



TOEFL.

TOEFL iBT – Reading Competency Descriptors

Competency Descriptors	TOEFL iBT Reading Score Levels (0–30)						
	1–5	6–10	11–15	16–19	20–23	24–27	28–30
I can understand major ideas when I read English.							
I can understand how the ideas in an English text relate to each other.							
When I read English, I understand charts and graphs in academic texts.							
I can understand English vocabulary and grammar when I read.							
When I read academic texts written in English, I understand the most important points.							
I can understand the relative importance of ideas when I read an English academic text.							
I can organize or outline the important ideas and concepts in English academic texts.							
When I read an academic text written in English, I can remember major ideas.							
When I read a text in English, I am able to figure out the meanings of words I do not know by using the context and my background knowledge.							
I can quickly find information that I am looking for in academic texts written in English.							
When I read academic texts in English, I can understand them well enough to answer questions about them later.							
I can read English academic texts with ease.							
I can read and understand texts in English as easily as I can in my native language.							

Likelihood of Being Able to Perform Each Language Task:



<50%
Very
unlikely



50–65%
Unlikely



66–80%
Borderline



81–95%
Likely



>95%
Very likely

1. Life at the chalk-face

Just how different is assessment in the classroom from the world of large-scale standardised assessment? Some believe that they are not only different paradigms, but exist in a state of conflict. Needless to say, it is standardised assessment that is seen as the villain. For example, Stiggins (2001: 12) argues that, while high-stakes tests may be motivating and challenging for the secure and able, for those who 'regard success as beyond their capacity' the outcome is usually demotivation and failure. Some would even go as far as to 'argue that teachers need help in fending off the distorting and de-motivating effects of external assessments' (Shepard, 2000: 7), even using 'the image of Darth Vader and the Death Star to convey the overshadowing effects of accountability testing' (Shepard, 2000: 8). The passion that testing arouses even calls up the rhetoric of the dark side of the force! What is evident in this battle between the practices of standardised testing and classroom assessment is that for the advocates of the latter, there is a sense of injustice, and a need to introduce 'bottom-up' practices that place the teacher in control (Shepard, 1995). The uneasy relationship between externally mandated testing and teacher assessment has been widely studied (Brindley, 1998, 2001), showing, in Rea-Dickins' (2008: 258) words, how 'the wider political context in which children are assessed may constrain desirable assessment practices.'

Even if we resist these external attempts to control what is done in the classroom, there will always be a place for the externally mandated standardised test. It is just that these tests do not do the kinds of jobs we want tests to do in the teaching and learning process. As we noted in Chapter 1, one area in which teachers have used standardised tests for learning is to improve motivation (although Stiggins would argue this is not always the case). Latham (1877: 40) saw why the teachers of his time liked tests: 'The value of examinations ... is far greater as an engine in the hands of the teacher to keep the pupil to definite work than as a criterion.' The analogy of the test as an engine to drive other goals is a powerful one. The technology of standardised testing has been developed in order to produce an engine that is capable of driving a meritocratic social system. Tests encourage learning because they are gateways to goals. In the classroom, however, we wish to devise engines that encourage learning, not only by motivating learners, but also by providing feedback on learning and achievement to both learners and teachers. If learning can also take place through assessment, we may have achieved the effect to which classroom testing aspires.

In this chapter we will consider two major approaches to classroom assessment:

Assessment for Learning and Dynamic Assessment. While these have much in common, they have a different theoretical basis. One is a highly pragmatic approach to classroom assessment, while the other is driven by sociocultural theory. In both cases we will focus upon the practice associated with the movement, although we will explain and critique the theory upon which each is based. We also briefly consider self- and peer-assessment, and portfolio assessment, as useful techniques in assessment for learning. We then look at the link between assessment and second language acquisition in order to see if there is a 'learning sequence' that can be used to inform assessment for learning. We argue that all current approaches to classroom assessment have grown out of criterion-referenced testing, which we describe in some detail. The practice of designing test specifications is the most important practical application of criterion-referenced testing, and so we devote a whole chapter to this later in the book. In this chapter, however, we look at the concept of *dependability* in criterion-referenced testing for the classroom, which is the counterpart of reliability in standardised testing. This provides the tools you will need to investigate the dependability of your own classroom assessments. We close the chapter by assessing the state of the theory underlying classroom assessment.

2. Assessment for Learning

The traditional approach to classroom assessment is sequential (Cumming, 2009: 91). Firstly, teachers establish educational goals and objectives. Secondly, they construct the activities and tasks that will move the learners towards those goals and objectives. Thirdly, they evaluate how well they have succeeded. Since the 1980s, however, there has been a strong interest in the role that assessment can have during the learning process, rather than just at the end of it. The work of Black and Wiliam (1998) in particular, and the 'Assessment for Learning movement' more generally, has had a great deal of impact on many educational systems around the world (Leung and Scott, 2009). While most externally mandated testing is summative, Black and Wiliam focused on formative assessment. The latter are tests or assessments used in the process of learning in order to improve learning, rather than at the end of a period of learning. Unlike tests that are imposed upon the schools, their function is to aid in the diagnosis of individual learning needs. Further, it is not useful in formative assessment to compare learners with one another. The work of Black and Wiliam was not only theoretical. In a large-scale project the practical classroom practices associated with assessment for learning were trialled in schools (Black *et al.*, 2003, 2004). Classes were selected to receive the assessment for learning 'treatment', so that outcomes could be compared with those of control classes at the end of the year. Unusually, the researchers reported the standardised effect size, rather than traditional significance statistics. This is the difference between the scores for the treatment group and the scores for the control group divided by the standard deviation. This takes into account differences in 'gains' as well as the distribution of the groups. An effect size of .3 was reported, which suggests that there are important effects associated with formative assessment.

