Plenary speeches

The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect

Fred Davidson University of Illinois at Urbana-Champaign, USA
fgd@uiuc.edu

Glenn Fulcher The University of Leicester, UK
gf39@le.ac.uk

This is a revised and combined version of two plenary addresses given at the conference The Implementation of the Common European Framework of Reference for Languages (CEFR) in UK Higher Education Institutions – Chances and Challenges, Queen Mary College, London, UK, 24 November 2006.

Language test development proceeds best when the test's effect is borne in mind, throughout the test development process. The authors discuss the flexible language of the Common European Framework of Reference for Languages (CEFR) and explore the pragmatic utility of such language to guide language test development. They select service encounters (e.g. airline ticket sales, open-air markets) as a sample language use domain to illustrate demonstrable weaknesses in the Framework. Using the CEFR Level A1 service encounter descriptor, suggested testing materials are shown in a versioned evolution of a proposed test specification. Provided that effect is kept in mind, the authors argue, the CEFR is actually a valuable – even an optimistic – starting point for language test development.

1. Effect-driven language test development and the Common European Framework of Reference (CEFR)

Language testers face two grand forces in the design decisions for their tests. On one hand, they can base the test on a model of language ability, such as that articulated by various scholars of applied linguistics over the years. This is MODEL-DRIVEN TEST DEVELOPMENT, in which the primary shaping force is the belief structure of the model under consideration. An alternative is to build a test upon its intended purpose or consequence – some model(s) of language ability may (and, we would argue, should) still shape the design of the test, but what really determines the test tasks is the effect they will have: on student learning, curriculum, educational policy, and so forth. We call this EFFECT-DRIVEN TEST DEVELOPMENT (Fulcher & Davidson 2007).

Effect-driven language test development is a pragmatic approach to testing. Pragmatism is a utilitarian or instrumental school of philosophy, rooted in the 'Pragmatic Maxim' of the philosopher Charles Sanders Peirce:

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object. (Peirce 1878: 146)

In this philosophy, the nature of a thing is its effect. For example, if a Ministry of Education (in some nation-state) develops a secondary-school foreign language exit exam, the pragmatic nature of the exam is the complex of impact it has upon its users, its test-takers, and the nation as a whole. If the exam becomes part of the evidence to admit (or deny) access to higher education, then that power defines it – to a large degree. The exam in question becomes what it does: its role is its meaning.

Peirce, John Dewey, and William James are associated with the early Pragmatic school of philosophy. In later years, Peirce disavowed the wider social

FRED DAVIDSON is Associate Professor of English as an International Language at the University of Illinois at Urbana-Champaign, USA. His interests and publications are in the areas of language testing, research design and statistical data management for applied linguistics, and the history and philosophy of educational and psychological measurement. He is author (with Brian Lynch) of Testcraft: A teacher's guide to writing and using language test specifications *(Yale, 2002).*

GLENN FULCHER is Senior Lecturer in the School of Education at the University of Leicester, UK. His main interests lie in the field of language testing and the philosophy of educational assessment, including validity theory, construct operationalization, and task design. He also holds interest in research methodology and statistics in testing and applied linguistics research, and has worked in the fields of second language acquisition, discourse analysis, lexis, CALL, teaching and methodology. He is editor (with Cathie Elder, Melbourne University) of Language Testing *and author (with Fred Davidson) of* Language testing and assessment: An advanced resource book *(Routledge, 2007).*

Lang. Teach. 40, 231–241. doi:10.1017/S0261444807004351 Printed in the United Kingdom © 2007 Cambridge University Press

application of the original idea of 'pragmatism'. He felt himself to be a logician first and foremost, and hence he re-named his view 'pragmaticism' although he used both terms in later writing (see, inter alia, Peirce/Turrisi 1903/1997). Whether attention to effect is a matter of tight logical reasoning or of wider social impact, we believe that such attention is fundamental to designing any good language test – including tests shaped by the Common European Framework of Reference (CEFR).

There is a growing belief that the Framework is THE system that describes what language learning is really like and what levels levels learners really pass through, and the system to which language tests can really be linked. The CEFR is being reified (Fulcher 2004). Reification is 'the propensity to convert an abstract concept into a hard entity' (Gould 1996: 27), and it is a fallacy into which grand measurement enterprises often fall. The fallacy seems borne of a desire for harmony. In this paper, we contend that reification of the CBFR is fallacious only if its users see it as inflexible. If, however, the Framework is seen as a series of guidelines from which tests (and teaching materials) can be built to suit local contextualized needs, then there is no fallacy because the Framework has not been reified.

Harmonization has taken on (at least) three forms within the European Union. The first is a matter of political identity. The CEFR is a European solution, authored by the Council of Europe – not the European Union itself, we acknowledge – and yet we now see the familiar claims that non-European tests are not based on the same rich constructs as the CEFR. The EU is embracing the CEFR regardless of the fact that it – itself – was not creator. The second meaning of harmonization is about the use of a model like the CEFR in practical policy. Irrespective of the problems that applied linguists may have in defining levels and writing prose descriptors for proficiency scales, for a system run by people who require centralized control, it is an easy step from requiring a standard (of language) for a particular level – a TESTING concern – to requiring a level for a particular test use – a POLICY concern. The third meaning of harmonization is the use of the CEFR as a tool of recognition of a particular test or curriculum through linkage to the CEFR. The need to be recognized further drives the process of institutionalization and reification, ensuring that bureaucrats and educators across Europe eventually see the world in the same way, and rank-order students according to their worth, whichever language they are learning, or for whatever purpose they are learning it.

