CHAPTER 9

RATING SCALE DEVELOPMENT: A MULTI-STAGE EXPLORATORY SEQUENTIAL DESIGN

EVELINA GALACZI AND NAHAL KHABBAZBASHI

CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

INTRODUCTION

The project chosen to showcase the application of the exploratory sequential design in second/ foreign (L2) language assessment comes from the context of rating scale development and focuses on the development of a set of scales for a suite of high-stakes L2 speaking tests. The assessment of speaking requires assigning scores to a speech sample in a systematic fashion by focusing on explicitly defined criteria which describe different levels of performance (Ginther 2013). Rating scales are the instruments used in this evaluation process, and they can be either holistic (i.e. providing a global overall assessment) or analytic (i.e. providing an independent evaluations for a number of assessment criteria, e.g. Grammar, Vocabulary, Organisation, etc.). The discussion in this chapter is framed within the context of rating scales in speaking assessment. However, it is worth noting that the principles espoused, stages employed and decisions taken during the development process have wider applicability to performance assessment in general.

We will start with a brief discussion of scale development and hope to illustrate the natural fit between the essence of scale development and the methodological possibilities offered by mixed methods research and the exploratory sequential design. We will then describe the studies which were undertaken at each stage of the scale development project and highlight key elements which shaped the development process. We will discuss their value and potential in scale development and the steps taken to ensure methodological rigour. Our discussion will also illustrate how the findings and insights from the different stages were used in an integrated and additive manner, which is a key consideration in mixed methods research. In the interest of space and clarity of focus, our discussion will be mostly methodological; a more detailed account of the project and its findings can be found in Galaczi, ffrench, Hubbard and Green (2011).

SCALE DEVELOPMENT AND A MIXED METHODS APPROACH

With some phenomena, such as the assessment of L2 proficiency, where decisions made are based on a complex web of interaction between a multitude of factors, such as, for example, the tasks, the raters, the interviewer, the background characteristics of the interlocutors (McNamara 1996), the gathering of data from a wide range of perspectives is instrumental in supporting the validity and trustworthiness of the decisions taken. Such is the case with assessment scales, which are the product of a complex process of reducing learner performance into finite and concrete categories, which capture the underlying construct.

Two key approaches have typically supported scale development over the past half a century: one is the expert-judgement approach (aptly termed the 'armchair approach' by Fulcher 2003), which is based solely on the expertise and intuition of experts who may be working collectively in committee or as individuals; the other approach is empirically informed, as shown in the work of Upshur and Turner (1995), Fulcher (1996), Knoch (2009), to name but a few, and can comprise quantitative or qualitative data or both. For example, Fulcher (1996) used discourse analysis and multiple regression in his analysis of fluency features in learner speech and in the generation of fluency scale descriptors. While rating scales of the past used to be based entirely on expert judgement, the empirically-informed approach to scale construction and validation has now become the norm in the L2 assessment community. A scale development approach grounded both in empirical data and expert judgement is also reflected in the procedures recommended in the influential Common European Framework of Reference (CEFR, Council of Europe 2001:205), which advocates the complementary use of both qualitative and quantitative methodologies in scale construction and the integration of findings. It advocates, in essence, a mixed methods approach.

The building of an empirical basis for the development of a set of assessment scales for large-scale high-stakes examinations is a complex undertaking, which requires a multi-dimensional approach in order to produce an instrument of the necessary rigour and validity. A mixed methods approach and its potential to address the task at hand from a range of different complementary perspectives is therefore a suitable methodological choice which also allows for a comprehensive understanding of a phenomenon by building on the strengths of the independent methods and minimizing their respective limitations (Sale, Lohfeld and Brazil 2002:50). As Greene, Caracelli and Graham (1989:256) note,

> all methods have inherent biases and limitations, so use of only one method to assess a given phenomenon will inevitably yield biased and limited results. However, when two or more methods that have offsetting biases are used to assess a given phenomenon, and the results of these methods converge or corroborate one another, then the validity of inquiry findings is enhanced.

Mixed methods research, however, is not just about convergence of findings. The use of data and empirical insights from a range of methods is well suited to addressing cases of divergence, since the different methods employed could potentially shed light on possible reasons for divergences and/or could inform further empirical investigations. In addition, a mixed methods approach can pinpoint cases of divergence which would have otherwise gone unnoticed (e.g. in the case of outliers in quantitative research which may just be removed from analysis, whereas a qualitative outlook can look in depth into those specific cases).

The potential complementary relationship of methodologies in a mixed methods study lays the ground for a symbiotic relationship where, as Aristotle noted many centuries ago, the whole is greater than the sum of its parts. Similar to the synergy of musical parts in the creation of a song, for example, where the overall effect is more dramatic than the effect of each of the parts played individually, the complementary integration of methods in a mixed methods design provides richer insights into the project at hand.

The premise of compatibility and complementarity in mixed methods research has not always gone unquestioned (as discussed in Chapters 3 and 4). Substantial debate has taken place in the mixed methods literature about the compatibility of quantitative and qualitative methodologies, aptly referred to as the 'paradigm wars' by Tashakkori and Teddlie (1998:3). The pragmatist approach, which reconciles the use of two different paradigms and advocates the eclectic use of any range of philosophies, methodologies and tools suitable to the project at hand, has now been accepted as the underlying theoretical paradigm for mixed methods research (Creswell 2014, see also Chapter 3). It provides researchers with the flexibility to opt for any approaches which would provide a usable and useful solution to an issue or task (Creswell, Plano Clark, Gutmann and Hanson 2003). A pragmatist approach conceptualises qualitative and quantitative viewpoints not as 'competing dualisms' (Onwuegbuzie and Johnson 2006:59) situated in dichotomous opposition, but as potentially complementary elements on a continuum. Such a pragmatist orientation finds a match with the purposes of large-scale 'real-world' projects, of which scale development is one example, that are driven by the very practical and powerful imperative to use all available approaches which work. There is, therefore, a natural fitness for purpose between mixed methods and scale development.

SCALE DEVELOPMENT AND THE EXPLORATORY SEQUENTIAL DESIGN

Empirically-based scale development is by nature a cyclical iterative process which moves in a spiral-like manner, with subsequent stages drawing on the findings of previous ones and leading to more in-depth and meaningful insights. Scale development is also a reductionist process which aims to distil the richness and complexity of learner language to a finite set of assessment categories (e.g. Grammar, Vocabulary, Interactive ability) and discrete competency statements within each category and ordered from low to high corresponding to a single score or set of scores along the proficiency continuum.

One possible starting point for reducing the complexity of language proficiency into a finite and concrete set of scale categories, descriptors and scores could be an empirical investigation of the views of key stakeholders, e.g. teachers, who would draw on the wealth of their experience. Such an approach was seen in the development of the CEFR (North 2000), where teachers' views about language proficiency at different levels were used in the generation of a bank of draft descriptors characterising different aspects of language proficiency. Another starting point could be learner performance data, as shown by Fulcher (1996) in his development of a Fluency scale. Whatever the approach adopted in scale development at the initial stage, it is typically characterised by an inductive exploration approach. Such an initial exploration stage which allows the complexity of language performance to be captured in the form of themes, codes, illustrative language or distinguishing linguistic features is in line with the constructivist paradigm, which espouses variable and multiple meanings and realities, acknowledges the complexity of views and data and is characterised by inductive interpretative generation of meaning. Findings from this stage can then be reduced to general categories and descriptors, which are embedded in the assessment scales. This subsequent stage is in line with the reductionist orientation of the post-positivist paradigm, where the intent is typically to reduce ideas into a final product/number/scale/survey, which captures the "truth". The nature of scale development, i.e. a progression from an inductive exploratory stage to a deductive measurement stage, finds a suitable match in the exploratory sequential design, which comprises an initial inductive qualitative discovery stage, in tune with the constructivist paradigm, followed by a reductionist, generalizable quantitative stage embedded within a post-positivist paradigm.

