian classes, which she has tried to do. But she is 57, an age when it is not easy to pick up a new language.

This vignette, based on a story in the *New York Times* (Levy, 2010), describes part of a test of spoken Estonian required since 2008 of teachers and other civil servants. In this small former Soviet republic on the Baltic Sea, the government has been mounting a determined campaign to elevate the status of its native language and to marginalize Russian, the tongue of its former colonizer. Public schools, where students have long been taught in Russian, are now linguistic battlegrounds and the test itself is a skirmish between Estonian and Russian. The test and Ms. Muravyova's experience illustrate two social dimensions of language testing: The first is the construct of language knowledge on which the test is built and on which test results are interpreted; the second is what happens to individuals, societies, and institutions when the test is used—the social consequences of assessment. In this chapter, I describe both of these dimensions and argue that, until quite recently, language tests have been built on incomplete knowledge of the social ground of language in interaction and the social consequences of language assessment.

The social life of language

Language testing is a branch of applied linguistics and applied linguistics grew out of earlier work by linguists. Linguists trace the history of their field to ancient India and usually to Pāṇini, who flourished

in the fourth century BCE. One of Pāṇini's main concerns was the correct pronunciation of the body of oral chants of ancient poems known as the *Veda* composed in an early form of Sanskrit that was already archaic by the fourth century. About a thousand years earlier in Shang Dynasty China, we have the earliest written records of a language that today we call Chinese. Ancient Chinese writing was found inscribed on animal bones and turtle shells used for divination. In those days when a ruler wanted advice, he would ask the spirits of his ancestors and other supernatural beings a question, which was written by court officials on an oracle bone or turtle shell. This was then heated, and court officials interpreted the pattern of cracks in the bone as an answer to the question.

In these two ancient examples, interpreting language written on bones and correcting the pronunciation of sacred oral chants was done by people whom today we would call linguists. Linguists today do much the same thing; that is, they take a record in some form and consider (and oftentimes correct) its oral form just like Pāṇini in ancient India, or they interpret the meaning of a written record that has been responded to in some way just like the court officials in Shang Dynasty China. Throughout the long history of their field, linguists have been bound by the physical form of the records that they consider data. When a society transitions from one physical form of language to another—from an oral culture to restricted forms of literacy, for instance—people often complain that the written record does not do justice to the hurly-burly of oral social interaction. In the oral culture of Anglo-Saxon England, when only a select few could read or write, there were many metaphors contrasting the written record with the spoken word. Written words were called 'mouthless speakers', 'dead lifegivers', and 'dumb knowledge-bearers'. To the Anglo-Saxons, the written record of interaction was dead, dumb, and limited—a thing. In contrast, spoken interaction itself is not a thing but a live discursive practice, which the technology of writing took and alienated from the world in which it was originally created. As the Anglo-Saxon scholar O'Brien O'Keeffe (1990: 54) wrote, 'The technology which preserves also kills'.

Today we can go beyond the remembrances of speech and written records that constrained ancient linguists and offended ancient readers. In the twenty-first century, we have still pictures, sound recordings, and movies and, thanks to these new records, our understanding of the forms and functions of language in human interaction—the social life of language—extends far beyond what we know about the forms that language takes in speech and writing. To revivify language, to put language back in its social context, a theory is needed which goes beyond records of disembodied and decontextualized records. The response is Practice Theory, developed by anthropologists (Bourdieu, 1977, 1990; Sahlins, 1981, 1985), sociologists (de Certeau, 1984; Giddens, 1984), and applied linguists (Erickson, 2004; Young, 2009), whose aim was to explicate the nature of social interaction in context. Practice Theorists study discursive practices like language tests; this involves understanding not only the production of meanings by participants as they employ the communicative resources at their command, but also how employment of such resources reflects and creates the processes and meanings of the community in which the test occurs. As Erickson (2004) wrote, although the conduct of talk in local social interaction is unique and crafted by local social actors for the specific situation of its use at the moment of its uttering, it is at the same time profoundly influenced by processes that occur beyond the temporal and spatial horizon of the immediate occasion of interaction. The goal of Practice Theory is to describe both the global context of action and the communicative resources that participants employ in local action. When the context of a practice is known and the configuration of communicative resources is described, the ultimate aim of Practice Theory is to explain the ways in which the global context and the local employment of resources are mutually constituted. Features of global context are described later in this chapter, but how does the global context influence the nature of the test? In other words, how does the global context impact social constructs in language tests?

Social constructs in language tests

The format of Ms. Muravyova's Estonian test could be called a conversation, one in which one party asks the questions and the other party responds. The aim of the test is to evaluate one party's knowledge of language. But what is knowledge of language? When language testers and score users interpret people's scores on a test, they do so by implicit or explicit reference to the construct on which the test is based. Almost all the constructs that underlie high-stakes language tests are theories of individual cognition that can be measured in one context (the test) and are stable enough to be ported to other non-testing contexts where the language is used.

This view of the independence of linguistic knowledge from context underlies the history of linguistics as far back as ancient India and China. In the early twentieth century Saussure (1983) stated the distinction between internal linguistics and external linguistics, which was recently summarized by Lantolf (2006: 74). Lantolf wrote, 'In essence Saussure drew a circle around language (Agar 1994: 41) and proposed that inside-the-circle language, the proper and exclusive domain of linguistic science, was restricted to the study of grammar and dictionary.' If the construct of cognitive ability underlying a language test involves linguistic knowledge that is independent of context—internalized language or I-language in Chomsky's (1986) formulation—then the testing context in which it is elicited is only important to the extent that it helps draw out underlying cognitive abilities (Chalhoub-Deville 2003: 371). According to this construct, the knowledge of Estonian grammar and vocabulary that Ms. Muravyova demonstrated in her conversation with the inspector is independent of the manner in which it was elicited, and her ability and knowledge can be generalized to another context in which she might be called upon again to speak Estonian.

