Raili Hildén

Multiple aspects of validity theory in the service of an oral proficiency assessment project

The article describes the theoretical orientation to a 3-year research project, HY-Talk, which focusses on the assessment of oral proficiency in foreign languages. The financial support from the Research Grants Committee of University of Helsinki was allocated specifically for the validation of five illustrative subscales of oral proficiency included in the new national core curricula for general language education in Finland (National Core Curriculum 2003; 2004). These address overall task management in terms of themes, texts and purposes, fluency, pronunciation, linguistic range and accuracy. Each of these is related to different competences utilised in speaking performance. Thus, the test combines competence and task based orientations to speaking assessment. In addition, the research activities will pay attention to language specific cultural determinants of the evolving oral proficiency. The dynamics of test-taking and student interpretation of the test task will also be explored.

The research consortium consists of experts in English, French, German and Swedish languages at the Faculty of Arts, along with experts in language education and assessment from the Faculty of Behavioral Sciences. The data will be collected from school and university levels and investigated in cooperation with professional researchers and students.

The article introduces three important orientations to validity: validity as scientific and interpretive inquiry, and as pragmatic argumentation. A number of links between past but still influential validity research and the HY-Talk study have been established, but closer attention is dedicated to formulating a set of research arguments in line with the pragmatic approach to validation. The major claim to be probed is that the oral proficiency scales currently included are reliable and valid tools for assessing the communicative oral proficiency of students in general language education. The claim needs to be supported by a set of grounding evidence and warrant statements derived from the data. On the other hand, the claim will be confronted with counterclaims and rebuttal data to challenge the conclusions. Specific research tasks assigned to individual researchers is generated from the overall argumentation frame.

Key words: Language assessment, validity, oral proficiency, performance assessment

1 Introduction

The article lays a theoretical foundation to a 3-year research project, HY-Talk, initiated at the University of Helsinki with a focus on the assessment of oral proficiency in foreign languages. The financial support from the Research Grants Committee of the university was allocated specifically for the validation of five illustrative subscales of oral proficiency included in the National core curricula (2003; 2004). These address overall task management in terms of themes, texts and purposes, fluency, pronunciation, linguistic range and accuracy. In addition, the research activities will pay attention to

language specific cultural determinants of the construct of oral proficiency and the dynamics of the test-taking process.

The research consortium consists of experts in the English, French, German and Swedish languages at the Faculty of Arts, along with experts in language education and assessment from the Faculty of Behavioral Sciences. The data will be collected from schools and university institutions and jointly investigated by professional researchers and students.

Since the general purpose of the project launched deals with validation, the first chapters of the article will offer a brief overview on the major strands of validity theory during the last decades. These will be summed up in a scheme that depicts the types of or approaches to validity that are addressed by our project.

2 Multiple layers of validity inquiry and their links to HY-Talk project agenda

2.1 Validity as scientific inquiry: The criterion Model

According to the earliest definitions, test validity simply meant that the test "measures what it purports to measure" (Kelly, 1927, p. 14). Traditional testing was not theory-driven in the current sense of the word, and both its reliability and validity were taken for granted (Davies, 2003, p. 356). Assessment practices were compatible with teaching practices dating back to the medieval tradition of teaching classic languages. Consequently, testing methods of language ability were targeted to detect linguistic knowledge rather than the ability to put it into use. (Spolsky, 1995)

There has been a long tendency in educational measurement to conform to the ideal of scientific inquiry in the field of natural sciences. The main goal of testing was therefore to determine the quantity and composition of latent traits, frequently cognitive in nature (McNamara & Roever 2006, p. 10). Validity was conceived as precise measurement of scores reflecting individual variables like personality traits, properties and skills (Kane, 2001, 320). The rapid development of statistical methods and programmes and the technology to promote their implementation accelerated particularly the scrutiny of reliability issues. In fact, the first attempts to map out the multifaceted terrain of validity were canalized through reliability studies, because reliability was assumed to be the necessary condition of validity. The assertion that it might not be a sufficient condition, however, was voiced later on.