COMPLEX EXPLORATORY SEQUENTIAL DESIGNS

As noted earlier in this discussion (and also in Chapters 3 and 4), the exploratory sequential design would typically occur in three stages: qualitative data collection/analysis followed by quantitative data collection/analysis (QUAL➔QUAN), with the second stage building on the findings of the first stage, and leading to a third stage which might involve the development of a measurement instrument or a procedure. This basic QUAL➔ QUAN progression, with stages following from one another in a linear fashion, is well suited to smaller-scale empirical endeavours, where the first stage would be exploratory, the second stage would result in instrument development and the third would involve the large-scale quantitative administration of the instrument.

In more complex and large-scale projects, such as the one discussed in this chapter, the basic exploratory sequential design may need modifications to ensure fitness-for-purpose with the complex nature of the project. As such, the design could take an iterative multi-stage character, where the basic 'QUAL→QUAN' unit is iteratively repeated with some variation. Creswell terms this approach 'multiphase or multistage mixed methods' (2014:228). A staged approach of this kind carries the overall features of the exploratory sequential design (i.e. very first stage is qualitative and very last stage is quantitative), but there may be nesting of other designs at some of the in-between stages. It is an advanced methods design where each stage stands on its own but the stages are connected sequentially. As discussed by Ziegler and Kang in Chapter 4, this is a way to address a large-scale project 'by breaking it into multiple smaller, semiautonomous, studies' and is the underlying design in the scale development project discussed here.

We now turn our attention to more detailed descriptions of each stage of the study. As we do that, we hope to provide both a descriptive account of what was done and a discussion drawing explicit links between the content of the project and the methodology adopted. We start with a brief discussion of the considerations, which were at the heart of the planning stage.

SNAPSHOT OF THE STUDY

Research goal	• Theoretically and empirically-supported development of a set of assessment scales for second/foreign language speaking tests
Design	• Advanced multistage scale development mixed methods design
Stages	• QUAL → Convergent QUAN & QUAL → QUAN Stage I Stage II Stage III
Qualitative data collection and data analysis	• Stage I: Conversation Analysis of learner speech • Stage I: Thematic analysis of examiner extended survey comments • Stage II: Verbal Protocol Study of examiner comments while using draft scales
Quantitative data collection and data analysis	• Stage II: Multi-Facet Rasch Analysis of descriptor performance • Stage III: Multi-Facet Rasch Analysis of rater, scale and assessment criteria performance
Methods level strategies*	• Connecting, Building, Merging
Interpretation level strategies*	• Data preparation: Reduction, Transformation • Data Analysis: Comparison • Data integration: Synthesis of findings at each stage through a matrix containing all flagged up issues/scale descriptors based on all recommendations from the QUAL and QUAN stages
Mixed methods value added	• Meta-inferences based on a range of qualitative and quantitative methodologies and a staged, additive approach
Conclusions	• Operational functioning scale developed

NOTE: SEE CHAPTER 4 FOR A MORE DETAILED EXPLANATION OF THESE MIXED METHODS RESEARCH STRATEGIES.

PLANNING STAGE: CONSIDERING COMPLEMENTARITY OF METHODS

During the planning stage of the project methodological tools needed to be critically considered, with careful attention given to their underlying assumptions, strengths and shortcomings. As Sandelowski (2003:329) notes,

> most studies in social and behavioural sciences … entail the use of more than one of something (e.g. investigators, participants, sites) for data collection. The mere use of more than one of some research entity in a study does not constitute a mixed methods study.

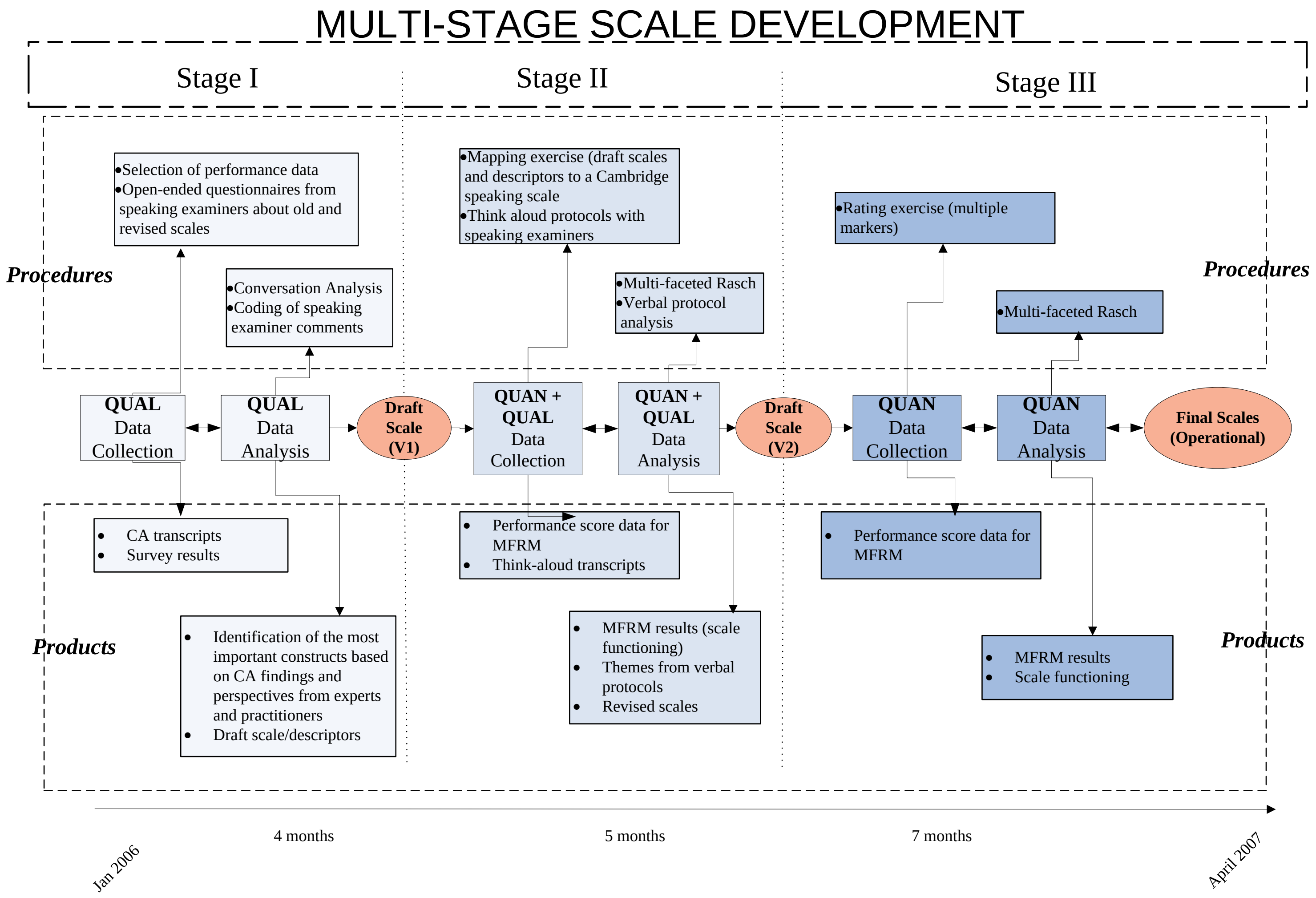
The methods selected in the project, therefore, needed to display 'complementarity' (Greene, Caracelli and Graham 1989:259) with the aim of increasing the meaningfulness and validity of the results. They were chosen not just as tools with different technical qualities, but as methodologies grounded in different paradigms which can provide data and interpretations which tap into different and equally important aspects of reality, and which can optimise strengths and minimise limitations in a symbiotic relationship.