What is missing here is a clear commitment to effect. Harmonization is model-driven test development, without full consideration of the effect that such tests have on their users. In this paper, we present an effect-driven approach to building tests based on the CEFR. We first discuss the CEFR and certain putative weaknesses it presents to test developers. At first glance, these weaknesses appear to stem from its vagueness and lack of focus, but in fact they actually allow latitude in local interpretation of the Framework, and that may be a good thing (as we have argued elsewhere, see Davidson & Fulcher 2006). We then illustrate a dynamic model of evolutionary test development through which CEFR-influenced test developers can build a language exam – provided that they keep the test's effect in mind. The vague and broad nature of the CEFR can be an advantage, for it allows the test developer to direct CEFR-inspired testing to a desired effect.

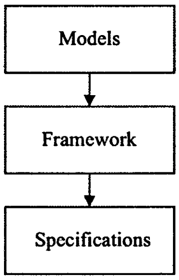


Figure 1 Levels of Architectural Documentation.

2. Testing architecture: models, frameworks and specifications

In any language education or use situation, at a most basic level, we need to make decisions about whether person 'y' has the language and communication abilities necessary to undertake activity 'z'. Constructs are selected from models, embodied in frameworks that relate constructs to contexts, and operationalized in test specifications that articulate purpose in practice (Chalhoub-Deville 1997; Fulcher 2004; Fulcher & Davidson 2007).

There are three levels of documentation in test architecture (Fulcher & Davidson 2007), represented in figure 1. Models, like those of Canale & Swain (1980), Canale (1983), Bachman (1990), Bachman & Palmer (1996) and Celce-Murcia, Dörnyei & Thurrell (1995) are the most general descriptions of communicative competence, or what it means to know and have the capacity to use a language. Other models, like the CEFR or the Canadian Language Benchmarks (Pawlikowska-Smith 2000) try to be encyclopedic and all-encompassing in a description of language abilities.

Our reading of the CEFR shows it to be non-purposive: it does not appear to discuss test effect. It does not detail – and perhaps this was an intent – particular contexts in which it would be used, and so lacks the necessary detail on which to build test specifications. This decontextualization has led to charges that the CEFR is inconsistent (and strictly speaking, it is), and thus that it is a weak tool on which to build tests (and strictly speaking, it is also, but only because it is not – itself – a bank of specifications).

3. Lack of specification for reading and listening in the CEFR

Criticism of the CEFR is not new, and it includes various forms of empirical enquiry. Alderson et al. (2006) report on a project to design reading and listening tests based on CEFR levels. This was done on the assumption that '. . . the CEFR in its current form may not provide sufficient theoretical and practical guidance to enable test specifications to be drawn up for each level of the CEFR' (ibid. 5).

The study identified areas in which the descriptors in the CEFR were inconsistent, where there were problems with understanding terminology (such as whether certain words were synonymous), or where terms lacked definition. It also found that there were gaps in the information needed for construct definition or the design of test specifications. The researchers were forced to create what they call a 'frame' which allowed identification of these problems, leading to the construction of a 'grid' that corrected the CEFR and filled in the gaps to describe TEXTS and ITEMS. In the next phase of research the grid was used by expert judges to match items and test specifications from existing tests to the grid, and hence to a CEFR level. Even with a finer-grained instrument, agreement between expert judges on matching tasks to CEFR levels only ranged from .49 to .78. The researchers found no significant association between text characteristics and CEFR level with the exception of vocabulary, and it proved impossible to distinguish between test specifications in terms of the grid or CEFR levels. These results are not surprising. Because the CEFR is so vast, it cannot detail purposive action about particular testing contexts. The judges' dilemma was this: they were trying to match things across a rather broad un-articulated gap.

4. Lack of specification for speaking in the CEFR: the case of service encounters

Another CEFR weakness can be seen in a particularly salient area of spoken language use: service encounters. We will explore this putative weakness in somewhat greater depth.

Service encounters are generally acknowledged to be of critical importance to language learners to 'get things done' in interaction with native and non-native speakers where a given target language is the common means of communication (McCarthy & Carter 1994: 24–27). Further, there is now significant evidence that failure to successfully negotiate service transactions in intercultural communication can lead to misunderstanding and hostility (Ryoo 2005: 81). Even in academic settings, it is now being recognized that service encounters on the campus should be tested as part of academic tests of English (Biber et al. 2002), and this has been embedded in the new TOEFL iBT listening as 'conversations in an academic setting' between students and the 'registrar, housing director, librarian, bookstore employee, departmental secretary, etc.' (ETS 2005: 10).

4.1 Service encounters and the CEFR: domain of use

The CEFR (Council of Europe 2001: 34) states that most descriptors that relate to transactions in the public domain occur at the B1 level, which is tied to the qualitative descriptions found in the Threshold Level document (van Ek & Trim 1991). These include the abilities to:

> make simple transactions in shops, post offices or banks; get simple information about travel; use public transport: buses, trains, and taxis, ask for basic information, ask and give directions, and buy tickets; ask for and provide everyday goods and services.

The CEFR provides the contexts in which these things may be undertaken within the public domain, as shown in figure 2.

The first point to note in this description is that 'goods' and 'services' are grouped together. No distinction is drawn between buying fish and chips or purchasing a new car, which are qualitatively different communicative transactions (Ylänne-McEwen 2004: 518f.), as these constructed extracts from McCarthy & Carter (1994: 63) demonstrate:

Customer: I'm interested in looking at a piece of cod, please.
Server: Yes madam, would you like to come and sit down.

Customer: A Ford Escort 1.6L please, blue.
Server: Right, £10,760, please.