By its emphasis on the mutual constitution of local resources and global context in interaction, Practice Theory provides a very different interpretation of the construct of language knowledge, an interpretation that is best approached by considering how communicative resources are employed in different contexts. In place of knowledge and ability, in Practice Theory the performance of a person-in-context is construed as a configuration of communicative *resources*—resources that a person employs together with others in a particular configuration. Assessment of a person-in-context is challenging, however, because from a person's performance on a test, not only do we wish to infer specific resources employed in the discursive practice of the test, but we also wish to know how the same person will perform in other practices. The challenge was eloquently expressed by Chalhoub-Deville and Deville (2005: 826):

> Evaluating test-takers' performance according to this model offers a conundrum. Generally speaking, we administer tests to, assign scores to, and make decisions about individuals for purposes such as selection, placement, assignment of grades/marks, and the like. If we view language as co-constructed, how can we disentangle an individual's contribution to a communicative exchange in order to provide a score or assess a candidate's merit for a potential position?

The conundrum can be solved by considering the relationship between test performance and the construct underlying a test proposed by Messick (1989, 1996) and revisited by Chapelle (1998) and Norris (2008). Chapelle distinguished between three perspectives on construct definition: a construct may be defined as a *trait*, as *behavior*, or as *some combination of trait and behavior*. In a trait definition of a construct, consistent performance of a person on a test is related in a principled way to the person's knowledge and speech production processes. That is to say, a person's consistent performance on a test is taken to index a fairly stable configuration of knowledge and skills that the person carries around with them—and which that person can

apply in all contexts. In contrast, in the definition of a construct as behavior, the consistent performance of a person on a test is related in a principled way to the context in which behavior is observed. That is to say, test performance is assumed to say something about a person's performance on a specific task or in a specific context, but *not* on other tasks or in other contexts—unless they can be shown to be related to the task or context that was tested.

Clearly, neither definition of a construct as trait or behavior is satisfactory for theories of language in use because, as Bachman (1990: 84) emphasized, communicative language ability consists of both knowledge and 'the capacity for implementing, or executing that competence' in different contexts of use. For this reason, it is desirable to consider the third of Messick's and Chapelle's definitions of a construct, which they refer to as the *interactionalist definition*. In an interactionalist validation of a test, a person's performance on a test is taken to indicate an underlying trait characteristic of that person and, at the same time, the performance is also taken to indicate the influence of the context in which the performance occurs. The interactionalist definition is, in other words, a way to have your cake and eat it: to infer from test performance something about both practice-specific behavior and a practice-independent, person-specific trait.

If an interactionalist definition is to allow test users to generalize from performance in one context to another—that is, from the discursive practice of a test to other practices—then what is needed is a theory that relates one discursive practice to another in a principled way. We need to know whether testees like Ms. Muravyova have the skill to mindfully and efficiently recognize contexts in which resources are employed and to use them when participating in different practices. If the purpose of her conversation with the inspector is to discover whether Ms. Muravyova is able to use Estonian when teaching, we really need to know whether the communicative resources that she displayed in her conversation are portable to another discursive practice: teaching high school biology and geography, practices which she currently does in Russian. Although the contexts of use are very different, it is an empirical question whether the same resources are required in the same configuration in the two contexts. To answer such an empirical question, McNamara (1997: 457) argued that what is needed is 'a close analysis of naturally occurring discourse and social interaction [to] reveal the standards that apply in reality in particular settings'. Such an analysis of discourse and social interaction is one objective of Practice Theory.

Language testing as discursive practice

An analysis of social constructs in discursive practice is characterized by three features (Young 2011). First, analysis of language in social interaction is concerned with communicative resources employed by persons in specific discursive practices rather than language ability independent of context. Second, it is characterized by attention to the co-construction of a discursive practice by *all* participants involved rather than a narrow focus on a single individual. Mehan (1982: 65) stressed the interactional nature of discursive practice when he wrote, '"Competence" becomes interactional in two senses of the term. One it is the competence necessary for effective interaction. Two, it is the competence that is available in the interaction between people.' Mehan's focus on interaction was taken up later by Kramsch (1986: 367), who wrote:

> Whether it is a face-to-face interaction between two or several speakers, or the interaction between a reader and a written text, successful interaction presupposes not only a shared knowledge of the world, the reference to a common external context of communication, but also the construction of a shared internal context or 'sphere of inter-subjectivity' that is built through the collaborative efforts of the interactional partners.

Intersubjectivity is the conscious attribution of intentional acts to others and involves putting oneself in the shoes of an interlocutor. Originating in the phenomenology of Husserl (Beyer 2007), intersubjectivity was first inferred empirically from studies of infant development by Trevarthen (1977, 1979). Examples include an infant's following the direction of an adult's direction of gaze when she points and recognition by an infant of transition-relevance moments in interaction with others.

Third, analysis of social interaction identifies a set of verbal, interactional, and nonverbal resources that participants employ in specific ways in order to co-construct a discursive practice. The resources employed by participants in social interaction are not the mouthless, dead, and dumb records that the Anglo-Saxons complained about. Communicative resources, like language, are embodied and include a participant's whole body, the physical presence and movement of the body, the muscles of the face, arms, and upper body—in particular gaze and gesture—and yes, of course, a participant's speech and writing.

When the communicative resources employed by all participants in a testing practice are specified, the problem of generalizability is resolved by identifying the particular configuration of resources that participants employ in a particular practice. Then, by comparing the configuration of resources in that practice with others, it is possible to discover what resources are local to that practice and to what extent those same resources are employed in different practices and how common resources are configured.

