



The Benefits and Challenges of Holistic In-house Task-based Language Learning and Assessment

Roger Nunn & John Thurman

Petroleum Institute, Abu Dhabi & Hokaido University, Japan

Abstract

In this paper, insights from the specialized learning and assessment literature are used to outline some important challenges and the proposed solutions in one academic context in which holistic in-house tasks and rating scales *that support teaching* were designed. The first challenge is to ensure that the design of tasks and rating scales is not just task-based but also construct-based. (Bachman, 2002, p.470). In this respect, holistic learning and the holistic nature of academic competence are key constructs. Secondly designers need to make sure the design is fully sensitive to all aspects of their own context (context validity, Weir 2005). Thirdly, specific areas of atomistic language ability that are relevant to the course need to be identified and included. (Bachman, 2002, pp. 470-471). Yet another challenge is to provide an adequate and varied sample of task responses through a variety of procedures. Arguably the most important challenge is to make sure that in-house assessment design enhances learning. In this respect, fully involving students in their own assessment can make an important contribution. In relation to construct validity one important holistic solution is to use “complex, skills-integrative and goal-oriented” assessment tasks (Norris et al., 2003, p.397) that match the teaching. Other relevant solutions involve designs of tasks that are sensitive to the relationship between academic competence and task performance and the relationship between task difficulty and task performance.

Introduction

Assessment literature is not frequently evoked in support of learning, but this paper will consider that learning and assessment are part of one holistic activity aimed at improving competence in any in-house context. Those institutions that have implemented some form of holistic teaching approach involving skills integration, project-based, task-based or content-based learning face concurrent challenges in the implementation of a more holistic form of assessment. Following the lead provided by Byrnes (2002), who addresses task-based assessment in relation to a content-oriented collegiate foreign language curriculum, this paper will therefore focus on the challenges of designing learning *and* assessment tasks for holistic in-house instruction.

The paper has two main purposes:

- to use insights from the specialized testing literature to outline the challenges that need to be addressed when designing in-house learning *and* assessment;
- to provide examples of how these challenges were provisionally addressed in one context to facilitate both the development *and* assessment of competent academic performance within a holistic project-based/task-based approach.

In the communication department of the Petroleum Institute in Abu Dhabi, a holistic project-based course has been designed for engineering undergraduate freshman students to cover a broad variety of different tasks that integrate not only language skills but also academic processes such as critical reading, basic research skills, cognitive processing, the ability to contribute to a team and to develop individually from a team process. A semester-long team-based project acts as a holistic task-based framework for learning study and research skills and developing academic communication and language ability. A variety of spoken and written tasks based on the needs of engineering students are assessed at different stages of the communication course. These vary from short drafts of meeting minutes to a full research report, and from individual- to team-drafted tasks.

Rather than reporting one experimental research project, my study addresses different local research episodes in the spirit of action research, as an attempt to find workable and above all *relevant* solutions to local in-house challenges. *Relevance* (Sperber and Wilson, 1995) has long been an important construct for my research into curriculum planning. (See Nunn 2006b for a full discussion of relevance, teaching and learning.) It will be referred to throughout this paper at different levels of application. According to Sperber and Wilson (1995) well-supported arguments or information can influence audiences to modify or even abandon former assumptions. At a macro-level, insights from specialist testers need to be adjusted to be relevant to in-house learning. At a micro-level, in the case of task-based learning, task fulfillment is assessed by determining that all parts of the task response are relevant and that all relevant aspects of the task as specified have been addressed.

Background Discussion: Teaching *and* Assessing Holistically

Samuda and Bygate (2008) suggest that the use of holistic tasks “has been one of the major focuses of educational debate over the last century” (p.5). For Samuda and Bygate a task is

holistic in the sense that “it requires learners to decide on potentially relevant meanings, and use the phonology, grammar, vocabulary and discourse structures of the language to convey these in order to carry out the task” (p.13). A holistic context-relevant approach is in harmony with ecological approaches to SLA. Within this approach, Toolan (2003) emphasizes the more general need to take the context into account in the following terms:

An ecology of anything has to be as holistic and inclusive an account of that thing as possible, rather than an account that must continually acknowledge post hoc the influence of factors previously denominated as ‘external’ or ‘contextual’ (p.123).

Fettes (2003) summarizes the ecological vision of linguistic activity as being in “active communication with its neighbours in the biological, social and human sciences, sharing and developing a holistic understanding of human thought, action, and ecological integration” (p.44). Viewed holistically, a language is a living organism that cannot be reduced to a fixed code to be acquired by a learner. Toolan (2003) differentiates between a code and *codification*, which he defines as an “after-the-fact assigning of value” (p.125). There may be some value in codification for educational purposes, but integrationists hold that “our experience in [...] various language tasks not only exposes us to variation, difference, fluidity and continual adjustment, but requires us to engage with and exploit this inherent variability, indeterminacy, and scope for refashioning” (pp. 126-127).

We take this view to have implications for assessment that supports learning as it negates views of a person’s linguistic competence as something finite or static. Competence is never a reified given entity waiting to be objectively measured, but more a negotiable potential that may lead to achievement at different levels according to relevant aspects of the environment.