The first influential definition of validity that was to persist a long while into the future was given by Cureton (1951), who characterized validity as indicating "how well the test serves the purpose for which it is used (Cureton, 1951, p. 621 as cited in Moss, Girard & Haniford, 2006, p. 113). The operationalisation of validity as the relationship between test scores and criterion scores on the target task that the test was intended to measure launched the criterion-based orientation towards validity investigation that is widely used still today. The criterion can be drawn from the actual test situation and operationalised as correlations between parts of the test with the overall score or other measures of the same trait, if available (concurrent validity). Alternatively, the criterion can be obtained

from future performances as parallels to the test score (predictive validity). The criterion approach was further elaborated by Cronbach and Meel (1955).

Criterion-based conceptualization of validity is subject to problems due to possible defects in the choice of variables. The quality of criterion variables was rarely questioned, although they were not inherently more truthful than the test score. (Kane, 2001, p. 320) Despite the acknowledged restrictions, criterion-based studies conducted by statistical means still belong to the core of validation procedures, albeit improved with more refined equipment for calculation. The basic idea is relatively unchanged in settings where test performance is compared with real-life performance (Cronbach, 1971) or in studies resorting to expert judgment in modeling a construct or qualities of a performance (Angoff, 1988).

In the HY-Talk context criterion-based validity is considered by comparing scores of the multiple dimensions of oral proficiency with each other and in relation to quantitative and qualitative student variables. The entire design is influenced by the expert judgment model and related statistical tools proposed by Angoff (1988).

A second aspect of validity is content-based validity, developed as an alternative and complementation to criterion-related validity. Content validity focused on obtaining a representative sample of the traits or performances that the test was targeted to measure (Fulcher & Davidson, 2007, p. 4). Carroll (1980, p. 67) suggested that content validity should be determined first by analyzing the communicative needs of the testees, and then by specifying the test content accordingly. The result of the test is thus interpreted in the light of its content, and sufficiently similar tests could be used as each other´s criterion (Ebel 1961).

There is a close link between the HY-Talk project content dimension and the description of the content dimension of the Common European Framework of Reference (CEFR, 2001). This document includes among other things a self-assessment grid (pp. 26 – 27) that, in turn, has been a point of departure for a selection of operationalised can do – statements developed for another tool of integration policy across Europe, the European Language Portfolio. The HY-Talk test tasks are derived from three sources: the CEFR illustrative scale descriptors and from a range of national ELP versions accessible at http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/portfolios.html , and from the Finnish ELP material, not yet accredited by the Council of Europe.

2.2 Validity as scientific inquiry: The Construct Model

The construct model of validity was introduced by Cronbach and Meehl (1955) to offer a more explanatory and theoretic approach than criterion and content validities. Theoretical models were considered to be composed of constructs and their connections in nomological networks, and researchers sought to confirm the existence of these networks by empirical observations (Kane, 2001, p. 321; Davies & Elder, 2005, p. 801). Constructs were defined in measurable terms, and the aim of the measurement was to clarify the structure of a construct by investigating its inner nomological links, and to

define its position in theory by establishing its relationships to other constructs. (Kerlinger & Lee, 2000, p. 40 as cited in Fulcher & Davidson, 2007, p. 7) In essence, validity studies aimed at identifying the fit between empirical observations and theoretical models. If the observations gathered were compatible with the model, the validity of the construct was confirmed. In negative cases, however, the reasons of incompatibility remained unclear. In language assessment this deductive view on validity was promoted by e.g. Lado (1961) and Davies (1977).

From the 1950´s to the late 1970´s, the different models of validity were employed as needed for the various validation purposes. The criterion-based approach was used for justifying admission and placement, while content-based validation pertained to especially achievement testing. During the period from the 1950's to the 1980's, which Moss, Girard and Haniford (2006) label as an era of validity as scientific inquiry, the study of validity conformed to the ideal of scientific orientation in theory building and methodology. Three salient principles of approaching validity dating back to that time period are appreciated still today: For the first thing, validity study was conceived as a multi-phased ongoing process (that of validation) grounded in theory as a point of departure. Certain dimensions were selected for closer investigation, and subsequent methodology was chosen to serve the measurement. The research process was guided by preset hypotheses that were tested against the observations obtained.
Secondly, the proposed interpretation of the test score and its consequences were specified and set as a hypothesis until it could be probed and evaluated. This was a substantial extension to the previous understanding of validation as related to the test itself or the test score. As Cronbach (1971) put it "It is not the test or the test score that is validated, but a proposed interpretation of the score". Thirdly, there was rising awareness directed towards considering alternative interpretations and challenging evidence in validity inquiry. (Kane, 2001, pp. 232 – 324)

