

The ALTE Can Do Project and the role of measurement in constructing a proficiency framework

NEIL JONES, SENIOR RESEARCH AND VALIDATION CO-ORDINATOR

Through the Framework Project ALTE members have classified their examinations within a common system of levels, with the aim of promoting the transnational recognition of certification in Europe. Part of this effort, the Can Do Project (see *Research Notes 2* for an introduction), aims at providing a comprehensive description of what language users can typically do with the language at each level, in the various language skills and in a range of contexts. The Can Do Project has a dual purpose: to help end users to understand the meaning of exam certificates at particular levels, and to contribute to the development of the Framework itself by providing a cross-language frame of reference.

This article provides an update on progress, and also attempts to draw some conclusions from a phase of the work which is nearing completion: the calibration of the individual Can Do statements on the basis of empirical data from self-report questionnaires.

Calibration means establishing the precise difficulty of each statement, in relation to a single scale, so that the Can Do statements become a yardstick against which any learner or any language exam can be measured.

Structure of the Can Do scales

The Can Do scales consist currently of about 400 statements, organised into three general areas: *Social and Tourist*, *Work*, and *Study*. Each area is sub-divided into a number of more particular concerns, e.g. the Social and Tourist area has sections on *Shopping*, *Eating out*, *Accommodation* etc. Each of these includes up to three scales, for the skills of *Listening/Speaking*, *Reading* and *Writing*.

Each such scale includes statements covering a range of levels. Some scales cover only a part of the proficiency range, as of course there are many situations of use which require only basic proficiency to deal with successfully.

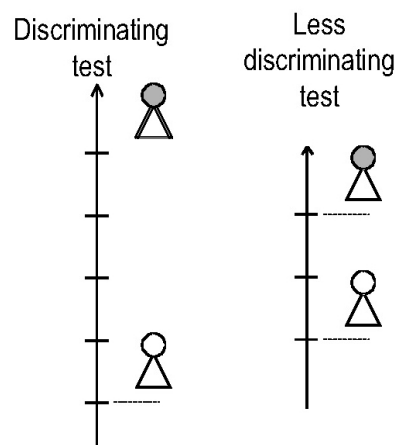
Measurement and judgement

The empirical work to make the Can Do scales into an instrument of measurement followed on from earlier work in which the Can Do statements were constructed and assigned to levels through a process of qualitative analysis, or judgement. The aim of the empirical study – to validate, improve and add precision to the scales – has been achieved, and yet an important conclusion of this work is that measurement and judgement are complementary and equally important aspects of constructing a proficiency scale.

Life provides enough illustrations of the shortcomings of judgement. The shortcomings of measurement are less apparent, and so they will be presented in what follows.

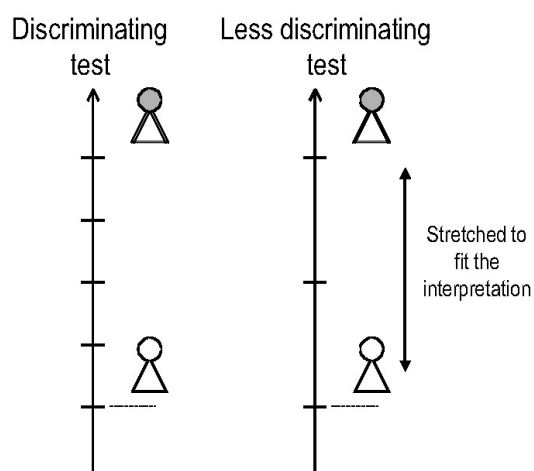
Let us begin by asking: how many levels of language proficiency are there? The ALTE Framework has five, or six if we include the embryo Breakthrough level. In this it agrees with the Council of Europe Common Framework, at least at one level of sub-division. Six is a reasonable number: large enough for putting learners into groups of practically comparable ability, and small enough to make distinctions of practical significance. But from a strictly measurement point of view, as many levels exist as a given measurement instrument is able to distinguish. A very short placement test might reliably distinguish just three levels, whereas a very long and time-consuming assessment might distinguish ten or more. In measurement terms, one unit on a proficiency scale means one reliably distinguishable shade of ability. Figure 1 shows two tests of varying discrimination: On one the two learners are separated by four units, on the other by only two.

Figure 1 : Two tests of varying discrimination



We might interpret this to mean that the pair on the right are closer in ability. But suppose we know that the pairs on the left and right are in fact the same people, who have taken both tests. The natural interpretation which we will probably wish to impose is that each learner has a single ability level, as shown in Figure 2. To fit the shorter scale to this interpretative framework we have to stretch it out.

Figure 2 : Two tests forced to fit an interpretative framework



Note that in this situation we have a basis for identifying the different discrimination of the tests, and bringing them into agreement with each other. But where our responses come from different groups of people (as with the Can Do data), it becomes much more difficult to distinguish between substantive differences in ability and differences in the precision of measurement of scales.

Why do tests (or Can Do scales) vary in their capacity to measure? Other things being equal, a longer test will always discriminate better. But other things are often not equal. It is a fact of assessment life that some things are more measurable than others. Thus scales that measure some aspects of language proficiency turn out shorter than others – the two figures above illustrate, for example, a situation actually observed in an oral interview procedure using separate scales for measuring grammatical accuracy and pronunciation. The pronunciation scale tends to be less discriminating. In this case it is human raters who are able to distinguish one aspect more finely than the other. In objective tests, it is the tendency of learners to respond uniformly to items – that is, to agree on what is difficult or easy, relative to their level – that makes for precise measurement.