At the same time that some in-house teaching has taken up more holistic paradigms, a similar more holistic trend can be found in recent testing frameworks that do not specifically focus on in-house assessment. Weir (2005), for example, provides a comprehensive socio-cognitive framework creating inclusive superordinates such as ‘context validity’ within which atomistic notions may be considered both independently and in connection to other interrelated factors. This moves away from the traditional and rather unproductive validity-reliability binary opposition. (Within Weir’s model reliability is subsumed under scoring validity.) It is this symbiotic relationship between the whole and the parts that can lead to some useful cross-

fertilization between specialist assessment practice in high-profile international tests and local in-house practices that focus on learning but may be under-developed in assessment. This holistic framework underlines the view that far more is relevant to assessment validity than was previously assumed.

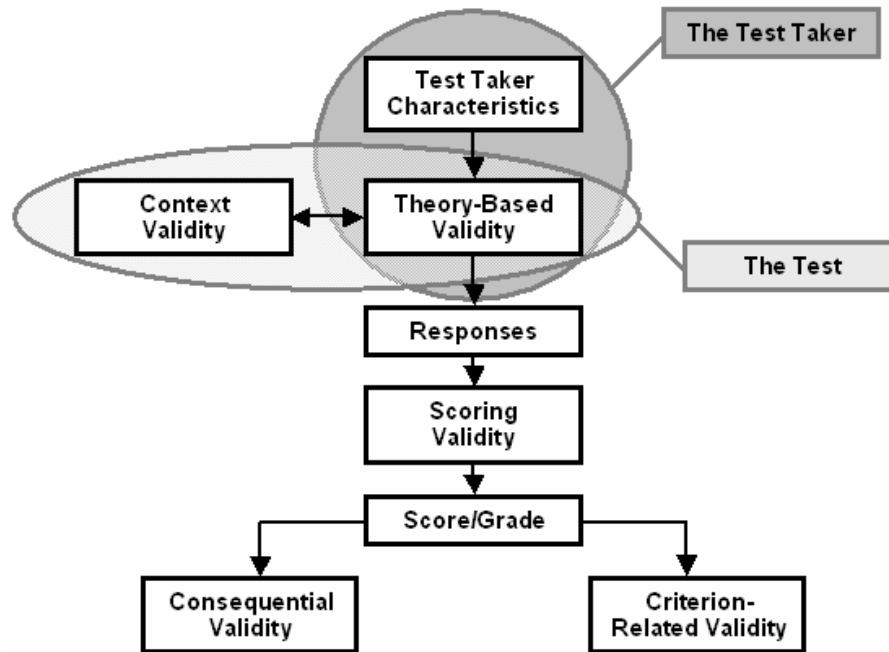


Figure 1 An Outline of the Socio-Cognitive Test Validation Framework (Weir 2005, p.2)

Scoring Validity

Scoring validity is used by Weir and Shaw (2005) as a super-ordinate for “all aspects of reliability” (p.5).

Scoring validity accounts for the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in their marking, are as free as possible from measurement error, stable over time. Consistent in terms of their content sampling and engender confidence as reliable decision making indicators. (Weir and Shaw, p.5)

Naturally, design features do influence both inter- and intra-rater reliability. The number of categories and the number of levels, the complexity of the concepts involved, the clarity of the wording, the distinctions between categories, for example, all influence the raters’ ability to be consistent.

Henning (1987) suggests using a Rasch model in the language assessment field partly

because of the ease of interpretation of the results and the low number of participants needed for the model to be operative (p. 116). To support in-house assessment, a simple Rasch model is preferred as the best means of considering atomistic traits of measurement within a holistic framework.

The Rasch approach is simple, not simplistic: The aim is to develop fundamental measures that can be used across similar appropriate measurement situations, not merely to describe the data produced by administering Test *a* to Sample *b* on Day *c*. Rasch modeling addresses itself to estimating properties of persons and tests that go beyond the particular observations made during any testing situation (Bond and Fox, 2007, p.143).

According to Yang and Kramer (2007), the Rasch model is a “mathematical probability model that allows for the investigation of dimensionality and the ordering of items on a measurement continuum” (p. 163). In this model, items and persons are measured on a common interval scale, and the estimates of persons and items are independent of one another, which makes for less subjective measurement.

The Rasch model transforms raw data into equal-interval scales through log transformations of raw data odds and probabilistic equations (Bond and Fox (2001, p.7). Through this, the measure of a person is separated from the scale to which he or she is measured. The person measure, called the ability estimate, is reported in logits. The entire sample is placed on a logit scale, which is an interval scale where the interval between two and three, for example, is the same distance as that between three and four. An important assumption of most Rasch models is that the data are fundamentally unidimensional. One way in which this assumption is often checked is through the inspection of Rasch fit indices. An item that does not fit the Rasch model may not be measuring the primary trait (Yang & Kramer, 2007, p. 163).

Bond and Fox (2001, p. 77) point out that items in a Likert response format are likely to vary in their difficulty, and the items are also unlikely to contribute to the construct equally (p. 85), an assumption of Likert scaling. In this case, data are regarded as ordinal data and the Rasch model is used to transform the ordered Likert categories into interval scales. Linacre (2004) suggested guidelines for the validation of rating scale categories. While not all guidelines are relevant for all data analyses (p. 276), the guidelines most important for verifying the validity of

the rating scale for this study are (a) the outfit mean squares, (b) the frequency of the observations for each response category, and (c) minimum and maximum step distances between the categories.

