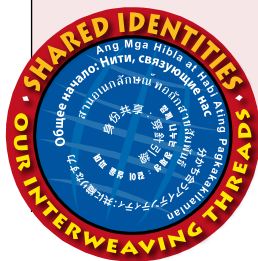


# Shared Identities: Our Interweaving Threads



## Evaluation in the curriculum development process

Brent A. Jones

*Kobe Gakuin University*

Martin Weatherby

*St. Thomas University*

### Reference data:

Jones, B. A., & Weatherby, M. (2009). Evaluation in the curriculum development process. In A. M. Stoke (Ed.), *JALT2008 Conference Proceedings*. Tokyo: JALT.

This paper outlines the curriculum development efforts for a tertiary-level business English program, with special emphasis on evaluation methodology and findings. A brief overview of each phase of the project is presented using a generic instructional design model. After visiting both progress and obstacles in these endeavors, we go on to highlight the crucial role of both formative and summative evaluation in the overall curriculum development framework, and offer our interpretation of preliminary feedback data as well as future directions.

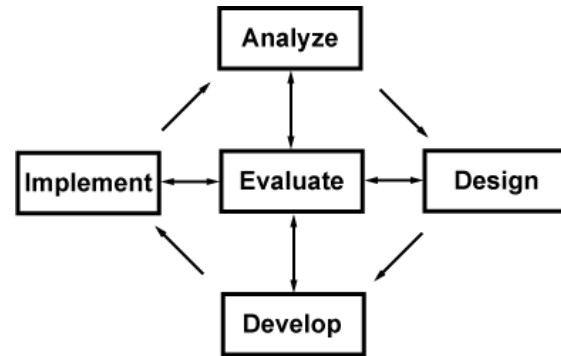
この論文では特にプログラム評価の方法論と研究結果に重きをおいて、大学レベルのビジネス英語カリキュラムの計画と開発について概要を述べています。カリキュラム開発の段階を一般的な教育デザインモデルを使って簡潔にあらましを述べた後、現状と問題点を踏まえ、カリキュラム開発全体のフレームワークにおける形式的で累積的な評価が大変重要であることを強調しています。今までの調査結果の一部に対しての解釈を将来の方向性を含めて示しています。

**H**ow are we doing? This deceptively simple question has been the driving force behind the evaluation phase of an ongoing curriculum development project. Our quest in these pages is to outline how our approach to evaluation is evolving. Specifically, we will put under the microscope our curriculum development efforts in a Business English program for undergraduates at a private university in western Japan, with special attention focused on the evaluation plan. Our discussion is organized as follows: First, we will provide a brief overview of the project, including our initial approach to evaluation. This is followed by an objective analysis of a subset of data together with suggested improvements for future

evaluation endeavors. We conclude with a short discussion of what we learned from this analysis and how we plan to proceed with the project.

### Overview

This project started in the spring term of 2005 and involved developing a Business English curriculum for second and third year students in the Faculty of Economics at a private university in western Japan. The first author has been researching instructional design (ID) as it relates to language education and decided to approach this challenge from an instructional design perspective. Specifically, a generic ID model involving Analysis, Design, Development, Implementation, and Evaluation (ADDIE) was selected as a framework for the project. We were attracted by the flexibility and simplicity of this oft-cited model, and at the same time recognized that these five elements are at the core of other influential ID frameworks. Whereas ADDIE is often depicted as a linear process in which each phase leads into the next, we soon realized the shortcomings of this approach and found that we were working on several phases of the project concurrently. We also realized that evaluation was needed throughout the process (formative evaluation) as well as during follow up (summative evaluation). Figure 1 is our preferred representation.



**Figure 1. The ADDIE instructional design model with evaluate at the center**

The four levels in the program are Elementary Business English I and II, and Intermediate Business English I and II. The original requirements for Elementary Business English were sophomore standing and a B grade (70%) or better in the general English course (*sogo eigo*). The number of students eligible for the class was so low that this second requirement was lowered to a passing grade (60%) or better from the 2006 academic year. The requirements for Intermediate Business English are junior standing and a B grade in Elementary Business English I and II. Classes for each of the four levels meet for 90 minutes once a week for fifteen weeks.