2 .3 Current Conceptions of Validity

2.3.1 Validity as interpretive inquiry : Messick

The representation of validity as an integrative constellation of all dimensions described above was acknowledged as the major vein of investigation due to the work of Messick. His seminal definition of validity, still prevalent in most of the validity studies is the following:
Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989, p. 13)
Messick´s model of validity links the content and criteria with the consequences of the particular assessment. The consequences (also termed consequential validity) refer to the values, usefulness, relevance and social consequences of test use. (Messick, 1989, p. 20) This integrated view of validity was taken up in the highly influential guiding documents of testing scholarship (Standards, 1985, p. 9; Standards, 1999, p. 11). Neither the space nor the scope of this article allow for an in-depth report of the Messick legacy in language testing research. There are, however, two strands that deserve to be mentioned:

the practical applications derived from the Bachman model (1990), and the evolving focus on the consequential aspects of assessment.

In the field of language testing the unitary model was promoted most effectively by Bachman (1990) and Bachman and Palmer (1996) who introduced test usefulness as the overall concept unifying five dimensions of test validity, namely, reliability, authenticity, interactiveness, practicality and test impact. Authenticity deals with the degree of similarity that test tasks share with target language use tasks. Interactiveness, on the other hand, refers to the internal processes that are evoked by the test task and its counterparts in real life. Practicality is about the practical constraints of test implementation. Test impact in out-of test contexts is studied from the perspective of washback on teaching, but in broader terms, impact also covers the social consequences as well as the ethical considerations of test use. (Bachman & Palmer, 1996, pp. 18 –19.)

The idea of validity study viewed as interpretative conclusions firmly grounded in performance data will be the leading principle of dealing with the HY-Talk data. The concepts introduced by Bachman have been discussed in project meetings and the dimensions of test usefulness will be addressed by some of our researchers. We have also found useful the approach suggested by Weir (2005), whose validity model essentially poses a re-arrangement of traditional validity types. Weir speaks about a priori and a posteriori validation. The former refers to construct validity put in action through task planning and test design, while the latter covers all the remaining types: reliability (termed scoring validity), criterion and consequential validity.

The second vein inspired by Messick´s model of validity is less practical and still at an emerging state. Nevertheless, the social, cultural and political aspects of validity evolved from consequential validity seem to become a new macroparadigm of language assessment research. The ethical quality of assessment instruments and the responsibility of their users have gained increasing attention at various levels of test development and implementation of assessment practices in a broad social context. (Lynch, 2001; McNamara & Roever, 2006; Shohamy, 2001). Ethical considerations of assessment as power issues are often imbued with postmodern critical theory, in language assessment literature most frequently cited from Habermas, Pennycook, Foccault and Fairclough. These contributions to validity theory are by no means unimportant to the assessment of spoken interaction, but in our case the broad social aspect is somewhat peripheral as the test deployed basically brings no consequences for the tested students. The major aim voiced by the project consortium is, however, to contribute to developing a prototype speaking test that could be implemented nationwide some time in the future and genuinely incorporated into high stakes school leaving reports. At that point of time the consequences can be studied properly from a large-scale social perspective. So far, we must accept a micro perspective to local interactions displayed in the samples.

2.3.2 Validity as pragmatic argument

Since the 1880´s there has been increased acknowledgement of validity theory as an evolving concept. What started as a firm belief in an ideal trait of an individual, moved

forward to recognize the interplay of underlying competences and the context of display. Conceptions of validity were further accompanied by issues of utility and generalizability, and ultimately pushed from the comfort zone of traditional psychometric qualities of reliability and construct validity (formulation by Bachman, 2005, p. 7). Influential in this shift were proponents of the consequences of tests, who advocated the inclusion of social and political reasons in test design and test use to be taken into account at each level of test development. It was increasingly admitted that validity is not solely absolute facts, but a process of interpretation (validation) is also needed to make the facts meaningful. Since there is no absolute answer to the validity question, understanding of the validity of test use for a particular purpose depend on the supporting evidence and the meaning we assign to that evidence. (Fulcher & Davidson, 2007, pp. 18 – 21.) Likewise, the relationship between theory and observation is not bipolar, but rather dialectic: "we see through our beliefs, and our beliefs change because of observation" (Fulcher & Davidson, 2007, p. 12).