As such, the development team decided that during Stage I different types of data would be gathered from two key stakeholder groups: speaking examiners, who would provide individual perspectives about various aspects of the scales, and speech performance data from video-recorded speaking tests, which would provide learner data. Speaking examiners - the primary users of assessment scales in test conditions - have valuable experience, which can usefully feed into decisions about assessment scales. Learner speech data, on the other hand, provides a look 'inside' speaking tests (van Lier 1989) and insights about what learners actually say and what features of their speech can be and need to be captured in scales and descriptors of their performance. During Stage II, the methods were selected again with the aim of complementarity in providing insights from different angles: the use of performance descriptors as discrete entities which are mapped to levels on a scale in one study, and in another study, the use of the descriptors, this time in their totality in the scales, but with a focus primarily on the examiner experience and not on the marks awarded. Stage III, which aimed at finalising the scales, was planned to include quantitative investigations which would allow for generalizable findings.

In addition to the complementarity of the selected methods, the robustness of the qualitative and quantitative data and analysis was addressed at the planning stage. This is a key consideration and challenge in mixed method research, since methodological criteria of quality differ based on the empirical paradigm guiding the study. The scientific requirements of objectivity, systematicity, validity and reliability are typically associated with a positivist paradigm and govern quantitative studies. Lincoln and Guba (1985), working in an interpretative research paradigm, propose the criteria of credibility, transferability, dependability and confirmability for qualitative enquiries. Adding to the latter, Richards (2009) suggests transparency. Despite the different terms used for the methodological criteria and irrespective of the methodological orientations of the underlying studies, every mixed method study should be based on methodological rigour or 'legitimation', a term suggested by Onwuegbuzie and Johnson (2006). The importance of methodological rigour was addressed at the planning stage and requirements were built into the different studies to ensure methodological quality and address validity concerns relevant to the project (see also Chapter 4 for a more comprehensive discussion of validity concerns specific to mixed methods research). For example, the analysis of learner language followed established Conversation Analysis procedures (Atkinson and Henritage 1984); the data collection and analysis of raters' verbalised thoughts when using the scales were closely guided by recommendations on carrying out Verbal Protocol investigations (Green 1998); the investigations using Multi-Facet Rasch Measurement ensured data set requirements of data connectivity were followed sets and that fit statistics fell within acceptable quality control limits (Myford and Wolfe 2003, 2004).

A further consideration at this stage was how to ensure that meta-inferences resulted from an *integration* of the different autonomous stages. This is arguably the most crucial and most challenging aspect of a mixed methods study. Attention has to be paid, therefore, to how to use the information from one stage in the other, so that the two stages 'are not discrete or just superficially sequential, but build on one another' (Creswell 2014:226). In formulating the research design supporting the development of the assessment scales it was important, therefore, to ensure that even though the studies were independent, their findings were nevertheless integrated in informing further steps. To support this goal, the group members responsible for the development of the scales independently considered each set of findings and recommendations and built in regular meetings to review, consider and discuss findings and suggest further revisions to the scales and next steps based on an integrated consideration of all available findings to date. A graphical display of the project is given in Figure 9.1.

FIGURE 9.1 GRAPHICAL DISPLAY OF THE RESEARCH DESIGN STAGES, PROCEDURES AND PRODUCTS


MULTI-STAGE SCALE DEVELOPMENT
Stage I
Stage II
Stage III
Procedures
Procedures
•Selection of performance data
•Open-ended questionnaires from speaking examiners about old and revised scales
•Conversation Analysis
•Coding of speaking examiner comments
•Mapping exercise (draft scales and descriptors to a Cambridge speaking scale
•Think aloud protocols with speaking examiners
•Multi-faceted Rasch
•Verbal protocol analysis
•Rating exercise (multiple markers)
•Multi-faceted Rasch
QUAL Data Collection
QUAL Data Analysis
Draft Scale (V1)
QUAN + QUAL Data Collection
QUAN + QUAL Data Analysis
Draft Scale (V2)
QUAN Data Collection
QUAN Data Analysis
Final Scales (Operational)
CA transcripts
Survey results
Identification of the most important constructs based on CA findings and perspectives from experts and practitioners
Draft scale/descriptors
Performance score data for MFRM
Think-aloud transcripts
MFRM results (scale functioning)
Themes from verbal protocols
Revised scales
Performance score data for MFRM
MFRM results
Scale functioning
Products
Products
4 months
5 months
7 months
Jan 2006
April 2007

STAGE I: SETTING OUT EMPIRICALLY-BASED DESIGN PRINCIPLES OF THE ASSESSMENT SCALES

GOALS FOR STAGE 1: DEVELOPMENT OF DRAFT SCALES

The first stage of the project aimed to gather information from a range of sources, including the academic literature and leading experts, speaking examiners, and learner speech, which would then support establishing the key design principles of the scales. The goal was to:

- Decide on the construct definition, i.e. the components of ability to be measured (e.g. Grammar, Vocabulary, Fluency, Pronunciation, Organisation, etc.)
- Decide on the weighting of each component of ability
- Decide on the distinctions in ability to be measured (i.e. the number of points in the scale)
- Produce draft competency statements in each assessment category at different proficiency levels.

In order to accomplish this goal the development team had to balance and reconcile two competing and opposing demands in scale construction: on the one hand, the need to develop scales which have construct coverage, accurately and comprehensively capture the constructs of interest and are therefore relatively long and detailed, and on the other hand, the need to consider usability and produce scales which can be used by examiners in real time and are therefore relatively short and succinct.

METHOD, DATA COLLECTION AND ANALYSIS FOR STAGE 1

Two key sources of data fed into this initial stage: learner speech and open-ended survey responses from speaking examiners. The details were as follows:

	Analysis of learner speech **(Lazaraton & Davis 2006, 2007)**	**Speaking Examiners' open-ended survey responses (Green 2006)**
	QUAL	**QUAL**
Objectives	• To identify discourse features associated with differently ranked performances based on 'thick' description of test performances • To identify salient features of candidate discourse at the CEFR	• To explore the opinions of speaking examiners about the use of the 'old' speaking scales used for Cambridge English speaking tests and their ideas about the development of the 'new' scales

	B1, B2 and C1 levels • To review the extent to which such features are captured by the current scales and elicitation procedures	
Data collection	• 32 speaking test performances of test-takers representing a range of L1s and ability levels	• 316 speaking examiners with a range of experience • Open-ended examiner feedback
Data analysis	• Conversation Analysis of micro-level conversational features	• Thematic analysis of extended examiner comments

In the study focusing on learner speech two researchers with expertise in Conversation Analysis of L2 learner speech undertook a Conversation Analysis of 32 learner performances (16 paired speaking tests) at different proficiency levels, ranging from CEFR B1 to CEFR C1, and representing different first languages (Lazaraton and Davis 2006, 2007). A key objective of this project was to identify salient features of learner discourse at the CEFR levels of interest, to review the extent to which such features are captured by the 'old' assessment scales (in use at the time of the project), and in the process to make suggestions for features of learner speech which should be included in the new assessment scales. Conversation Analysis was chosen as the methodology to use (as opposed to, for example, just an analysis of lexico-grammatical features) due to its suitability to provide insights not just about lexico-grammatical features of learner speech, but also their interactional ability, which is an important feature of the Cambridge English exams. In other words, Conversation Analysis allows investigations to go beyond words and sentences to also consider interactional aspects of speech, which is an important feature of the Cambridge English speaking exams. Conversation Analysis was, therefore, suitable for an in-depth construct validity investigation which addressed a range of features of learner speech, including the interactional abilities.

