Pergamon

S0346-251X(96)00029-2

System, Vol. 24, No. 3, pp. 375–389, 1996

0346-251X/96 $15.00 + 0.00

DEMYSTIFYING MATERIALS EVALUATION

J. T. ROBERTS

Department of Language and Linguistics, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, U.K.

INTRODUCTION

The role of and even the necessity for (commercially available) foreign language teaching materials have, of course, not remained unchallenged—see, for example, Allwright (1981). However, this paper rests on the assumption, shared *inter alia* with O'Neill (1982), Matthews (1985), Sheldon (1988) and McDonough and Shaw (1993), that in many if not most foreign language teaching situations, commercially-produced materials, and pre-eminently amongst them "the textbook for the course", represent the fundament on which teaching and learning are based, or, at the least, may most conveniently be supported. The reasons may often be theoretically unimportant, though imperatives in practice. Among them are: pressure of time on teachers (such that they cannot develop their own materials); the uncertain language competence of teachers (such that it is better that they do not produce their own materials), the greater "slickness" and perhaps, by that token, appeal to learners, of established publishers' offerings; the need for a yardstick of progress, both for learners and for others looking in on the situation (e.g. inspectors, parents), insofar as working from one end of a textbook to the other is tantamount, or may be interpreted as tantamount, to this. Reasons such as these, which between them are classifiable under both of Allwright's (1981) *Deficiency* and *Difference* views, will bring little comfort to those who see the way forward in language teaching as lying in innovative management techniques. The premise here, however, is that failing a revolution in teaching and other relevant conditions, the centrality of commercially-produced materials, and again, of the commercially-produced "textbook for the course", will persist.

Kelly (1969: 258ff) reminds us, writing of the European context, that at one time the only person in the language classroom who would have owned a textbook was the teacher,[1] and that it was probably not until the late 16th century that individual pupils could buy their own. Pupil-ownership then spread rapidly during the 17th and 18th centuries, when publishers began to pirate, plagiarise and hawk aggressively for trade, and "the cost of unofficial versions was as low as consistent with minimum legibility" (Kelly 1969: p. 260). Even so, "...these early texts had a publishing life of over a century" (*ibid.*). Further, we

may be sure that both want of literacy and of means confined book-ownership to the relative few, and that even if Comenius "realized how necessary it was to have the same textbook in the hands of every pupil in the class" (*ibid.*), he was not thinking of the quantities of textbooks involved in universal education; nor would he have credited the choice of foreign language teaching materials on the market today. As an indicator of this choice, the study of Goodman and Takahashi (1987) is not here cited for the first time: they found, in the middle years of the last decade, 28 U.S. publishers offering between them a total of 1623 ESL textbooks. While one might speculate that ESL is exceptionally well provided-for, the variety of textbooks on offer for the teaching of other major languages can also be relatively vast, and confusing. Here, then, is the impetus for attempts made in recent years to *systematise* materials evaluation.

By "systematise" is meant to render efficient and accurate by contrast with less formal approaches to the task, for example, flicking through textbooks or listening at random to tape-recorded units and making a few notes as one goes along. Admittedly, faced with a choice of two or three texts, it may be that an evaluator or a group of evaluators does not need any form of "technical apparatus" to assist a judicious decision, but once the choice widens considerably, it is evident that the difficulty in maintaining evaluative *consistency* will increase. It would be surprising if one evaluator could preserve consistency in the examination of, say, 10 plausibly competing textbooks without a formal set of criteria to refer to, but quite incredible if consistency were achieved in the evaluation of 100 textbooks by a team of five evaluators, all believing that they were applying the same evaluative measures, but without agreed and explicitly stated guidelines. Indeed, one is tempted to say that in such circumstances, consistency would just be impossible. Initiatives to develop systematic approaches are therefore indispensable.