The reason for this research was a desire to 'raise standards', by which Black and Wiliam meant improving the educational achievement of learners. They argued that the most important way in which standards can be raised is to gather information through assessment that can be used to modify or change teaching and learning activities to better meet the needs of learners. It was also claimed that the process of assessment, including self-assessment, could improve motivation and self-esteem, leading to additional learning gains. Achieving motivation within the classroom is, they argue, more associated with all learners gaining a sense of achievement, rather than encouraging the comparison that inevitably occurs when extrinsic awards are involved. Therefore, one of the most important practices they implemented in their trials was providing only feedback on work, rather than grades or scores. The findings showed that this was particularly beneficial to less able students who achieved much more than they otherwise would have done. It has to be acknowledged that this particular practice can raise larger problems for language teachers. In many institutions – particularly schools – there is an expectation from the learners, their parents and the school management that teachers will grade each piece of work. If a piece of work is not given a grade, the teacher is perceived to be failing in one of their most important tasks. However, research has shown that in a classroom context if a grade is given, a learner will probably pay very little attention to the feedback, however useful it might be. Teachers therefore face the uphill struggle to convince managers, learners and their families that giving grades is not always good practice.

Motivating learners is part of instilling a 'culture of success', where all participants feel that through active participation they can achieve more than they previously thought possible. Most important is giving feedback 'about the particular qualities of his or her work, with advice on what he or she can do to improve' with no comparison to the work of others. What the research discovered was that a number of simple practices led to significant levels of improvement. Firstly, in feedback to all tasks, teachers should try to make learners aware of what they have learned. It should be 'descriptive' rather than 'evaluative', so that it is not perceived negatively (Rea-Dickins, 2006: 168). Secondly, learners need to know what aspects of their performance can be improved and, critically, how they can make that improvement. It is the process of understanding what the goal is, where the learner is now and how they can move towards the goal. Thirdly, the researchers recommended that time for the learner to digest and respond to teacher feedback should be planned into the learning time so that the learners can start to develop metacognitive awareness of their own learning processes.

The practical steps that we have discussed so far have been designed to improve learning. There are also a number of practical teaching practices that support improved learning through assessment. The one that has received the most attention is questioning. Traditionally, teachers spend a lot of class time asking questions. Language teachers have known this for a long time from the discourse studies of classrooms that identified the Initiation–Response–Feedback (IRF) patterns (Sinclair and Brazil, 1982). What Black and Wiliam discovered was that teachers in the classrooms they observed did not leave sufficient time after a question for learners to think about what was being

asked. Rather, if there was no immediate response, teachers would provide the answer themselves or move on to the next topic. The practical recommendation for this most basic form of classroom assessment is to frame questions that do not require the simple repetition of facts, but are more open-ended. Once a question has been asked, teachers should allow longer 'wait times' for the learners to think, and respond. This could very well involve learners discussing the question for a period of time before formulating an answer.

The other critical component of formative assessment is the choice and design of classroom tasks, and how the classroom is managed for learners to undertake these tasks. Designing tasks that engage the knowledge, skills or abilities that we are trying to teach is a complex process. It requires a sound knowledge of the subject area, discussion with colleagues, and a certain amount of technical know-how that we discuss in detail in Chapter 5. Classroom tasks frequently look very different from the kinds of items and tasks that appear in standardised tests. The main reason for this, as we have seen, is the requirement that standardised tests have many items in order to achieve reliability. The response to each item is a piece of information that is used to construct a picture of the test taker's ability. This is not a requirement for classroom assessment, where there is time for much more open-ended tasks that take considerable time to complete. The context of the assessment makes a great deal of difference. Tasks that involve group and pair work are particularly useful in the language classroom, providing the opportunity for production and learning from interaction. They create the opportunity for collaborative learning, in which language learning takes place through language use. As learners become aware of the communication problems that they face in achieving goals, they begin to focus on what they need to acquire next. Swain (2000: 100) put it this way:

There are several levels of noticing, for example, noticing something in the target language because it is salient or frequent. Or, as proposed by Schmidt and Frota, in their 'notice the gap principle', learners may not only notice the target language form, but notice that it is different from their own interlanguage. Or, as I have suggested, learners may notice that they do not know how to express precisely the meaning they wish to convey at the very moment of attempting to produce it – they notice, so to speak, a 'hole' in their interlanguage.

Assessment of the gap between what is now possible and the goal in language learning has been a central theme of second language acquisition research in recent years, with a particular emphasis on the kinds of tasks that encourage 'noticing the gap' (Bygate, Skehan and Swain, 2001),

3. Self- and peer-assessment

Another important component of helping learners to develop a clear picture of the goals of their own learning compared to their current performance is self- and peer-

assessment. Black and Wiliam were among the first to recommend that learners be given the criteria by which teachers (or examiners) judge the quality of work. In many cases, these may have to be simplified for learners. Alternatively, after being shown model samples of language performance, learners may be asked to produce their own rating criteria in groups, and use these to peer-assess each other's work. Following Frederiksen and Collins (1989), this is seen as introducing 'transparency' to the criteria for success, and which Shepard (2000: 12) sees as a basic principle of 'fairness'. The practice of self- and peer-assessment using transparent criteria is designed primarily to assist the development of the awareness of the 'gap' between what is being produced now, and the target performance, thus improving learning.

In order for self- and peer-assessment to work well, it is essential that classroom time be spent on training learners to rate their own work, and the work of their colleagues. This can take significant amounts of time, and the kinds of techniques are not dissimilar to those of rater training to use rating scales (see Chapter 7). Research has shown that, without substantial experience of applying the criteria to work samples, self-assessments can fluctuate substantially (Ross, 1998a; Patri, 2002), but that with training it can be dependable for short periods of time (Ross, 2006). Peer-assessments tend to be much more stable, although they may be more lenient than assessments made by teachers (Matsuno, 2009).

While consistency may be a virtue, dealing with self- and peer-assessment purely in these terms is rather to miss the point. Oscarson (1989: 2) famously argued that 'the question of subjectivity does not necessarily invalidate the practice of self-assessment techniques in language testing and evaluation and, furthermore ... self-assessment may be motivated by reasons that go beyond mere evaluation'. He saw the primary value in the introduction of a shared responsibility between learners and teachers for deciding what constituted 'good' work. This, he contended, led to improved learning through raising awareness of the quality of writing or speech, and establishing a goal-orientation in study. Oscarson recommended the practical devices of getting learners to keep records of their work, and their own perceptions and ratings of how their work improves and develops. This may involve a diary, or a continuous assessment card like the one illustrated in Figure 3.1.

