li174
Linguistic Insights
Studies in Language and Communication

Chihiro Inoue

Task Equivalence in Speaking Tests

Peter Lang

li174

This book addresses the issue of task equivalence, which is of fundamental importance in the areas of language testing and task-based research, where task equivalence is a prerequisite. The main study examines the two 'seemingly-equivalent' picture-based spoken narrative tasks, using a multi-method approach combining quantitative and qualitative methodologies with MFRM analysis of the ratings, the analysis of linguistic performances by Japanese candidates and native speakers of English (NS), expert judgements of the task characteristics, and perceptions of the candidates and NS. The results reveal a complex picture with a number of variables involved in ensuring task equivalence, raising relevant issues regarding the theories of task complexity and the commonly-used linguistic variables for examining learner spoken language. This book has important implications for the possible measures that can be taken to avoid selecting non-equivalent tasks for research and teaching.

Chihiro Inoue works at CRELLA at University of Bedfordshire, UK. Originally from Japan, she has keen interests in language testing, second language acquisition, and English education. She holds a PhD in Applied Linguistics from Lancaster University, and has been involved in various research projects on the development and validation of language tests in different countries.

Task Equivalence in Speaking Tests

Linguistic Insights

Studies in Language and Communication

Edited by Maurizio Gotti,
University of Bergamo

Volume 174

PETER LANG
Bern · Berlin · Bruxelles · Frankfurt am Main · New York · Oxford · Wien

Chihiro Inoue

Task Equivalence in Speaking Tests

PETER LANG
Bern · Berlin · Bruxelles · Frankfurt am Main · New York · Oxford · Wien

Bibliographic information published by die Deutsche Nationalbibliothek
Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available on the Internet at ‹http://dnb.d-nb.de›.

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from The British Library, Great Britain.

Library of Congress Cataloging-in-Publication Data

Inoue, Chihiro
Task equivalence in speaking tests / Chihiro Inoue.
pages cm – (Linguistic insights : Studies in language and communication, ISSN 1424-8689 ; v. 174)
Includes bibliographical references.
ISBN 978-3-0343-1417-6
1. Oral communication–Ability testing. 2. Language and languages–Ability testing. 3. English language–Spoken English–Examinations. 4. English language–Study and teaching–Japanese speakers. I. Title. II. Series: Linguistic insights ; v. 174.
P53.6.I56 2013
418.0071–dc23

2013035898

ISSN 1424-8689 pb.
ISBN 978-3-0343-1417-6 pb.
ISSN 2235-6371 eBook
ISBN 978-3-0351-0564-3 eBook

Hochfeldstrasse 32, CH-3012 Bern, Switzerland
info@peterlang.com, www.peterlang.com, www.peterlang.net

Printed in Switzerland

Contents

Acknowledgements

This book is a revised version of my PhD thesis, which was submitted to Lancaster University in July 2011. I am most indebted to my supervisor, Dr. Judit Kormos, for her unwavering support throughout my PhD study. My heartfelt thanks also go to Prof. Charles Alderson, who was my co-supervisor for the first two years of my PhD, for helping shape a solid foundation for the thesis.

The members of the Language Testing Research Group and Second Language Learning and Teaching Research Group at Lancaster deserve my sincerest appreciation. I would especially like to thank my raters, Zahra Al-Lawati, Karen Dunn, Tania Horák, Janina Iwaniec, Gareth McCray, Geoff Shaw-Champion, Hiroko Usami, and Lynn Wilson, for sparing time from their MA and PhD research.

Without the cooperation from staff and students at Tokyo University of Foreign Studies, this book would never have been possible. I am very grateful to the students who participated in this study, and to the teachers who helped coordinate the various phases of data collection: Prof. Masashi Negishi, Prof. Asako Yoshitomi, Prof. Yukio Tono, Dr. Naoyuki Naganuma, Mr. Yoji Kudo, and Mr. Naoyuki Kiryu. Receiving such great support from my home university was an enormous encouragement.

My sincerest gratitude also goes to Mr. Hirano at ALC Press, who granted me permission to access the recordings for the NICT JLE Corpus, without which this book would not have been feasible.