Nonetheless, there are problems in all this because the efforts at providing systematicity have hardly been confluent. On the one hand, Tucker (1975),[2] Williams (1983), Cunningsworth (1984), Matthews (1985), Sheldon (1988) and McDonough and Shaw (1993), for example, offer lists of criteria or *checklists* which, though they have certain points in common, are in general differently informed and motivated. On the other, Tucker proposes a system which elicits arithmetic *scores*, while, say, Matthews and McDonough and Shaw do not advocate a scoring system at all. In between Tucker, with his purely *quantitative* approach, and those interested only in *qualitative* evaluation come Cunningsworth and Sheldon, who invite both quantitative and qualitative (i.e. descriptive) responses. Further, Tucker has a *weighting* system which is taken up by D. Williams and Cunningsworth but which otherwise seems not to attract emulation. Breen and Candlin (1987) put forward an inventory of questions which is markedly more learner-centred than the other "instruments". And whereas Tucker moves "straight in" with his one-phase evaluation, Matthews, McDonough and Shaw, and Breen and Candlin, for instance, propose two-phase evaluations, though all for different reasons. Little wonder, then, from a certain point of view, that Low (1989:153) remarks: "The assessment of language teaching materials, even when supplemented, as it should be, by empirical studies, remains...something of a 'black art'".

Yet, understandable as Low's scepticism, or, conversely, conviction that cabbalism is afoot, may be, the simple fact is that materials evaluation is not a "black art" at all, and

can be totally demystified. What is required is to supply a proper perspective on the matter, to clear up terminological obfuscations and to fetch a priest[3] to some of the red herrings floating around in the literature. To state that these requirements are met exhaustively here would be to claim too much, but what follows is intended at least as a step in the right direction.

MATERIALS EVALUATION AS A TOTAL PROCESS

Up to now the term "materials evaluation" has perhaps been used as though it implied only some sort of "armchair" application of criteria to sets of materials. If it appears in the present context to mean this, there is a sense, as we shall see, in which such an interpretation relates to everyday reality, and we shall, in fact, be drawing the conclusion that it is a valid interpretation. But it does not correspond to the imaginable ideal. Such an ideal—a "total" evaluation process—is modelled as in Fig. 1 in order, as it were, to *reculer pour mieux sauter.*

The model assumes (and since we are talking here about an ideal, we will for the moment take for granted the complete scrupulousness and dedication of all involved at all stages between conception and adoption) that the evaluation process begins not even as late as the moment at which the materials designer types the first plan, but within a short time of the conception of the first germ of an idea for a set of materials for a certain target population of learners, in just the same way as the first flush of elation following an idea for an article, for instance, is soon tempered by doubt and self-questioning. Given, however, that self-criticism can be overcome, a proposal will sometime later be made to a publishing company, which will make its own initial evaluation of the project. If this is encouraging, the materials designer will then suffer more self-doubt leading to modifications and corrections throughout the writing process, but, again assuming the vanquishing of self-torture, the final draft will go off to the publishing house. The publishers will now make a further assessment in the light of all variables important to them, and, if this assessment is in the main positive, will call for piloting of the materials on a sample of the target population. This piloting on "real learners" will lead to the first definitive decision (DECISION STAGE 1),[4] at which the project may be deemed unsuccessful, or entirely successful, or successful but with reservations entailing further modifications. Thus, even in the PRE-PUBLICATION STAGE, evaluation will have been going on perhaps for some considerable time before the materials enter the public domain.

It is at the point at which materials meeting publishers' criteria enter this public domain that educator–evaluators will subject them to their own scrutiny. The first step will be to conduct precisely an "armchair" or "pencil and paper" evaluation aimed at identifying materials which are plausibly not only appropriate for a particular population of learners, but better than materials heretofore to hand. Here, published reviews may be referred to as well as the opinions of colleagues, and so on. Then, materials passing the "pencil and paper" test are subjected to a classroom trial, once more on "real learners", but this time, one's own. At the end of the classroom trial comes the *summative evaluation*, which will demand the pooling of all available information: most saliently, the results of the experiences of teachers and learners, but also, perhaps, the judgements