When doing multi-faceted assessment only the Rasch model is mathematically available (McNamara, 1996, p. 257). Multi-faceted assessment has been chosen as the most appropriate theoretical tool for considering rater reliability: it can be used to examine the interaction between judges, students and each of four traits on the rating scales. According to McNamara, “this measure of the candidate’s ability results from an automatic adjustment of the candidate’s raw score to take account of what is known about the influence of these facets” (p. 128). Advantages of using a multifaceted analysis are that the effects the facet has on the raw score given to a candidate can be rated, that we can examine the bias between facets, and that fit statistics, which show scores that significantly departs from the expected responses.

The Challenges of Holistic Task-based Design – Assessment and Learning

Bachman (2002, pp. 470-471) discusses four aspects of test-design that represent the most relevant challenges. In my view, these are equally relevant to learning. Firstly (*challenge 1*), Bachman recommends that we should "ensure that test-design is both construct-based and task-based" (p.470). Bachman argues that focusing only on task specifications results in intractable problems of generalizability and extrapolation. Fulfillment of a task by achieving the stated purpose in terms of an outcome or final product is open to over-interpretation unless it can be demonstrated that facets of the test are based on valid pedagogical constructs. There are many ways at arriving at a competent-looking product, so factors such as the amount of modeling provided also needs to be considered. In the worst case scenarios assessed responses may even have been rote-learnt.

Bachman next refers to task characteristics "that are uniquely attributable to facets of the particular assessment being designed" (p. 470). In this respect, Bachman (*challenge 2*) warns against adopting an external framework, encouraging in-house designers to "select those characteristics that are most relevant to their own particular testing situation".

A third point (*challenge 3*) underlined by Bachman is the need to define "specific

areas of language ability to be assessed" (pp. 470-471). In relation to challenge 2 above, we might also underline that this challenge involves defining the most relevant aspects relating to holistic rather than atomistic concepts such as the systemic characterization of modality. Modality has linguistic realizations that help express the important holistic academic function of expressing an academic author's level of commitment to the truth of claims. Such claims need to be based on the strength of evidence provided.

Finally Bachman recommends using a "variety of procedures to collect information about test performance" (p.471). Norris et al. (2003, p.397) suggest that assessment tasks that "require language use for their accomplishment" should be (1) "complex, skills-integrative and goal-oriented." They also state that specific criteria for task accomplishment should be determined by specialist informants. But they also argue that, in addition to task-specific criteria, a different set of more global "criteria for holistic examinee abilities" (p.397) are also required across performances. Relevant examinee abilities can only then be determined with respect to both "particular tasks" and "a domain of tasks" (p.397) (*challenge 4*).

Norris et al.(2003) consider that the most relevant focus is on the construct of task performance and that we need to consider the relationship between task difficulty in terms of cognitive processing and task performance (*challenge 5*). Bachman (2002, p.466) makes an essential link between task difficulty and context, in particular the test-taker's ability. Both Bachman and Norris et. al. (2003) consider the three aspects of difficulty proposed by Skehan (1998, p.88): code complexity, cognitive complexity and communicative stress involved in the demands of processing and accomplishing the task components. It can be noted that in any assessment, these three factors are not directly attributable to tasks per se. For Bachman, code complexity "is uniquely a characteristic of test tasks themselves" (p.466). Bachman goes on to argue that cognitive complexity and communicative stress are not characteristics that can be unilaterally assigned to task. They are rather "functions of the interaction between the test-taker and the task" (p. 466).

However, as even code complexity involves predicting the language needed to accomplish the task, we might also suggest that task participants may find ways of compensating for gaps in their language ability, so even code complexity is not uniquely a feature of the task. Difficulty therefore has to be assessed holistically by considering

the interactions between all of the components involved in the task. Task conditions *and* the personal characteristics of test-takers inevitably intervene.

Task design: Covering an Adequate Sample (Solution 1)

One implication of these challenges is that the potential for extrapolation from one task is limited because it lacks content validity. Davies et al. (1999) characterize content validity as “a conceptual or non-statistical validity, based on a systematic analysis of its test content” (p.34). The purpose of such an analysis - “to determine whether it includes an adequate sample of the target domain to be measured” (p.34) – is clearly important when a holistic view of language and language learning is a central tenet. Leather and van Dam (2003) underpin the holistic nature of an ‘ecological approach’ stating the following premise:

The premise that most clearly characterizes an ecological approach to language acquisition is that language behavior always involves more than can be captured in a single frame or script (p.13).

Course assessment needs to be applied to a comprehensive and varied sample of students’ work. Selecting a broad range of related tasks within the domain allows students to exercise choices from a broad range of knowledge and skills in relation to both the tasks themselves and their own competence.

Robinson and Gilabert (2007) make a “fundamental pedagogic claim” based on their cognition hypothesis: “that pedagogic tasks should be *designed* and then *sequenced* for learners on the basis of increases in their cognitive complexity” (p.162). Robinson and Gilabert (2007, p.164) provide a detailed rationale of task classification. While the outcome of our background discussion leads to the premise that *perceptions* of task difficulty and complexity will vary across individual performers and task-performance contexts, the following simplified list of criteria has been considered most relevant for task design in our context. (See also column 4 in figure 2 below.)