Still less is there any attempt to distinguish between purchasing goods, and obtaining services that are 'less tangible' (Coupland 1983: 464f.), and where the nature of the transactional exchange is less restricted and more variable. Rather, what we have in the CEFR is an unstructured, incomprehensive list of things that language users might want to get done in a range of contexts. Any, or none, of these might be relevant to a particular testing situation. Nor is there any suggestion that particular task types might be linked to these situations. Readers of the CEFR are merely asked to consider what task types might be relevant (Council of Europe 2001: 54):

> Users of the Framework may wish to consider and where appropriate state:
>
> - the communicative tasks in the personal, public, occupational and/or educational domains that the learner will need/be equipped/be required to tackle;

Domain	Public
Location	Public spaces: street, square, park. Public transport. Shops, (super)markets. Hospitals, surgeries, clinics. Sports stadia, fields, halls. Theatre, cinema, entertainment. Restaurant, pub, hotel. Places of worship.
Institutions	Public authorities. Political bodies. The law. Public Health. Services, clubs. Societies. Political parties. Denominations.
Persons	Members of the public. Officials. Shop personnel. Police, army security. Drivers, conductors. Passengers. Players, fans, spectators. Actors, audiences. Waiters, barpersons. Receptionists. Priests, congregation.
Objects	Money, purse, wallet. Forms. Goods. Weapons. Rucksacks. Cases, grips. Balls. Programmes. Meals, drinks, snacks. Passports, licences.
Events	Incidents. Accidents, illness. Public meetings. Law-suits, court trials. Rag-days, fines, arrests. Matches, contests. Performances. Weddings, funerals.
Operations	Buying and obtaining public services. Using medical services. Journeys by road/rail/ship/air. Public entertainment and leisure activities. Religious Services.
Texts	Public announcements and notices. Labels and packaging. Leaflets, graffiti. Tickets, timetables. Notices, regulations. Programmes. Contracts. Menus. Sacred texts, sermons, hymns.

Figure 2 Interactional communication with the public domain (Council of Europe 2001: 48f.).

- the assessment of learner needs on which the choice of tasks is based.

It is therefore not at all surprising that we have no indication of PERFORMANCE CONDITIONS that might apply to any of these situations. Performance conditions are 'specific conditions that give us the purpose of communication, setting/place, audience, topic, time constraints, length of task, assistance allowed, etc.' (Pawlikowska-Smith 2000: ix). Such conditions need to be outlined, for they are an important step to detailing desired test effect, and so again, we see the need for detailed test specifications.

4.2 Service encounters and the CEFR: quality of performance

Likewise, the CEFR provides little guidance about the quality of test task performance. Transactions are classed as one example of spoken interaction by the CEFR (Council of Europe 2001: 73) and have an attached 'illustrative scale' for ranking performance (ibid. 80). The descriptors used at each scale level use 'can-do' statements to define the levels. This is illustrated in figures 3 and 4.

When analysing this scale we face a number of problems. The first is that some of the descriptors refer to specific situations, while others do not. Level B2, for example, refers to getting a traffic (parking?) ticket, damaging property, and dealing with being blamed for an accident. Other levels give an indication of types of situations, but are far less specific.

Secondly, where a specific situation is mentioned, it is not necessarily referred to in other descriptors. Dealing with travel agents is specifically mentioned in Level B1, and although travel is referred to at other levels, this particular context is not duplicated. It is difficult to know whether this is something that a learner can suddenly 'do' at level B1, or whether this is just a stage in acquiring the communicative language skills for this context.

B2 (and above):	Can cope Can negotiate a solution Can outline a case Can state clearly Can explain
B1	Can deal with Can cope Can make a complaint
A2	Can deal with Can get information Can ask (e.g. price) Can get Can give Can state what is wanted Can order
A1	Can order Can handle (numerical information)

Figure 3 Sample CEFR 'can-do' descriptors for transactions.

Thirdly, the descriptors seem to mix participant roles within a single level. At A2, for example, the leaner can 'ask for and provide' goods and services, implying that they would be able to function as a shopkeeper or travel agent, as well as a procurer of goods and services. Would this imply that at level B2 the learner could take on the role of an agent in a citizen's advice bureau and explain to a client how to seek compensation, as well as ask for compensation?

Fourthly, the distinction between levels is not at all clear, often referring to a vague notion of 'complexity' of the transaction. For example, at level B1 learners can deal with 'most' transactions and

	TRANSACTIONS TO OBTAIN GOODS AND SERVICES
C2	As B2
C1	As B2
B2	*Can cope linguistically to negotiate a solution to a dispute like an undeserved traffic ticket, financial responsibility for damage in a flat, for blame regarding an accident. Can outline a case for compensation, using persuasive language to demand satisfaction and state clearly the limits to any concession he/she is prepared to make.*
	Can explain a problem which has arisen and make it clear that the provider of the service/customer must make a concession.
B1	*Can deal with most transactions likely to arise whilst traveling, arranging travel or accommodation, or dealing with authorities during a foreign visit. Can cope with less routine situations in shops, post offices, banks, e.g. returning an unsatisfactory purchase. Can make a complaint. Can deal with most situations likely to arise when making travel arrangements through an agent or when actually travelling, e.g., asking passenger where to get off for an unfamiliar destination.*
A2	*Can deal with common aspects of everyday living such as travel, lodgings, eating, and shopping. Can get all the information needed from a tourist office, as long as it is of a straight-forward, nonspecialised nature.*
	Can ask for and provide everyday goods and services. *Can get simple information about travel, use public transport: buses, trains, and taxis, and give directions, and buy tickets.* *Can ask about things and make simple transactions in shops, post offices, or banks.* *Can give and receive information about quantities, numbers, prices, etc.* *Can make simple purchases by stating what is wanted and asking the price.* *Can order a meal.*
A1	*Can ask people for things and give people things.* *Can handle numbers, quantities, cost, and time.*

Figure 4 Illustrative scale for transactions.

situations, as well as 'less routine' situations. But there is no indication as to what kinds of 'less routine' situations a learner might not be able to deal with, and no definition of 'less', 'more' and 'most'. A2 is characterized by 'common', 'everyday', 'simple', and 'straightforward' transactions, but the reader is left to infer what these presumably 'more routine' transactions might be.