I cannot thank my family enough for their love, support, and understanding over the years. Most of all, I would like to thank my husband, Rob, for his unwavering love and faith in me. I dedicate this book to him.

1. Introduction

1.1 Rationale of the Book

Achieving high proficiency in the English language has become increasingly important in Japan as it is considered essential for Japanese people in order to participate in today's globalised world where English is used as a common international language (The Ministry of Education, Culture, Sports, Science and Technology [MEXT], 2003). In response to this situation, in 2003 MEXT has launched a large-scale action plan for better English education which aims to improve its Course of Study as well as curricula, teaching methods and teacher training, and to promote international exchange programmes in high schools so that the Japanese will acquire more communicative English proficiency with stronger productive skills, especially in speaking. Accordingly, it is of no doubt that there need to be test tasks which can reliably measure the English speaking proficiency of Japanese learners.

What is vital for reliable English proficiency tests is to have equivalent forms, i.e. comparable test versions to give to a number of candidates over the years so that meaningful comparisons of scores are possible while maintaining test security. Nevertheless, establishing evidence of equivalence among different test forms or at the task level, especially in productive tests, are rarely carried out by test administrators (Weir, 2005: 250), which seriously threatens not only the reliability but also the validity and fairness of tests. Moreover, the same problem applies to previous studies on task complexity in task-based research, where equivalence of tasks is a prerequisite but has seldom been demonstrated (Weir & Wu, 2006). This issue clearly deserves further exploration. Focusing on the spoken narrative tasks which are frequently used in English tests in Japan, this book seeks to explore how evidence of equivalence of speaking tasks might be established and to examine which variables can be used in establishing

such evidence. In turn, it is hoped that a better understanding of task design for producing more reliable speaking tests can be achieved.

1.2 Spoken Narrative Tasks

1.2.1 Definition of Spoken Narrative

According to Labov (1972: 360), a narrative can be minimally defined as "a sequence of two clauses which are temporally ordered". Based on the analysis of hundreds of stories told in natural conversation by informants from various backgrounds, Labov identified six core features of a more fully-developed narrative: *abstract* (summarising the story briefly before the narrative begins), *orientation* (setting the time, place, characters and situation), *complicating action* (telling the events in the story), *result* or *resolution* (telling what happened at the end), *evaluation* (indicating the point of the story) and *coda* (concluding the narrative) (Labov, 1972: 363-70). Labov's framework has been highly influential in the field of sociolinguistics (Holmes, 2003: 118) and is also frequently cited in studies of second language narrative development (e.g. Liskin-Gasparro, 1996; Verhoeven & Strömqvist, 2001; Montanari, 2004).

Whilst naturalistic data (i.e. data obtained from everyday language use) are collected by sociolinguists, the narratives in second language development are often elicited artificially by prompts such as silent-movie clips and picture books. Spoken narrative tasks, which refer to sequences of a small number of pictures (i.e. 4, 6 or 8 pictures in this book), can be classified as one such form of elicitation prompt for a narrative. In particular, in the fields of language testing and second language acquisition, spoken narrative tasks are administered in order to elicit a relatively long monologue so that the language elicited can be of a certain length that then provides an adequate sample of performance. The next section briefly reviews the use of this task type in these two fields of research.

1.2.2 Spoken Narrative Tasks in Language Testing

In language testing, spoken narrative tasks refer to tasks based on picture sequences that candidates are asked to describe orally in a single time frame (Luoma, 2004: 144). More specifically, Luoma notes that candidates should demonstrate their control over the following essential features of a narrative: setting the scene, identifying the characters and referring to them consistently, identifying the main events, and telling them in a coherent sequence. In the light of the narrative features mentioned above by Labov (1972), candidates may include orientation, complicating action and resolution in a coherent manner in their narration.