Definitions of verbal, interactional, and non-verbal resources and examples of how participants configure resources in various interactive practices were provided in Young (2008, 2009, 2011) and are revisited here. *Verbal resources* include the register of the practice, defined as a recognizable repertoire of pronunciation and lexicogrammar that occurs with high frequency in certain practices, the combination of which is associated with a specific activity, place, participants, or purpose. In addition, verbal resources that participants employ create certain kinds of meaning in a practice. In Systemic Functional Grammar, Halliday (1994) identified verbal resources as instantiating *ideational*, *interpersonal*, and *textual* metafunctions. The ideational metafunction is how language mediates participants' construction of their experiences of the external physical and biological world, their own internal thoughts and feelings, and the logical relations among them. The interpersonal metafunction, as its name implies, is how language mediates speakers' and writers' personal and social relationships with other participants such as their interlocutors and readers. The textual metafunction is how participants build sequences of discourse, organize the discursive flow, and create cohesion and continuity within a linear and time-dependent flow. In every communicative exchange, participants employ all three metafunctions, although they may be configured in different ways in different practices. As Schleppegrell (2004) showed for the language of schooling, these three metafunctions are realized by various aspects of linguistic substance, and their configuration defines communicative registers and genres.

Interactional and non-verbal resources that participants use to construct a discursive practice include the selection and sequential organization of actions, the turn-taking system that participants use to manage transitions from one speaker to another and the ways in which participants repair interactional trouble. Sequences of actions realized in speech or nonverbally have been investigated most rigorously within the framework of Conversation Analysis developed in the 1960s. Conversation analysts have observed that certain acts occur in ordered pairs and that the production of the first act of a pair gives rise to participants' expectations of the production of the corresponding second act of the pair in an adjacent turn by a different speaker. Schegloff and Sacks (1973) showed how this expectation works with question-and-answer in American English, but it applies equally to other sequences of two adjacent utterances produced by different speakers, including greeting–greeting, offer–acceptance/refusal, and call–response. Act

sequences longer than adjacency pairs are not generally recognized by conversation analysts, but the study of discursive practices of the same type does in fact show that more than two actions occur quite regularly and that they usually occur in a sequence.

The system of turn taking in conversation was first described in detail by Sacks *et al.* (1974), who answered two basic questions about turn taking: How is the next speaker selected? And how do participants know when to end one turn and when to begin another? The question of next speaker selection was answered by Sacks *et al.* by means of the following algorithm.

1. If the current speaker selects the next speaker, then that party has the right and obligation to speak.
2. If no next speaker is selected, then self-selection may (but need not) occur. The first starter acquires rights to the turn.
3. If no next speaker is selected, then current speaker may (but need not) continue.

Sacks *et al.* answered the second question about transitions between turns by invoking the notion of the turn-constructional unit or TCU. Such a unit may be a unit of the lexicogrammar, of intonation, or a pragmatic unit (a complete idea) and, as Ford and Thompson (1996) pointed out, these units often coincide to make a complex TCU. In any given conversation, the moment at which a transition between speakers occurs is not necessarily at the boundary of a TCU, but speakers are able to predict when a boundary is forthcoming and are therefore able to project the completion of the TCU. Transitions between speakers occur at places when participants project the completion of the TCU, projecting not only the form of the next word but also the completion of larger lexicogrammatical, intonational, and pragmatic units. Prediction is thus an important part of what recipients do when listening to talk in progress and the place in an ongoing turn when participants are able to project the completion of the TCU is called a transition-relevance place or TRP. That is to say, participants do not necessarily take a turn *at* a TRP but, if they do, then they are more likely to do so at a TRP than elsewhere.

Repair is the treatment of trouble in talk-in-interaction. Trouble can be anything in talk to which participants in interaction orient as problematic. One participant may use a word that is misunderstood or misheard by another participant; one participant may realize that a phrase that they have just used is less preferable than another phrase. Although the source of trouble is often a word or phrase, it may be anything to which participants orient as repairable. Thus, the absence of the second pair part of an adjacency pair may elicit an apology or, when a listener projects a TCU and takes a turn while the current speaker wishes to continue, this may be oriented to as an interruption. In many cases, however, the source of trouble in a repair is a choice of words or phrasing and, in understanding repair, conversation analysts have focused on two questions about the participants in the repair: In whose turn did the trouble occur? And who initiated the repair sequence? Beyond the participants, the analysis focuses on the sequence of actions in the repair. For example, a repair is called an *other-initiated self repair* if the repair is initiated by a different participant from the one in whose turn the trouble source occurred, and the repair is completed by the same participant in whose turn the trouble source occurred. Excerpt 1 is an example of an other-initiated self-repair in Estonian between a government official and a client who needs information.

Excerpt 1

Pardon? (Gerassimenko, Hennoste, Koit and Rääbis 2004)

1 Client: aga kallis see tööluba on. (0.5)
How much does this work permit cost?

2 official: Kuidas
Pardon?
3 Client: kallis tööluba on.
How much does the work permit cost?
4 official: ei, tö- tööluba ei ole vaja.
No work permit is needed.

In Excerpt 1, Client's question in line 1 is indexed as a source of trouble by the following 0.5-second pause and by Official's repair initiation in line 2. Client concludes the repair by repeating the question in line 3. Client is the other who initiated the repair that is completed by Official (the self). Three other kinds of repair are classified in a similar way according to the participant who initiates the repair and the participant who completes the repair as *other-initiated other repair, self-initiated self repair* and *self-initiated other repair.*

In addition to these, one further resource identified by Levinson (1992) is the way in which participants construct boundaries of a practice. In order for participants to establish mutual orientation to how what they say is creating a context in which the meaning of what they say can be interpreted (what Gumperz 1982, 1992, 1995, called *contextualization cues* in a discursive practice), participants must distinguish it from other practices in which contextualization cues are functioning differently. This is done by means of locating the boundaries of the practice—the opening and closing verbal or non-verbal actions in the sequence of a practice. Not all practices begin and end abruptly and, in fact, boundaries of a practice may be vague, may be negotiated, or may be resisted by one or more participants; nonetheless, boundaries are essential for participants to establish mutual orientation to meaning.