Recent developments in validation and validity theory are pragmatic in nature. This is understandable considering their capacity to integrate theoretical and practical elements into a cohesive whole, and above all current validity arguments also imply alternate hypotheses and disagreement as an essential part of an open discussion. (Fulcher & Davidson, 2007, pp. 18 – 21.) One of the most promising openings to conduct validation study in this line of research is proposed by Kane, Crooks & Cohen (1999) and additionally elaborated by Kane (2006) and Bachman (2005).

The validity argument rests on the assumption that the interpretations assigned to assessment scores are said to be valid to the extent that these interpretations are supported by appropriate evidence. A second premise is that the evidence supporting the interpretation needs to substantially outweigh any evidence against the proposed interpretation. The core of validation is, therefore, collecting supporting evidence for the inferences, and to convince the stakeholders of the power of the supporting evidence to outweigh competing interpretations. It is of vital importance that the interpretation be stated explicitly and as clearly as possible by laying out the inferences in the interpretive argument and the assumptions on which they depend (Kane, Crooks, & Cohen, 1999, p. 6).

The validity argument as defined by Kane, Crooks, & Cohen (1999) is particularly suitable for performance assessment, because the intent of performance assessment, as opposed to "objective" paper-and-pencil tests, is to focus attention on a broadly defined and valued type of performance, of which the performances elicited by the assessment tasks are instances. This type of assessment is labeled as "direct", although every performance assessment task unavoidably is artificial and constrained in many ways. Nevertheless, if the test tasks are chosen carefully to reflect a principled set of features shared by the target task in real life, inferences can be drawn from the observed performance to the target variable. Given that the test performance belongs to a set of tasks in the target domain, there are three phases critical to the chain of inference linking the observed performance to the expected performance in the target domain. (Kane, Crooks, & Cohen, 1999, p. 6).

Once students have accomplished the test task, their performance is judged, yielding an observed score. This stage is called *Scoring*, and for this particular step to be acceptable as a starting point for further validation effort, the test context needs to be in consonance with the intended score interpretation (i. e. free from technical or other impediments). Apart from the test situation itself, we need appropriate scoring rubrics that are consistently applied across raters and performances. In practical argumentation effort, alternative interpretations are considered. In particular, a critical review of the scoring rubrics, the scoring procedures, and the procedures for administering the assessment are likely to be involved. (Kane, Crooks, & Cohen, 1999, pp. 9 - 10).

The second phase of establishing a validation argument, is *generalization* implying an inference from the observed score to the universe score, defined over performance in a set of similar or exchangeable tasks in real life outside the test. A statistically justified generalization would require a random or at a minimum, a representative sample from the universe of generalization. In complex performances, however, this is not always feasible. The level of consistency is investigated by reliability studies that have indicated certain problems pertaining to performance assessments. (Kane, Crooks, & Cohen, 1999, p. 10). In oral proficiency assessment, for instance, substantial problems in terms of variation have been reported concerning numerous dimensions of task type, interlocutor effect and rater bias (Fulcher & Márquez Reiter, 2003; Bachman, Lynch & Mason, 1995; Chalhoub-Deville, 1995).

Alternative interpretations with the aim of challenging the grounds of generalising beyond the task performance typically address sample size or representativeness of the sample, as well as a range of sources of invariance (tasks, raters, administration, context etc.) Serious doubts on any of these might undermine the overall argument. Consistency of rating, and subsequent power of generalization, are typically decreased by complex tasks involving several alternatives to choose among. The condition of generalization can be improved by restricting the number of critical task features, but this brings along the drawback of limited authenticity. Reliability can customarily also be strengthened by increasing the number of independent observations, but since performance tasks often require substantial amounts of time and resources, this might not be the first choice of the test designers. What Kane, Crooks and Cohen (1999, p. 10) propose, is increased standardization of sets of task features (instead of single features) and raising the level of rigor in administration procedures.