To help readers better understand the scope of the project, we will offer a brief description of each phase. A more detailed description can be found in Jones (2007a).

## Analysis

Our first task was to get a better understanding of the context, including target-language proficiency levels, general knowledge, learning styles, attitudes, and expectations. Specifically, we wanted to find out as much as possible about entry-level competencies as well as exit-level targets. Our job would then be to figure out how to best bridge the *performance* gap. An existing Business English program was in place, so we were able to begin rapid-prototyping and testing various activities and task-chains right away. At the same time, this gave us access to student input and feedback.

Analysis also included (1) a review of relevant literature, especially in the areas of English for specific purposes (ESP), vocabulary acquisition, and communication strategies, (2) interviews with colleagues, administrative staff, business consultants, and the students, (3) classroom observation notes, (4) survey instruments with follow up interviews, and scores from the Test of English for International Communication (TOEIC). Most students had two recent scores on this test, one each from the beginning and end of their freshman year. We soon found however that the average scores on these tests were so low that they were of limited value. Overall though we gained a fairly good understanding of where the students were in their language-learning endeavors and where we wanted them to be. We were thus better prepared to begin the design phase.

## Design

Most of our efforts in this phase were aimed at writing clear performance objectives and compiling lists of design

decisions and guiding principles. We felt these would provide a firm foundation for developing instructional materials and streamline the future development process, but at the same time afford us some amount of flexibility in terms of handling a wide range of content, language difficulty, and teaching as well as learning styles. Example objectives include students demonstrating the ability to (1) read and respond to online postings of the teacher or classmates, (2) compose short email messages using a predetermined format, (3) speak about a product for two minutes, and (4) summarize the main points of mini-lectures on business topics. Our preliminary lists of design decisions and guiding principles are included in Appendix 1.

One of the main challenges we are confronted with is having only 90 minutes of class time each week. To increase the amount of exposure to the target language we decided early on that we would try to make use of a learning management system (LMS) and self-access materials. We also decided to develop 3-week modules looking at specific industries (e.g., travel, fashion, music) and business issues (e.g., ethical business, sustainability, marketing). We felt that this would help establish a sense of continuity from week to week and boost relevance and thus motivation. Another decision was to require students to complete a term project in each course. The idea was to help students develop a sense of self-reliance and responsibility by working on an individual project in spring and then focus more attention on communication and social skills via a group project in the fall term.

## Development

Development has proceeded along several lines. Our main efforts in this phase have included:

1. Writing syllabi and teacher manuals for each of the courses,
2. Compiling a database of vocabulary test questions targeting high frequency words as well as words from the academic words list (Coxhead, 2000),
3. Organizing and piloting 3-week modules as discussed above,
4. Developing handouts and other materials for various in-class activities (e.g., Reading Skills PowerPoint activity) and the term projects,
5. Preparing the LMS and online activities and materials (e.g., mini-lectures converted to QuickTime movie clips),
6. Recording podcast episodes and developing related activities,
7. Writing mini-quizzes and practice materials related to the course textbook, and
8. Preparing sample reports and presentations for the term projects.

A major part of our evaluation plan was to get student feedback on activities and materials. This was done through informal questionnaires that were presented via dictation practice. Survey results for the various activities are discussed below. A detailed description of the activities is

beyond the scope of this paper, but readers are directed to Jones (2007b) for an outline of two activities, a Reading Skills PowerPoint activity and a podcast listening activity (2007b).

## Implementation

One of our first endeavors was to begin piloting the 3-week modules. Follow up questionnaires revealed a need to simplify material and allow for more time for individual components. We attempted to modify these modules, but eventually opted to move them to the Intermediate Business program.

The LMS we decided on was Moodle. The university was already using two other systems, but we judged that Moodle offered more functionality and flexibility than the other two combined. The main two Moodle modules we have used are Forums and Wikis. As mentioned above, we were concerned about the limited class time and wanted to increase exposure to the target language. Forums have been used for disseminating information, gathering feedback and promoting student-to-student interaction. Wikis have been used by students to compile reports and other project-related materials. One other module that we used quite extensively was Quizzes, where we uploaded practice exercises made using Hot Potatoes software.