Despite these problems, the CEFR can still be used as a heuristic that allows for creative development of test specifications – and eventually tests. The can-do statements in the scale can trigger healthy localized debate and can even cause discovery. Only that way can we obtain the detail needed for further test development (Davidson & Fulcher 2006). We believe that such a journey begins with a realistic appraisal of the weaknesses, gaps, and vague scope of the CEFR in the first place.

4.3 Filling the gaps in the CEFR: the discourse structure of service encounters

In basic service encounters, it is possible to focus on the statements at a given level, and investigate what it would actually mean to test an ability to do these things within a meaningful communicative context. With reference to A1, 'Can ask people for things and give people things' and 'Can handle numbers, quantities, cost and time' Hasan (1985) provides an excellent starting point for fleshing out the CEFR illustrative scales to make test development possible.

4.4 Hasan's model of spoken language and its context

Hasan (1985) argued that a description of the field, tenor and mode (Halliday 1985: 12) of a text provides a definition of its CONTEXTUAL CONFIGURATION which allows us to link the realization of specific utterances to their social context. As Hasan argues, within specific contexts of language use, we can see the intimate relationship between language and context when we are able to describe text structure in terms of:

1. **What** elements **must** occur;
2. **What** elements **can** occur;
3. **Where must** they occur;
4. **Where can** they occur;
5. **How often** can they occur.

Hasan (1985: 56; bold type in the original)

In service encounters, the field of discourse is the institutionalized social activity of buying and selling, although this is much more complex and less institutionalized when services rather than products are involved (Ylänne-McEwen 2004). Tenor is defined by the participant roles. In service encounters the vendor and customer are involved in hierarchic dyadic interactions, in which the customer is the more powerful speaker. Social distance varies, however, depending upon the length of the relationship between the participants. Familiarity is therefore a key variable. The mode of discourse is speech, in which the language is largely ancillary to (or accompanying) an act of selling. Once again, however, in more complex (socially expanded) service encounters the language also serves to establish relationships.

Hasan (1985) analyses texts that conform to this contextual configuration, and identifies nine discourse elements that we might expect to see in service encounters: the greeting (G), the sale intention (SI), the sale request (SR), the sale compliance (SC), the sale enquiry (SE), the sale (S), the purchase (P), the purchase closure (PC), and the close or finis (F). SE, SR and SC are iterative and may be repeated during an interaction. Of the nine elements, SR, SC, S, P and SC are obligatory, and therefore define the genre of the text, thus linking the realization of generic structure potential to a particular situation. Other elements are optional. We also recognize that elements must occur in a certain sequence. Hasan (ibid. 64) reduces this to the formula:

$$[(\mathrm{G}).(\mathrm{SI})\hat{}]\,[(\overleftarrow{\mathrm{SE}}.)\,\{\mathrm{SR}\hat{}\mathrm{SC}\hat{}\overleftarrow{\}\hat{}\mathrm{S}}\hat{}]\mathrm{P}\hat{}\mathrm{PC}(\hat{}\mathrm{F})$$

In this formula round brackets indicate optional elements, and the caret indicates required sequence. A dot indicates that there is more than one option in the sequence, but this is restricted by the square brackets. Thus, (G) may come before or after (SI) if both are realized, but they must both come before any other element. An arrow indicates that an element may be iterative. The braces around SR and SC with an arrow indicate that if there is iteration both elements must occur within a single iteration.

The excerpt below from Hasan (1985: 59) illustrates some elements of this formula. If a test intends to measure language ability for service encounters, then one option available to the test designer would be to follow Hasan's formula. And any learner who completes a task by realizing the obligatory elements of a service encounter would have demonstrated basic discourse competence within this domain. This will therefore act as the first criterion in a judgment about whether a learner is capable of successfully engaging in service encounter interaction. For the test designer, the challenge is to design test tasks that elicit the evidence upon which such a judgment could be based.

SR = [Can I have ten oranges and a kilo of bananas please?]
SC = [Yes, anything else?
No thanks.]
S = [That'll be dollar forty.]
P = [Two dollars.]
PC = [Sixty, eighty, two dollars. Thank you.]'
Excerpt from Hasan (1985: 59, bracketing in the original)

We might add that the SE element with Hasan's model also encompasses a wide range of potential interactions, including those that establish and maintain relationships even in the most simple of service exchanges. These are normally recognizable as shifts in tenor as speakers temporarily share experiences or ideas as equals before completing a transaction (Ylänne-McEwen 2004), such as talking about the weather (Coupland & Ylänne-McEwen 2000). It is therefore possible, even at lower ability levels, to test the extent to which learners are able to select or use appropriate speech acts to establish rapport, as we will show below.