The data used in this study were chosen from learner speaking test performances used for examiner training and standardisation and were thus among the most reliably marked performances available. While this was a qualitative study, systematic principles of sample selection usually associated with quantitative studies were followed and a stratified sample was selected from the set of recorded speaking tests used for examiner training and standardisation. The selected test takers were representative of the population and covered a range of L1s and proficiency levels i.e. weak, average and strong.

All the data were transcribed using Conversation Analysis transcription conventions (Atkinson and Heritage 1984, ten Have 1999), which attempted to capture features of the interaction such as pausing, turn taking and speech overlap. Initial transcriptions were compared to audio played in real time, and finally to the video. At times the researchers referred to the printed task materials to help understand the words or concepts the learners were trying to convey.

The data analysis in this study was an iterative process, following the 'data exploration strategy' outlined by ten Have (1999) and Lazaraton (2002), in which the researchers scrutinised the transcriptions for relevant and interesting features of the learner talk as they relate to the different assessment categories in the scales. The analysis was carried out independently by the two researchers, and then comments were considered together, as a quality check for inter-coder agreement. Once the transcription notes were discussed and agreed on, the learner performances were arranged from high to low scores, so that patterns within and across a score level could potentially emerge.

The data collection, transcription and analysis procedures followed established Conversation Analysis procedures in order to ensure methodological rigour of the findings (ten Have 1999). They were in line with a qualitative orientation to research and provided an emic perspective and a look 'inside' speaking tests (van Lier 1989).

The second key study during this stage of the project consisted of the collection and analysis of examiner survey comments about the performance of the scales in use at the time and their views on potential changes. The data collection involved a questionnaire distributed to speaking examiners exploring their use of the current assessment scales. Questionnaires were returned from 316 examiners and the responses were used to inform the ongoing development of the new scales. The thematic analysis of the open-ended examiner comments provided a rich set of data and recommendations for features of the new scales.

Results for Stage 1

The stage 1 research studies yielded a set of general recommendations about the scales and specific suggestions about wording of the performance descriptors and assessment categories. Data integration – a key element of mixed method research - was based on the additive insights from the quantitative and qualitative findings from the studies in this stage, and resulted in decisions about assessment categories in the scale, the number of scale points, the specific wording of the descriptors.

The Conversation Analysis led to specific recommendations about the wording of the performance descriptors. In the interest of space, only a few illustrative examples based on the findings/recommendations from the analysis are presented. For example, the analysis pointed to the need for more precision in the language used in the scales:

> The descriptors for the PET Speaking Test rating scale are extremely general and abbreviated. Although the descriptors seem reasonably accurate, they aren't very precise.

> Precise scores appear to be based upon a set of exemplars, i.e., the training videos, and the examiner training process more generally. The rating scale seems better suited for use as a ‘reminder’ or mnemonic device for examiners, rather than a catalog of features that distinguish levels of performance on the PET (Lazaraton and Davis 2007:91).

The analysis also allowed for specific recommendations to be made regarding the wording of the descriptors. For example, regarding the CEFR B1 descriptors for Grammar and Vocabulary, the authors noted:

> Given the degree of inaccuracy we saw at the 3.0 level, it might be useful to extend the description to read: *Shows a good degree of control of simple grammatical forms, **but errors may obscure meaning**.* Similarly, the 5.0 band descriptor might read: *Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms, **although errors may occasionally obscure meaning*** (Lazaraton and Davis 2007:34).

Most importantly, there was convergence of findings between the questionnaire and the analysis of learner talk on some issues relevant for the assessment scales. To take the Interactive Communication assessment criterion as an example, the Conversation Analysis indicated the importance of including an assessment of the interactional ability of learners in the construct underlying the rating scales, and the examiner comments corroborated this inference, as seen in the joint displays below.

Conversation Analysis	**Examiner survey**
Learners at different ability levels manage interaction differently and therefore a scale should include concrete competency statements which discriminate across these different levels of ability.	‘Interactive Communication is, I think, the essence of what we are trying to assess, and the other criteria contribute to this’
(Lazaraton and Davis 2006:xxx).	‘Interactive Communication is the main reason for students learning a language’
	‘Walking dictionaries are rightfully worthless if

	they lack communication skills'. (Examiner ID Green 2006: xxx)

The Conversation Analysis also indicated that the concept of 'turn-taking' in the scales is too vague. The examiners comments indicated the same, noting that the criterion of how well a learner interacts can be difficult to judge.

Conversation Analysis	**Examiner survey**
'All candidates seem to follow the norms of turn-taking, so this feature does not seem useful for distinguishing performances at the 3.0-5.0 level' (Lazaraton and Davis 2006:66).	'Interpreting sensitivity [to turn taking] requires a more subjective assessment. I feel much more comfortable with the objectivity of other [assessment] criteria.' (Examiner ID Green 2006: xxxxx)

The inclusion of 'hesitation' in the Interactive Communication was noted in the Conversation Analysis and also emerged as a theme in the examiner feedback:

Conversation Analysis	**Examiner survey**
We do feel that there is the potential for a confounding effect between hesitation and the discourse management subcategory, because if a candidate hesitates excessively, coherence may suffer. If hesitation is a feature of relevance to *both* categories, then scores in the two categories are not independent (Lazaraton and Davis 2006:70).	'Hesitation is difficult to decide if it belongs in Interactive Communication or somewhere else, e.g. insufficient vocabulary.' (Examiner ID Green 2006: xxxxx)

The results from the examiner questionnaire provided insights into a wide range of issues, going beyond learner language, such as, for example, aspects of the scales which examiners find difficult to award scores for, sub-criteria which examiners find most or least difficult to awards scores for, or which they find most/least useful. For example, the survey indicated that the main aspects of the scales which cause difficulty for examiners are the 'Range' aspect in the Grammar and Vocabulary scale, the 'Adequacy' aspect in the Discourse Management scale, the 'Turn taking' and 'Hesitation' aspects of the Interactive Communication scale, and 'Stress' and 'Rhythm' in the 'Pronunciation' scale. The examiner feedback also indicated that examiners favoured a splitting up of the single Grammar and Vocabulary assessment category into two at the higher CEFR levels. In the words of one examiner:

> If a candidate has good vocabulary but poor grammar, it's difficult to decide on a score. Candidates' vocabulary resources do vary enormously and with the present scales it is not possible to give enough credit to the candidates with a wide vocabulary. [Examiner ID 244, Exam: BEC Vantage CEFR B2] (Green 2006:23).

This was, in short, the process which characterised the integration of findings at this initial qualitative stage in order to produce a set of meta-inferences forming the basis for the next stage.

Decisions made at end of Stage 1

The qualitative approach during the first stage allowed issues to be explored in depth from two different perspectives – learner speech and examiner views. This initial qualitative stage allowed the project to start off with an inductive exploratory and descriptive approach following the principles of qualitative research, and to provide an empirical basis for the decisions taken at the end of this stage. These decisions related to general features of the scales, such as: (i) the assessment categories, (ii) the sub-categories, (iii) exceptions to the general design principles, (iv) the weighting applied to each category, and (v) the number of scale points. They also referred to principles about wording of the performance descriptors, as given in the CEFR (Council of Europe 2001).

STAGE II: A COMPONENTIAL ANALYSIS OF ELEMENTS OF THE ASSESSMENT SCALES

GOALS FOR STAGE 2: INVESTIGATION OF FUNCTIONING OF THE DRAFT SCALES (VERSION 1) AND PRODUCTION OF REVISED DRAFT SCALES (VERSION 2)

The second stage of the project aimed to investigate the functioning of the draft performance descriptors from two perspectives: comparing the *observed* vs *intended* level of each descriptor, and exploring the use of the draft scales by raters in real time. (Examples of the draft descriptors can be seen in Tables 9.1 and 9.2). The aim was to establish difficulty parameters and Multi-Facet Rasch Measurement (MFRM) proved useful for this purpose. The second study had a predominantly qualitative approach (with some quantification of codes) and aimed to gather extended feedback from examiners while they were using the draft scales.