In summary, then, we can describe a discursive practice by specifying the ways in which participants avail themselves of the verbal resources of register and modes of meaning together with the interactional and non-verbal resources of action selection and sequencing, turn taking, repair, and boundary construction. Taken together, these six resources are the fundamental building blocks of intersubjectivity in conversational interaction and the means by which participants craft their local identities as social actors in oral and literate practices. In order to describe local construction of identity, Goffman (1979, 1981) developed the concepts of *participation framework* and *footing.* For Goffman, the identity of a participant in interaction can be animator, author, or principal: animator is an individual engaged in the role of utterance production; author has selected the sentiments being expressed and the words in which they are encoded; and principal's position is established by the words being spoken, whose beliefs have been told and who is committed to what the words say. Three important corollaries of Goffman's theory are: (a) not all participants are necessarily physically present in the interaction—as McNamara (1997: 459) wrote, there are others 'whose behavior and interpretation shape the perceived significance of the candidate's efforts but are themselves removed from focus'; (b) an individual's identity may change from moment to moment throughout the interaction—recognized by Goffman as changes of footing and much expanded by Bucholtz and Hall (2004) as tactics of intersubjectivity; and (c) the participation structure of the practice is the configuration of the identities of all participants, present or not, official or unofficial, ratified or unratified and their footing in the practice.

If we follow McNamara's call for a close analysis of naturally occurring discourse in order to compare it with the discourse of a testing practice, we will see to what extent the communicative resources employed and their configurations are similar or different. Such a comparison was carried out by Young and He (1998) in their collection of studies analyzing the spoken discourse of language proficiency interviews. The studies collected by Young and He addressed a simple comparative question: Is a language proficiency interview an instance of natural conversation? Most

studies answered the question in the negative, pointing out that the system of turn-taking and goal orientation of language proficiency interviews reflects the institutional context in which they are embedded, while in ordinary conversation, topics and turns are neither prescribed nor proscribed by a specific speech activity and none of the participants has a predefined role in managing the conversation. Such a comparison may be extended to compare the test of spoken Estonian that Ms. Muravyova took and the discourse of classroom interaction in an Estonian high school. A corpus analysis will identify the features of Estonian vocabulary, grammatical structures, and pronunciations that occur with frequency in both practices. A systemic functional analysis will identify how participants in the test conversation and the classroom make ideational, interpersonal, and textual meanings, and which participants make which meanings. A conversation analysis of both practices will reveal how participants select and sequence social actions, how they manage the turn-taking system including transitions from one speaker to another and the ways in which participants repair interactional trouble. A similar analysis will describe how participants orient to boundaries and transitions in the classroom and in the oral test.

An examination of these social constructs in tests and in the classroom also reveals how the physically present participants—Ms. Muravyova, her students, and the inspector—construct their identities in interaction, in particular how they co-construct relative power. Power is not only the institutional control of people by a powerful group, nor is it just a mode of thought control, nor does knowledge imply liberation. For Foucault (1978; Foucault and Gordon, 1980), power is exercised in every social interaction and its insidiousness lies in its very ordinariness. At his most explicit, Foucault wrote, 'Power is everywhere; not because it embraces everything, but because it comes from everywhere' (1978: 93). In discourse, power often involves controlling and constraining the contributions of non-powerful participants and the system of allocation of turns in conversation is a particularly effective means of doing so. However, powerful participants are not alone in the exercise of power. Power in discourse is co-constructed by all participants—both the powerful and the non-powerful. For instance, the use of a technical register by one participant constructs power if other participants do not challenge that power by expressing their lack of comprehension. Another example is in Ms. Muravyova's language proficiency interview. She understood and answered all the questions; in other words, the inspector controlled talk by means of allocating the next turn to her, thus constraining her right to speak. Non-powerful participants co-construct power by accepting the constraints imposed upon them. The student who doesn't understand a word may decide to search for the meaning in a dictionary or to ask somebody to explain it; and testees in a language proficiency interview may simply accept the fact that their discursive contributions will be limited.

But the deployment of communicative resources by persons in local interaction does not create these results alone. Power is also created by the system—organized social situations and political institutions that create enduring identities for individuals as testers and testees, as teachers and students, as officials and clients, and expectations for their roles in society. As I argue in the following section, language—more specifically, language testing—is the construction and reflection of these social expectations through actions that invoke identity, ideology, belief, and power.

The social consequences of language tests

After reviewing a variety of nationally mandated tests of language proficiency, McNamara and Roever (2006) concluded that through language tests, political goals affect language learning and the lives of testees at every level. Such influences of the global on the local are to be expected in Practice Theory, the ultimate aim of which is to explain the ways in which the global context and the local employment of resources are mutually constituted.

Mutual constitution means that the actions of a local practice are influenced by the global context *and* the global context is influenced by local actions. In Ms. Muravyova's experience, the influence of the global context is painfully clear: The government of Estonia required her to take Estonian language lessons. It was the government that mandated that she be tested, it was the government that established a National Examinations and Qualifications Centre to administer the tests, it was this body that designed the Estonian language proficiency examinations, it was by this body that the language inspector was trained and paid, and it was this inspector who examined Ms. Muravyova. There is a long political trail before their twenty-minute conversation, and their actions have a long history, a history that goes back long before the 1995 Estonian Law on Citizenship that decreed the test. Modern Estonian is a descendant of one or possibly two of the original Baltic–Finnic dialects. As long ago as the first century CE, the Roman historian Tacitus mentioned a language he called *aestii* (the Estonians' name for their language is *eesti*), but the language has had a long and difficult history with strong early influences from German and Finnish and, in the twentieth century, from Russian. When Estonia was invaded and occupied by the Soviet Union in World War II, the status of the Estonian language changed to the first of two official languages (Russian being the other one). In the second half of the 1970s, the pressure of bilingualism intensified, resulting in widespread knowledge of Russian throughout the country. The Russian language was termed the language of friendship of nations and was taught to Estonian children as early as kindergarten. Although teaching Estonian to non-Estonians in schools was compulsory, in practice, learning the language was often considered unnecessary. The collapse of the Soviet Union led to the restoration of the Republic of Estonia's independence in 1991, and Estonian went back to being the only official language in Estonia.