Often, criticisms are made of the spoken narrative tasks in tests for their lack of authenticity; it is almost impossible to imagine a real-life situation where a person has to tell a story based on a picture sequence. Nevertheless, the use of this task type may be defended on the ground that narrative is a part of the information routine of reporting, which is a common type of discourse in everyday life (Weir, 2005: 148-149). Besides, in exchange for lower authenticity, constraining the content of narration by pictures can lead to greater reliability. As the pictures control the content of the story for all candidates, so comparisons of performances can be relatively unaffected by background or cultural knowledge, provided that the pictures used are culturally unbiased (Weir, 2005: 148). In addition, this task type is well suited to lower-level candidates because "telling simple stories is one of the first things that they are able to do in a second language" (Fulcher, 2003: 70).

To benefit from these advantages, there are a number of speaking tests which utilise narrative tasks, for example: Test of Spoken English,[1] English Language Skills Assessment[2] and Test in Practical English Proficiency.[3] However, little evidence for the comparability of narrative tasks in different test versions can be found in published research. Moreover, the comparability of different test versions is

1 Administered by Educational Testing Service, USA.
2 Administered by London Chamber of Commerce and Industry Examinations Board, UK.
3 Administered by the Society for Testing English Proficiency, Japan.

seldom demonstrated by testing organisations (Weir & Wu, 2006: 169), although it is vital for any language test to ensure meaningful comparisons of scores across a number of administrations whilst maintaining test security. This issue is discussed further in Section 2.3.

1.2.3 Spoken Narrative Tasks in Task-based Research

In the field of second language acquisition, especially in task-based research, spoken narrative tasks are of a "well-established and frequently researched task type" (Albert & Kormos, 2004: 286). A number of researchers have utilised them to examine the effects of manipulating task administration conditions and/or task characteristics on candidates' performance, including Robinson (2001), Skehan and Foster (1999), Bygate (1999), Ortega (1999) and Yuan and Ellis (2003), to name but a few. These studies are part of an effort to justify and assist in the pedagogic use of tasks by determining how certain task characteristics affect L2 performance, so that teachers can decide which tasks to implement in their classrooms according to their teaching goals (Skehan, 1998: 97). The underlying rationale for this line of research is that tasks with certain characteristics and/or administration conditions will impose varying processing loads, which may then direct the attention of L2 learners to different aspects of language use (see Section 2.5.1 for a more detailed review).

In order for the results and implications of these studies to be valid, it is obvious that the tasks used in a study must be comparable, except for the particular task characteristics or conditions in question. Otherwise, any differences observed in performance cannot be credibly attributed to the task characteristics or conditions in question; they may have been caused by unintended and uncontrolled inherent differences between tasks. Nevertheless, very few such studies have provided evidence of the comparability of tasks beforehand (Weir & Wu, 2006: 169). In fact, many of them do not reveal the actual tasks or the source of where the tasks were obtained. The lack of this important piece of information, as well as the lack of comparability evidence for tasks, can cast doubts on the reliability and validity of the findings of such research.

1.3 Terminology

When discussing comparability, several different terms are used by researchers. It appears that 'comparability' is very comprehensive (e.g. Bachman, 1990: Luoma, 2004) and is used to refer to the comparability of levels or scores on tests by different test organisations, as well as of the versions and forms of the same test. The term 'equivalence', on the other hand, is used in a narrower sense and refers only to the latter. This is because 'equivalent' test versions or forms, by definition, must be designed based on the same specifications (Alderson, Clapham & Wall, 1995). Furthermore, if taken by the same candidates, the test versions or forms should yield the same mean scores and standard deviations, and correlate equally with a third measure of the construct (Associations of Language Testers in Europe, 1998: 144). Considering that the spoken narrative tasks in this book will draw on the same construct of describing events based on a cartoon strip, it is the 'equivalence' of tasks that this book will attempt to establish.

In order to investigate the equivalence of spoken narrative tasks, this book employs the theories of task complexity to discuss the task design and its effects on learner performance. 'Task complexity' is an umbrella term advocated by Robinson (2007), and it refers to the linguistic and cognitive demands of tasks. Linguistic and cognitive demands are, respectively, labelled as 'code complexity' and 'cognitive complexity' by Skehan (1998). A detailed discussion of the terms can be found in Section 2.5.1.3.