Cognitive complexity

- Number of skills needed
- Evaluative, inferential reasoning demands
- Integration of skills
- Familiarity with the task

- Predictability and fixity of the task structure

Level of language sophistication needed

- Specific language requirements that cannot be avoided
- Stylistic, or genre considerations

Task conditions

- Time allowed for the task
- Size of the deliverable
- Team or individual work
- Support given to complete it
- Communicative stress potentially caused by the task (such as spontaneous public spoken performance)
- Importance of the results

Different but inter-related genres of academic communication need taking into consideration. In the Petroleum Institute context, an initial task specification was drafted for a semester-long project for each of eight written and spoken deliverables that led to the production of a full research report and presentation. These have been classified in order of difficulty (figure 2 below), 1 being the most difficult, 8 the least by all five teachers and by six students. For each group the total figures were used to make an average score. For example, three teachers chose the source evaluation as most difficult, one as second most difficult and one as third most difficult. This was therefore the most difficult with a total of 8 and an average of 1.6. In agreement with Bachman's (2002) view of perceptions of task difficulty, the students had very different perceptions to the four instructors. There was also a range of answers within each group, but students tended to agree more, with unanimous agreement on the most difficult task, for example.

Task-Deliverables <i>In order of use</i> Team-written unless stated	Order of Difficulty as assessed by 6 students	Order of Difficulty as assessed by 5 teachers	Relevant Criteria
Minutes of Meetings	order & average 8 (7.5)	order & (average) 7 (6.8)	Short and relatively simple, low grade weighting, 2 nd draft graded.
Memorandum of Understanding	5 (5.66)	6 (4.83)	Unfamiliar, but relatively short and simple, low grade weighting, 2 nd draft graded.
Source Evaluation (individual task)	7 (6)	1 (1.6)	Requires critical reading and evaluative reasoning, sophisticated use of modality. Individually written. Awarded grades lowest of all tasks.
Research Proposal	2 (2.66)	5 (4.2)	Proposal for a semester-long project, some time pressure, synthesis skills in background reading section, research knowledge and organization needed.
Progress Discussion	6 (5.83)	8 (7.4)	Format familiar, spoken and informal, support from teacher.
Full Written Research Report	1 (1)	2= (3.4)	Longest, most complex document requiring multiple skills, sophisticated language and reasoning, good organization and genre awareness. High grade weighting on first draft.
Presentation Storyboard	3 (2.83)	2= (3.4)	New and unfamiliar task, requiring a combination of media creativity and planning, visual and verbal skills.
Multi-media Research Presentation	4 (4)	2= (3.4)	Communicative stress of public performance, complex task requiring multiple skills, sophisticated speaking ability, but some familiarity with presenting.

Figure 2 Providing Adequate Coverage – 8 Research Project Tasks

We can note from this micro-investigation that an activity like the source evaluation can be identified as one extremely relevant focus for the next course, given the low grades obtained and the difference in perceptions between students and teachers. It appears that students under-perform because they remain unaware of the difficulty. In contrast, grades awarded for the most difficult task according to students, the full team-written report, were high. Another point that arises here is that a source evaluation task is needed early in terms of the research project, but in terms of cognitive complexity studies (Robinson and Gilabert,

2007), it should be introduced later. The fact that the order of use conflicts with the order of difficulty has pedagogic implications in terms of the amount of support provided by the teacher and the need to modify the task itself.

In addition to these team-based research tasks, several non-research related individual tasks are learnt and assessed at different stages of the course, with the expectation that a team process will also impact on individual performance. Only one of these, the source evaluation to support the team's literature review, is directly linked to the research project. All other individual tasks are essay-style, single-drafted free-writing questions performed under time restrictions. (Approximately one hour.) These include a diagnostic writing reflecting on teamwork in the previous course, a mid-semester test answering a question on a seminar topic such as effective listening, a reflective writing task, repeating the diagnostic task, in which students reflect on teamwork at the end of the current course and a final test repeating the mid-semester task on a different seminar topic. It is assumed that students' holistic competence cannot be fully assessed by any of these tasks alone. In combination, they provide an adequate sample from which holistic competence can be developed and within which teachers and students can identify the most relevant individual focus for this course.

Providing a Holistic Definition of Academic Competence (Solution 2)

As a construct "is generally defined in terms of a theory" (Davies et. al. 1999, p.31), the view of competence that underlies the learning and its assessment needs to be made as explicit as possible. The construct validity of a test is defined by Davies et al. as "an indication of how representative it is of an underlying theory of language learning" (p.31), in this case, a holistic task-based approach that aims to develop academic competence.

There are so many areas related to academic competence that are holistic in nature that 'holism' itself can be usefully viewed as a guiding philosophy of education. (For an earlier discussion of holism in relation to designing integrated task-based units of learning, see Nunn, 2006, 2007.) Constructs to consider are (1) the holistic nature of language use itself, as expressed in fields such as systemic linguistics, pragmatics and discourse analysis (2) the holistic nature of competence, (3) the interlocking nature of SLA principles (See for example the ten SLA principles of instructed language learning in Ellis, 2005), (4) the range of genres that students need to be familiar with and the frequent need to integrate a broad range of linguistic and extra-

linguistic skills (5).

Competence is always re-negotiated in every context where it is evoked (*challenge 2*), but assessment is more transparent when an effort is made to specify as explicitly as possible the view of competence that underlies the assessment rubrics. One important aspect of holistic academic competence proposed in Nunn (2007) is that it is only partially available to any individual, and only partially exploited in any single context.