We decided to implement two different term projects, one in spring and one in fall. For the spring project students individually researched small and medium-sized local businesses and introduced their findings. With the fall term project, students worked in small groups (of 3 or 4) to

develop their own small-business plans and present their findings via PowerPoint. Term projects were implemented from the fall of 2006 and course teachers completed the first full cycle in the academic year of 2007. Some of the adjustments that were made included having students (1) prepare and present individual projects via a poster presentation, (2) embed their recorded voices in group-project PowerPoint presentations which were then uploaded to the Moodle site, and (3) evaluate the work of their peers in other classes for both projects. These adjustments were made to address time constraints and broaden the target audience for presentations.

Implementation of the various activities and materials has been staggered over the past several years. The Reading Skills PowerPoint activity and podcast listening activity were first introduced in the fall term of 2006, and were subsequently fine-tuned based on student feedback as well as our own observations.

### Evaluation

In developing our evaluation protocol, we looked to experts in various related fields, including performance training and human resource development (HRD). One of the best references we found was the classification scheme put forward by Kirkpatrick (1994). Table 1 outlines the levels and types of evaluation. This classification scheme provided us with a framework for building our evaluation plan. One point stressed by Kirkpatrick is that effective evaluation needs to proceed through each level.

**Table 1. Kirkpatrick (1994) evaluation classification scheme**

Level	Type of evaluation
1	<b>Reaction</b> evaluations – measure how those who participated in the program react to it.
2	<b>Learning</b> evaluations – can be defined as the extent to which participants change attitudes, improve knowledge, and/or increase skill as a result of the program.
3	<b>Behavior</b> evaluations – are defined as the extent to which change in behavior has occurred because the participant attended the training program.
4.	<b>Results</b> evaluations – are designed to determine the final results that occurred because the participants attended the program.

Another valuable resource in developing our evaluation plan that may be off the radar of most language-teaching professionals was the work of Phillips (1997) in the field of HRD. His advice was found to be so valuable that we compiled job aids for teachers in our program based on his ideas in the following areas: Evaluation Myths, Why Change?, Purposes and Uses of Evaluation, Complete Results-Based HRD Model, and Calculating Return on Investment (see Appendixes 2-6).

Based on a review of relevant literature, we developed an evaluation plan that included the following components: (1) standard course evaluations (administered by the institution), (2) individual student grades, (3) questionnaires/surveys for specific activities or materials, (4) Moodle logs, (5) samples of in-class and online work, and (6) teachers' in-class notes. Findings from numbers (2) and (3) are the focus of the remainder of this paper and are discussed in more detail in

the following section. We generally found that students were satisfied with the overall learning experience but did not make extensive use of the online components. Both teachers and students reported moderate or higher levels of language-learning progress and achievement of course objectives. Students judged some of the activities and materials as being beyond their proficiency levels but still rated them as being moderately to highly effective and enjoyable to the point where they would recommend them to their peers. The remainder of the paper deals mainly with an analysis of results from components (2) and (3) above.

### Analyzing our evaluation plan

The issue of the role of the students in course evaluation is clearly an important one for the course instructor/curriculum developer. From a rating of course objectives in the introduction, through dictation activities eliciting qualitative data registering student reactions to course components, to a post instruction survey instrument on the perceived utility of course components, the students are involved throughout the process.

The results of 12 scored assignments comprised of homework, classwork, tests, and quizzes have been tabulated by the researchers (total top score for the 12 assignments is 230), with seven assignments of 10 points each, one assignment of 15, two assignments of 20, one of 40, and one of 65. Some of these assignments are scored as binary choices (complete or incomplete), some are on an interval scale (seven out of ten correct on a quiz), and others are scaled on an interval scale using more subjective criteria (70 percent for a poster assessment). While providing

percentages for each assignment informs the students of their progress at each step, a more systematic assessment of the program for diagnostic purposes is attained for the instructor with the use of a Rasch analysis of objective measurements, using a partial credit model. The data from each assignment is then analyzed together, with interval data receiving a scale score based on standard deviation. This allows the instructor to analyze the contribution of the data in each assignment to the overall variation, so that items in assignments in an existing course can be retained, modified, or discarded in subsequent administrations of the course.