Today this could take other forms, such as a digital audio or video diary, or an online blog in which samples of work and a commentary are saved side by side. This naturally leads on to the use of portfolios, where students collect together samples of writing, or digital copies of speech, into a collection of their work. However, it may also contain reading and listening texts, with an assessment of how well they were understood, and reactions to them. A portfolio represents a wide sample of the work of a particular student to show what they can do with the language. It can be assessed by themselves, their colleagues, the teacher, and even parents. Genesee and Upshur (1996: 100) see the primary benefits of portfolio assessment in conjunction with self- and peer-assessment to be collaboration, inclusiveness, involvement and responsibility, in both learning and assessment. This, they believe, leads to increased motivation.

CONTINUOUS ASSESSMENT CARD		Name <i>Peter Anderson</i>		
Test No →	1	2	3	
Type of test and date	<i>Interview 21 January</i>	<i>Role-playing tasks 19 February</i>	...	
Self assessment	<i>'I thought I could answer about half of the 10 questions satisfactorily. Weak on pronunciation'</i>	<i>'Went very well. But there were a few words and phrases I didn't remember (Important?)'</i>	...	
Test result	7/10	<i>Good</i>	...	
Comments (by teacher or learner)	<i>'Slight under estimation Pronunciation not too bad' (Teacher) 'Better than I thought' (Student)</i>	<i>'You sounded a bit blunt, perhaps (Teacher) 'Must practise polite phrases' (Student)</i>	...	

Fig. 3.1. Continuous assessment card (Oskarson, 1989: 6)

4. Dynamic Assessment

Assessment for Learning in the classroom is therefore premised upon the belief that activities should focus upon making the learner aware of the gap between current abilities and performance levels, and the target or goal that the learner wishes to achieve. Dynamic Assessment (DA) makes the same assumption, but is built upon the work of Vygotsky. Based in sociocultural theory, it provides what advocates claim is 'a new understanding that cognitive abilities are developed through socially supported interactions' (Shepard, 2000: 7). From Vygotsky, DA takes the notion of the Zone of Proximal Development (ZPD) to describe the gap between the learner's current stage of development and the next stage of development. This differs from learning for assessment, as the learner is not necessarily shown what the final target performance is, but shown the gap to the next level of development in a sequence of acquisition or learning. The second difference is in the conceptualisation of the role of the teacher. Rather than 'just' a provider of feedback, teachers are 'mediators'. Lantolf and Poehner (2008a: 273) say:

In DA, assessment and instruction are a single activity that seeks to simultaneously diagnose and promote learner development by offering learners mediation, a qualitatively different form of support from feedback. Mediation is provided during the assessment procedure and is intended to bring to light underlying problems and help learners overcome them.

Mediation is about intervening in the learning process in a way that aids learners to modify their use of language or communication, so that they constantly improve. In terms of the kinds of tasks that learners are given, DA holds that teachers should use activities that learners cannot complete independently, so that mediation is required. The ZPD is then defined as the gap between what the learners can do unaided, and what they can do with assistance (Lantolf, 2009: 363). The nature of the mediation is also important. It can be of two types. If it is 'interventionist', the mediator standardises the mediation, so that it is common across learners. Indeed, this kind of intervention could be provided by a computer in what was traditionally known as programmed learning. DA practitioners, however, recommend 'interactionist' mediation, in which the mediator interacts with each learner depending upon the ongoing assessment of the current stage of the individual's development. It is this 'interaction' that provides DA with the rationale for the use of the word 'dynamic' in its title.

The three methods most closely associated with DA are the 'graduated prompt', 'testing the limits', and the 'mediated learning experience' (Lantolf and Poehner, 2007: 53). The first two are interventionist techniques, and the latter an interactionist technique. In the 'graduated prompt', the mediator creates a task with a graded series of questions to ask a learner who has problems completing a task. The questions start from the most implicit to see if a learner can overcome a difficulty through guided thinking, to very explicitly focusing on the nature of the problem. These prompts are prepared in advanced, and not varied. In 'testing the limits', learners are given feedback on their performance on a task, and then asked to verbalise the problems they feel they have faced, and what they will try to do to overcome them. This technique requires a teacher to work with a single student on a task, and to provide whatever scaffolding is necessary to enable the learner to complete it successfully. The preferred 'mediated learning experience' is a one-to-one interaction in which the mediator interactively helps the learner move toward the next stage of learning through scaffolding attempts to communicate.

Each technique can also be used in a 'cake' or 'sandwich' approach. In the 'cake' approach, mediation takes place after each item or task, and so can only really be used with individuals. On the other hand, the 'sandwich' approach involves mediation at the end of a test or series of activities, and so can also be used with groups.

Whichever combination is used, during the process the teacher notes the extent of mediation necessary in order to evaluate the current level of the learner. This information is used to select the next task. Lantolf and Poehner (2007: 68–69) provide an example of how an interventionist mediation might occur with reference to an item from a language aptitude test (see Figure 3.2).

If the examinee's first attempt to complete the pattern is incorrect, s/he is provided with the following implicit hint: 'That's not correct. Please, think about it once again.' If the second attempt is also unsuccessful, the examiner offers a more explicit hint: 'That's not correct. Think about which rows are most relevant to the one you are trying to complete.' In this case the first row is not relevant ... If the third attempt fails to produce the correct response, the examiner offers an even more explicit hint: 'That's









	haba
	talo
 	talo lata breda
 	talo roto breda
 	?

Fig. 3.2. An item from an aptitude test

not correct. Let's look at rows three and four.' At this point, the examinee's attention is drawn to several important pieces of information: heart is talo; that the language has words (lata and roto) that indicate relative horizontal and vertical position of objects; that the subject or topic of a sentence is given first; that sentence in the third row most likely means "the heart is above the square'.

If the learner still doesn't get the answer correct, it is given by the mediator.

Examples of the much freer interactionist techniques are provided (Lantolf and Poehner, 2007: 72–73), in which an examiner helps a learner to select the correct verb ending.

Example 1

S: **Jugué al tenis* [I played tennis]

[the correct form for the third person is *jugó*]

E: *Jugué o jugó* [I played or she played?]

S: *Jugó* [She played]

Example 2

E: Very good. And here you said, what did she do?

S: *Comí* [I ate]

E: *Comí o comió?* [I ate or she ate?]