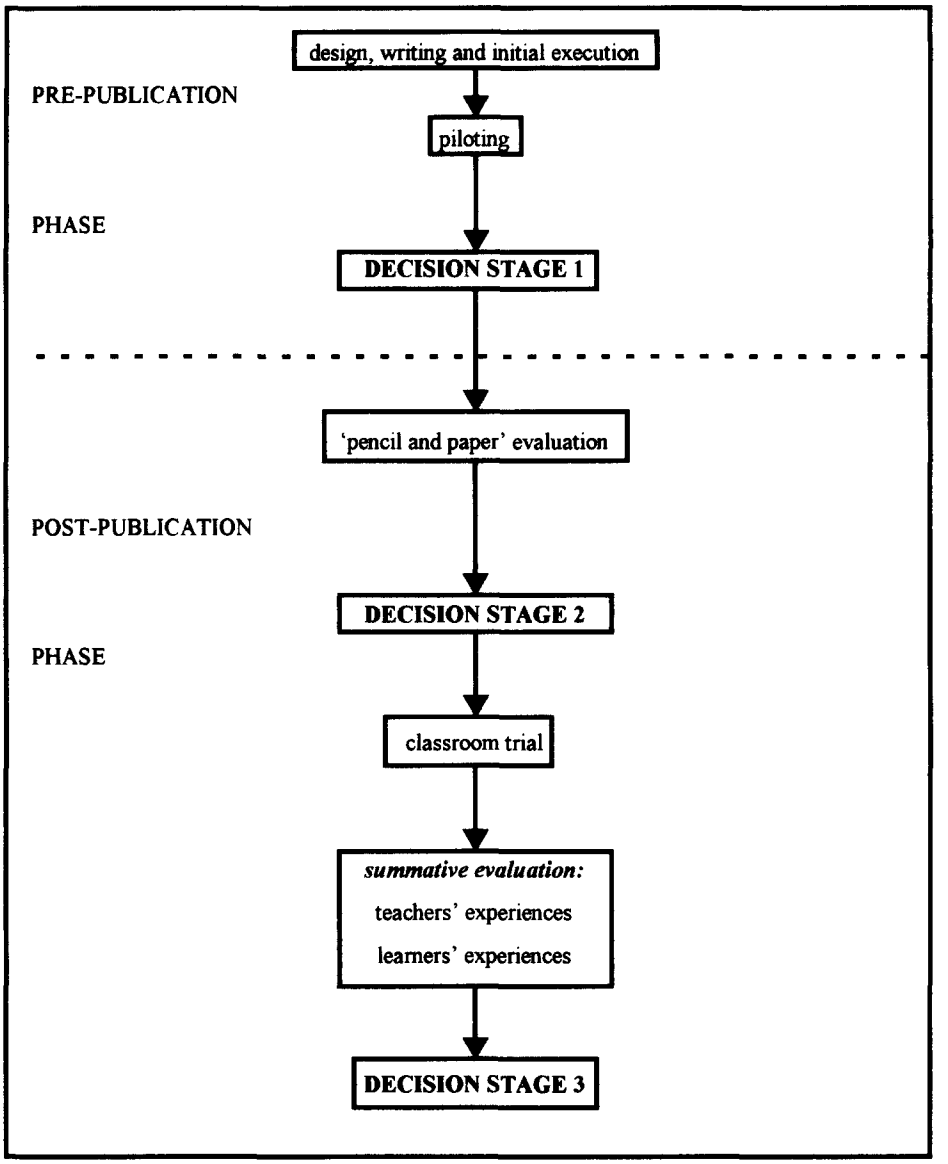

Fig. 1. Materials evaluation as a "total process".

expressed in further reviews appearing in the meantime. Now all is ready for DECISION STAGE 3, at which the set of materials in question is either adopted for the foreseeable future, or else rejected.

THE PRE-PUBLICATION PHASE

What will perhaps seem unusual about the model is that it commences not with the POST-PUBLICATION PHASE, as does, say, the model of McDonough and Shaw (1993: 75: Fig. 4.1), but includes a PRE-PUBLICATION PHASE allowing for the evalu-