1. Competence in academic communication encompasses various **interlocking components** of usable **knowledge** and the **skills** and **abilities** needed to put these into practice within an academic **community**. The main components are **generic, pragmatic, discursal, strategic, interpersonal** and **linguistic**. Competence includes **skills** in areas related to both **written** and **spoken** language and certain **adaptive, cognitive and strategic skills** such as the ability to read a text critically in order to extract relevant information that is supported by adequate evidence (an important critical thinking skill) to support and inform original research. **Creativity** is also a characteristic of competence. The sum of these and other components amounts to something very large and only certain aspects will be called upon in any one **context**.
2. Individual competence is always **partial** and subject to **compensation** and development both for local and global use. Total competence is beyond the range of any individual, group of individuals, or indeed of any single community, but competent users and members of communities will legitimately **compensate** for weakness in one area with skill or knowledge in another.

Figure 3 Aspects of Holistic Academic Competence (modified from Nunn, 2007)

Allowing for Compensation (Solution 3)

Fettes (2003) argues, “individuals never really ‘acquire a language’ in the sense of being able to reproduce the whole system in all its dynamic complexity” (p.37). Coping with partial elements of a larger system that is “pre-existent and external to any individual agent” (p.37) is therefore the norm. Languages only exist within and between people so this larger system is not to be seen as a fixed entity.

The problem with atomistic *assessment* is that any particular atomistic test may identify a (rare) gap in the knowledge of a competent student who is able to compensate when engaged in a holistic task. Less competent students may also be able to score highly in an atomistic test without having the ability to apply this knowledge to a holistic task in combination with other aspects of competence. A holistic view therefore implies that to produce relevant and applicable results in any assessment context, students need to be given a range of tasks that provides the

opportunities to both display all aspects of their competence and to compensate for those aspects they do not possess.

Providing Task-familiarity (Solution 4)

It has already been pointed out that some tasks are required early within a particular course, where they would be best introduced later in terms of familiarity and complexity. Task familiarity is important because students doing a new task are confronted with too many cognitive and organizational demands to function at their optimum level. Students are therefore rarely assessed on a completely unfamiliar task, but neither are they overly primed on a limited number of tasks, as they need to demonstrate that they can transfer their competence across tasks and genres. In our context, each research-related task is supported by a task specification document. A task specification needs to find a balance between supporting students with useful guidelines for tasks, in relation to a genre that might be new to some students, without providing an inflexible format that must be followed blindly by all teams regardless of particular characteristics of their own project.

Task familiarity is also addressed by using some previous examples of student-drafted documents, which are not models to imitate, but rather examples of similar documents produced for a different project for students to engage with and improve upon. Such reconstructive modeling activities are used to provide scaffolding for the least familiar tasks. Exercises are designed from competent examples of previous students' project work, which both practice language and create familiarity with the task type.

Designing Rating Scales to Highlight a Holistic Task-Category (Solution 5)

Having briefly considered the design of tasks, the second purpose of this study is to consider the design of rating scales to assess performance on these tasks. Hyland, (2004) provides a scoring rubric for an argumentative essay which illustrates a three-part classification: format and content (40%), organization and coherence (20%), sentence construction and vocabulary (40%). The highest level for format and content is worded as follows:

Fulfills task fully; correct convention for the assignment task; features of target genre mostly adhered to; good ideas/good use of relevant information; substantial concept use; properly developed ideas; good sense of audience (Hyland, 2004, p.176).

Given our current focus on holistic tasks, a separate category is proposed for holistic task fulfillment to avoid overloading one category and to allow for both a holistic overview of the task addressed and a separate evaluation of detailed argumentation within and between sentences at paragraph level. One category is holistic in scope looking at the whole piece, the other focuses more atomistically on detailed argumentation, on the expression and support of particular arguments within sub-sections. The example below contrasts the two categories at the ‘excellent’ level of the scales for a research report written in small teams.

Holistic task fulfillment	Content: Detailed argumentation within paragraphs
<p>Task fully understood. The writers demonstrate full awareness of the holistic task requirements, as outlined in the task specification but also clearly define the task in terms of their own research project.</p> <p>Virtually all parts of the task are fully addressed.</p> <p>Virtually all of what is addressed is relevant.</p> <p>All references to source reading is fully acknowledged and used appropriately.</p>	<p>Paragraphs contain an adequate quantity of relevant and accurate information.</p> <p>Virtually all statements are consistently and qualitatively well supported with evidence and/or sound argumentation.</p> <p>'Evidence' can include: examples, facts, primary and secondary data, reference to, quotation from, authoritative sources.</p> <p>The writers demonstrate the ability to express a fully appropriate level of confidence in the evidence using modality.</p>

Figure 4 Holistic Task Fulfillment Opposed to Content as Detailed Argumentation

While it is highly relevant to highlight task-fulfillment within a task-based approach, it is important to emphasize that practising 'holism' is not an excuse for neglecting detailed content or linguistic knowledge, as expressed in *challenge 4*. (See Nunn, 2006 for a detailed description of the design of task-based units as an attempt to address this issue and Nunn 2010 for a previous version of rationale for combining holistic and atomistic rubrics.)