The term 'task difficulty' is used in this book to refer to the logit values for tasks calculated by MFRM analysis. Where there is a need to use this term differently so that concepts employed by other researchers can be appropriately introduced, this is clearly explained, e.g. 'task difficulty in the framework by Robinson (2001)[4]' or 'perceptions of task difficulty'. The terms for other characteristics of tasks are introduced and defined in Sections 2.5.1.3 and 2.5.2.

4 Robinson's task difficulty refers to the affective variables in candidates (discussed further in Section 2.5.4). However, they are labelled as 'candidate factors' in this book to avoid confusion.

1.4 Organisation of the Book

This book consists of seven chapters. Following this introductory chapter, Chapter 2 presents a review of relevant literature in language testing and task-based research. Drawing on the frameworks of validity by Messick (1989; 1996) and of contextual factors in speaking assessment by McNamara (1996), Skehan (1998) and Bachman (2004), the aspects of spoken performance to be examined and controlled for in order to establish equivalence are identified. Then, by reviewing previous studies on equivalence in language tests and theories of speech production, attention and task-related factors, operationalisation for the variables is sought. It is concluded that equivalence should be evidenced in terms of ratings adjusted by MFRM analysis, the perceptions of candidates and native speakers of English, expert judgements, and elicited narrative performances characterised in the areas of fluency, accuracy, complexity, and idea units. Chapter 3 describes a series of pilot studies based on tasks from a speaking test in Japan. It describes vital methodological implications for the main study after trialling several variables, collecting expert judgement and native speaker performance data, and selecting appropriate tasks for the main study. The need is also identified to conduct a validation study of the variables to examine the performances elicited. Chapter 4 presents the methodology of the main study, and reports on the instruments, rater training, and methods of analysis in detail. Chapter 5 shows the results for the research questions which address the aspects of spoken narrative performance that are examined. Chapter 6 discusses and synthesises the findings in light of the relevant literature. Chapter 7 considers the implications for the design of spoken narrative tasks and the implications for theories of task complexity, in addition to outlining the limitations of this study.

2. Review of the Literature

2.1 Introduction

In this chapter, drawing on literature in the fields of language testing and task-based research, relevant previous research is reviewed for the purpose of identifying what needs to be considered as evidence of equivalence in spoken narrative tasks. The first half of this chapter mainly handles previous studies in language testing research, discussing relevant aspects of validity and contextual factors of speaking assessment that should be controlled for (Section 2.2), related research on the equivalence of test forms and tasks (Section 2.3), and the methodological implications for this study. The latter half of the chapter summarises task-related research and explores how relevant aspects of validity can be operationalised in this study. It includes a review of models of speech production (Section 2.5.1) and a discussion of relevant task characteristics (Section 2.5.2) and linguistic variables to examine different aspects of spoken narrative performance, such as complexity, fluency and accuracy (Section 2.5.3), as well as task-specific variables (Section 2.5.4). Reviewing the variables for linguistic performance leads to the selection of appropriate rating scales (Section 2.6) for candidates' performance. Finally, the research questions are presented at the end of the chapter (Section 2.7).

2.2 Theoretical Framework

This section reviews the theoretical frameworks on which this book is based. Firstly, it explains the models of speaking assessment by McNamara (1996), Skehan (1998) and Bachman (2002) in order to conceptualise the relevant factors involved in researching speaking

tasks. Secondly, with a view to demonstrating the evidence for the equivalence of spoken narrative tasks, Weir's (2005) socio-cognitive framework for validating speaking tests is reviewed.

2.2.1 Models of Speaking Assessment

When designing a study on speaking tasks, one needs to consider what constitutes a person's speaking proficiency. The model of language proficiency which is most frequently referred to in the current field of language testing is Bachman and Palmer's (1996) model of communicative language ability (Luoma, 2004: 97). This model is based on the work of Bachman (1990), who reorganised the components of one's communicative competence, drawing on earlier frameworks by Hymes (1972) and Canal and Swain (1980), as well as an empirical study by Bachman and Palmer (1982).