For the current study, a Rasch analysis of the post instruction survey instrument was conducted using the software program WINSTEPS. The survey consisted of a single page of 20 items, five items each in the four categories of: Textbook and Listening Tasks, Project, the Moodle Site, and Other Issues, the last category consisting of items on dictation activities, shadowing activities, online movie clips, random seat assignment, and whether the course could be recommended to other students. There are six response choices in a Likert scale, from 6 to 1: strongly agree, agree, slightly agree, slightly disagree, disagree, strongly disagree. In keeping with pedagogical goals, the survey was in the target language and was anonymous.

However, Trochim and Donnelly (2008) recommend translating survey instruments, taking care of differences in nuance, and note that language and cultural issues can confound an analysis and be a threat to validity. Nunan (1992, p. 232) defines the term survey, as data collection "... without attempting to manipulate the phenomenon/variables under investigation," so protection against the threat of

differences in language proficiency confounding the data collection must be considered. Even though the language learners had just completed a course with a central objective of being entirely in the target language, any assessment of the course should be conducted through the use of an equivalently translated instrument in their native language using back translation.

The sample for the analysis comes from 3 intact classes (N=105). The mean raw score of the data, 90.9, is considerably higher than a mean of 70, which would be the expected mean if the 20 items (with choices from one to six) were evenly endorsable for agreement/disagreement out of a minimum of 20 and a maximum of 120. Table 2 divides the summary statistics for the Rasch analysis into two facets: measures of person ability (in this case, willingness to endorse agreement) and item difficulty (difficulty regarding endorsement of agreement). The objective measurements of the Rasch analysis center the data at a measure as close to zero as possible for both facets, with fit estimates as close to one as possible, and the fit statistic is the infit. These results confirm that the students tended to over-endorse these items, and the model error for the person measure was fairly large at 0.27.

**Table 2. Summary statistics for person and item facets (n = 105)**

	raw score	count	measure	error	fit	reliability
person						
mean	90.9	19.9	0.97	0.27	1.02	0.88
SD	12.4	0.4	0.87	0.07	0.51	
item						
mean	4.77	104.5	0	0.11	0.99	0.87
SD	31.5	1.1	0.34	0.01	0.42	

The structure of the response categories (Table 3) is examined to determine whether it makes sense to collapse or retain categories in the rating scale, which would ensue if responses in any category were underrepresented. Objective measurement is primarily concerned with the fit of empirical data to a model, and does not regard the data itself as sacrosanct (there are many reasons, psychological and practical, why a respondent may not answer every item to the best of his or her ability). In this case, there are more than 10 responses for each category, and the structure shows monotonicity, meaning each category measures increases from the lowest category (1) to the highest (6) (Bond & Fox, 2001, p. 162). Fitness is good with all categories obtaining outfit estimates less than 2.0, for a well-functioning six-category rating scale, (Bond & Fox, 2001, p. 163).

**Table 3. Category structure of responses (1 = strongly agree, 6 = strongly disagree)**

category	count	measure	outfit mnsq	category measure
1	33	-0.19	1.14	-2.49
2	58	-0.01	1.22	-1.37
3	234	0.23	1.01	-0.63
4	586	0.58	0.8	0.19
5	728	1.14	0.93	1.41
6	451	1.81	1.03	3.18
missing	10			

Finally, the item measures and fitness estimates (Table 4) are examined. Again, the measures are centered on zero, with items found more difficult to endorse in increasing positive logits and items easier to endorse in increasing negative logits. For the subsections of the test, the student questionnaire survey shows more difficulty to endorse the less traditional, more innovative assignments and greater ease to endorse more traditional assignments. The following measures are the means for each subsection: Moodle (1.4), Other Issues, i.e. online movie clips, shadowing, dictation, etc. (-0.18), Project (-0.5), Textbook and Listening Tasks (-0.72). For the fit statistics, Bond and Fox (2001, p. 179) provide a rule of thumb for interpretation of the best fit of Likert scale survey data as the range from (0.6 - 1.4). In Table 3, misfitting items (M3, O3, M4, T1, and O1) are marked with an asterisk, with comments following the table. One quarter of the 20 items do not fit the model, requiring them to be modified on any further administration of the instrument and excluded from the current one.