The pyramid (figure 5 below) expresses the relationship between categories. The broader macro-structural elements are rated in the first two categories, rating (1) the extent to which the different components and generic aspects of the task are fully addressed and the extent to which all sections are relevant to the task, and (2) the overall organization of the report including the development of the argumentation and the coherence of the macro-structure between sections such as the discussion of methodology and the results.

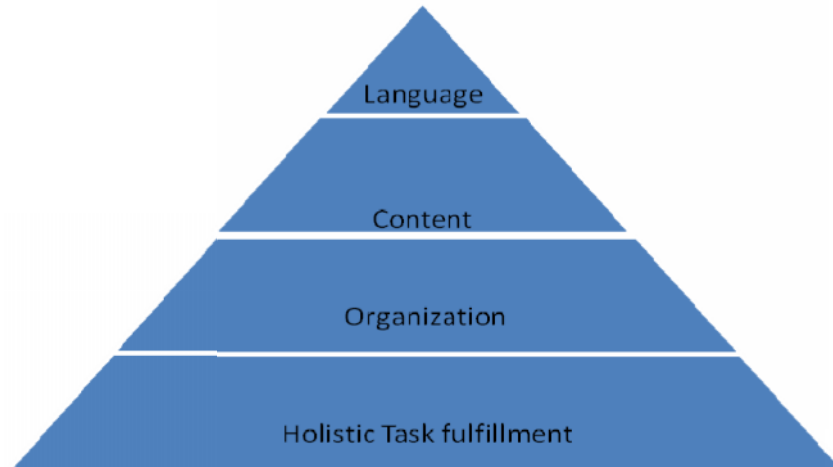


Figure 5 The Assessed Components of Holistic Task Fulfillment

The more detailed argumentation within subsections (paragraphs) and the linguistic choices within and between utterances are rated in the third and fourth categories (figure 6 below). In the task/genre category a whole paragraph or section may be irrelevant to the task. In the content category, which emphasizes argumentation at paragraph level, a sentence may be irrelevant although the subsection is highly relevant to the overall task. Different areas of competence are thus identified within the task framework to provide usable feedback both in detail and in relation to the whole task. This responds to *challenge 3*, the solution being to provide descriptions of particular language and content areas highlighted for a course.

Content : Detailed Argumentation Within Paragraphs	Language
<p>Paragraphs contain an adequate quantity of relevant and accurate information.</p> <p>Virtually all statements are consistently and qualitatively well supported with evidence and/or sound argumentation.</p> <p>'Evidence' can include: examples, facts, primary and secondary data, reference to, quotation from, authoritative sources.</p> <p>The writers demonstrate the ability to express a fully appropriate level of confidence in the evidence using modality.</p>	<p>Style and expression highly developed and engaging, with a consistent level of formality fully appropriate to the task.</p> <p>Sophisticated and appropriate use of a broad repertoire of grammar and vocabulary.</p> <p>Excellent coordination between sentences within paragraphs.</p> <p>Clarity and accuracy are of a high standard. Evidence of thorough, careful proofreading. Correct spelling and punctuation.</p>

Figure 6 Atomistic Content within a Holistic Writing Task

The four-part framework used above remains constant across task assessments to provide stability and comparability, but the wording of the rubrics varies according to the specified focus of the task. For example, the content and language scales (figure7 below) are used for a multi-media spoken presentation.

Content : Detailed Argumentation Within Presentation Sections	Language
<p>Information is fully adequate in quantity, relevance and accuracy.</p> <p>Virtually all statements are consistently and qualitatively well supported with evidence and/or persuasive argumentation.</p> <p>(“Evidence” can include: examples, facts, primary and secondary data, reference to, quotation from, authoritative sources but may also be in the form of visuals such as charts, images.)</p>	<p>Very good pronunciation and delivery. Fluency and clarity are combined to make the talk enjoyable to listen to. Excellent use of non-verbal communication (eye contact, gesture, etc.)</p> <p>Sophisticated and appropriate use of a broad repertoire of grammar and vocabulary. Excellent coordination between utterances.</p> <p>Text in slides/images/graphics are readable and error free.</p>

Figure 7 Rewording According to Task – A multi-media Presentation Task

Providing Students (and Teachers) with Common and Relevant Goals (Solution 6)

A rating scale provides students with a realistic goal by describing the performance just above their present level. It is reasonable to assume that in graded courses in institutional learning, in subjects which draw on so many different aspects of competence, what is seen to be emphasized for evaluation is more likely to be learnt. To be seen as relevant, in-house holistic assessment criteria need to be perceived as a central part of the learning process. Rating scales cannot provide a complete description of the competence criteria highlighted for particular courses. They can only act as a summary of the skills and knowledge that underlie competent performance in relation to a particular theoretical and practical framework. They help provide meaningful reports of all assessed deliverables to stakeholders alongside task descriptions and recorded samples of work produced. They also guide the teaching process as they are based on task specifications of both assessment tasks and pedagogical tasks, providing teachers (and students) with descriptions of the criteria used to define competence in relation to particular tasks. By summarizing different levels of competent performance, they also help teachers and students set themselves achievable goals. This can create a kind of solidarity between teachers

and students who share the goal of developing the competence that can lead to improved performance. By training students in the use of rating scales, it is possible to involve them in both peer- and self-assessment. The scales discussed above have been extensively trialed with students as well as being used by instructors.