S: *Comió* [She ate]

E: *Comió* [She ate]

One of the first reactions to examples like these is to wonder just what the difference is between 'mediation' and what teachers normally do in classrooms to get learners to notice their mistakes. Teachers assume that noticing and correcting leads to learning, even if acquisition may require several instances of noticing over a considerable period of time; research supports this assumption (Long and Robinson, 1998). We therefore have to ask why such an elaborate sociocultural theory is necessary to explain how teachers are constantly assessing in the language classroom. But perhaps this is to miss the point. Perhaps the contribution of DA is to show that teachers are in fact constantly assessing, and that this assessment is part of the learning process. The additional layers of theory are perhaps attempts to explain what already happens, although to what degree this is successful is a question that has not been thoroughly investigated. As Rea-Dickins (2006: 166) correctly asks: 'But what, exactly, constitutes an assessment or an assessment event? How are these differentiated from, say, a teaching event? Is it even possible to distinguish between them? Teachers, so it seems to me, may engage in a continual process of appraising their learners.'

For practical purposes, however, the fact that most of DA has to occur with individuals is problematic. Despite the references to group DA in the 'sandwich' approach, many teachers around the world are faced with large classes where this kind of mediation is not possible. These restrictions upon the use of such techniques should be acknowledged.

What is important to recognise, however, is that adopting DA as the paradigm of choice in classroom assessment makes assessment an entirely local practice. Growing as it does from sociocultural theory, DA practitioners argue that the meaning of the assessment is contextually dependent, where 'language teaching and learning need to be conceived as integrally interactive, jointly constructed, and negotiated processes between teachers and learners, which cannot be prescribed or predicted by general curriculum policies' (Cumming, 2009: 93). The reason it is important to understand the claim is that anything learned from DA is only meaningful in context; with these particular participants, these particular tasks, at this particular time. It cannot have, and does not claim to have, any generalisable meaning beyond the instance of occurrence. As a theory, it therefore embodies the essence of a postmodern social constructivism that restricts meaning to the moment.

5. Understanding change

What is not in dispute is that learning for assessment generally, and DA in particular, are more concerned with change than with stability. Indeed, the whole purpose of 'feedback' or 'mediation' is to cause change. If the intervention or interaction is successful, learning takes place and the learner is no longer the same person. In standardised testing, on the other hand, we assume that the learner's state will remain stable at least for a period of time. Some large-scale testing programmes issue certificates that are recognised for the purpose of university entrance for a period of two years – the period over which little language gain or loss is expected to occur. And the purpose of these tests

is certainly not to cause change. This observation is not in dispute, and so it has been recognised for some time that it is not possible simply to 'map' the validity criteria of standardised testing onto formative assessment (Taylor and Nolan, 1996; Teasdale and Leung, 2000). The idea of calculating a correlation coefficient between the scores of an assessment given at two different times, as we did in Chapter 2, would seem bizarre to DA practitioners. As we have seen, the claims made for DA are that significant changes can be seen over relatively short periods of time. Lantolf and Poehner (2008a: 280) see the difference this way:

Both psychometric concepts [reliability and validity] are built on a foundation that privileges the autonomous individual as the site from which performance and development emerge. DA, on the other hand, is built on a foundation which privileges the social individual, or as Wertsche (1998) puts it 'person-acting-with-mediational-means.' It also must be remembered that DA is not an assessment instrument but is instead a procedure for carrying out assessment.

All the contextual features of DA that lead to change would be classed as threats to reliability in standardised testing. Lantolf appears to have fewer problems with traditional notions of validity. He focuses specifically on predictive validity, as the purpose of assessment in DA is to predict (and assist) learner development from their current stage to the next. This does not appear to be problematic, as long as the claims made by those who practise DA are never generalised, and the measure of the validity of the practice is the extent to which a learner moves from his current stage to the target stage. Or, as Poehner (2008: 3) puts it, 'validating the activity of teaching-assessment requires interpreting its impact on learner development'. When it comes to other aspects of validity, DA practitioners have more serious problems. Discussing the definition of learner competencies, for example, Poehner says: 'In some sense, this is akin to two artists arguing over who has more accurately rendered an autumn landscape, with neither noticing that the seasons have changed' (2008: 9). I think that this argument misses the point. We need to be able to define competencies, skills and abilities even if they change over short periods of time; but this lack of interest in definition shows that DA is not at all concerned with anything but change itself.

Nevertheless, even this weak notion of validity as successful change is not without its problems, as we cannot be sure that the intervention is the cause of the learning. This is because, by definition, there can be no comparison with what any other learner is doing, or even with what a given learner might have been doing if he had not been doing DA. The kinds of control groups used by Black *et al.* (2003) would be meaningless in DA. We therefore do not know whether the learner would make more, less or essentially the same progress under different conditions. But this is the price to be paid for the relativism that comes with social constructivism.

This is not, of course, to say that classroom assessment should adhere to the same validity criteria as standardised tests. We have seen that it is a very different paradigm, with different rules. To this extent I agree with Moss (2003; see a discussion in Fulcher and Davidson, 2007: 192–202) that evaluation has to be in terms of pedagogical deci-

sions made, and their effectiveness. And this has to be done in a paradigm where 'the context is part of the construct' (2007: 25). Validity evidence, however, will not have the power of generalisability beyond the case study.