The choice of both a quantitative and qualitative methodology at this early stage was felt to be important and the project moved from descriptors on paper to their use by examiners in real time. At the same time, it was important to establish whether the descriptors were able to reliably distinguish between candidates at different ability levels. Therefore, statistical evidence for the functioning of the descriptors which goes beyond subjective individual perceptions was important as well. As we noted earlier, such complementarity of methods is an important characteristic of mixed method research and was a fundamental feature of all stages of the project.

METHOD, DATA COLLECTION AND ANALYSIS FOR STAGE 2

The two studies which formed the basis of this stage focused on mapping the draft descriptors against the Common European Framework of Reference (CEFR, Council of Europe 2001) and the Cambridge English common scale, and on examiner use of the draft descriptors.

	Mapping of the draft descriptors (Green 2006)	**Examiner use of draft descriptors (Hubbard 2006)**
	QUAN	**QUAL**
Objectives	• To provide evidence for the validity of draft descriptors by mapping against the Cambridge English and CEFR scales	• To explore the use of the draft descriptors and scales by experienced speaking examiners
Data collection	• 31 examiners and 64 draft descriptors	• 8 examiners and 12 test performances
Data analysis	• Multi-Facet Rasch Analysis of descriptor 'difficulty'	• Verbal Protocol Analysis

The mapping study involved 31 speaking examiners who represented a cross section of the examiner hierarchy, but with a strong representation of the most senior and experienced members of the examiner cadre. They were divided into four groups and each received a set of 20 of the 64 new descriptors. Each set had 8 descriptors which overlapped with other sets in a linked design. This was a practical solution which ensured that examiners were not overwhelmed by 64 descriptors, but also met the requirements of MFRM which necessitates a linking of data through overlapping items. An example of a marks collection grid used for this study can be seen in Appendix 9.1.

The examiners, working independently, were asked to indicate which level on the Cambridge English common scale they believed each descriptor represented. The data collection, therefore, included examiner ratings for each descriptor. The ratings were analysed through a Multifaceted Rasch analysis using FACETS (Linacre 2006). FACETS takes into account the relative harshness of the examiners to generate a 'fair average' score for each descriptor in which any differences in the severity levels of examiners rating different sets of items are taken into account by the programme and the raw scores are adjusted accordingly.

The data analysis focused on the relative 'difficulty' of the descriptors as estimated through FACETS based on the examiner ratings in comparison to the difficulties intended by the scale developers. Examiner consistency in the interpretation of descriptors was also investigated. This study was quantitative in orientation, as at this stage it was important to start gathering findings, which have generalizability and are based on the views of a group of examiners and on a powerful statistical procedure.

The second study involved a verbal protocol study and think-aloud procedure related to the assessment process in real time. Its goal was to investigate the comprehensibility and applicability of the scales. Eight examiners, representing all levels of the speaking examiner hierarchy, were selected from four different major test-taker regions. The examiners used the draft descriptors to assess a set of recorded test performances at different ability levels, and verbalised their thoughts. In order to reflect assessment procedures in real time, there was no pausing of the recordings. Data on examiner reactions to using the draft scales were also collected through an open-ended questionnaire administered after the marking.

The analysis involved a transcription of the examiner verbal protocol comments and coding for key themes. The coding system involved a set of categories: 'Grammar and Vocabulary', 'Discourse Management', 'Pronunciation', 'Interactive Communication', 'Assessment Comment/Decision', 'Other'. The coding scheme had been piloted in an earlier study (Hubbard, Gilbert and Pidcock 2006) and the categories had been found to work well in capturing examiner comments relating to the assessment categories, to their assessment decisions and any other comments. A thematic analysis of the extensive examiner comments was also carried out. This study was qualitative in orientation, as it focused on individual experiences when using the scales.

Sampling was an important consideration, and similar to the sampling approach during the previous stage, purposive sampling was used, since the study necessitated the use of speaking performances which display a range of abilities and first language backgrounds, with the view of ensuring representation of weak/average/strong learners and different geographic locations.

Sampling examiners to participate in the two studies was also based on purposive sampling, with the aim to achieve representativeness of examiners at different levels of experience and with different positions within the speaking examiner hierarchy.

A further important question at this stage was whether the sample for the qualitative and quantitative stage should be the same. Creswell (2014) cautions against this, since the qualitative sample would typically be much smaller than the quantitative sample needed to generalise from the sample to the population. In line with this recommendation, care was taken to draw samples from the same population of examiners and test taker performances, but to use different individuals for both samples and also to ensure quality of sampling through stratified sampling based on relevant criteria, e.g. experience for examiners, range of proficiency for test takers.

RESULTS FOR STAGE 2

The findings at this stage included statistical measures for the perceived CEFR level of each descriptor and extended rater feedback on the usability of the scales and descriptors. The integration of these two sets of findings provided the basis for making any changes to the descriptors. Meta-inferences from a joint consideration of the mapping of the draft descriptors and the verbal protocol feedback of the examiners were drawn, with special attention focused on those descriptors which did not map as expected and/or were flagged by the verbal protocol data.

The quantitative scaling exercise provided evidence of broad agreement between the intended levels of the performance descriptors and the examiner ratings and showed that the descriptors could reliably be separated into seven levels according to difficulty. The statistical output also identified a small number of performance descriptors where the examiner ratings contradicted the intended levels. Table 9.1 provides an example of the types of findings which were generated and which served as the basis for further decisions. It gives the intended level of the descriptor in column 1 and the empirically observed Rasch difficulty values in ascending order (the ‘Fair Average’) in Column 3. The criterion Interactive Communication has been chosen as an example.

TABLE 9.1 INTERACTIVE COMMUNICATION DRAFT DESCRIPTORS AND CEFR MAPPING

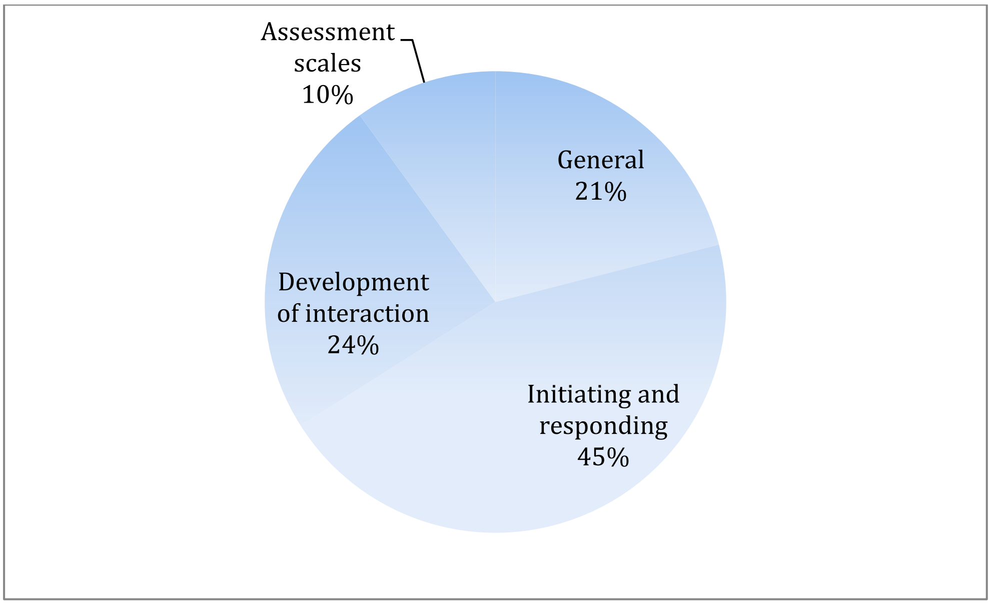
Intended Level	*Descriptor*	*Fair average*
A1	Has considerable difficulty answering questions or responding to descriptors.	1.15
A1	Requires extensive prompting.	1.15
B1	Maintains and closes simple exchanges.	1.76