Bachman and Palmer's (1996) notion of language ability includes constituents of competence, such as knowledge about the language (language knowledge) as well as the capability to implement the knowledge for use (strategic competence). Language knowledge includes organisational (grammatical and textual) and pragmatic (functional and sociolinguistic) competencies. Strategic competence is a collection of dynamic strategies (goal-setting, assessment and planning) which are utilised when one engages in communication: estimating the task goal and planning what to say and how to say it, while drawing on necessary language knowledge as well as topical knowledge to complete the task.

The concept of language ability is a primary part of candidate characteristics which also incorporate personal characteristics (such as gender, age, L1 and L2 proficiency), topical knowledge, and affective schemata (i.e. emotional attitudes to the topic of a task). Bachman and Palmer (1996: 62) argue that performance should be understood as resulting from a complex interaction between candidate characteristics and task characteristics, as these two sets of characteristics are considered to affect performance greatly.

While Bachman and Palmer's model has contributed immensely to conceptualising an underlying structure of language proficiency (Luoma, 2004: 101), it has been criticised for focusing too much on

the individual candidate (Chalhoub-Deville, 1997: 5). McNamara (1996) drew our attention to the contextual factors that influence a candidate's score or rating in speaking assessment. In addition to the task (characteristics) that Bachman and Palmer (1996) noted, McNamara listed not only the test tasks, but also interlocutors, rating scales and raters as additional elements of contextual factors, as summarised in Figure 2.1, below.

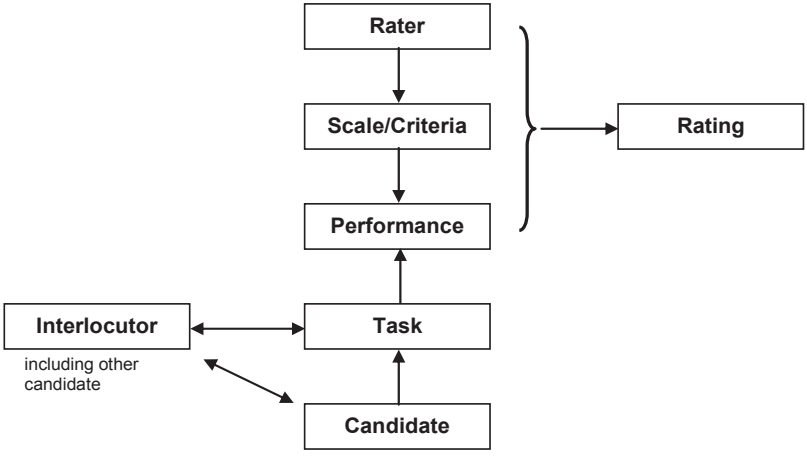

Figure 2.1. Contextual Factors in Speaking Assessment (from McNamara, 1996: 86)

Starting from the bottom in Figure 2.1, a candidate speaks to or with an interlocutor (or interlocutors in the case of a paired or group oral test) on a test task. The performance elicited by the task is rated according to the rating scale(s) or criteria by trained raters, who finally produce a final rating or score for the candidate. McNamara's model has been a very influential framework when organising research (Skehan, 1998: 170), which has led to numerous studies on how different contextual factors may influence spoken performance. Such studies have researched the effects of, for example, different candidate characteristics such as gender (O'Sullivan, 2000), personality (Berry, 2004), interlocutors (Brown, 2003), tasks (Fulcher, 1996b) and raters (Weigle, 1998), to mention but a few.

While recognising the influence that McNamara's model has had in the field of language testing, Skehan (1998) has, nonetheless, argued for its further expansion in order to account for how individual candidates engage with performing a task. Skehan divided task factors into task qualities and task conditions, and incorporated *competence* and *ability for use* as the two factors that influence a candidate, as presented in Figure 2.2. His notion of ability for use "goes well beyond the role of strategic competence [i.e. by Bachman and Palmer (1996)], and draws into play generalised processing capacities and the need to engage worthwhile language use" (Skehan, 1998: 171).