Given the recent history of Estonia, the government's desire to strengthen and disseminate the Estonian language by requiring teachers and other civil servants to be proficient in it is understandable but it is clearly an example of political goals affecting language testing. In other societies, too, political goals of ruling elites have had significant influence on examinations. Miyazaki (1976) described the long history of the imperial Chinese examination system, which allowed people from all walks of life to enter the prestigious and powerful imperial civil service. At the time it was instituted over fourteen hundred years ago, the system was designed to weaken the power of the hereditary aristocracy at court by requiring aristocrats and commoners to compete on equal terms for positions in imperial service. From a twenty-first-century perspective, the system may appear equitable but it was not designed to be so, for the amount of preparation time required for candidates to study the classical texts on which examinations were based required a degree of economic support simply not available to poor people. Fulcher (2004, 2009) provided many other examples, in societies both ancient and modern, of attempts by political elites to gain control over the education system through testing and to establish norms. Tests are powerful means of political control and tests are effective largely because, as Foucault (1995: 184) wrote, they are a generally accepted 'normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them'.

The normalizing gaze of tests affects the lives of individuals like Ms. Muravyova and at the same time it has been effective in changing the status of languages in multilingual societies and, consequently, the power and prestige of their speakers. Shohamy (2006: 95–98) listed three ways in which language policy objectives are achieved by language tests. Tests are instrumental in (1) determining the prestige and status of languages (and thus maintaining the power of speakers of prestigious language varieties); (2) standardizing and perpetuating language correctness (and thus maintaining the subordinate status of speakers of non-standard varieties); and (3) suppressing language diversity (in favor of speakers of the prestigious standard variety).

Language tests preserve the prestige of the national language

In societies that Fulcher (2009) termed collectivist, the identity and value of individuals is equated with their membership in a collective unit such as a state, a nation, or an institution. One way in which membership is maintained is through use of a common language, and it follows that it is in the interest of the collective that a common language be preserved. Just as when the Estonian Soviet Socialist Republic formed part of the Soviet Union, it was in the interest of the collective to uphold Russian as the national language, in the independent Republic of Estonia today, it is in the interest of the collective to develop Estonian as the national language. Such a collectivist dynamic is in tension with a different political philosophy, termed individualism by Fulcher, which was described by Locke (1690: ¶95) as radically different: 'Men being ... by nature all free, equal, and independent, no one can be put out of this estate, and subjected to the political power of another, without his own consent.' Tensions between collectivism and individualism abound in the plurilingual, multicultural, highly mobile societies of the early twenty-first century. According to Shohamy (2006), a collectivist ideology drives the requirement in American public schools for students to be tested in English in order to graduate and to be tested, again in English, for admission to higher education. Taking an individualist stance, Canagarajah (2009) argued that in India, communities had developed local varieties of English to the extent that the language had now become 'Plurilingual English' and these varieties were the ones most appropriate for local schooling and testing.

Language tests help maintain standards

In language tests, norms of lexicogrammar and style are enforced by the evaluation of a response to a norm as either correct or incorrect. In the case of languages such as English, which are used in and among many different communities around the world, there are many varieties, both regional and international and both nativized and non-nativized. The role of language tests has generally been to maintain the international standard variety at the expense of regional standards or nativized varieties. As Lowenberg (1993) demonstrated in discussing the Test of English for International Communication (TOEIC), a widely used test of Standard English for international communication developed by the US-based Educational Testing Service, lexicogrammatical or stylistic variants prevalent in regional varieties of English are considered errors on TOEIC. Examples of TOEIC "errors," which Lowenberg argued are acceptable to educated speakers of non-native varieties of English, are the following italicized elements taken from TOEIC tests:

> His proposal met with a lot of *resistances*.
> Chemicals in the home *they* should be stored out of the reach of children.
> We *discussed about* the problem until a solution was found,
>
> *(Lowenberg, 1993: 102)*

The collectivist norms used to evaluate test performance on TOEIC are those of the group that uses standard American, British, or Australian English, no matter whether different groups adopt different norms. The normative gaze of the collective is internalized by speakers of non-dominant varieties of the standard language as a subject position that requires them to accept personal responsibility for the communication problems that they encounter. In prescriptive grammars of English, examples abound of non-standard English "errors" such as *irregardless*, *you could of got one*, and *I could care less*. Those are all "errors" that were supposedly committed by native speakers of English, but the problematic status of the norms of Standard English becomes,

like, ginormous in plurilingual contexts or when English functions as a lingua franca, as Seidlhofer, Breiteneder, and Pitzl (2006) have described.

Language tests suppress diversity

Perhaps most problematic for language assessment in multicultural settings are differences in discourse pragmatic norms between a socially dominant group and less dominant groups. Such differences are most often found in the contexts in which directness and volubility are evaluated positively and those in which the same degree of directness and volubility are evaluated negatively. Second language pragmatics is often assessed by discourse completion items such as Excerpt 2, in which verbal action is required.

Excerpt 2
Apology (Röver 2005: 130)

> Ella borrowed a recent copy of *Time* magazine from her friend Sean but she accidentally spilled a cup of coffee all over it. She is returning the magazine to Sean.
> Ella: _____
> Sean: "No, don't worry about replacing it, I read it already."

The pragmatic ideology that Excerpt 2 promotes is that Ella should say something related to her action and she should promise to replace the damaged magazine because of the rejoinder from Sean. In other words, the response of the party who has damaged the possession of another is entirely satisfied *verbally*. No action is required except verbal action, and yet there are occasions when a physical action may be more appropriate and more welcomed by the injured party than any words, although such a test item promotes the pragmatic ideology that verbal action is sufficient.

A critical analysis of language testing practices thus brings to the forefront the social dimension of language testing including speaker subject positions, lexicogrammatical norms, transcultural pragmatic conventions, and many other aspects of societal ideology. Practice Theory proposes that the practices of language testing occur in contexts that are much broader than the testing practice itself, including not only the designers and takers of a particular test, but also the purposes for which the test is designed, the purposes for which people take the test and the ends to which the results of the test are put. McNamara and Roever (2006) have stressed the importance of these broader political questions because, they argue, the requirement to distinguish between *them* and *us* has increased in intercultural societies and in a world of cross-border migration. Distinguishing between them and us is famously recorded in tests such as the password used by American defenders of the Bataan Peninsula against the Japanese in World War II. Stimpson (1946: 51) recounted an Associated Press dispatch from the Bataan front. The Americans discovered an infallible way to distinguish friendly troops from Japanese who attempted to pass the sentries at night dressed in American or Filipino uniforms:

> They simply pick a password with numerous *l*'s, such as *lollapalooza*. Sentries challenge approaching figures and if the first two syllables of *lollapalooza*, for instance, should come back as *rorra*, they open fire without waiting to hear the remainder.