The third span to continue the chain of inference is called *extrapolation* from the universe score (assigned for expected performance in the universe of tasks similar to or exchangeable with the test task) to the target score, defined over the target domain. The target domain is broader and generally less well-defined than the universe of generalization. In educational contexts, especially in general education, the target domain may be very large both in terms of current setting (everyday life) and temporal determinants (adult life in the future). The degree of certainty will depend on how similar the universe of generalization is to the target domain. In the case of simulations, carried out in isolation of the target domain the link the from universal score to the target score is

potentially weaker than in tasks completed in an authentic setting, such as a work place. (Kane, Crooks, & Cohen, 1999, p. 10) Since it is rarely possible to check the comparability against real life samples, test designers are customarily advised to ascertain that test performance will require approximately the same kinds of knowledge and skill as the critical real life performance.

Akin to most educational occasions the project at hand resorts to simulations as test tasks. These are designed as type tasks (Van Avermaet & Gysen, 2006) that attempt to combine a broader range of features shared by both pedagogical tasks in learning contexts and real life language use tasks in the teenagers´ out-of-school life. The purpose of this procedure is to draw a principled stratified sample from the target domain including many different kinds of tasks (Kane, Crooks, & Cohen, 1999, p. 10) The speaking tasks deployed in the project are intended to include one or more tasks from specific, standard categories of tasks so as not to restrict the universe of generalization too much, but instead to allow for reasonable level of extrapolation to the target domain. Generalization is the necessary condition of extrapolation to occur, even if it is not sufficient by itself. "No matter how authentic the tasks and how carefully they are evaluated, the intended interpretation in terms of the target domain fails if the generalization step fails." (Kane, Crooks, & Cohen, 1999, p.5)

Alternative interpretations will most readily threaten the legitimacy of inference to target scores because of the dissimilarity between the universe of generalization and the target domain. Too narrow a task may not allow for extrapolation over a reasonable set of tasks in the target domain, but complex high-fidelity tasks may be too complicated to administer and score, and therefore the number of tasks included in the test will necessarily be low. To balance between the various stages of inference Kane, Crooks, & Cohen (1999) suggest the following option:

We can strengthen the third inference (extrapolation) at the expense of the second inference (generalization) by making the assessment tasks as similar to those in the target domain as possible, or we can strengthen the second inference at the expense of the third by employing larger numbers of tasks, possibly with somewhat lower fidelity. (Kane, Crooks, & Cohen, 1999, p. 11)

Recently the interpretive argument described above has been extended with an additional link leading from the target score interpretation to decisions based on the use of the test. The final stage of interpretation is labeled Utilization, and it clearly echoes the socio-cultural views on assessment as social and political enterprise dealt with in previous chapters. The complete process of interpretation presents links in an *assessment use argument* (Kane 2004) that consists of *an interpretive argument*, on one hand, and *a validity argument*, on the other. The validity argument approximately covers the traditional selection of validity aspects addressed as early as in the psychometric era of scientific inquiry. The interpretive argument is more of a novelty, and there is certain discrepancy among language testing experts on how far the utilization component of a validity argument is to range over decisions of social and political nature (Bachman, 2005, p. 28).

3 Validity in the HY-Talk study

3.1 Overview of validity considerations of the project

A brief history of validity approaches is presented in Table 1, where the shaded areas depict the adequacy of the particular item to HY-Talk project agenda. Among the most traditional kinds of validity reliability and criterion-related validity will unavoidably be considered. Messick is not directly addressed, whereas Bachman is prominent, and obviously also Weir. We miss the chance of observing e.g. ethical considerations due to the pilot nature of the test, but as far as possible, external matters will draw our attention in the principled validation work based on pragmatic argumentation. Even there, the validity argument will be the preferred focus over the use argument.