The Can Do statements as a measurement instrument

It is central to the idea of using the Can Do statements to define a framework of levels that people will agree on what is difficult or easy. If users of a foreign language, irrespective of what that language is, or of their own language, or their educational or professional background, agree on how they rank tasks by difficulty, then there is a basis for using such statements to describe levels in a way that will support precise measurement, and generalise well across a variety of situations of use.

A proficiency framework defined briefly and vaguely will generalise to every situation but be of no practical use. Conversely, a framework defined in great detail is unlikely to generalise across all learners or situations of use. Clearly there are limits to generalizability, and the empirical work on the Can Do project has been useful in enabling us to explore these limits. In the end our aim remains to construct a useful, practical descriptive framework.

The Can Do data collected so far have shown a number of effects where groups of people disagree to a greater or lesser extent on what they find easy or difficult. The next sections review some of these effects, looking first at issues associated with the text of the statements themselves, and then at issues associated with particular groups of respondents.

Textual effects

The Can Do statements exist in 13 languages, and it is not surprising that some *translation* effects were found at an early stage. Statements which were unexpectedly hard or easy when presented in a particular language were studied, and in some cases this could be linked to the translation. Another predictable effect concerned the *orientation* of statements. Negatively worded statements (Cannot Do) performed badly for higher-level respondents, who found it unnatural to endorse low-level statements negatively worded. Such statements were re-worded positively or removed.

Other effects were unexpected but interpretable. For example, *detailed exemplification* of a task tended to make it more difficult than predicted.

All these effects could be corrected to the extent that they were problematic. Somewhat more difficult to deal with were systematic differences in discrimination. This issue arose during a study to equate the Can Do scales to the Council of Europe Common Framework. Several scales from the Framework document (Council of Europe 1996) were included in versions of the Can Do questionnaires and response data collected. Although there was close agreement between self-ratings on the two types of scale, the CE statements were found to define a significantly longer scale. This was because the CE scales consisted of six detailed composite statements, each epitomizing one level, whereas the Can Do scales in their deconstructed form consisted of a larger number of short atomic statements. A satisfactory equating of the two scales required a qualitative study – that is, the exercise of judgement.

Person effects

Effects found for groups of respondents are particularly interesting, as they indicate the limits to which Can Do statements generalise across learners and situations of use.

Demographic effects studied included *age*, *background* and *profession*.

Respondents included a limited number between the ages of 13 and 18. This age group tended to respond in ways which were inconsistent with the responses of older people. This is not surprising, as the Can Dos chiefly concern ability to operate in an adult world, and refer to tasks which children of school age would have had no experience of.

It was found that a person's occupation or professional status might affect their ability to use a foreign language in particular situations. Thus for example, employers found it significantly easier than employees at middle or junior level to deal with situations likely to arise in a hotel, restaurant, a bank, or while travelling. This is hardly surprising. More unexpected was that employers found it significantly easier to understand a photo-copier or fax machine. What this appears to reflect is a different understanding

of what the Can Do statement actually means (the employer probably has never tried to understand the instructions to fix a paper jam or change the toner cartridge).

Grouping respondents by *target language*, an interesting contrast was found between what learners of English and French find relatively hard or easy.

Whatever their overall level, learners of French seem likely to be relatively more confident of their receptive language skills (e.g. *CAN understand the general outline of a guided tour...*). Learners of English on the other hand are relatively more confident of their active communicative skills (e.g. *CAN participate in casual conversation over the phone with a known person on a variety of topics*).

It is interesting to consider what this might indicate about people's reasons for studying foreign languages or perhaps approaches to teaching different languages.

A noticeable effect concerned *ability* level. In self-report data respondents at lower proficiency levels tend systematically to overrate their ability. That is, they have a different understanding of "can do".

Constructing a language proficiency framework: quantitative and qualitative aspects

The empirical, statistically-based approach to validating the Can Do statements has been useful in calibrating the individual statements and constructing the individual scales. It has also been useful precisely because it identifies issues with how people understand and use assessment scales. We have found evidence of a range of group effects such as age, proficiency level, or area of language use, that may affect understanding of a scale and of the meaning of Can Do level descriptors. Thus there are scales where there is good agreement as to what is hard or easy, and scales where agreement is less. Consequently, (as explained above) some scales measure more precisely than others.

Some exercise of judgement becomes necessary in order to impose a single frame of reference on the different scales. The approach followed was based firstly on a close analysis of the text of each statement, in order to identify tasks which are very similar in different areas of use (Social and Tourist, Work and Study). These were posited to be of similar difficulty. Secondly, reference was made to the ALTE levels originally assigned to statements (restricting attention to those statements which had not been edited during the textual revision). The correlation between original and empirically found level was high, and so it could be assumed that overall these assigned levels could be used to anchor the scales to each other. From these two sets of observations, a separate linear transformation was found (that is, a formula for "stretching" the scale) for each language skill within each area of use. After applying these the textual analysis was repeated. Some apparent anomalies remained in several of the scales from the Study area of use, and these were individually rescaled to bring them into line.

A subsequent step has been to select from the individual statements and construct Can Do scales consisting of composite level statements. Statements were selected both for their content and for their statistical properties. As far as possible statements about which respondents disagreed were excluded. These composite statements are used in the computer-based Can Do self-assessment tool, and will also be exploited in validation activities currently being planned.

Reference:

Council of Europe (1996): *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference*. CC-LANG (95) 5 rev IV, Strasbourg, Council of Europe.

See also Jones (2001): Appendix D – ALTE Can Do Statements, in Council of Europe (2001): *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.