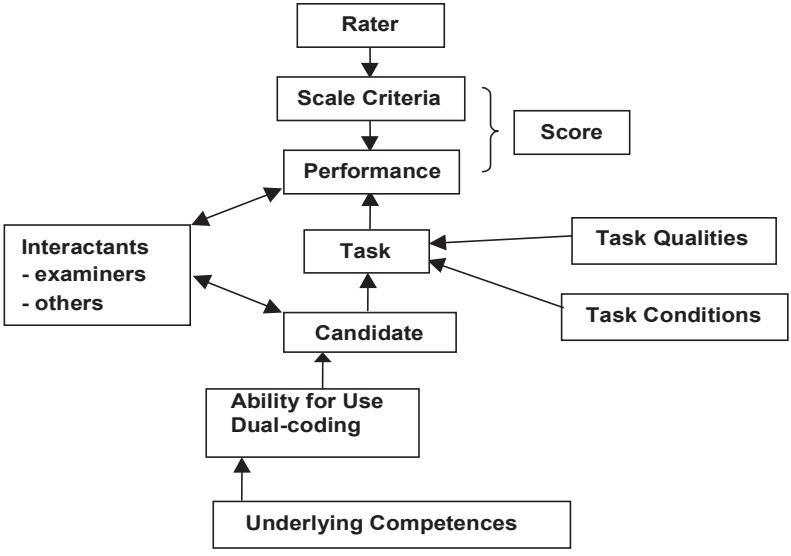

Figure 2.2. An Expanded Model of Speaking Assessment (from Skehan, 1998: 172)

More recently, Bachman (2002) modified Skehan's model, putting more emphasis on the dynamic interaction between the contextual factors, and reorganised candidate factors and task factors, as shown in Figure 2.3. Bachman argues that:

> Candidates, who will differ in their underlying competencies and ability for use, may find tasks with different qualities and conditions differentially difficult to perform. Different candidates will find different examiners and other interactants differentially easy or difficult to interact with. Different raters may apply the scale criteria differently to different performances, so that they may be differentially lenient or severe. (Bachman, 2002: 466)

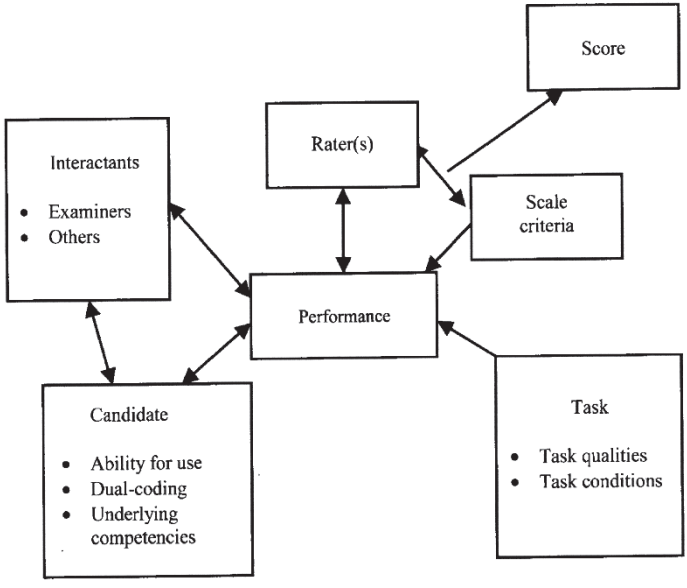

Figure 2.3. Bachman's Model of Interacting Factors in Speaking Assessment (2002: 467)

With such complex interactions between the factors which influence spoken performance and scores or ratings, it is evident that these factors must be strictly controlled for if the equivalence of tasks is to be investigated; thus tasks should be administered by the same interviewer to the same candidates whose performances are then rated by the same raters using the same rating scales. Regarding task factors, Bachman (2002: 469) recommended conceptualising tasks as sets of characteristics and clearly distinguishing between the features inherent in tasks, the attributes of candidates, and the interactions between the

characteristics of candidates and tasks. This issue is discussed in detail in Section 2.5.1.3.