Table 1. Approaches to validity inquiry addressed in the HY-Talk project (shaded areas)

Period/ proponent	**Internal considerations (microlevel)**	**External considerations (macrolevel)**
Pre-theoretic era	No articulated theory base	
Cronbach & al. 1955		
	Reliability	
	Content validity	
	Criterion-related validity	
Messick (as cited in McNamara & Roever 2006, p.14)	Score content and meaning	Score use and consequences
Bachman 1990	Test usefulness	
Bachman & Palmer 1996	Construct validity	Impact
	Reliability	
	Authenticity	
	Interactiveness	
	Practicality	
Shohamy & al. 2001	Critical language testing	
Weir 2005	A priori validation	
	A posteriori validation→	
Kane 2004	Assessment use argument	
	Validity argument	Interpretive argument
Bachman 2005	Assessment argument	
	Assessment validity argument	Assessment utilization argument

3.2 Validity as argumentation as a special focus of the HY-Talk research design

Validity as argumentation, substantially inspired by the work of Toulmin (2003) and further elaborated by Kane (2006), builds on a relatively simple architecture of basic

logical reasoning. The main components of an argument are claims, data, backing, warrants and rebuttals that can be completed by a few additional modifying categories. The **claim** is the conclusion of the argument that we seek to establish.

Example: "John´s oral proficiency in English is at CEFR level B1."
John is not entitled to enter a university program where CEFR level B2 required.

Data consist of information on which the claim is based, such as the responses of test takers, live or recorded. (Toulmin, 2003, p. 90; Bachman 2005, p. 9)

Backing is an assurance of the warrant to be justified, for instance theory, prior research or evidence collected specifically for the validation process (protocols of validation sessions, records of retrospection etc.) (Bachman 2005, p. 10; Fulcher & Davidson 2007, p. 165)

The categories of data and backing are treated slightly differently by Fulcher and Davidson (2007, p. 164 – 165), who combine both categories under **Grounds,** which they define as "the facts, evidence, data or information we have available to support the claim".

A **warrant** is a general statement, a proposition that links the data to the claim thus justifying the inference based on the data.

Bachman suggests subdividing warrants for a utilization argument into four types. Type 1 warrant is about the relevance of the argument to the decision to be made. In essence, this type of warrant addresses the extent to which the ability assessed is a relevant part of the task in the target language use (TLU) domain. Type 1 warrants also concern the degree of correspondence between the characteristics of the assessment task and those of the TLU task. (Bachman, 2005, p. 18) Relevance oriented warrants are grounded in traditional categories of content and construct validity in the first place, but also in authenticity in more recent terms.

Type 2 warrant is about the utility of the score-based interpretation for making the intended decision. The usefulness of a test type, for instance, is weighted against a test of a different kind used as a criterion to establish the practical value of the backing. (Bachman 2005, p. 19) This reasoning touches upon the issues of practicality and even consequential validity.

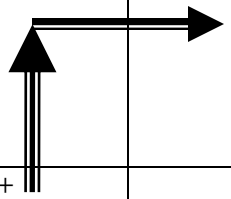
Type 3 warrant is about intended consequences in the sense that the intended decisions will be beneficial to the individuals, organizations or to the society at large. It provides a basis for using a particular assessment as a basis for making decisions (Bachman, 2005, p. 19), and brings us to the core of consequential validity and increasingly stronger emerging issues of fairness.

Type 4 warrant is about how sufficient the information is that the assessment or the test provides for decision-making. The concept of sufficiency links to content coverage and

construct validity and the relationships between language related and other competences in the performance on which the decision is based. (Bachman, 2005, p. 21) Language proficiency is seldom a sufficient condition of hiring employees, while in school settings, displayed language ability alone, may well suffice for a high grade, despite obvious problems with getting along with school mates.

Rebuttals are statements implying alternative explanations or counterclaims that challenge the intended conclusion, the warrant. The rebuttals correspond to potential sources of invalidity, basically due to either construct irrelevant variance or construct under representation (Messick 1989 as cited in Bachman, 2005, p. 10). As a matter of fact, each type of warrant can have a counterpart among the suite of rebuttals. Rebuttals are supported by **rebuttal data**, which is evidence introduced to support, weaken, or reject the alternative explanation (Bachman, 2005, p. 10).