ation of materials by those who produce them: materials designers and publishers. While it may be true that the incentive for the latter is sometimes an easy profit, it would seem unreasonable to suppose that *no* material designers and *no* publishers ever took the utmost care in assessing the instructional value of their products, and more unreasonable still not to admit the theoretical possibility that painstaking evaluation *could* precede the release of these products. Therefore, any model of the "total" process must include this phase. At the same time, it has to be recognised that, at least in a free market economy, the interests of course designers and publishers on the one hand, and those of educational evaluators, teachers and learners on the other, will be rather differently focused.[5] Thus, much as it would be comforting to believe that the interests of both sides did not grossly conflict and that the situation were not inherently "adversarial" but enrichingly symbiotic, it is accepted that this is often not so. It is therefore understandable if evaluators with a stake in education rather than publishing adopt the stance exhorted in *caveat emptor*, as they should, indeed, because it is their task not only, in the best case, to choose the most appropriate materials from amongst those testifying to quality of production, but also, in the worst, to protect learners from materials designers and publishers who have cut corners for a quick return. However, the ideal with regard to the PRE-PUBLICATION PHASE having now been proposed as fairly[6] as possible, and this phase having been included in the model in order that a full perspective might be supplied and the range of meaning of "materials evaluation" explored, it must be left to materials designers and publishers to determine how far they wish to adhere to or pursue the ideal, and the rest of these remarks will concentrate on the POST-PUBLICATION PHASE, in which it falls to educators to assess the efforts of materials producers.

THE POST-PUBLICATION PHASE

Where this POST-PUBLICATION PHASE is concerned, the model will, of course, represent an ideal distant from reality for many, if not most. For a start, some education systems provide for no consultation at all of teachers, let alone learners, on their opinions of materials. Some education systems allow teachers a choice of materials, but from a shortlist, or, more poignantly, a short list. In other systems, the ministry of education employs an evaluator who selects above the head of teachers, perhaps without consulting them at all, and who may play the role of censor more than that of evaluator. Sometimes, the ministry of education commissions materials and there is no chance of a post-production evaluation before they pass into the classroom. If such situations are regretted by those caught up in them, the model itself can only help by providing a view of how things could be otherwise. In the meantime, for there to be any connection at all between the model and reality, some minimal freedoms have to be assumed, i.e. that there is in the system an evaluator or team of evaluators empowered to assess and select any from all materials designed for the teaching of whatever language is relevant; perhaps, optimally, that the evaluators would be teachers, acting in consultation with learners.

However, even where maximal freedoms exist, there are still at least two major problems affecting the implementation of the model, as follows:

1. An ethical problem. Should one run classroom trials on materials in which one does not have faith, using "real learners"? Note that the moral dilemma here is different from

that, say, of observing that materials adopted in good faith a week or two ago have been superseded within that time by something very much superior, or of registering over several months that the materials selected so carefully on the basis of a checklist and reviews do not in practice seem to work very well. That is an accident of life with which one has to live, and is not at all the same proposition as prescribing for learners materials in which one does not *yet* have complete confidence, unless one happens to elieve that the end justifies the means.[7] Of course, "classroom trials" could be interpreted loosely—trying out the odd unit, leaving one or two copies of promising-looking materials around for learners to inspect in their own time, and so on, but such informal procedures do not equate with what Low (1989) terms "empirical studies". This means full-scale trialling over a significant period of time, perhaps on learners preparing for a crucial examination. And the ethical problem is accompanied by practical ones. Firstly, if one is to trial several competing sets of materials on one group of learners, other materials will have appeared on the market before the trial is over. Secondly, if, in order to accelerate trialling, one distributes competing sets of materials between different groups of learners, the major difficulty associated with empirical work—controlling the variables—increases dramatically, and especially if different teachers are involved. The undertaking, if carried through seriously, is more in line with the work of a research establishment than that of a school in which teachers are already heavily burdened.

2. An ineluctable problem for many, if not most, institutions: money. How many schools and universities respecting copyright laws can afford to buy multiple sets of published materials for classroom trials without any commitment to using them in the longer term, indeed, to put it graphically, in cognisance of the fact that at the end of the trials 30 or so expensive textbooks may have to be thrown on the rubbish-tip?

These two problems alone determine why, commendable as materials evaluation as a "total" process might be in theory, and "supplemented...by empirical studies" as it should be in the best of worlds, the evaluation of materials as a generalisation does ultimately, in the POST-PUBLICATION PHASE, hinge upon the "armchair" or "pencil and paper" evaluation. Consequently, though this type of evaluation should never be regarded as the whole of a possible story, it is important to make it work well. At the crux of it must, then, both despite and because of the various conflicting factors cited above, figure the *checklist*.