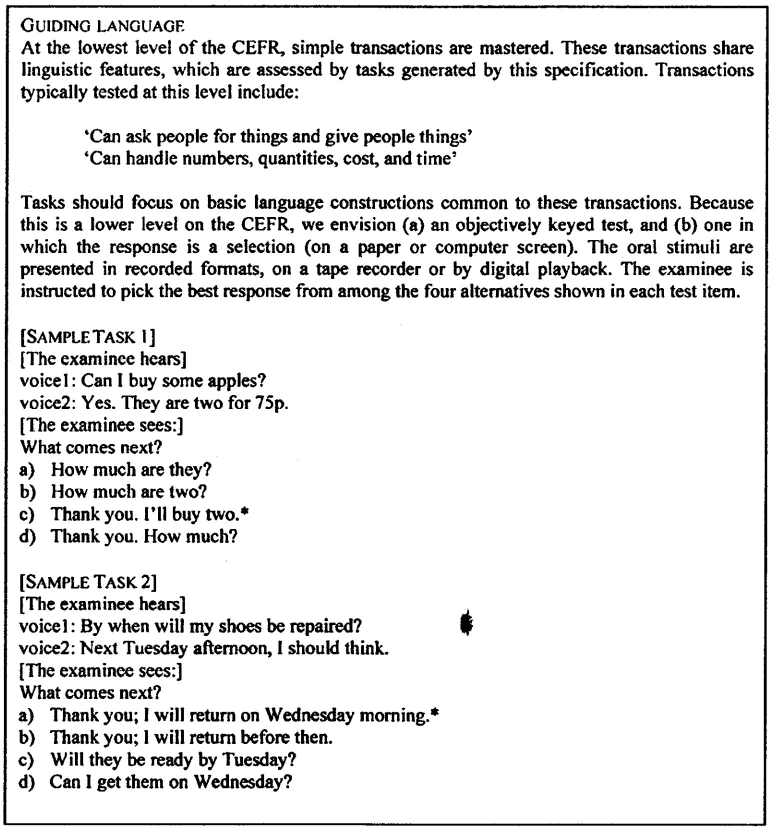
5. Dynamism in test development

Hasan's model, above, suggests that test designers should go through various decisions as a test task is developed. The CEFR does not take the test designer through these types of design decisions; instead, the test designer has to do his or her linguistic and communicative homework – the test developer has to study sources such as Hasan. We see nothing in the CEFR that prohibits a test developer from doing the study and reasoning we have done here, and more importantly, the CEFR does not constrain how that homework may – or may not – shape the intended effect of the test. The test developer is free to build an effect-driven test and to fairly claim anchorage to the CEFR (Davidson & Fulcher 2006).

The study and homework on service encounters outlined here has not (quite) yet yielded a test. Some additional work is needed to move to an operational test. This last step is a test specification, which is a generative document that details how to produce actual test tasks. And it is in the specification that effect must be most clearly articulated. In order to build an effect-driven test specification, we need a dynamic, evolutionary, critical dialogue amongst a team of test developers. We would like to illustrate this dynamism with a very simple example: a task built on the A1 descriptor.

5.1 Dynamic effect-driven spec development: an example

A minimalist test specification has two elements: sample tasks and guiding language. The samples are actual questions or statements or prompts as intended



GUIDING LANGUAGE
At the lowest level of the CEFR, simple transactions are mastered. These transactions share linguistic features, which are assessed by tasks generated by this specification. Transactions typically tested at this level include:

> 'Can ask people for things and give people things'
> 'Can handle numbers, quantities, cost, and time'

Tasks should focus on basic language constructions common to these transactions. Because this is a lower level on the CEFR, we envision (a) an objectively keyed test, and (b) one in which the response is a selection (on a paper or computer screen). The oral stimuli are presented in recorded formats, on a tape recorder or by digital playback. The examinee is instructed to pick the best response from among the four alternatives shown in each test item.

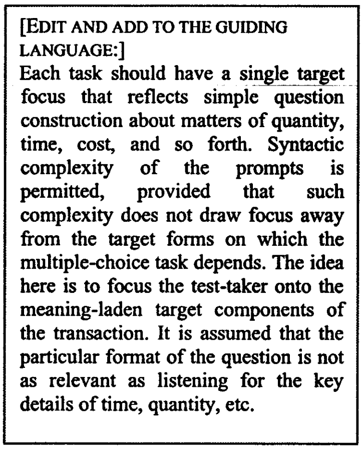
[SAMPLE TASK 1]
[The examinee hears]
voice1: Can I buy some apples?
voice2: Yes. They are two for 75p.
[The examinee sees:]
What comes next?
a) How much are they?
b) How much are two?
c) Thank you. I'll buy two.*
d) Thank you. How much?

[SAMPLE TASK 2]
[The examinee hears]
voice1: By when will my shoes be repaired?
voice2: Next Tuesday afternoon, I should think.
[The examinee sees:]
What comes next?
a) Thank you; I will return on Wednesday morning.*
b) Thank you; I will return before then.
c) Will they be ready by Tuesday?
d) Can I get them on Wednesday?

Figure 5 Version 0.10 (A1 service encounter spec).
Note: In the sample items, an asterisk (*) indicates the intended correct choice, or 'key'.

for the operational test. The guiding language is exposition and justification to help test writers build equivalent test versions. For more discussion on test specifications, see Fulcher & Davidson (2007, especially chapters A4, B4, and C4) and Davidson & Lynch (2002), and the references that they cite.