Results for Scoring Validity in One In-house Context (*Challenge 6*)

A team of three student raters were employed and trained to rate six writers on a simple 50-minute free-writing essay task. An important focus of this study is the separate use of a task category. It was noted that total agreement was reached by all three raters in the task category on all of the scripts except one, in which only one rater was one point apart. The results of the Rasch analysis indicate that there are large gaps between the items, Language, Content, Organization, and Task (Figure 8). This is usually not the case in standard testing situations as the items might normally be expected to group around the midpoint, but it is consistent with the situation in this context. For example, Language has a logit score of 2.53. This means that the item is very difficult for the students. Assuming that all items are weighted equally, this item is having the most serious consequences for the final score in comparison with the other items.

In the trait measurement report, Task is almost a mirror of Language and is rated as very easy, with a logit score of -2.72, suggesting that students have fulfilled the holistic task requirement better than the language requirement. Again, more samples may help even the scores out and possibly center them on the scale. Assuming it is a rating rather than an achievement issue, this is possible as the raters become more familiar with the rating scales. For the rating overall, the rating scale itself was reliable with a score of .93 (figure 8).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N Traits
59	18	3.3	3.25	2.53	.45	.71	-.9	.72	-.8	1.33	4 Language
69	18	3.8	3.84	.54	.46	1.31	.9	1.31	.9	.66	3 Content
73	18	4.1	4.08	-.35	.48	.99	.0	.88	-.1	1.06	2 Organiz.
82	18	4.6	4.71	-2.72	.57	.58	-1.0	.48	-.3	1.34	1 Task
70.8	18.0	3.9	3.97	.00	.49	.90	-.2	.85	-.1		Mean Count:4
8.3	.0	.5	.52	1.88	.05	.28	.8	.30	.7		S.D.

Figure 8 Traits Measurement Report

RMSE (Model) .49 Adj S.D. 1.82 Separation 3.67 Reliability .93
 Fixed (all same) chi-square: 54.2 d.f.: 3 significance: .00

Inter-rater Reliability

Using Rasch Analysis, the raters were grouped together in the minus logit area (Figure 9). However, the number of scripts in this sample is small and it is reasonable to assume that the raters would even out more with a larger sample. While rater reliability remains a major challenge in this and other holistic assessments, according to Wright (2001), this does not mean that the scale is unreliable but rather that, with all its potential for assessing various levels of proficiency, it is not being fully utilized with the participants in this study. Wright also suggests that there is a correction available using the logit scores and the standard errors. In fact, rather than being unreliable, the raters fit the model well and there was no misfitting rater.

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Infit S.E.	Outfit MnSq	Estim. ZStd	Exact Obs %	Agree. Exp %	N Raters			
91	24	3.8	3.83	-1.16	.41	1.42	1.3	1.28	.7	.62	41.7	56.7	2
95	24	4.0	4.01	-1.85	.42	.57	-1.5	.55	-.9	1.40	58.3	58.7	1
97	24	4.0	4.11	-2.21	.43	.75	-.7	.72	-.3	1.23	58.3	58.0	3
94.3	24.0	3.9	3.98	-1.74	.42	.92	-.3	.85	-.2	Mean (Count: 3)			
2.5	.0	.1	.12	.43	.01	.36	1.2	.31	.7	S.D.			

Figure 9 Raters Measurement Report

RMSE (Model) .42 Adj S.D. .10 Separation .23 Reliability .05
 Fixed (all same) chi-square: 3.2 d.f.: 2 significance: .20

Using classical test theory to assess inter-rater reliability (figure 10), the results confirm that there was an acceptable rate of agreement.

		Rater1	Rater2	Rater3
Rater1	Pearson Correlation	1.000	.865(*)	.805
	Sig. (2-tailed)	.	.026	.053
	N	6	6	6
Rater2	Pearson Correlation	.865(*)	1.000	.799
	Sig. (2-tailed)	.026	.	.057
	N	6	6	6
Rater3	Pearson Correlation	.805	.799	1.000
	Sig. (2-tailed)	.053	.057	.
	N	6	6	6
* Correlation is significant at the 0.05 level (2-tailed).				

Figure 10 Rater Reliability – Classical Test Theory

The writers (figure 11 below) separated well in this sample, as did the items. As with the raters, there were no misfitting items or participants. This indicates that the rating scale is working well in this study. As for the rating scale, Linacre (2004) suggests that no category should be more

than five logits distant from a neighboring category, but that there be at least one logit of a step distance between the categories for a rating scale with five categories. For this data set, there were no occurrences where a category was more than five logits apart from the neighboring category and where there was less than one logit between the ratings. This shows that the rating scale separated the writers efficiently.

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrim	N Writers
56	12	4.7	4.79	3.10	.71	1.30	.7	.95	.3	.76	4
55	12	4.6	4.69	2.62	.67	.43	-1.4	.34	-.6	1.58	2
48	12	4.0	4.00	.07	.57	1.56	1.2	1.57	1.2	.45	3
47	12	3.9	3.92	-.24	.56	1.21	.6	1.19	.5	.77	5
39	12	3.3	3.23	-2.62	.55	.45	-1.6	.43	-1.6	1.59	1
38	12	3.2	3.14	-2.93	.56	.63	-.8	.63	-.8	1.38	6
47.2	12.0	3.9	3.96	.00	.60	.93	-.2	.85	-.2	Mean (Count: 6)	
7.0	.0	.6	.64	2.31	.06	.44	1.1	.43	1.0	S.D.	