Thus far, drawing on the models of speaking assessment, this section has shown that strict control of contextual factors is indispensable when conducting a study on the equivalence of speaking tasks. Following on from this, the next section reviews a validation framework for speaking assessment by Weir (2005) with a view to gaining insights into the types of evidence necessary to demonstrate task equivalence.

2.2.2 Validity

Because speaking tasks, which are to be proven equivalent, should be designed to represent the same construct, demonstrating their equivalence involves collecting evidence for the validity of the tasks in question. Discussing validity in language testing means ascertaining whether or not a particular test measures what it is intended to measure (Lado, 1961: 321). What language testers intend to measure by their tests is a construct, which is "a theoretical conceptualisation about an aspect of human behaviour that cannot be measured or observed directly" (Ebel & Frisbie, 1991: 108). Examples of constructs include intelligence, motivation, anxiety, attitude and reading comprehension. Messick (1996) offered a widely accepted definition of construct validity in educational assessment as follows:

> Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment. (Messick, 1996: 245)

Accordingly, Messick identified six aspects of validity that should be evidenced to support a validity argument for a test: content validity, structural validity, external validity, consequential validity, substantive validity, and generalisability. Tailoring these aspects to the field of language testing and taking into account the complex nature of speaking assessment described above, Weir (2005; updated in O'Sullivan & Weir, 2011) reorganised and elaborated the concept of

validity, and presented a framework for validating speaking tests as shown in Figure 2.4, below.

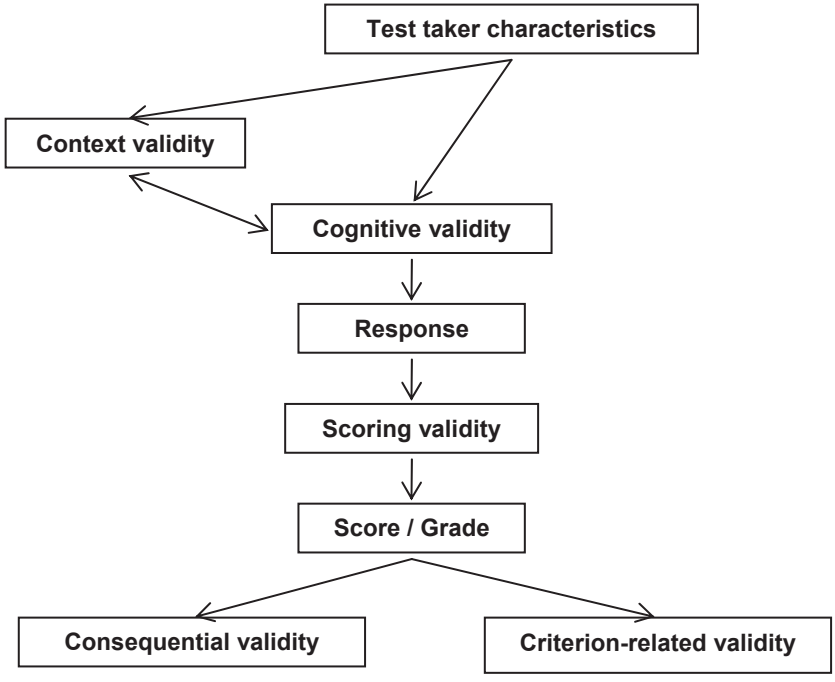

Figure 2.4. A Socio-Cognitive Framework for Validating Speaking Tests by Weir (2005)

Starting from the top, *test taker characteristics* refers to the characteristics that the candidates have when taking a test, e.g. age, gender, nervousness, background knowledge, and experience. *Context validity*, which encompasses Messick's notions of content validity and generalisability, concerns the relevance and representativeness of the test task (including its administration conditions such as response format, time constraints and interlocutor(s)) in relation to the construct. This aspect is often examined by expert judgement of the task content, as well as by analysing the language elicited (Weir, 2005). Previous studies which have looked at this aspect of different test versions for equivalence are reviewed in Section 2.3. In addition, although under