6. Assessment and second language acquisition

Perhaps one of the most intractable problems associated with DA is the notion that the mediator, who is always a person with more knowledge (the teacher) who can act as a guide, is able to identify and describe both the present state of the learner and the next level of development. It assumes that learning is a progression along a known pathway. This implies a strong link between assessment and second language acquisition (SLA) theory. The kind of theory that is required is a model of SLA that 'includes two dimensions: 1) development, which [is] regular and predictable, and 2) variation, which is largely the result of individual differences' (Bachman, 1998: 190). However, SLA does not provide us with a theoretical model that can be used to construct the kind of progression required to describe the current stage of an individual learner, or the most likely next step on the path. The closest that SLA research can offer is Krashen's (1981) natural order hypothesis and Pienemann's acquisition-based procedures for assessment (Pienemann and Johnston, 1986; Pienemann, Johnston and Brindley, 1988). The problem is that regular and predictable development has only been described for the very limited area of morpheme acquisition in English, and word order in German and English. Extensive research has shown that there is a reasonably stable implicational hierarchy, in which certain forms tend to be learned before others as if they were the building blocks of an interlanguage grammar. Pienemann *et al.* (1988) report these (in acquisitional sequence) as:

Structure

1. Single words, formulae
2. SVO, SVO?
3. Adverb Preposing
4. Do fronting
5. Topicalisation
6. NEG + V (don't)
7. Pseudo-inversion
8. Yes/No-inversion
9. Particle shift
10. V-'to'-V
11. 3rd-SG-s
12. Do-2nd
13. Aux-2nd
14. Adv-ly

Example

- How are you?
*The tea is hot?
*Yesterday I work
*Do he work?
This I like
*He don't eat meat
Where is my purse?
*Have he seen it?
*He turn the radio on
We like to sing
She comes home
They did not buy anything
Where has he seen you?
They spoke gently

- | | |
|------------|---------------------------|
| 15. Q-tag | It's expensive, isn't it? |
| 16. Adv-VP | He has often heard this |

Explanatory notes

Adverb preposing: In English some, but not all adverbials, may be placed in sentence initial position.

Topicalisation: The placement of objects or subordinate clauses in sentence initial position, such as 'Because I feel ill, I can't work.'

Pseudo-inversion: In wh-questions with a copula, the subject and copula must be inverted.

Yes/No-inversion: In questions to which the answer is 'yes'/'no', the modal or auxiliary comes to sentence-initial slot.

Particle shift: The verb and preposition of a phrasal verb are split.

Do-2nd and Aux-2nd: In main clauses the auxiliary and the model are in second position in positive sentences and wh-questions.

In acquisition studies it was found that these features were acquired in five discrete stages that could be used in a speaking test to place a learner in an acquisitional level (Pienemann *et al.*, 1988: 228):

Stage 1: Single words and formulae

Stage 2: SVO, plural marking

Stage 3: Do fronting, topicalisation, adverb preposing, Neg+V

Stage 4: Pseudo-inversion, Yes/No-inversion

Stage 5: 3rd-SG-s, Aux-2nd, Do-2nd

Pienemann *et al.* (1988: 221) argue: 'If the teachability of grammatical forms is constrained by the learner's current stage of language development, and furthermore if this development is the same for all learners, then teaching and by extension, testing, can be geared to what is currently learnable by profiling the learner's present state of development.' This became known as the 'teachability hypothesis' (Larsen-Freeman and Long, 1991: 282), which puts constraints on what can be learned next, and also predicts what will be learned next. The problem for language testing is that we normally do not wish to restrict ourselves to testing grammatical structures.

How does DA cope with this problem? Lantolf (2009: 357–358) explicitly rejects any view of second language acquisition that posits a universal process that is regular and predictable. He also rejects any general learning theory, like Piaget's, that posits a developmental sequence. Rather, Lantolf claims that DA prioritises action, so that 'effective instruction must precede and indeed lay down the path for development to follow' (2009: 358). This implies that it is the mediator who is able to influence the acquisitional path of the learner, so that the next stage is decided by the current intervention. This position is theoretically unsatisfying. If we are unable to make predictions about acquisition based on the theory, there is no way, even in principle, of observing any changes that might bring the theory into question. So, while SLA offers only limited findings that

can help us establish an acquisitional sequence, DA appears to abandon any hope that it may exist in favour of a view that if we only use the recommended technique, anything is possible.

This raises a crucial question about the relationship between theory and data. We have observed that the kinds of interventions recommended by DA are not dissimilar to what most language teachers would see as regular teaching practice. Further, no one is questioning that these interventions are highly likely to lead, however slowly, to language acquisition. The problem is in explaining why. Whenever we observe phenomena like these exchanges leading to learning, we attempt to create explanations (theories). The value of these theories lies in whether they are capable of predicting what will happen under certain conditions in new contexts. In other words, it should be testable. DA appears to be based on a theory that is not testable, but is self-validating in each new context. While the practice may be useful, the theory may be just so much unnecessary baggage.

7. Criterion-referenced testing

The insights that have led to Assessment for Learning and DA come from criterion-referenced testing and assessment (CRT). Whether we decide to use these integrated approaches to classroom assessment or traditional linear classroom testing, one of the key features of classroom testing is that test takers are not compared with each other. As Stiggins (2001: 10) puts it, the change is 'from merely sorting students to ensuring attainment of specific competencies'. If there is a 'score' at all, its meaning is not derived from the distribution of scores. Furthermore, we do not expect a set of scores to be normally distributed. If the purpose of assessment for learning is to improve performance on tasks or any kind of test that is a measure of what has been learned, we expect (and hope) that most of the learners will do well. The kind of distribution that we wish to see is *negatively skewed*, as shown in Figure 3.3.

When this happens, the technology of standardised testing fails. The engine no longer runs in the way predicted, and the statistics that we discussed in Chapter 2 can no longer be used. Those statistics depend on the assumptions of normal distribution, and good discrimination. We have neither of these when assessment for learning is working well.

This recognition led to the evolution of the second paradigm in assessment, which was named 'criterion-referenced testing' by Glaser (1963: 519), and he described it as follows:

Achievement measurement can be defined as the assessment of terminal or criterion behaviour; this involves the determination of the characteristics of student performance with respect to specified standards.