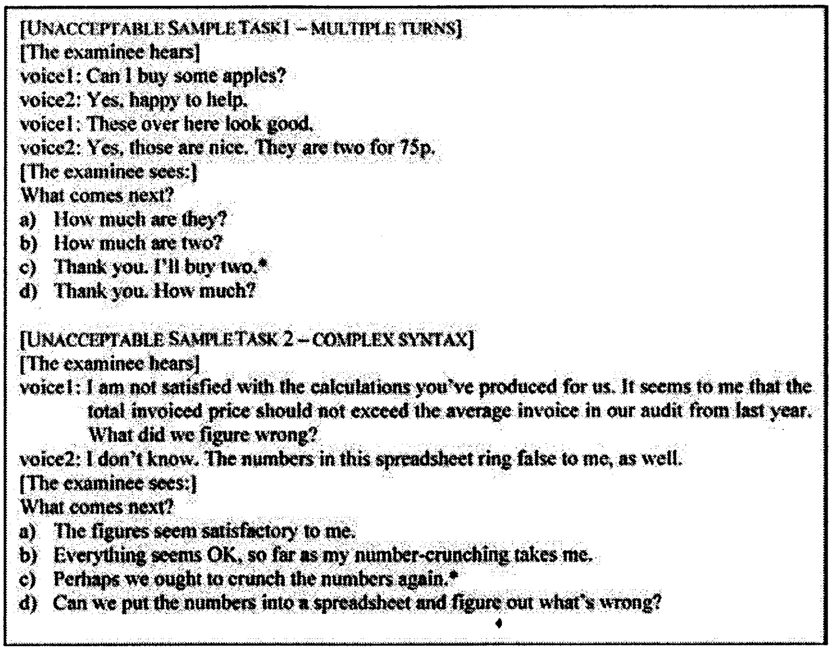
Let us envision a very simple service encounter in a target language, such as what might happen at an open air market. A specification for this situation might look like that shown in figure 5. We see immediately that the Sample Task 2 item is more complex than the first. Under effect-driven testing, we can ask ourselves: would such syntactic complexity actually happen out in a real open-air market? Clearly, the answer is 'yes', and so: how do we make such a task still fit at A1? One way is to focus the task on a specific target. We could accomplish this by editing the guiding language, as shown in figure 6.



[EDIT AND ADD TO THE GUIDING LANGUAGE:]
Each task should have a single target focus that reflects simple question construction about matters of quantity, time, cost, and so forth. Syntactic complexity of the prompts is permitted, provided that such complexity does not draw focus away from the target forms on which the multiple-choice task depends. The idea here is to focus the test-taker onto the meaning-laden target components of the transaction. It is assumed that the particular format of the question is not as relevant as listening for the key details of time, quantity, etc.

Figure 6 Guiding language: first edit (version 0.15).

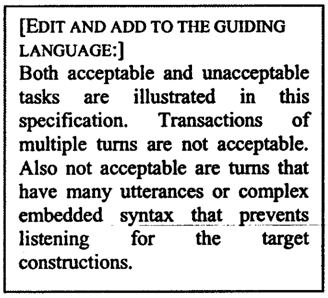
Note that we are starting to resolve the vastness of the two claims in the CEFR at A1 which we quoted in our specification: 'Can ask people for things and give people things' and 'Can handle numbers, quantities, cost, and time'. By focusing on a target element, we are operationally defining such vague phrases. Let us say this takes us to Version 0.15 of our specification. What is next? Effect-driven development of test specifications is often done in a group of colleagues, and we have noticed a natural human tendency to suggest negative cases: to suggest test tasks that would not work. Suppose we imagine sample task items which our test development group does not like, such as those shown in figure 7. Our group feels that the first is not good because it includes multiple turns. The group dislikes the second because it presents syntax that is entirely too complicated – it seems to violate the 'focus' principle articulated just above, which suggests the change to our specification also shown in figure 7.



[UNACCEPTABLE SAMPLE TASK1 – MULTIPLE TURNS]
[The examinee hears]
voice1: Can I buy some apples?
voice2: Yes, happy to help.
voice1: These over here look good.
voice2: Yes, those are nice. They are two for 75p.
[The examinee sees:]
What comes next?
a) How much are they?
b) How much are two?
c) Thank you. I'll buy two.*
d) Thank you. How much?

[UNACCEPTABLE SAMPLE TASK 2 – COMPLEX SYNTAX]
[The examinee hears]
voice1: I am not satisfied with the calculations you've produced for us. It seems to me that the total invoiced price should not exceed the average invoice in our audit from last year. What did we figure wrong?
voice2: I don't know. The numbers in this spreadsheet ring false to me, as well.
[The examinee sees:]
What comes next?
a) The figures seem satisfactory to me.
b) Everything seems OK, so far as my number-crunching takes me.
c) Perhaps we ought to crunch the numbers again.*
d) Can we put the numbers into a spreadsheet and figure out what's wrong?

Figure 7 Unacceptable sample tasks.
Note: In the sample items, an asterisk (*) indicates the intended correct choice, or 'key'.



[EDIT AND ADD TO THE GUIDING LANGUAGE:]
Both acceptable and unacceptable tasks are illustrated in this specification. Transactions of multiple turns are not acceptable. Also not acceptable are turns that have many utterances or complex embedded syntax that prevents listening for the target constructions.

Figure 8 Guiding language: second edit (version 0.20).

Test specifications often benefit from such 'negative examples', and it would be wise to alter the guiding language as shown in figure 8, so that users of the specification clearly understand that some samples are intended to illustrate what is not wanted. This is effect-driven testing, because the test development team is calling attention to something they do NOT wish to see. Note that the newly added guiding language specifies a means by which we can decide whether syntax is too complex – does it interfere with the focused listening for the target construction? This would take us to new guiding language and version 0.20, given in figure 8.