Figure 11 Writers Measurement Report

RMSE (Model) .61 Adj S.D. 2.23 Separation 3.66 Reliability .93
 Fixed (all same) chi-square: 81.4 d.f.: 5 significance: .00

Conclusions

It has been argued in this paper that a holistic approach requires awareness of a broad range of potentially relevant aspects of learning and assessment before selecting the most relevant to focus on in local contexts. In contrast to external examining, in many in-house university assessment contexts, highly qualified subject teachers have the ultimate and exclusive right to determine their own grading. Ideally, participants, including students, will have been given the opportunity to influence and modify the description of these criteria. The criteria are a developmental tool and periodical modifications should be seen as a normal part of the process.

In our context, once a holistic framework has been established, it becomes possible to focus on more atomistic aspects of the system for more detailed investigation, such as rating scale design, task difficulty and scoring validity. An atomistic focus is not an end in itself, it is not opposed to holism, it is rather included within a holistic focus. The most relevant foci for particular learners at a particular stage of learning will always vary. No holistic course can focus on all aspects of competence, but attention to atomistic detail may be a very relevant focus at certain stages of the course. One aspect identified in our context is the need to regularly address language ability throughout the course. Once the more holistic categories of the framework have been established with learners, it is then possible to focus in detail on particular weaknesses for

special attention as a regular feature of a holistic course.

References

- Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19, 453-476.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Byrnes, H. (Ed.) (2006). *What kind of resource is language and why does it matter for advanced language learning? An Introduction*. In *advanced language learning. The contributions of Halliday and Vygotsky* (pp. 1-28). London: Continuum.
- Byrnes, H. (Ed.) (2006). *Advanced language learning. The contributions of Halliday and Vygotsky*. London: Continuum.
- Byrnes, H. (2002). The role of task and task-based assessment in a content-oriented collegiate FL curriculum. *Language Testing*, 19, 419-37.
- Davies, A., A. Brown, C. Elder, K. Hill, T. Lumley, T. McNamara (1999). *Dictionary of language testing. Studies in language testing 7* Cambridge: CUP.
- Ellis, R. (2005). Principles of instructed language learning. In P. Robertson, P. Dash and J. Jung (Eds). *English language learning in the Asian context* (pp. 12-26). Pusan: The Asian EFL Journal Press.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Boston, MA: Heinle & Heinle.
- Hyland, K. (2004). *Genre and second language writing*. The University of Michigan Press: Ann Arbor.
- Leather, J. & J. van Dam (2003). *Ecology of language acquisition* Dordrecht: Kluwer.
- Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith & P. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.
- McNamara, T. (1996). *Measuring second language performance*. Harlow, UK: Longman.
- Norris, J., J. Brown, T. Hudson, W. Bonk. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language*

- Testing*, 19, 337-346.
- Nunn, R. (2006a). Designing holistic units for task-based learning. In *Study of second language acquisition in the Asian context*, P. Robertson & R. Nunn (Eds.), Asian EFL Journal Press, Pusan, Korea, pp. 396-420. (Also available at http://www.asian-efl-journal.com/site_map_2006.php)
- Nunn, R. (2006b). The Pragmatics of Cooperation and Relevance for Teaching and Learning. *Asian Linguistics Journal*, 1, 5-16.
- Nunn, R. (2007). Re-defining communicative competence for international and local communities. *The Journal of English as an International Language*, 2, 7-49. (Available at: http://www.eilj.com/2007_Index.php)
- Nunn, R. (2010). Rubrics, relevance and task-based oral performances. In A. Jendli and C. Coombe (Eds.), *Developing oral skills in English: Theory, research and pedagogy* (pp. 263-289). Dubai: TESOL Arabia Publications.
- Randall, M. with B. Thornton (2001). *Advising and supporting teachers*. Cambridge: Cambridge University Press.
- Robinson, P. & R. Gilabert. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL - International Review of Applied Linguistics in Language Teaching*, 45 (3), 161-176.
- Samuda, V., and M. Bygate. (2008). *Tasks in second language learning*. Basingstoke: Palgrave Macmillan.
- Shaw, S. and C. Weir. (2008). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.
- Sperber, D. and Wilson, D. (1995). *Relevance: communication and cognition* (2nd Ed). Oxford, England: Blackwell.
- Toolan, M. (2003). An integrational linguist view. In Leather, J. & J. van Dam, *Ecology of language acquisition* (pp.123 -129) Dordrecht: Kluwer.
- Vygotsky, L. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Weir, C. (2005). *Language Test Validation: an evidence-based approach*. Oxford: Palgrave.
- Weir, C. and S. Shaw. (2005). Establishing the validity of Cambridge ESOL writing tests: towards the implementation of a socio-cognitive model for test validation. *University of*

Cambridge ESOL Examinations Research Notes, 21, 10-14.

Wright, B. D. (2001). Separation, reliability, and skewed distributions. *Rasch Measurement Transactions*, 14(4), 786.

Yang, C. L., & Kramer, G. A. (2007). Using Rasch analysis to construct a clinical problem-solving inventory in the dental clinic: A case study. *Journal of Applied Measurement*, 8(2), 161-174.