The first thing to note in this quotation is the use of 'standards'. We have already noted that this has multiple meanings in the language testing literature, and here it is interpreted interchangeably with 'criteria' in real-world performance. The principle is that

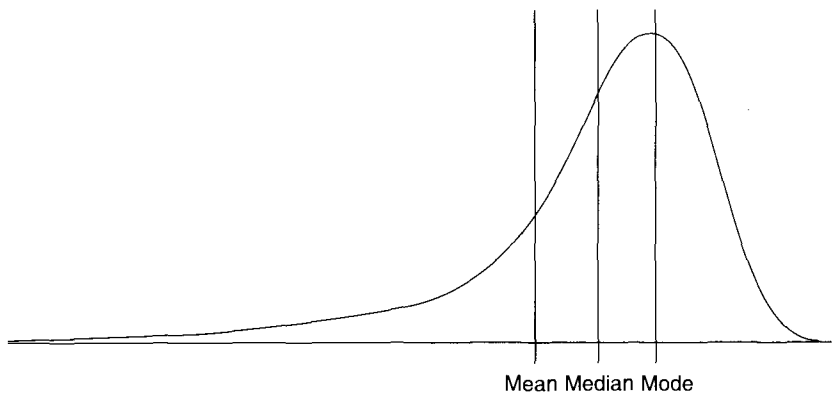


Fig. 3.3. A negatively skewed distribution

if we can describe the target performance, and stages along the route to that performance, we can assess where a learner is on the trajectory. This is the same principle that we see in Assessment for Learning and DA, but there is an assumption that we can specify a 'continuum of knowledge acquisition' (Glaser, 1963: 519). The fact that Glaser's work was done in relation to the use of new technologies in programmed learning indicates the fact that this progression was seen as a linear 'building-block process', in a behaviourist learning model (Shepard, 1991; Glaser, 1994b: 27). However, what was completely new was the focus upon the description of what was to be learned, and what came to be called 'authentic assessment' (Glaser, 1994a: 10). This was essentially a new interest in the content of tests that had not been taken so seriously in earlier large-scale, standardised tests. While recognising the role played by closed response items in standardised tests, the criterion-referenced testing movement also saw that they were not the most efficient way of representing real-world performances in tests (Frederiksen, 1984). Glaser (1994b: 29) overtly argued:

as assessment departs from the confinement of multiple-choice testing, freer formats will enable many of the processes and products of learning to be more apparent and openly displayed. The criteria of performance will be more transparent so that they can motivate and direct learning. The details of performance will not only be more open for teacher judgement but will also be more apparent to students so that they can observe and reflect on their own performances and so that they can judge their own level of achievement and develop self-direction. If this occurs, in an appropriate social setting in the classroom, then students along with teachers can observe one another and provide feedback and guidance as they learn to help and receive help from others. In this scenario, one can ask: In such classroom assessment, where do the performance criteria reside?

The challenge is in producing such criteria, a topic to which we turn in Chapters 4 and 7. However, we should note here that criterion-referenced testing is no longer linked to

behaviourist theories of language learning. Secondly, as Shepard (1991: 5) points out, within this model, testing is not used to 'drive instruction'. Rather, the test is used to aid and monitor instruction. It is in the service of teaching, rather than being its master. Or as Latham (1877: 8) would put it: 'It makes all the difference whether the teaching is subordinate to the examination or the examination to the teaching.'

In criterion-referenced testing the teaching comes first, and the results of the tests are used to make decisions about learners and instruction. As Popham and Husek (1969: 3) pointed out, criterion-referenced testing was therefore not the tool of choice for selection purposes. As we have already seen, there is no expectation of discrimination, or large standard deviations. If, as they argued, the meaning of any score 'flows directly from the connection between the items and the criterion', the critical feature of criterion-referenced testing is the *test specification*, which describes the nature of the items and the rationale for their use in the test. It is in the test specifications (see Chapter 5) that the link between test and the real world is established (Popham, 1994) which has been called 'item-objective congruence' (Hambleton, 1994: 23).

8. Dependability

In classroom testing we wish to know whether the results of the assessment are dependable. This concerns whether an estimate of a learner's current stage would change if a different teacher conducted the assessment, or if a different (but comparable) task was used. Dependability is the criterion-referenced correlate of reliability in standardised testing.

The traditional methods of investigating reliability that we considered in Chapter 2 cannot be used, as the lack of variance 'would lead to a zero internal consistency estimate' (Popham and Husek, 1969: 5). Cronbach's alpha would always be very low. Rather, we need to turn to estimates of the consistency or dependability of decisions. These are the kinds of decisions that teachers make when they decide that a learner has or has not achieved a certain level. This decision is sometimes called a 'mastery/non-mastery' decision, or even 'pass/fail' (Brown and Hudson, 2002: 151). Alternatively, we may have three or more levels, each indicating a stage in the learning process. Each level would normally be carefully described, and in some cases a *cut score* on the test would be established to place learners into levels (see Chapter 8 for a discussion of cut scores and how to decide where a cut score should be placed on a test). These are referred to as 'absolute decisions', defined as 'one in which we select or reward test takers on the basis of their level of knowledge or ability, according to some pre-determined criteria' (Bachman, 2004: 11).

The most common ways of calculating dependability are known as the 'threshold loss agreement approaches'. These require the same test to be given twice, just as in calculating test-retest reliability. The purpose is to calculate whether learners are consistently classified as 'masters' or 'non-masters'.

The first approach that we discuss is called the agreement coefficient, or P_o (Bachman,

2004: 200), and is very easy to calculate. The teacher gives a test or assessment twice, and on each occasion the decision to class a learner as a 'master' or 'non-master' is recorded. Table 3.1 sets out the results using a fictional group of 60 learners.

Classification		2nd administration		Total
Classification		Master	Non-master	
1st administration	Master	41 (A)	6 (B)	47 (A + B)
	Non-master	5 (C)	8 (D)	13 (C + D)
Total		46 (A + C)	14 (B + D)	60 (N)

Table 3.1 A classification table

With this information, the calculation of the agreement coefficient is now very simple:

$$P_o = \frac{A + D}{N}$$

With our sample data, this translates into:

$$P_o = \frac{41 + 8}{60} = .82$$

In other words, there is an 82 per cent agreement in the classification of students across two administrations.

As this calculation only requires information about classification of learners on two administrations, it can be used independently of the type of assessment used. It could be a normal classroom test, a collaborative communication task, or a piece of writing. The problem normally comes with conceptualising giving the task twice, because this is not something that teachers would normally do. But this is an illusory problem for the teacher. In most syllabuses we return to skills and subjects in a cyclical manner. We know that learners do not acquire language ability without repetition and practice. Teachers therefore design multiple tasks to practise the same skills, or use similar linguistic forms. For example, when we teach the skill of skim reading to get the gist of a text, we are likely to use multiple texts, each with its own prompts. If we have two such tasks that have been 'generated' by the same target skill (from a task specification – see Chapter 5), and the texts are equally difficult in terms of structure, vocabulary, length and cognitive load, we can treat them as equivalent.