**Table 4. Items in measure order with fitness statistics (\*denotes misfitting items)**

item	measure (error)	infit	item prompt
M2	0.79 (0.1)	0.96	I regularly accessed the Moodle site throughout the term.
O4	0.59 (0.1)	1.37	The online mini movies helped me understand what we were doing.
M4	0.4 (0.1)	*1.59	Our team used the wiki to prepare our business plan.
T4	0.29 (0.1)	0.77	The topics in the textbook were interesting.
M1	0.26 (0.1)	0.81	The Moodle site is helpful for studying business English.
P5	0.13 (0.11)	0.7	The project was an enjoyable way to study.
P4	0.09 (0.11)	0.67	The project helps students learn business concepts.
M3	0.07 (0.11)	*1.94	The activities for units 6 & 7 helped me prepare for the mini-quizzes.
T2	-0.03 (0.11)	0.71	The textbook helped me develop my English language skills.
P1	-0.04 (0.11)	0.7	The project helped me develop my English language skills.
T3	-0.05 (0.11)	0.77	The level of the vocabulary is appropriate.
O3	-0.08 (0.11)	*1.88	Changing seats each week is a good idea.
M5	-0.12 (0.11)	1.12	The teacher should continue to use Moodle with Business English classes in the future.
P3	-0.15 (0.11)	1.33	Each member contributed to completion of the project.



item	measure (error)	infit	item prompt
O5	-0.17 (0.11)	0.83	I would recommend Business English to my friends.
O2	-0.22 (0.11)	0.62	Shadow talking is useful for my language studies.
O1	-0.3 (0.12)	*0.56	The warm up dictation practice is useful for my language studies.
T5	-0.36 (0.12)	0.61	The listening tasks were useful for the study of business English.
P2	-0.53 (0.12)	1.32	Our group worked well together.
T1	-0.57 (0.12)	*0.58	The textbook is useful for studying business English.

Note: Items are in four categories: M=Moodle, O=Other, P=Project, T=Textbook and Listening.

Interpretation of fitness is likely the most difficult aspect of objective measurement with which the uninitiated must become familiar. The three items with infit estimates over 1.4 show too much variation (i.e., students who otherwise found similarly weighted items easy inexplicably found these difficult, and vice versa). The misfit is between the measure of person endorsability and item endorsability:

(1.94) The activities for unit 6 & 7 helped me prepare for the mini quizzes.

(1.88) Changing seats each week is a good idea.

(1.59) Our team used the wiki to prepare our business plan.

The other two misfitting items show too little variation; the observed cutoff between the zone where the probability that a student will endorse and not endorse (agree) with the prompt is too clearly delineated for the model, which

parses for variation. These items are over-structured and thus inefficient at indicating variation:

(0.58) The textbook is useful for studying business English.

(0.56) The warm-up dictation practice is useful for my language studies.

It was determined that an anonymous collection of this survey data would promote validity by avoiding the threat caused by social desirability. No one wants to give information that may make the provider or the receiver (in this case, the teacher who has yet to issue their grades!) look bad, as Trochim and Donnelly (2008, p. 123) note. Objective measurement provides detailed information about the variation in the data as a step prior to statistical analysis, and it is extremely useful to the researcher to identify its source. The collection of this data may be more fruitfully accomplished by a colleague or member of the administrative staff where feasible, with the student-participants informed of the purpose of the research and assured that their survey will not be accessed until final grades are submitted.