Many more recent and less fatal ways of distinguishing friend from foe are recorded in the language assessment of immigrants, asylum seekers, and those who wish to become citizens. The political context of language testing is just as pertinent in widespread language testing enterprises

resulting from the No Child Left Behind (NCLB) Act of 2001 in the United States and the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR) (Council of Europe 2001). Both of these frameworks are designed to achieve collectivist policy goals and do so, as Foucault (1995: 184) predicted, by combining 'the ceremony of power ..., the deployment of force and the establishment of truth.'

In the case of NCLB, the policy was designed to improve education for all by allowing communities to distinguish between schools where students do well on tests from schools where students perform poorly, and to direct financial resources to those schools with good testing results and, over the long term, to sanction those with consistently poor results. In the early years of the twenty-first century, NCLB came to be seen as the poster child for assessment-based intervention in language education by the Republican Party; in fact, central political control over education in the United States is neither recent nor partisan. NCLB was actually a reauthorization of the Elementary and Secondary Education Act (ESEA), which had its roots in the 1960s during the administration of Lyndon Johnson, a Democrat, and his Great Society. The first mention of language education was discussion of bilingualism in a reauthorization of the ESEA in the 1970s. Harbingers of NCLB were seen in the reauthorization debates in the early and mid-1990s. The movement to control language education began during a Democratic administration, and legislators on both sides of the aisle have long seen that being strong on education helps get votes. NCLB is part of an evolutionary trajectory toward greater state control over education, a movement that continues under the present Obama administration, which is continuing along the path with 'Race to the Top' designed to spur reforms in state and local public education.

McNamara and Roever (2006) summarized the assessment procedures involved in NCLB as statewide tests of reading/language arts and mathematics in grades 3–8 and at least once during high school. The subjects tested and the grades tested are mandated by NCLB, but states develop their own testing instruments. Aggregate results are reported at the school, district, and state level for the entire student populations at the different grade levels. Speakers of English as a second language, however, are one of the four groups of students whose scores are disaggregated from the population. Each school is required to make *adequate yearly progress* in all parts of the assessment: scores of its entire student body and all its disaggregated subgroups on both reading/language arts and mathematics. Evans and Hornberger (2005) argued that the consequences of NCLB for English as a Second Language (ESL) learners have been mixed. On the one hand, added attention has been paid by school districts to their ESL students and small increases in funding of ESL programs have resulted because of the recognition in NCLB of ESL learners as one of the disaggregated groups, who must also show adequate yearly progress. On the other hand, because students' achievement in foreign languages is not assessed in the NCLB framework, bilingual education programs are disappearing in the push to quickly develop students' proficiency in English. In addition, Rosenbusch (2005) reported that NCLB has resulted in a decrease in instructional time for foreign languages, especially in schools with high minority populations. In effect, NCLB has contributed to the hegemony of English in schools and in US society.

The CEFR was initially designed to facilitate the recognition of language credentials across national boundaries, and the framework promulgated one particular theory of language knowledge and the establishment of a progressive set of standards. It has rapidly become institutionalized throughout Europe but, as Fulcher (2004: 260) has argued, the impact of the CEFR and the adequacy of its underlying construct are in need of debate because 'For teachers, the main danger is that they are beginning to believe that the scales in the CEFR represent an acquisitional hierarchy, rather than a common perception. They begin to believe the language of the descriptors actually relates to the sequence of how and what learners learn'.

In both the American and European cases, the establishment of a particular assessment framework has had a very significant effect on teaching and learning.

Because language assessments like these serve the purpose of distributing scarce resources such as jobs, higher education, and financial support to those who desire them, the question of how to distribute those resources fairly is by no means academic. In recent years, the critical theory of Foucault as applied to language testing by Shohamy and Fulcher has set out quite clearly the goals both overt and covert of policy makers and the role of language tests in achieving those goals. No matter how policy is implicated in language tests, no matter whether power is exercised in the service of ends that we admire or ends that we abhor, knowledge of power and how power is exercised is liberating for testees and test designers. It is only by taking this critical perspective on language testing and by implementing proposals such as those by Shohamy (2001, 2004) for democratic assessment and by Fulcher and Davidson (2007) for effect-driven testing that those involved in language testing can assume full responsibility for tests and their uses.

Chapter summary and a look ahead

In recent years, the social dimensions of language testing have received the attention that they deserve largely thanks to the new technologies of recording human interaction that erased the circle around language, which had characterized the linguistic theories of Saussure and his predecessors. Once language is seen as greater than vocabulary and grammar and human meaning making is seen as embodied, theories of language can be expanded to embrace how persons employ communicative resources in social interaction. A theory of the mutual constitution of language and social context permits us to revivify language and to reject a construct of language knowledge that interprets test performance in a disembodied, decontextualized context.

In Practice Theory, social constructs in language tests are seen as a configuration of communicative resources employed by all participants in a test, not just the testee. It is the configuration of communicative resources rather than language knowledge or strategic abilities that characterizes a test as a discursive practice, and the configuration of communicative resources is a means by which participants in a test create local identities and discursive power. Through close analysis of discourse and social interaction, a person's performance in a testing practice can be compared with practices outside the testing room. Testing practice can, however, never be removed from the global context in which powerful elites design and administer language tests and interpret test results. The political context of a newly independent Estonia and the desire of the Estonian government to establish the country's independence from its colonial ruler by language planning and testing have an effect on the lives of teachers like Ms. Muravyova. In communities around the world, the same linguistic battles are fought between local languages and the languages of colonial rulers, between standardized languages and local varieties, and between national languages and the languages of immigrants. The role of language testing in these battles is to maintain the power of speakers of prestigious language varieties and to maintain the subordinate status of speakers of non-standard varieties. It is only recently that critical language testing has made clear the relationship between language and social context, the social consequences of assessment, and the power of language tests.