A2	Answers questions and responds to simple descriptors.	1.92
A2	Engages in the interaction, but occasionally needs additional prompting to keep up the exchange of information.	2.58
B1	Keeps the interaction going with minimal prompting.	3.04
B2	Initiates and responds appropriately.	3.67
B2	Maintains and develops the interaction without support.	3.77
C1	Develops the interaction and/or negotiates an outcome.	3.89
C2	Collaborates with (an)other speaker(s) to widen the scope of the interaction.	4.38
C1	Initiates and responds appropriately, linking his/her own contributions to those of (an)other speaker(s).	4.38
C2	Interacts with ease, linking his/her own contributions to those of (an)other speaker(s).	4.98
C2+	Collaborates with (an)other speaker(s) to develop the interaction fully and effectively.	5.32
C2+	Interacts with ease by skilfully interweaving his/her contributions into the conversation.	6.62

The findings in Table 9.1 show that the rank order and clustering of the descriptors and the range of levels covered (from A1 to C2+) was broadly in line with the scale developers' intentions and provided evidence for the validity of the descriptors. Some cases of divergence from the intended levels were also observed (shaded in the Table). For example, towards the lower end of the scale, 'Maintains and closes simple exchanges' (B2) was rated easier than the two A2 descriptors, indicating that the distinction between 'maintains exchanges' and 'keeps the interaction going' may need clarification as both appear to refer to the same aspect of interaction, although originally intended to discriminate between different levels of interaction ability. Towards the higher end of the scale, 'Collaborates with (an)other speaker(s) to widen the scope of the interaction' (C2) is placed at the same level as 'Initiates and responds appropriately, linking his/her own contributions to those of (an)other speaker(s) (C1), again indicating that the distinction between the descriptors needed to be made more explicit. Quantitative findings of this sort, which focused on each descriptor in isolation, formed the basis of decisions about further revisions to the descriptors.

The qualitative verbal protocol findings explored the use of the scales as a whole in real-time conditions. In addition to the thematic analysis of rater protocol comments, the verbal protocols were also converted to quantitative codes in order to examine the extent to which different assessment categories from the scales were used (Figure 9.2). The coding and quantification of the extended verbalisations of the examiners when using the scales provided information about what assessment criteria examiners pay attention to and whether there is a balance among the criteria they focus on. For example, Figure 9.2 indicates the key components in the 'Interactive Communication' assessment category and confirms the decisions taken at the end of Stage 1 to include 'initiating', 'responding' and 'developing the interaction' as key features of this criterion.

FIGURE 9.2 VERBAL PROTOCOL CODES: INTERACTIVE COMMUNICATION, CEFR B2 LEVEL



The extended examiner comments gathered during this stage of the project provided additional insights about the wording of the descriptors. For example:

> "I found that using 'control' rather than 'accuracy' forced assessments to look at grammatical forms over a number of utterances rather than just focusing on individual mistakes." (Grammar & Vocabulary)

> "The removal of 'incoherent' and 'coherent' and their replacement with notions of 'repetition' and 'digression' are easier to match to performance." (Discourse Management)

> "A judgment on intelligibility is easier to apply." (Pronunciation)

> "I like the reference in Interactive Communication to 'linking contributions to those of other speakers'. This is useful and immediately comprehensible." (Interactive Communication)

The need for greater clarity also emerged as a theme:

> "What constitutes 'a good degree of control', 'limited control'?"
>
> "The 'range' aspect was quite difficult to judge."
>
> "What is meant by 'some complex grammatical forms'?"

An important final step during this stage was the integration of the statistical quantitative findings and the qualitative verbal protocol findings. This allowed the scale development team to focus attention on specific descriptors and wording choices. The statistical results signalled that there was an issue with the wording of some of the descriptors, and the extended examiner feedback provided insights into the cause of some of those issues and how they might be addressed (e.g. through a revision of the wording of the descriptors and/or through examiner training). This was done through a variation of a joint display (see Chapter 5), which included the problematic descriptors and compared relevant findings from the examiner verbal protocols and the mapping study.

Decisions made at end of Stage 2

The qualitative and quantitative findings at this stage confirmed that the criteria, sub-criteria and wording of the descriptors were generally appropriate and applicable in real-time assessment. The findings also provided guidelines about further revision. The examples in Table 9.2 illustrate the change in wording which initial and final draft descriptors went through during stage 1 and 2 of the project:

Table 9.2 Examples of initial and revised draft descriptors for Interactive Communication

CEFR level	Initial draft (version 1)	Draft (version 2)

A2 Band 3	Answers questions and responds to simple descriptors.	➔	Maintains simple exchanges, despite some difficulty.
	Engages in the interaction, but occasionally needs additional prompting to keep up the exchange of information.	➔	Engages in the interaction, but occasionally needs additional prompting and support to keep the interaction going.
A2 Band 1	Has considerable difficulty answering questions or responding to questions.	➔	Has considerable difficulty maintaining simple exchanges.

Findings at this stage also led to the development of a supporting Glossary of Terms which a practical aid for all users of the scale, defining and exemplifying some of the terms used in the scales. For example:

Development of the interaction

Actively developing the conversation, e.g. by saying more than the minimum in response to the prompt, or to something the other candidate/interlocutor has said, or by proactively involving he other candidate with a suggestion or question about further developing the topic (e.g. What about bringing a camera for the holiday? Or Why's that?)

STAGE III: OPERATIONAL ANALYSIS

GOALS FOR STAGE 3: INVESTIGATION OF OPERATIONAL FUNCTIONING OF THE SCALES (VERSION 2) AND FINALIZATION OF THE SCALES

Two quantitative studies formed the final stage of the project and had the overall aim of confirming the soundness of the scales, assessment criteria and descriptors as a whole prior to their use in rater training/standardisation and in operational live test conditions. The need to investigate the scales as they would be used in live conditions was well suited to the adoption of a quantitative methodology and carrying out a marking trial with a large enough number of examiners and test performances to allow for generalizability of the findings.

The studies had dual aims: the first marking trial aimed to gather statistical measures about the performance of the scales and also to allow a comparison between the 'new' and 'old' scales. Such a comparison was necessary, since it was important to ensure that the new scales had not changed the standard of the tests they are used with. The aim of the second marking trial was to provide a final check of the functioning of the scales, and additionally to produce benchmarked test performances exemplifying the different points on the new scales and feed into examiner training and standardisation purposes.

METHOD, DATA COLLECTION AND ANALYSIS FOR STAGE 3

The two trials at this stage involved gathering of multiple marks on video-recorded tests, and the use of Multi-facet Rasch measurement as the analysis tool. Rasch measurement was used because of its value in diagnosing issues in the use of rating scales through statistical measures of all facets involved, e.g. raters, learners, assessment criteria (Linacre 2006, Myford and Wolfe 2003, 2004). The data collection in both studies consisted of raters individually assigning marks to video-recorded speaking tests. The raters involved speaking examiners with extensive rating experience.