## Discussion

The above findings provide at least three main areas for improving our evaluation plan. First, to gather more meaningful data we need to reconsider our strategy of binary grading in assessment of students. The binary scoring on certain assessment items (either zero or full points) limits the usefulness of data. The original intention was to measure mastery as outlined in Mager (1997) and Shrock and Coscarelli (2000), i.e. whether or not students could

demonstrate the target skill. In practice, these scores often showed only that the students had or had not submitted an assignment. More meaningful data could be obtained if we had a scoring matrix for all items. This would give us an interval scale for these items that would allow for Rasch analysis, thus allowing us to evaluate better what is working in the program and what needs improvement.

Another issue that deserves our attention is the language used for questionnaires and other evaluation instruments. Specifically, we need to reconsider our initial plan of using the target language and treating these events as learning opportunities. As mentioned above, variability in language proficiency puts us at risk of confounding the data we collect via these instruments. A happy middle ground might be reached by including the target-language and first-language versions side by side.

The issue of anonymity for survey instruments and collection methods also needs to be reconsidered. Obviously we want students to respond as candidly as possible, but conducting surveys anonymously limits the value of data collected in that it cannot be correlated with other variables such as assessment results. We will heed the suggestions to invite a third party to conduct these surveys and explain that grading will be completed before the results are looked at.

## Conclusion

So, how are we doing? Overall, this systematic approach to curriculum development seems to be working. We are closing in on an efficient and effective combination of in-class and online components for our Business

English program. Again, we view instructional design and curriculum development as a cycle and will continue to revisit each phase and adjust as needed. The improved evaluation plan will help in these endeavors. This paper has argued for a systematic approach to course development and provided some examples of how this is being done in a Business English context. Granted, an ID approach to course development like that mentioned above does require quite a bit of effort and time. Still, the potential benefits make this a worthwhile effort and the return on investment should increase as our program matures (i.e., greater returns will come from less investment of time and energy). Greater accountability is being demanded of classroom teachers, materials developers, curriculum specialists, and other related professionals. One way we can meet these demands and the changing environment is to expand the view of our role as language teachers. Focusing more attention on customer (student) satisfaction should put us in a better position to deal with the changing environment in our field and raise the level of individual and collective professionalism. And isn't that the direction we want to be heading?

**Brent A. Jones** has been teaching and tutoring ESL/EFL for two decades, first in Hawaii and then in Japan and other parts of Asia. He has advanced degrees in Language Education and Instructional Systems Technology from Indiana University, School of Education. He is currently in charge of the Business English program at Kobe Gakuin University.

**Martin Weatherby** has taught EFL in Japan for over 19 years, and teaches a seminar in pragmatics and courses

in communication, TOEIC preparation, and Extensive Reading at St. Thomas University in Amagasaki, Japan. He has a M.Ed. degree from Temple University Japan and is working on a dissertation in pragmatic development in the doctoral program in Teaching English to Speakers of Other Languages at TUJ. He has recently been working on survey assessment and objective measurement using Rasch analysis.

### References

- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-38.
- Jones, B. (2007a). Instructional design in a business English context. In K. Bradford-Watts (Ed.), *JALT2006 Conference Proceedings*. Tokyo: JALT.
- Jones, B. (2007b). Digital components of a business English program. In *Japan Society for Educational Technology, Research Report of JSET Conferences, 07-4*, 107-114.
- Kirkpatrick, D. (1994). *Evaluating training programs: The four levels*. San Francisco: Berrett-Koehler.
- Mager, R. F. (1997). *Preparing instructional objectives: A critical tool in the development of effective instruction* (3rd ed.). Atlanta: CEP Press.
- Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.
- Phillips, J. J. (1997). *Handbook of training evaluation and measurement methods* (3rd ed.). Boston: Butterworth-Heinemann.
- Shrock, S. A., & Coscarelli, W. C. C. (2000). *Criterion-referenced test development: Technical and legal guidelines for corporate training* (2nd ed.). Silver Spring, MD: International Society for Performance Improvement.
- Trochim, M. K., & Donnelly, J. P. (2008). *The research methods knowledge base* (3rd ed.). Mason, OH: Cengage Learning.