Table 2. Validity argumentation scheme for interpretation of the HY-Talk project data (adapted from Fulcher & Davidson, 2007, 164 – 174; Bachman, 2005)

		Claim = decision to be made
		The illustrative scales of descriptors of oral proficiency included in the national core curricula for language education enable sufficiently valid conclusions on students´ oral proficiency in general school education in Finland.
Grounds: Warrants (W) + Backing data		
Assessment-based interpretation: The data gathered by the project support the rationale of the scale for oral proficiency included in the national core curricula for language education.		**Qualifiers** based on Rebuttals (R) + Rebuttal data
Warrants (W) ↑ (since…)		Rebuttals (R) ↑ (unless…) Construct-irrelevant variance/ construct under-representation
1. The critical dimensions included in the scale are relevant indicators of oral performance. (relevance) 2. The tasks used to elicit student performance correspond to pedagogic tasks and target language use tasks of students at the age of		1. The dimensions included in the scales are marginal or irrelevant as indicators of oral performance (relevance counterclaim) 2. The tasks used to elicit student performance correspond inadequately to pedagogic tasks or TLU tasks of students. Moreover, the link to the scale descriptors may

general education. (utility) 3. The critical trait dimensions detected in performances display a logical progression across the steps of subsequent scales and in relation to the overall scale for oral proficiency. (intended consequences) 4. Reliability of assessments based on the scale and the tasks to elicit performances is found to be high enough. (sufficiency)		be weak. (utility counterclaim) 3. Variability detected in the critical trait dimensions is not related in a consequent manner to the bands of subscales or the overall rating. (counterclaim against intended consequences) 4. Reliability of assessments is not stable, but varies too much across tasks, raters or languages, or is caused by intervening variables or inadequate evidence base. (sufficiency counterclaim)
Backing ↑ Based on Assessment performance and associated data		Rebuttal data↑ Based on Assessment performance or other sources
1. Theoretical models of communicative oral proficiency and theory of oral testing include the dimensions proposed. (support to W1) 2. The tasks were derived from CEFR based can do – statements written for general school education in Finland and a number of other European countries. Rater and test taker feedback confirm the perceived authenticity of the tasks and appropriateness of administration. (support to W2) 3. Empirical analyses of the performance data gathered in the project support the progression across each of the scale in particular, and in relation to the overall scale of oral proficiency. The empirical indicators corresponding to the cut-off scores set for each criterion scale fit the theoretical and empirical model selected for the purpose. (support to W3)		1. Alternative models of oral communication challenge the construct applied along with the traditional quality dimensions. (support to R1) 2. The task selection is undermined by up-to-date scholarship, need analyses mapping school-aged students´ target language use, or rater or/and test taker feedback. (support to R2) 3. Statistical evidence shows that the overall rating of oral proficiency displays low correlations with ratings on the more specific criteria of speaking performance. (support to R3) 4. The statistical reliability evidence reveals instability in terms of raters, tasks, languages or undefined sources of invariance. (support to R4) 5. Analyses of student records, session protocols or any other source reveal rebuttal data that does not fit the predetermined criteria. This type of data render additional insights into an emergent construct of oral proficiency as perceived and displayed by students as they interpret the test tasks for performance.

4. Statistical reliability evidence confirm sufficient level of consistency across raters, tasks and languages, and interlocutors (support to W4) 5. Statistical reliability evidence from previous studies corroborates initial stability of the scales, both in terms of internal consistency and in relation to corresponding scales of the CEF. (support to W4)		

The scheme presented above can only be a tentative one, because treating validity from the angle of pragmatic argumentation is a dynamic enterprise. Appropriate evidence and counter-evidence may bring forth a need to modify any of the warrant and rebuttal statements, at any point of the course of study. As it looks now, however, most research questions that the HY-Talk consortium intends to address can be derived from the generic framework of argumentation.

There is forthcoming work on e.g. interlocutor effect on performance (W4), cultural issues across languages (W4) and theoretically oriented accounts on the construct of oral proficiency in test settings (W1). We will also collect test taker and rater feedback to shed light on their perceptions (W2). Our most laborious empirical effort addresses the quality and cut-off scores of the subscales. It is expected that several research papers will be published in the next few years.

Acknowledgments

I wish to thank Dr Sauli Takala, the President of European Association for Language Testing and Assessment (EALTA), and Dr Heini-Marja Järvinen from the University of Turku, for their valuable comments on my draft and suggestions for improvement.