It has been argued that there is a problem with the agreement coefficient. Whereas Cronbach's alpha can range from 0 to 1, P_o can never be 0 because just by chance, cells A and D in Table 3.2 would have a positive entry (Brown and Hudson, 2002: 169–170). This chance factor in the assessment can be calculated easily from the data in Table 3.1.

The formula for the chance factor is:

$$P_{\text{chance}} = \frac{(A + B) * (A + C) + (C + D) * (B + D)}{N^2}$$

For our example, this would be:

$$P_{\text{chance}} = \frac{(47) * (46) + (13) * (14)}{3600}$$

$$P_{\text{chance}} = \frac{2162 + 182}{3600}$$

$$P_{\text{chance}} = \frac{2344}{3600} = .65$$

This tells us that 65 per cent of the 82 per cent classification agreement between the two teachers could have occurred by chance alone. In order to correct the agreement coefficient for this chance element, we normally use the kappa coefficient, which is easily calculated as follows:

$$k = \frac{P_o - P_{\text{chance}}}{1 - P_{\text{chance}}}$$

We can fill this in from the calculations that we have already made:

$$k = \frac{.82 - .65}{1 - .65}$$

$$k = \frac{.17}{.35} = .49$$

The rule of thumb to interpret Kappa is:

.01-.20	slight agreement
.21-.40	some agreement
.41-.60	moderate agreement
.61-.80	substantial agreement
.81-.99	very high agreement

The figure in our example shows moderate agreement, and can be used as an indication that the working definition of 'mastery', or the tasks, may not be as stable as they might be to achieve higher levels of dependability.

Threshold loss agreement approaches estimate consistency in classification, and it does not matter what kinds of tasks or activities are used in making the classification. There is no requirement that there is a 'score', only a decision; but it is assumed that the tasks used measure the same constructs and are of roughly equal difficulty. Another approach to estimating consistency is squared-error loss agreement. Instead of looking at just the consistency of classification, this approach takes into account the degree of mastery or non-mastery, rather than just the classification (Brown and Hudson, 2002: 193–197). However, it also means that it must be possible to get a range of scores on the test; as such, it can only be used with more traditional tests that contain many items, and the items must be scored right or wrong. *Partial credit* is not possible. The most useful statistic, which is very easy to calculate by hand, is Phi Lambda (written $\Phi\lambda$). The formula for this statistic is:

$$\Phi\lambda = 1 - \frac{1}{K-1} \left(\frac{\bar{X}_p (1 - \bar{X}_p) - S_p^2}{(\bar{X}_p - \lambda)^2 + S_p^2} \right)$$

The symbols in this formula mean:

K	number of items on the test
\bar{X}_p	mean of the proportion scores
S_p	standard deviation of the proportion scores
λ	cut-point expressed as a proportion

All of these elements are familiar from Chapter 2, with the exception of λ . This is the 'cut score', or the score on the test over which a teacher will judge learners to be masters. In order to illustrate the use of this statistic, I will assume that a teacher creates a ten-item test of skimming for the gist of a passage, which she gives to fifteen learners. This is for illustrative purposes only. We would normally prefer to use more items than this. We will also assume that the cut score has been established at 6, using one of the procedures described in Chapter 8. The purpose of the test is to assess whether this skill has been acquired in the reading classes. The results are presented in Table 3.2. An entry of 1 in a cell indicates that the item has been answered correctly, while an entry of 0 indicates that the response is incorrect. These are added up in the column headed 'total' for each learner, giving each person's score. The final column is the proportion correct for each learner, which is the total correct, divided by the number of items (ten). The mean and standard deviation of the proportions are then calculated, in the same way that we learned to calculate the mean and standard deviation in Chapter 2.

Learner	Item number										Total	Proportion correct
	1	2	3	4	5	6	7	8	9	10		
1	1	1	1	1	1	1	1	1	1	1	10	1.0
2	1	1	1	1	1	1	1	1	1	0	9	0.9
3	1	1	1	1	1	1	1	1	0	0	8	0.8
4	1	1	1	0	1	1	1	1	1	0	8	0.8
5	1	1	1	0	1	1	1	0	1	0	7	0.7
6	1	1	1	1	0	1	1	1	0	1	7	0.7
7	1	1	1	1	0	1	1	0	1	0	7	0.7
8	1	1	1	1	0	0	1	1	0	0	6	0.6
9	1	1	1	1	0	1	1	0	0	0	6	0.6
10	1	1	0	1	0	1	1	1	0	0	6	0.6
11	1	1	0	1	0	1	1	0	0	0	5	0.5
12	1	0	1	1	0	1	0	0	0	0	4	0.4
13	1	1	0	1	0	1	0	0	0	0	4	0.4
14	1	1	1	1	0	0	0	0	0	0	4	0.4
15	1	0	0	0	0	0	0	0	0	0	1	0.1
											\bar{X}_p	.61
											S_p	.23

Table 3.2 Results of a reading test

These figures can now be plugged into the formula:

$$\Phi\lambda = 1 - \frac{1}{10 - 1} \left(\frac{.61(1 - .61) - .23^2}{(.61 - .60)^2 + .23^2} \right)$$

$$\Phi\lambda = 1 - \frac{1}{9} \left(\frac{.24 - .23^2}{.00 + .23^2} \right)$$

$$\Phi\lambda = 1 - .11 \left(\frac{.19}{.05} \right)$$

$$\Phi\lambda = 1 - (.11 * 3.8)$$

$$\Phi\lambda = .58$$

With the cut score at 6, we have a moderate value of dependability. This could be increased by using a different cut score, but if the cut score has been established on substantive grounds, it is much more appropriate to go back to investigating whether the test really assesses the construct in a satisfactory way. This may involve reviewing

test specifications, changing or increasing the number of items, and reviewing decisions relating to the cut score.

Just as we looked at the standard error as one of the most important statistics in estimating the reliability of a standardised test, we can also calculate its equivalent for a criterion-referenced test. This is called the *confidence interval* (CI) around a score (Brennan, 1984), which allows us to see whether learners near the cut score may be below or above just by chance. The formula for the confidence interval is:

$$CI = \sqrt{\frac{\bar{X}_p (1 - \bar{X}_p) - S_p^2}{K - 1}}$$

The meaning of these symbols is the same as in the calculation of $\Phi\lambda$ above, with one exception. The standard deviation is calculated with N (rather than $N - 1$) in the denominator. The formula is therefore slightly different from that given in Chapter 2:

$$SD = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

For our data, this actually makes no difference at all to S_p^2 , but with a larger number of items or learners, there may be differences.