## Appendix 1

### *Preliminary design decisions and guiding principles*

#### *Design decisions*

- A. Clear performance objectives will be established at both the macro (curriculum) and micro (task/activity) levels (see, for example, Mager, 1997).
- B. Criterion-referenced test items will be developed to clearly measure progress and performance (see, for example, Shrock & Coscarelli, 2000).
- C. The curriculum will target development of all four language skills (reading, writing, listening, and speaking) but will concentrate more effort on improving the receptive skills of reading and listening at earlier stages and productive skills later in the program.
- D. All components will have the underlying goal of increasing familiarity and confidence with high-frequency words in the English language.
- E. Attention will be focused on improving language competencies (including communication strategies), social skills and business competence.
- F. Non-native varieties of English will be respected.
- G. Efforts will be directed at raising cross-cultural awareness.
- H. Efforts will also be focused on nurturing positive language learning attitudes and beliefs.
- I. Individualized instruction will be implemented whenever possible.

- J. A repository of self-access materials will be developed to supplement face-to-face meetings.

#### *Guiding principles*

The following list of guiding principles will be the foundation of our development stage endeavors:

- A. Effective and efficient use of existing resources.
- B. Balance between face-to-face meetings and self-access materials.
- C. Balance between concept learning and procedural learning.
- D. Activities and materials that appeal to various learning styles.
- E. Activities and materials that are both relevant and intrinsically motivating.
- F. Teaching methodology based on accepted and emerging theories of learning.
- G. Activities and materials that promote success and boost confidence.
- H. Get students active within the first five minutes of any encounter.
- I. Include non-native varieties of English.

## Appendix 2

### *Evaluation myths (Phillips, 1997)*

1. I can't measure the results of my training effort.
2. I don't know what information to collect.
3. If I can't calculate the return on investment, then it is useless to evaluate the program.
4. Measurement is only effective in the production and financial arenas.
5. My CEO does not require evaluation, so why should I do it?
6. There are too many variables affecting the behavior change for me to evaluate the impact of training.
7. Evaluation will lead to criticism.
8. I don't need to justify my existence, I have a proven track record.
9. Measuring progress toward learning objectives is an adequate evaluation strategy.
10. Evaluation would probably cost too much.

## Appendix 3

### *Why change? (Phillips, 1997)*

1. It just makes good economic sense.
2. Accountability is an important trend.
3. Increased scrutiny of HRD budgets.
4. Pressure from the top to make a contribution.

5. Peer pressure from HRD professionals.
6. Self-satisfaction.
7. More information is available.
8. Professionalism.
9. Survival.

## Appendix 4

### *Purposes and uses of evaluation (Phillips, 1997)*

1. To determine success in accomplishing program objectives.
2. To identify the strengths and weaknesses in the HRD process.
3. To compare the costs to the benefits of an HRD program.
4. To decide who should participate in future programs.
5. To test the clarity and validity of tests, cases, and exercises.
6. To identify which participants were the most successful with the program.
7. To reinforce major points made to the participants.
8. To gather data to assist in marketing future programs.
9. To determine if the program was the appropriate solution for the specific need.
10. To establish a database that can assist management in making decisions.

**Appendix 5****Complete results-based HRD model (Phillips, 1997)**

- 1 Conduct a needs assessment and develop tentative objectives.
- 2 Identify purposes of evaluation.
- 3 Establish baseline data, if available.
- 4 Select evaluation method/design.
- 5 Determine evaluation strategy.
- 6 Finalize program objectives.
- 7 Estimate program costs/benefits.
- 8 Prepare and present proposal.
- 9 Design evaluation instrument.
- 10 Determine and develop program content.
- 11 Design or select delivery methods.
- 12 Test program and make revisions.
- 13 Implement or conduct program.
- 14 Collect data at proper stages.
- 15 Analyze and interpret data.
- 16 Make program adjustments.
- 17 Calculate return on investment.
- 18 Communicate program results.

**Appendix 6****Calculating return on investment (ROI) (Phillips, 1997)**

$$\text{ROI (\%)} = \frac{\text{Net Program Benefits}}{\text{Program Costs}} \times 100$$