The overall objectives, data collection and analysis features of this stage were as follows:

	Marking trial 1 (Galaczi 2007a)	**Marking trial 2 (Galaczi 2007b)**
	QUAN	**QUAN**
Objectives	• To provide statistical evidence of the functioning of the scales in terms of level of learner discrimination, examiner severity/ agreement/ consistency, assessment criteria separation/consistency, and scale points • To compare the marks awarded using the new scales with the 'old' scales • To provide recommendations for examiner training	• To provide statistical evidence of the functioning of the scales in terms of level of learner discrimination, examiner severity/agreement/consistency, assessment criteria separation/consistency, and scale points • To provide benchmark test performances for examiner standardisation
Data collection	• 12 raters, 32 test performances (full tests and test parts)	• 28 raters, 96 test performances (full tests and test parts)
Data analysis	• Multi-Facet Rasch Analysis of the performance of the raters, learners and assessment criteria	• Multi-Facet Rasch Analysis of the performance of the raters, learners and assessment criteria

In general, before raters are asked to mark any test performances, it is important to ensure that they are standardised in their interpretation of the scale. The most effective way to do this is to standardise raters with benchmarked test performances which exemplify the different points on the scales. In the first trial this was not possible, since no benchmarked performances existed yet. The standardisation of the raters was accomplished through familiarisation with the scales and the Glossary of Terms. During the second study the examiners were standardised with benchmarked performances which had been produced based on the first marking study. Once again, we see how different stages in the exploratory sequential design additively build on one another and how one stage could not have been completed without the previous one. We also see the importance of rigour in the research design and data collection to ensure validity of the findings (e.g. ensuring a familiarisation stage before the data collection).

The size of the data sets for this stage met the criteria for a quantitative study which needs to have the power of generalizability. A total of 32 speaking test performances (some as full tests and others as test parts) were rated in the first study by 12 raters; the performances displayed a range of ability levels, e.g. at the CEFR B2 level, weak, average and strong learners were included. A total of 96 test performances (some as full tests and some as test parts) representing learners at different speaking ability levels were rated by 28 raters in the second study. In both studies each rater gave between 3 and 5 marks per candidate. The analysis was carried out separately for each CEFR level.

The analysis involved calculating descriptive statistics and a multi-facet Rasch analysis (using FACETS, Linacre 2006). The facets were: test taker, examiner and assessment criteria; logit and Infit Mean Square measures were provided for each facet. In line with Multi-facet Rasch measurement (Myford and Wolfe 2003, 2004), the analysis focused on:

- Rater harshness/leniency, as seen in the logit measure for each rater and the rater separation strata. If harshness/leniency differences between raters proved to be small, this would provide evidence for the adequate functioning of the scale;
- Rater consistency, as seen in the outfit and infit mean square values. If few raters displayed inconsistency or central tendency (i.e. restricting their scores to the middle of the scale), this would provide evidence for the adequate functioning of the scales;
- Test-taker separation, as seen in the test-taker separation ratio. A higher separation ratio, i.e. a large spread of test takers, would provide evidence for the ability of the scales to discriminate between test takers at different ability levels;
- Difficulty of the assessment criteria, as seen in the logit measure for each one. If differences in difficulty between the criteria proved to be small, this would provide evidence for the equal role played by each criterion in contributing to a final score.

Results for Stage 3

The results from the first study provided measures for the functioning of the assessment criteria and rater severity and consistency. Table 9.3 provides an example of the type of findings, which informed this stage of the project.

Table 9.3 FACETS measures (from Galaczi et al 2011:231)

	CEFR A2 level	**CEFR B1 level**	**CEFR B2 level**	**CEFR C1 level**	**CEFR C2 level**
Number of data points used for estimation	252	748	831	832	560
Learner discrimination					
Number of learners	12	24	24	24	12
Spread (logits)	-1.46 to 5.22	-1.73 to 4.35	-1.61 to 3.00	-1.84 to 3.84	-1.50 to 1.77
Separation strata	9.45	10.56	9.21	10.27	12.12
Rater separation and consistency					
Number of raters	11	19	22	22	14
Spread (logits)	-1.39 to 1.89	-.92 to 1.11	-.95 to .76	-.94 to 2.19	-.64 to .59
Separation strata	4.8	4.0	3.7	6.6	4.3
Raters with infit mn. sq. > Mean + S.D.	1 (none critically)	2 (1 critically)	4 (1 critically)	5 (1 critically)	1 (1 critically)
Raters with infit mn. sq. < (Mean - S.D.)	1	1	1	3	2
Assessment criteria separation and consistency					
Spread (logits)	-.51 to .5	-.29 to .36	-.38 to .29	-.16 to .37	-.38 to .26
Criteria with	None	None	None	PR (marginally)	PR (marginally)

infit mn. sq. . > Mean + S.D.					
Criteria with infit mn. sq. < (Mean - S.D.)	None	None	None	GR (marginally)	LR (marginally)

In terms of the discriminatory power of the scales, the findings in Table 3 indicate that they were found to adequately separate test takers into different ability levels. The results also indicated that there were different levels of rater harshness/leniency (as seen in the logits spread and separation strata in Table 9.2). This was an expected finding, since rater variability is an inevitable part of the rating process, even with highly trained raters. The results showed that the rater severity differences were within acceptable parameters, which was evidence that raters were interpreting the scales in similar ways and, by extension, that the scales were performing at an acceptable level. The results also generated measures of the consistency of the raters. Rater inconsistency is a cause for concern since it could potentially indicate scales which are not applied consistently and are therefore not reliably used. The small number of inconsistent raters suggested that cases of inconsistency were idiosyncratic, and not indicative of an inherent issue with the scales. The results also indicated that the assessment categories were similar in difficulty, indicating that, as intended, they were contributing equally to the overall assessment.

Decisions made at end of Stage 3

The results at the end of Stage 3 provided justification for the finalisation of the scales and sign off for operational release.

Conclusion

The aim of this chapter has been to demonstrate the application of a multistage exploratory sequential scale development mixed methods design in L2 assessment research. We have shown the natural fit between the nature of scale construction and this particular mixed method design. Stage I of the project cast a broad net in order to tap into examiner perceptions and analysis of speech and generate inductive findings. Stage II adopted a narrower focus and looked at individual aspects of the scales, both through a discrete 'clinical' approach, which focused on each descriptor individually, and through an in-depth consideration of examiner experiences when using the draft scale and descriptors in real time. Stage III broadened the empirical focus by using a larger sample of examiners and learner performances and by using the scales in their entirety. Each stage drew on the findings of the previous one in an additive fashion. Such an approach, which capitalises on complementary strengths, enables meta-inferences based on integration of findings and seeks to counterbalance methodological shortcomings, provided the final product of the project – the set of assessment scales – with a strong validity argument about their robustness.

We hope to have also demonstrated in this chapter the necessity for careful consideration of a range of issues at the planning stage, which in turn would support the rigour and systematicity underlying the final outcome or product. Careful planning leads researchers to weight different possibilities during the multiple stages of data collection and analysis, which ultimately leads to a more rigorous final product.

Every study needs to be evaluated against criteria of empirical quality and validity. This was discussed conceptually at the beginning of this chapter and it would be fitting at this stage to close by addressing the empirical quality and validity of the mixed methods project discussed here. In Chapter 4 Ziegler and Kang provided a useful list of concerns for researchers to address regarding the validity of a mixed methods investigation. Their list, even though more suited to basic mixed methods designs and very detailed, will serve as a useful taxonomy of key validity areas to address (Table 9.4).

TABLE 9.4 VALIDITY CONCERNS AS ADDRESSED IN THIS PROJECT

Validity concern	How the research team has addressed the concern
Internal validity: Design quality (multi-stage exploratory sequential design)	A constructivist paradigm espousing variable and multiple realities and focusing on individual 'voices' and 'stories' was complemented by a post-positivist paradigm which reduces ideas into a final product/number/scale/survey. Scale development is suited to a progression from an inductive exploratory stage to a deductive measurement stage, which is reflected in the exploratory sequential design. The large-scale nature of the project further necessitated an advanced design which broke the process into semiautonomous stages. Established procedures were followed within each stage and study, e.g. Conversation Analysis procedures were applied during the initial exploration of learner language, Verbal Protocol norms were followed during the exploration of examiner views when using the scales, data assumptions about linking were taken into account for the quantitative studies employing Rasch analysis.