Finally, let us assume that some members of our test development team agree on a particular effect necessary for this test: it must value rapport and contribute to a belief that rapport is an essential element of service encounters. If the test did not value rapport, it might send out 'false positives' – test takers whose results appear strong but who, once they are in a real language-use situation, would not contribute to a positive affective climate of a service encounter. Some additional guiding language ensues in version 0.25, shown in figure 9, along with some additional samples – in this case, alternative distracters.

The entire contiguous Version 0.25 of our specification is presented in the Appendix to this paper.

5.2 Versioning in dynamic effect-driven test development

Our use of effect-driven test specifications has taught us to keep track of the various changes made over time. As a test development team builds its examinations, it should archive the previous versions of test specifications. We have adopted a versioning system to keep track of this in several projects, and we have also adopted the mindset that 'Version 1.0' is the operational test: we allow ourselves to debate, to dissent, to explore, to create, and to trial our tasks over time, all the while polishing our guiding language and sample tasks, and only launching the test when all parties agree that it is appropriate to do so.

Li (2006) illustrated the value of tracking test specifications across versions. She isolated several key design debates in the creation of an EFL test for the airline industry, and she showed how the team resolved each debate. Following Davidson & Lynch (2002: 9), she built 'audit trails' for the evolution of the test's specifications. Audit trails are an extremely useful source of evidence to argue a

[EDIT AND ADD TO THE GUIDING LANGUAGE:]
Distracters are permitted that test rapport. Consider the alternative version of Sample Task One. Note the change to (d) in which a somewhat more rude response is presented – while technically accurate in terms of focused listening, the more-rude choice (d) violates an expectation of politeness for the encounter, and it is therefore considered to be a wrong response.

[SAMPLE TASK 3 – AN ALTERNATIVE VERSION OF SAMPLE TASK 1 INTENDED TO ASSESS RAPPORT]
[The examinee hears]
voice1: Can I buy some apples?
voice2: Yes. They are two for 75p.
[The examinee sees:]
What comes next?
a) How much are they?
b) How much are two?
c) Thank you. I'll buy two.*
d) Gimme two.

Figure 9 Guiding language: third edit (and a new sample test item) (version 0.25).

test's validity. In our short, illustrative example above, we could audit the test developer's debates and show how several things have evolved (and, probably, will continue to evolve), notably the balancing act of syntactic complexity against targeted listening, and the inclusion of rapport.

6. Closing remarks

Our A1 service encounter specification is nowhere near an operational Version 1.0, and in fact, calling it 'Version 0.25' at this stage may be a bit optimistic. For one thing, we have not yet tried it, and our experience as test developers has taught us – often painfully – that nothing beats a tryout. No matter how secure test developers feel about their design decisions, and no matter how much they feel that their vision of effect has guided their work, once the test is piloted then new and surprising response data will shape the work in significant ways.

We heartily encourage readers of this article to take the specification beyond 0.25, massage it, alter it, reverse some of our decisions, and make the specification their own. There is much yet to be done, but at no stage in our development (nor, we think in that which our readers may pursue) does this specification deviate from the CEFR.

The best way to guide debate (as a specification evolves) is to ask this question: is our test evolving to the desired effect? The various changes to our sample task specification here are generally guided by a simple question: what will the test mean if we do (or do not do) a particular design decision? Over time, sample tasks and guiding language serve to record our consensus. At different locations and varying contexts, this effect-driven evolution will also differ. What works as a service-encounter test at one language school may (probably will) not work at another.

Our view of effect-driven testing may seem time-consuming. Peirce (undated: 4f.) spoke of: 'the process in which the mind goes over all the facts of the case, absorbs them, digests them, sleeps over them, assimilates them, dreams of them, and finally is prompted to deliver them in a form'. We trust the creative energy of the human mind – and we trust, in particular, the creative energy of a group of minds as they collaborate and debate, and agree on how to build a test so that its effect is what the group desires. We have observed this process many times, and it is remarkably swift and organic – far faster than the plodding discourse of this academic paper may make it out to be. It is also empowering. We encourage our readers to try effect-driven evolutionary development of test specifications as a way to anchor any test to the CEFR, or for that matter, to any external framework or model.

Appendix: Version 0.25 of the CEFR A1 Service Encounter Specification, in contiguous form

Note: In the sample task items, an asterisk (*) indicates the intended correct choice, or 'key'.

GUIDING LANGUAGE
At the lowest level of the CEFR, simple transactions are mastered. These transactions share linguistic features, which are assessed by tasks generated by this spec. Transactions typically tested at this level include:

'Can ask people for things and give people things'
'Can handle numbers, quantities, cost, and time'

Tasks should focus on basic language constructions common to these transactions. Because this is a lower level on the CEFR, we envision (a) an objectively keyed test, and (b) one in which the response is a selection (on a paper or computer screen). The oral stimuli are presented in recorded formats, on a tape recorder or by digital playback. The examinee is instructed to pick the best response from among the four alternatives shown in each test item.

Each task should have a single target focus that reflects simple question construction about matters of quantity, time, cost, and so forth. Syntactic complexity of the prompts is permitted, provided that such complexity does not draw focus away from the target forms on which the multiple-choice task depends. The idea here is to focus the test-taker onto the meaning-laden target components of the transaction. It is assumed that the particular format of the question is not as relevant as listening for the key details of time, quantity, etc.

Both acceptable and unacceptable tasks are illustrated in this spec. Transactions of multiple turns are not acceptable. Also not acceptable are turns that have many utterances or complex embedded syntax that prevents listening for the target constructions.

Distracters are permitted that test rapport. Consider Sample Task 3: the alternative version of Sample Task 1. Note the change to (d) in which a somewhat more rude response is presented – while technically accurate in terms of focused listening, the more-rude choice (d) violates an expectation of politeness for the encounter, and it is therefore considered to be a wrong response.