What are the consequences of this newfound understanding of the social context of language testing? What debates will still be raging and which issues will be unresolved say five or ten years from now? As I gaze into the crystal ball of the future of language testing, I observe three events. One is an image of two psychometricians, experts in the field of educational measurement, sitting in front of a computer monitor scratching their heads as a waterfall of data pours down the screen. I interpret this image to mean that the attempt to measure the rich social

context of language will produce so much quantitative data that new means must be developed to analyze and understand it. I also hear audio from a satellite that picks up sounds from the places over which it flies. The sounds are familiar, almost like English, but while I can understand the audio from the satellite as it flies over some countries, the audio coming from other parts of the world I can't understand, although some words sound like English. I interpret this to mean that world languages such as English will continue to spread as they are doing today in China and Korea, but local varieties will diverge further and further from what a native speaker of standard English can understand today. Finally, I see an image of a band of brown-clad brothers and sisters, guardians of a sacred code of ethics by which they wish all language testers should live. Their code is displayed in a temple of shining marble and the brown-clad band recites it every day. They are, however, alone in their recitations and their temple is empty.

Further reading

Davies, A. (ed.), (1997). *Ethics in Language Testing* [special issue of *Language Testing* 14, 3]. The articles in this special issue were presented in a symposium on the ethics of language testing held at the World Congress of Applied Linguistics in 1996. In ten articles, well-known scholars of language testing address the role of ethics (and the limits of that role). in professional activities such as language testing. The authors discuss language testing as a means of political control, the definition of the test construct, the effects of language tests on the various stakeholders who are involved, and criteria for promoting ethicality in language testing.

McNamara, T. F. and Roever, C. (2006). *Language Testing: the social dimension*. Malden, MA: Blackwell. This book focuses on the social aspects of language testing, including assessment of socially situated language use and societal consequences of language tests. The authors argue that traditional approaches to ensuring fairness in tests go some way to addressing social concerns, but a broader perspective is necessary to understand the functions of tests on a societal scale. They consider these issues in relation to language assessment in oral proficiency interviews and to the assessment of second language pragmatics. They argue that traditional approaches to ensuring social fairness in tests go some way to addressing social concerns, but a broader perspective is necessary to fully understand the social dimensions of language testing.

Shohamy, E. (2006). *Language Policy: hidden agendas and new approaches*. London, UK: Routledge. Shohamy illuminates the decisions surrounding language policy and tests and emphasizes the effects of these decisions on different groups within society. Drawing on examples from the United States, Israel, and the UK, Shohamy demonstrates different categories of language policy, from explicit use by government bodies and the media, to implicit use where no active decisions are made. She also reveals and examines the mechanisms used to introduce language policy, such as propaganda and even educational material. Her critical exploration of language policy concludes with arguments for a more democratic and open approach to language policy and testing, suggesting strategies for resistance and ways to protect the linguistic rights of individuals and groups.

Young, R. F. (2009). *Discursive Practice in Language Learning and Teaching*. Malden, MA: Wiley-Blackwell. Young sets out to explain Practice Theory and its implications for language learning, teaching and testing. He examines the consequences of considering language-in-interaction as discursive practice and of discourse as social action. Discursive practice is the construction and reflection of social realities through language and actions that invoke identity, ideology, belief, and power. The ultimate aim of Practice Theory is to explain the ways in which the global context affects the local employment of communicative resources and vice versa. In chapters 5 and 6, Young uses Practice Theory to take a new look at how the employment of communicative resources in a specific discursive practice may be learned, taught, and assessed.

References

Agar, M. (1994). *Language Shock: Understanding the Culture of Conversation*. New York: Morrow.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Beyer, C. (2007). Edmund Husserl. In *Stanford Encyclopedia of Philosophy*. plato.stanford.edu/entries/husserl (accessed 18 September 2010).

Bourdieu, P. (1977). *Outline of a Theory of Practice*. R. Nice (trans.). Cambridge, UK: Cambridge University Press.
—— (1990). *The Logic of Practice*. R. Nice (trans.). Stanford, CA: Stanford University Press.
Bucholtz, M. and Hall, K. (2004). Language and identity. In A. Duranti (ed.), *A Companion to Linguistic Anthropology*. Malden, MA: Blackwell.
Canagarajah, S. (2009). The plurilingual tradition and the English language in South Asia. *AILA Review* 22: 5–22.
Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing* 20: 369–83.
Chalhoub-Deville, M. and Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah, NJ: Erlbaum.
Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds), *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge, UK: Cambridge University Press.
Chomsky, N. (1986). *Knowledge of Language: its nature, origin, and use*. New York, NY: Praeger.
Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.
de Certeau, M. (1984). *The Practice of Everyday Life*. S. Rendall (trans.). Berkeley, CA: University of California Press.
Erickson, F. (2004). *Talk and Social Theory: ecologies of speaking and listening in everyday life*. Cambridge, UK: Polity.
Evans, B. A. and Hornberger, N. H. (2005). No child left behind: repealing and unpeeling federal language education policy in the United States. *Language Policy* 4: 87–106.
Ford, C. E. and Thompson, S. A. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff and S.A. Thompson (eds), *Interaction and Grammar*. Cambridge, UK Cambridge University Press.
Foucault, M. (1978). *The History of Sexuality*, vol. 1, R. Hurley (trans.). New York, NY: Pantheon.
—— (1995). *Discipline and Punish: the birth of the prison*, 2nd Vintage edn, A. Sheridan (trans.). New York, NY: Vintage.
Foucault, M. and Gordon, C. (1980). *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977*, C. Gordon *et al.* (trans.). New York, NY: Pantheon.
Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly* 1: 253–66.
—— (2009). Test use and political philosophy. *Annual Review of Applied Linguistics* 29: 3–20.
Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment*. London: Routledge.
Gerassimenko, O., Hennoste, T., Koit, M. and Rääbis, A. (2004). Other-initiated self-repairs in Estonian information dialogues: solving communication problems in cooperation. Paper presented at the Association for Computational Linguistics Special Interest Group Workshop on Discourse and Dialogue, Boston, MA, April 30–May 1.
Giddens, A. (1984). *The Constitution of Society: outline of the theory of structuration*. Berkeley, CA: University of California Press.
Goffman, E. (1979). Footing. *Semiotica* 25: 1–29.
—— (1981). *Forms of Talk*. Philadelphia, PA: University of Pennsylvania Press.
Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge, UK: Cambridge University Press.
—— (1992). Contextualization and understanding. In A. Duranti and C. Goodwin (eds), *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge, UK: Cambridge University Press.
—— (1995). Mutual inferencing in conversation. In I. Marková, C. Graumann and K. Foppa (eds), *Mutualities in Dialogue*. Cambridge, UK: Cambridge University Press.
Halliday, M. A. K. (1994). Systemic theory. In R. E. Asher and J. M. Y. Simpson (eds), *The Encyclopedia of Language and Linguistics*, vol. 8. Oxford, UK: Pergamon.
Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal* 70: 366–72.
Lantolf, J. P. (2006). Re(de)fining language proficiency in light of the concept of languaculture. In H. Byrnes (ed.), *Advanced Language Learning: the contribution of Halliday and Vygotsky*. London, UK: Continuum.
Levinson, S. C. (1992). Activity types and language. In P. Drew and J. Heritage (eds), *Talk at Work: Interaction in Institutional Settings*. Cambridge, UK: Cambridge University Press.
Levy, C. J. (2010). Estonia raises its pencils to help erase Russian. *New York Times*, June 8: A6.