References

American Education Research Association, American Psychological Association & National Council on measurement in Education. (1985) *Standards for Educational and Psychological Testing.* Washington DC: Authors.

American Education Research Association, American Psychological Association & National Council on measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington DC: American Education Research Association.

Angoff, W. H. (1988). Validity: An Evolving Concept. In H. Wainer & H. I. Braun (Eds.) *Test Validity* (pp. 19 – 33). Hillsdale, NJ: Lawrence Erlbaum.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Bachman, L. F., Lynch, B. & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12, 238 – 257.

Bachman. L.F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly*, 2(1), 1–34.

Carroll, B. J. (1980). Specifications for an English language testing service. In J.C. Alderson & A. Hughes (Eds.) *Issues in Language Testing. ELT Documents 111*. London. British Council, 66 – 110.

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

Chalhoub-Deville, M. (1995). A Contextualized Approach to Describing Oral Language Proficiency. *Language Learning, 45* (2), 251–281.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (ed.) *Educational Measurement* (pp. 443 – 507).Washington, DC: American Council of Education,

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281 – 302.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621 – 694). Washington, DC: American Council on Education.

Davies, A. (1977). The construction of language tests. In J. P. B. Allen & A. Davies (Eds.) *Testing and experimental methods* (pp. 38 – 104). The Edinburgh Course in Applied Linguistics. Vol 4. Oxford: Oxford University Press.

Davies, A. & Elder, C. (2005). Validity and validation in Language Testing. In E. Hinkel (Ed.) *Handbook of Research in Second Language Learning and Teaching* (pp. 795 – 813). Mahwah, NJ: Lawrence Erlbaum.

Davies, A. (2003). Three heresies of language testing research. *Language Testing, 20* (4), 355 – 368.

Ebel, R. L. (1961). The Relation of Scale Fineness to Grade Accuracy. *Journal of Educational Measurement, 6* (4), 217-221.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment. An advanced resource book*. Abington & New York: Routledge.

Fulcher, G. & Márquez Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20* (3), 321 – 344.

Kane, M. D. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement, 38* (4), 319 – 342.

Kane, M. D. (2004). Certification testing as an illustration of argument-based validation. *Measurement: interdisciplinary Research and Perspectives, 2*, 135 – 170.

Kane, M. D. (2006).Validity. In R. L. Brennan, (Ed.), *Educational Measurement* (4th edition), (pp. 17 – 64). Westport, CT: Praeger.

Kane, M., Crooks, T. & Cohen, A. (1999). Validating Measures of Performance. Educational Measurement: Issues and Practice 18 (2), 5 – 17.

Kelly, J. P. (1971). A Reappraisal of Examinations. *Journal of Curriculum Studies, 3* (2), 119 – 127.

Kerlinger, F. N. & Lee, H. B. (2000). *Foundations of Behavioral Research.* (4th ed.). Orlando, FL: Harcourt Brace.

Lado, R. (1961). *Language testing: the construction and use of foreign language tests*. London: Longman.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18* (4), 351 – 372.

McNamara, T. & Roever, C. (2006). Language testing: *The Social dimension*. Language Learning: Monograph Series. London: Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement.* (3dh ed.). NY: Mc Millan, 13 – 103.

Moss, P. A., Girard, B. J. & Haniford L. C. (2006). Validity in Educational Assessment. In J. Green & A. Luke (Eds.) *Review of research in education,* 109 – 162.

National core curriculum. (2003). *National Core Curriculum for General Upper Secondary Education Intended for Young People*. Helsinki: Finnish National Board of Education.

National Core Curriculum. (2004). *National Core Curriculum for Basic Education* Helsinki: Finnish National Board of Education.

Shohamy, E. (2001). *The power of tests: a critical perspective on the uses of language tests.* London: Longman.

Spolsky, B. (1995). *Measured Words. The development of objective language testing.* Oxford: Oxford University Press.

Toulmin, S. E. (2003). *The uses of argument.* Cambridge: Cambridge University Press.

Van Avermaet, P. & Gysen, S. (2006). From needs to tasks. In K. Van den Branden, K. (Ed.) *Task-Based Language Education. From theory to practice* (pp. 17 – 46). Cambridge: Cambridge University Press.