Internal validity: Interpretive rigour	Consistency between meta-inferences was observed through the high level of convergence of the findings from the qualitative and quantitative investigations. Both qualitative and quantitative findings were integrated additively to form meta-inferences. Interpretations, findings and recommendations were discussed within the development team which included members with a range of theoretical and practical knowledge about scale development and the different methodologies employed.
External validity	The samples of learner test performances in each relevant study were selected to represent a range of background variables, such as first language, age, gender, and as such to be representative of the population and to have wider applicability. The use of a large-scale final phase which drew on a varied sample of learners contributed to the generalizability of the assessment scales. The sample of examiners used in the study was selected to be representative of examiners at different levels of the speaking examiner hierarchy, thus supporting the transferability of inferences to the examiner cadre as a whole. High transferability to other contexts of assessing speaking ability, since the final product – the scales – is theoretically based, empirically-supported and generic in nature.

The amount of effort which was needed for the completion of this project also highlighted some general caveats which need to be noted. The successful completion of the project necessitated a large team, which contributed different areas of expertise. It is important to note, therefore, that a key limiting factor in the use of this design is the need for a team with a wide range of research expertise to be involved in the process. The relevant methodological literature consistently identifies this design as a challenging one, since both quantitative and qualitative skills are needed, much more so than in an explanatory sequential design (Creswell 2014, Hesse-Biber and Johnson 2013). As Creswell and Zhou note in Chapter 3 of this volume, 'a large repertoire of research skills' are needed when carrying out an advanced mixed methods design and collaboration between individuals with diverse methodological skills is essential.

A further caveat is the considerable length of time needed for the completion of this design from start to finish, especially in the case of complex multi-stage designs. In the case of the project discussed in this chapter, the time frame covered 18 months, which is a substantial period of time which needs to be a factored in during the initial planning stage. These limitations have considerable resource implications, which are not to be underestimated.

Despite these practical issues, we believe that a multi-stage exploratory sequential scale development design is exceptionally well-suited to scale development projects as its sequence of methods and potential for additive insights leads to a synergetic relationship between the individual parts - a relationship which allows the whole to be more than the sum of the parts.

REFERENCES

Atkinson, J.M., & Heritage, J. (Eds.). (1984). *Structures of social action: Studies in conversation analysis*. Cambridge: Cambridge University Place.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Thousand Oaks, CA: Sage.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(208-238).

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman/Pearson Education.

Galaczi, E. D. (2007a). *Main Suite and BEC assessment scales revision: Marking trial Dec 2006*. Internal Cambridge English Language Assessment report.

Galaczi, E. D. (2007b). *Main Suite and BEC standardisation videos: Multi-Facet Rasch Analysis*. Internal Cambridge English Language Assessment report.

Galaczi, E. D., ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: a multiple-method approach. *Assessment in Education*, 18(3), 217-237.

Ginther, A. 2012. Assessment of Speaking. In C. Chapelle (Ed.) *The Encyclopaedia of Applied Linguistics*. Wiley & Sons.

Green, A. (1998). *Verbal Protocol Analysis in Language Testing Research: A Handbook*. Cambridge: Cambridge English/Cambridge University Press.

Green, A. (2006). *Main Suite Speaking Test Modifications Questionnaire to Oral Examiners*. Internal Cambridge English Language Assessment report.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(3), 255-274.

Hesse-Biber, S., & Burke Johnson, R. (2013). Coming at things differently: Future directions of posisble engagement with mixed methods research. *Journal of Mixed Methods Research, 7*(2), 103-109.

Hubbard, C., Gilbert, S., & Pidcock, J. (2006). Assessment processes in Speaking Tests: A pilot verbal protocol study. *Cambridge ESOL Research Notes, 24*, 14-19.

Knoch, U. (2009). Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale. *Spaan Fellow Working papers in Second or Foreign Language Assessment*, 7, 21-46.

Lazaraton, A. (2002). *A Qualitative Approach to the Validation of Oral Language Tests* (Vol. 14). Cambridge: Cambridge University Press.

Lazaraton, A, & Davis, L. (2006). *An analysis of candidate language output on the PET Speaking test standardisation videos*. Internal Cambridge English Language Assessment report.

Lazaraton, A, & Davis, L. (2007). *An analysis of candidate language output on the FCE Speaking test standardisation videos*. Internal Cambridge English Language Assessment report.

Linacre, M. (2006). *Facets Rasch Measurement computer programme*. Chicago: Winsteps.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

McNamara, T. (1996). *Measuring Second Language Proficiency*. London: Longman.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research*, 1(1), 48-76.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using Multi-Facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using Multi-Facet Rasch measurement: Part 2. *Journal of Applied Measurement, 5*(2), 189-227.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools,* 13(1), 48-63.

Plano Clark, V. L., & Creswell, J. W. (2008). *The Mixed Methods Reader*. Thousand Oaks, CA: Sage.

Richards, K. (2009). Trends in qualitative research in language teaching since 2000. *Language Teaching, 42*(2), 147-180.

Sale, J. E. M., Lohfeld, L. H., & Brazil, K. (2002). Revisiting the quantitative-qualitative debate: Implications for mixed methods research. *Quality and Quantity, 36*, 43-53.

Sandelowski, M. (2003). Tables or tableaux? The challenges of writing and reading mixed methods studies. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of Mixed Methods in Social and Behavioral Research* Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining Qualitative and Quantitative Approaches.* Thousand Oaks, CA: Sage.

ten Have, P. (1999). *Doing Conversation Analysis*. London: Sage Publications.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49, 3-12.

van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23*, 480-508.

Appendix 9.1

Example marks collection grid (from Green 2006)

What is your name?	
In which country do you usually examine?	

Instructions

1. Please look at the list of descriptors below. Can you match each one to a test level?

2. Based on the assessment scale bands, please select the **lowest** performance level (A1/KET band 1; A2/KET 3; B1/PET 3; B2/FCE 3; C1/CAE 3; C2/CPE 3; C2+/CPE 5) that you feel each descriptor relates to, and enter this in the 'Level' column.

3. For example if you feel that the descriptor '*can give a short simple description of events'* first describes learners at KET 3 level, you would enter A2/KET3 next to that descriptor in the 'Level' column below.

4. Several descriptors from the same scale (e.g. grammar and vocabulary) may appear at the same level (e.g. there could be up to six or seven descriptors at the B1/ PET3 level).

5. All of these descriptors relate to Speaking test performance

Set 3	**Level**
Maintains control of a wide range of grammatical forms.	
Uses vocabulary with flexibility and ease to express meanings.	
Uses vocabulary with flexibility to express meanings.	
Individual sounds are generally articulated clearly, with appropriate use of weak forms.	
Word stress is accurately placed.	
Has a good degree of control of simple grammatical forms, and attempts some complex grammatical forms without obscuring meaning.	
Uses a wide range of grammatical forms.	
Limited but effective control of phonological features at both word and utterance levels.	
Contributions are relevant and there is a clear organization of ideas.	
Collaborates with (an)other speaker(s) to widen the scope of the interaction.	
Vocabulary is adequate to talk about everyday situations.	
Shows only limited control of a few grammatical forms.	
Initiates and responds appropriately, linking his/her own contributions to those of	

(an)other speaker(s).	
Interacts with ease by skilfully interweaving his/her contributions into the conversation.	
Maintains control of a range of grammatical forms.	
Connects sequences of ideas using basic cohesive devices.	
Contributions are varied and always relevant.	
Initiates and responds appropriately.	
Is intelligible, despite strain for the listener.	
Is intelligible throughout.	