Locke, J. (1690). *A Essay Concerning the True Original, Extent, and End of Civil Government.* http://jim.com/2ndtreat.htm (accessed 18 September 2010).

Lowenberg, P. H. (1993). Issues of validity in tests of English as a world language: whose standards? *World Englishes* 12: 95–106.

McNamara, T. F. (1997). Interaction in second language performance assessment: whose performance? *Applied Linguistics* 18: 446–66.

McNamara, T. F. and Roever, C. (2006). *Language Testing: the social dimension.* Malden, MA: Blackwell.

Mehan, H. (1982). The structure of classroom events and their consequences for student performance. In P. Gilmore and A. A. Glatthorn (eds), *Children in and out of School: Ethnography and Education.* Washington, DC: Center for Applied Linguistics.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement,* 3rd edn. New York, NY: American Council on Education and Macmillan.

—— (1996). Validity of performance assessments. In G. W. Phillips (ed.), *Technical Issues in Large-Scale Performance Assessment.* Washington, DC: US Department of Education, Office of Educational Research and Improvement.

Miyazaki, I. (1976). *China's Examination Hell: the civil service examinations of imperial china.* C. Schirokauer (trans.). New York, NY: Weatherhill.

Norris, J. M. (2008). *Validity Evaluation in Language Assessment.* New York, NY: Peter Lang.

O'Brien O'Keeffe, K. (1990). *Visible Song: transitional literacy in Old English verse.* Cambridge, UK: Cambridge University Press.

Rosenbusch, M. H. (2005). The No Child Left Behind Act and teaching and learning languages in U.S. schools. *The Modern Language Journal* 89: 250–61.

Röver, C. (2005). *Testing ESL Pragmatics: development and validation of a web-based assessment battery.* Frankfurt, Germany: Peter Lang.

Sacks, H., Schegloff, E. A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50: 696–735.

Sahlins, M. D. (1981). *Historical Metaphors and Mythical Realities: structure in the early history of the Sandwich Islands Kingdom.* Ann Arbor, MI: University of Michigan Press.

—— (1985). *Islands of History.* Chicago, IL: University of Chicago Press.

Saussure, F. de (1983[1916]). *Course in General Linguistics,* R. Harris (trans.). London, UK: Duckworth.

Schegloff, E. A. and Sacks, H. (1973). Opening up closings. *Semiotica* 8: 289–327.

Schleppegrell, M. J. (2004). *The Language of Schooling: a functional linguistics perspective.* Mahwah, NJ: Erlbaum.

Seidlhofer, B., Breiteneder, A. and Pitzl, M.-L. (2006). English as a lingua franca in Europe: challenges for applied linguists. *Annual Review of Applied Linguistics* 26: 3–34.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing* 18: 373–91.

—— (2004). Assessment in multicultural societies: applying democratic principles and practices to language testing. In B. Norton and K. Toohey (eds), *Critical Pedagogies and Language Learning.* Cambridge, UK: Cambridge University Press.

—— (2006). *Language Policy: hidden agendas and new approaches.* London, UK: Routledge.

Stimpson, G. (1946). *A Book About a Thousand Things.* New York, NY: Harper.

Trevarthen, C. (1977). Descriptive analyses of infant communicative behaviour. In H. R. Schaffer (ed.), *Studies in Mother-Infant Interaction: Proceedings of the Loch Lomond Symposium, Ross Priory, University of Strathclyde, September, 1975.* London, UK: Academic Press.

—— (1979). Communication and cooperation in early infancy: a description of primary intersubjectivity. In M. Bullowa (ed.), *Before Speech.* Cambridge, UK: Cambridge University Press.

US Congress (2002). *No Child Left Behind Act of 2001.* Public Law 107–10 (8 January). United States Congress.

Young, R. F. (2008). *Language and Interaction: an advanced resource book.* London, UK: Routledge.

—— (2009). *Discursive Practice in Language Learning and Teaching.* Malden, MA: Wiley-Blackwell.

—— (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning,* vol. 2, London, UK: Routledge.

Young, R. F. and He, A. W. (eds), (1998). *Talking and Testing: discourse approaches to the assessment of oral proficiency.* Amsterdam, The Netherlands: Benjamins.