[Sample Task 1]
[The examinee hears]
voice1: Can I buy some apples?
voice2: Yes. They are two for 75p.
[The examinee sees:]
What comes next?
a) How much are they?
b) How much are two?
c) Thank you. I'll buy two.*
d) Thank you. How much?

[Sample Task 2]
[The examinee hears]
voice1: By when will my shoes be repaired?
voice2: Next Tuesday afternoon, I should think.
[The examinee sees:]
What comes next?
a) Thank you; I will return on Wednesday morning.*
b) Thank you; I will return before then.
c) Will they be ready by Tuesday?
d) Can I get them on Wednesday?

[Unacceptable Sample Task 1 – multiple turns]
[The examinee hears]
voice1: Can I buy some apples?
voice2: Yes, happy to help.
voice1: These over here look good.
voice2: Yes, those are nice. They are two for 75p.
[The examinee sees:]
What comes next?
a) How much are they?
b) How much are two?
c) Thank you. I'll buy two.*
d) Thank you. How much?

[Unacceptable Sample Task 2 – complex syntax]
The examinee hears:]
voice1: I am not satisfied with the calculations you've produced for us. It seems to me that the total invoiced price should not exceed the average invoice in our audit from last year. What did we figure wrong?
voice2: I don't know. The numbers in this spreadsheet ring false to me, as well.
[The examinee sees:]
What comes next?
a) The figures seem satisfactory to me.
b) Everything seems OK, so far as my number-crunching takes me.
c) Perhaps we ought to crunch the numbers again.*
d) Can we put the numbers into a spreadsheet and figure out what's wrong?

[Sample Task 3 – an alternative version of Sample Task 1 intended to assess rapport]
[The examinee hears]
voice1: Can I buy some apples?
voice2: Yes. They are two for 75p.
[The examinee sees:]
What comes next?
a) How much are they?
b) How much are two?
c) Thank you. I'll buy two.*
d) Gimme two.

Acknowledgements

The authors gratefully acknowledge the assistance of the editors and reviewers in crafting this paper.

References

Alderson, J. C., N. Figueras, H. Kuijper, G. Nold, S. Takala & C. Tardieu (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly* 3.1, 3–30.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. & A. S. Palmer (1996). *Language testing in practice*. Oxford: Oxford University Press.

Biber, D., S. Conrad, R. Reppen, P. Byrd & M. Helt (2002). Speaking and writing in the University: A multidimensional comparison. *TESOL Quarterly* 36.1, 9–48.

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (ed.) *Issues in language testing research*. Rowley, MA: Newbury House, 333–342.

Canale, M. & M. Swain (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1.1, 1–47.

Celce-Murcia, M., Z. Dörnyei & S. Thurrell (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics* 2, 5–35.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing* 14.1, 3–22.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching and assessment.* Cambridge: Cambridge University Press. <http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.>. Accessed 01 February 2007.

Coupland, N. (1983). Patterns of encounter management: Further arguments for discourse variables. *Language in Society* 12, 459–476.

Coupland, N. & V. Ylänne-McEwen (2000). Talk about the weather: Small talk, leisure talk, and the travel industry. In J. Coupland (ed.), *Small talk.* London: Longman, 163–182.

Davidson, F. & G. Fulcher (2006). Flexibility is proof of a good 'Framework'. *The Guardian Weekly* 175.22, 'Learning English', p. 5.

Davidson, F. & B. K. Lynch (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven & London: Yale University Press.

ETS [Educational Testing Service] (2005). *TOEFL iBT Tips: How to prepare for the next generation TOEFL test and communicate with comfort.* Princeton, NJ: ETS.

van Ek, J. A. & J. L. M. Trim (1991). *Threshold Level 1990.* Cambridge: Cambridge University Press.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly* 1.4, 253–266.

Fulcher, G. &, F. Davidson (2007). *Language testing and assessment.* London & New York: Routledge.

Gould, S. J. (1996). *The mismeasure of man.* London: Penguin.

Halliday, M. A. K. (1985). Context of situation. In Halliday & Hasan, 3–14.

Halliday, M. A. K. & R. Hasan (1985). *Language, context, and text: Aspects of language in a social-semiotic perspective.* Victoria, Australia: Deakin University Press.

Hasan, R. (1985). The structure of a text. In Halliday & Hasan, 52–69.

Li, J. (2006). *Introducing audit trails to the world of language testing.* M. A. thesis, University of Illinois.

McCarthy, M. & R. Carter (1994). *Language as discourse: Perspectives for language teaching.* London: Longman.

Pawlikowska-Smith, G. (2000). Canadian Language Benchmarks 2000. Ontario: Centre for Canadian Language Benchmarks.

Peirce, C. S. (undated). Lecture I of a planned course. Ms. 857: 4–5. <http://www.helsinki.fi/science/commens/terms/abduction.html> accessed 1/2/2007.

Peirce, C. S. (1878) How to make our ideas clear. In E. C. Moore (ed.), 1998, *The essential writings of Charles S. Peirce.* New York: Prometheus Books.

Peirce, C. S. & P. A. Turrisi (1903/1997). *Pragmatism as a principle and method of right thinking* (The 1903 Harvard Lectures on Pragmatism). Albany, NY: State University of New York Press.

Ryoo, H-K. (2005). Achieving friendly interactions: A study of service encounters between Korean shopkeepers and African-American customers. *Discourse and Society* 16.1, 79–105.

Ylänne-McEwen, V. (2004). Shifting alignment and negotiating sociality in travel agency discourse. *Discourse Studies* 6.4, 517–536.