We can therefore calculate CI as follows:

$$CI = \sqrt{\frac{.61 (1 - .61) - .23^2}{10 - 1}}$$

$$CI = \sqrt{\frac{.24 - .05}{9}}$$

$$CI = \sqrt{.02} = .14$$

This figure tells us that if an individual took the reading test a number of times, their score may go up or down by .14 proportion score points (or 14 per cent of raw score), around 68 per cent of the time (Brown and Hudson, 2002: 186–187). This is very important information when making decisions about learners whose score falls near to the cut score. For our reading test with a cut score of 6 this would be any score between 4 and 8; we could only be fairly certain that learners with scores of 1–3 had not mastered the skill, and learners with scores of 9–10 had. In fact, additional information is needed to make decisions about anyone who has a score within the CI of the cut score. Once again we have discovered that in assessment we need to develop strategies to deal with uncertainty.

9. Some thoughts on theory

Shepard (2000: 10) argues: 'I believe we should explicitly address with our teacher education students how they might cope with the contesting forces of good and evil assessment as they compete in classrooms to control curriculum, time, and student attitudes about learning.' It is probably not useful to see testing as quite so black and white. All forms of testing and assessment are socially constructed activities to achieve certain goals. There is a very important role for classroom assessment, and the integration of assessment with learning. Unlike Shepard (2000: 6), I would argue that there is also a place for tests that do not require teachers to make judgements about their own students. We know that sometimes teachers are influenced by factors other than the knowledge, skills and abilities of some students. This is inevitable when people work together in a learning environment over an extended period of time. This extended contact is essential for learning; it creates the social learning context. It also means that the learning context differs from teacher to teacher, and school to school. When testing or assessment is being used for high-stakes purposes, or where learners are being compared with each other across contexts, there is a case for using external tests. This does not mean that teachers should be excluded. They are stakeholders in the process. It is important that they are consulted and included in decision-making processes.

Sometimes the use of externally mandated tests can protect teachers. If they are personally responsible for high-stakes decisions, they are open to the accusation of personal bias. And, whether teachers or test designers like it or not, politicians are going to use test scores for more than informing learning and teaching. For example, Obama (2006: 161) has an insight into the good that testing and assessment can do in the classroom when he calls for 'meaningful, performance-based assessments that can provide a fuller picture of how a student is doing'. But he is also concerned with the statistics of failure, which come from national and standardised test scores.

Throughout our history, education has been at the heart of a bargain this nation makes with its citizens: If you work hard and take responsibility, you'll have a better life. And in a world where knowledge determines value in the job market, where a child in Los Angeles has to compete not just with a child in Boston but also with millions of children in Bangalore and Beijing, too many of America's schools are not holding up their end of the Bargain.

(2006: 159)

The fact is that there never was a time when testing was not high stakes, when it was not used to select individuals for 'a better life'. The scores have always had an economic value, even though, as Latham (1877: 6) says, 'people are hardly aware of how thoroughly the educational world is governed by the ordinary economical rules'. This is the reason for treating formative classroom assessment as a different paradigm. Its role is to aid learning, not to make high-stakes decisions. To create awareness of learning goals

and stages of development, not to make awards or to certify achievement, not to monitor time and materials use, not report scores to external authorities.

Assessment for Learning and DA have much to offer the teacher. Both provide practical advice that has been found to lead to improved learning. Nevertheless, they differ in their theoretical underpinnings. Assessment for Learning attempts to combine the lessons learned from research in large-scale testing with a sensitivity to context. It does not adopt a strong constructivist stance, and so practitioners can conduct research to show that its methods are more successful (under some circumstances, with certain types of learner) than other methods. It is prepared to live with probabilistic statements of success. It is a position that is essentially experiential. The various practices have been seen to work in many (but not all) contexts, and to have benefited less able learners in particular. There is no strong theoretical claim to support the recommended practices, making it a pragmatic approach to what works in the classroom.

DA is different, because it is based upon sociocultural theory. Indeed, the jargon of DA is frequently impenetrable on first encounter with the literature. When it comes to looking at the examples of DA practice, they seem to differ little from what most language teachers would do anyway; and in many respects the techniques look less innovative than those of Assessment for Learning. This raises the question of why the associated theory is necessary to explain the evidence.

Perhaps the most serious problem for DA is that it does not appear to have any apparatus for rejecting alternative hypotheses for what is observed in case studies. Each case study is presented as a unique, non-generalisable event. What happens in each instance of DA involves the contextual interpretation of the participants who co-construct temporally bound meaning. Poehner (2008: 12) follows Luria (1979) in calling this 'Romantic Science'. Unlike regular science, this means that they 'want neither to split living reality into its elementary components nor to represent the wealth of life's concrete in abstract models that lose the property of the phenomena themselves'.

Indeed, this is not science. It is poetry, and art. It is an attempt to appreciate the whole as a piece, rather than create variables that can be investigated. There is a place for poetry and art. However, this is a fundamental flaw in DA. The purpose of any theory is to explain generalities in observable phenomena. But DA says that there are no generalities. In substantive theories systematic effects of interventions should be predicted and tested. This is one way to investigate the validity of theory. The effects may change according to the presence or absence, or degree of presence, of contextual or individual variables. These can be listed in the theory as moderating the predicted effects of the intervention or interaction on the outcomes. Only in this way can we know whether the theory has explanatory adequacy. DA denies this to us. Instead, it appeals to us to appreciate the holistic meaning of the event; to savour the landscape. The fundamental contradiction in DA is that it attempts to claim substantive theoretical justification, but abandons the need for it.

Formative classroom assessment is undertheorised. The rationales can be drawn from sociocultural theory or from SLA, but where it has been attempted it has been found that the needs of less able learners are

has provided much needed guidance in the
evaluation of classroom assessment has grown up in the form of criterion-referenced
testing and assessment. The teachers' formative assessment toolbox is far from empty.