THE CHECKLIST

Some differences between various checklists in the literature are already mentioned above. There are many more: for example, that Tucker's list incorporates parameters considered critical in the heyday of Audiolingualism, whereas if one examines the others in chronological order of publication, the influence of communicative methodology becomes progressively apparent. Again, in terms of numbers of specific criteria or questions,[8] Tucker has 18, Williams 28, Cunningsworth 52, Matthews 19, Breen and Candlin 34, Sheldon 17 and McDonough and Shaw 28. At least, this is how things seem on the surface, but in scrutinising the lists more minutely, one discovers that some of them con-

tain criteria or questions entailing several "evaluative steps". One of Tucker's criteria, for instance, under *Grammar*, is "adequacy of drill model and pattern display", which invites the evaluative steps: assess the drill model, assess the pattern display. Cunningsworth's 52 criteria arguably involve 116, and, Breen and Candlin's 34 criteria, 56, evaluative steps. Indeed, all the checklists cited, however elegant some of them look, contain more than meets the eye where it comes to the labour they impose upon users.

However, the question of users now brings us to the central issue: it is easy to see that people unfamiliar with materials evaluation, but beginning to read up on the topic, may say to themselves: "Some of these checklists are outdated, but which of the others should I choose?". Demystification is urgently needed here, if nowhere else. The answer is, essentially, "none of them". The proper users of the checklists alluded to are: Tucker, Williams, Cunningsworth, Matthews, Breen and Candlin, Sheldon and McDonough and Shaw, and/or their colleagues working in the same environment, respectively. In justice, it must be added that virtually all the checklist designers quoted stress that they are offering examples, since it is unlikely that any two teaching/learning situations will correspond exactly. As Sheldon (1988: p. 242) puts it: "We can be committed only to checklists...that we have had a hand in developing, and which have evolved from specific selection priorities". In other words, checklists in the literature should be regarded as illustrative and suggestive only, and never as decretory. While some of the criteria they embody may be relevant to one's own teaching/learning situation, perhaps their most valuable aspect is that they stimulate thought about the *system* of evaluation and the *modus operandi* to be adopted.

There follows a number of considerations to which published checklists give rise.

The context of evaluation

In view of the remarks made above about the specificity of teaching/learning situations, the starting point for any evaluation must be an analysis of the context in which materials are to be employed. Matthews (1985:203f), for example, has already explicitly proposed what one might call a "pre-evaluation phase", "defining your own teaching situation", and the variables he identifies as critical are: *syllabus, time available, students' age, students' interests, students' background, class size, students' level.* Here it is suggested that a more comprehensive list might be constituted as follows:

- *Learners*: age, stage in learning, enabling and disabling factors, interests and motivation, preferred learning styles.
- *Teachers*: teaching competence and experience, competence in the language, preferred teaching styles.
- *Aims*: of the course and of the learners.
- *Syllabus and (if any) prescribed methods*: constraints imposed.
- *Examinations and/or tests*: constraints imposed and "backwash effect".
- *Cultural and related factors*: acceptability or non-acceptability of values conveyed in materials in given cultural and social contexts.
- *Practical factors*: time available for teaching, presence or absence of homework, size of classes, availability of hardware to implement materials, the teaching and learning environment, etc.

Even this longer list might not, of course, be sufficient to cover the variables pertinent to some situations. For instance, one could envisage the addition of *educational factors* if it were of concern that language teaching materials should contribute through their content to general or moral education. What is indisputable, however, is that there is no point in even beginning to look at materials until one is clear about the exigencies to which they must respond. This leads directly to the next consideration:

Good and bad materials

There are undoubtedly materials which are bad by any standards—shoddily produced, error-ridden, unduly prescriptive, based on dubious methodologies, etc.—and which one hopes nobody would waste money on. There are also materials which one describes as "good" because they are carefully designed, helpful, informative, attractive, durable, and so on. However, once one starts thinking about the relationship between materials and particular learners, then one has to be more careful with the word "good" because now it has to take on a different connotation: "good for my learners in my situation and in the light of the constraints applying to them and me". Thus, evaluation is directed not so much towards the selection of "good" materials as measured by some absolute standard, but of *appropriate* ones. It follows from this that while designers of checklists might like to leave room for evaluators to expostulate over absolutely useless materials, the questions or criteria should aim at eliciting the way in which a given set of materials and users of them would interact. Not then "Is the storyline interesting?", but "Is the storyline likely to appeal to our first-year students?". This, again, leads to the next consideration:

The meaning and framing of criteria and questions

Examination of checklists in the literature will show that the meaning of no small number of criteria or questions is not immediately transparent to anyone coming to these checklists "cold", for which reason their designers feel compelled to talk readers through all or parts of them. In the case of more comprehensive instruments—Cunningsworth, Breen and Candlin—this is perhaps not quite so true, but the price seems to be length and complexity, and even if the meaning of the questions is at some level apparent, the motivation for them does not always emerge from the checklist itself. Cunningsworth's 12th checklist point, for example, is formulated thus: "Is the language process assumed to be essentially (a) inductive (b) deductive (c) a combination of both?". Possible answers are: yes, no, no; no, yes, no; no, no, yes; yes, yes, yes (much depending on the interpretation of "essentially"). But what is one to make of any of these answers in the abstract? Is there just one right one, or does it all depend upon the situation, or upon a definite view of the language learning process on the part of the evaluator?

This sort of problem underlines the fact that starting out from the "local" context is really the only fruitful way to proceed, since designers of evaluation instruments and evaluators working together can establish that they "speak the same language" and can ensure that when criteria and questions are entered into a checklist, all concerned share an understanding of what these criteria and questions mean, and what the consequences of possible responses to them will be for a given set of materials. (A model for the dialogue which might precede the production of a definitive checklist is offered in Williams, 1981.)

Even so, questions should still be sensible and leave as little room for subjective interpretation as possible, or else a full "key" should be provided, especially where responses are elicited by "headings" such as "coverage of grammar" rather than by questions. Messih Daoud (1977) for example, includes in his checklist aimed at textbooks the wondrous question (which he does not explicate): "Do the sentences gradually increase in length to suit the growing reading ability of the pupils?". If this question is worth asking at all, he should, of course, have asked about complexity and hypotaxis versus parataxis, and have supplied some indicator of comparative "length". And his question: "Do illustrations create a favourable atmosphere for reading and speech by showing realism and action?" is basically unanswerable, because, apart from the difficulty of determining what "realism" and "action" are, severe doubts surround the suggested causality between these elements and "a favourable atmosphere for reading and speech". One can easily conceive of illustrations portraying what might be supposed to be "realism" and "action" combined, for example, with bloodshed and gruesomeness (e.g. Poussin's picture "The Martyrdom of St Erasmus"[9]) and which would procure nausea rather than a "favourable atmosphere". Compare this with Matthews, whose checklist contains the terse criterion "Illustrations", but who, in his "talk-through" or "key" asks:

> What pedagogic purpose do the illustrations have? Or are they mainly intended as decoration?... Are the illustrations, whether colour or black and white, pleasing to the eye or are they grotesque? Do they portray clearly what is intended pedagogically or do they serve to confuse?

Subjectivity versus judgement

People (typically students of Applied Linguistics) looking at checklists for the first time and noting the wide discrepancies between them, as well as the more weird and wonderful formulations sometimes proposed, often react by protesting that the whole exercise of materials-evaluation is subjective, and in this subscribe to the "black art" point of view. It therefore has to be emphasised again, firstly, that looking at the checklists of others without knowing about the precise conditions of their work can be quite uninformative if one is dredging the literature for criteria rather than systems, and, secondly, that, where specific criteria are concerned, there is a difference between subjectivity and *judgement*. The former is arbitrary and has no rational defence, whereas the latter is based on reason, experience and intimate knowledge of a particular teaching/learning situation, and can be articulately defended. While there are indeed checklists in the literature which attest to subjectivity, any conscientious compiler of a checklist, qualitative or quantitative, exercises judgement in selecting the parameters to ask about (and not to ask about) and, whatever the means of elicitation adopted, judgement, not a subjective impression, is again being invited from evaluators using the checklist. Worthwhile checklists cannot be created by sitting down and dreaming up "a few points it seems interesting to ask about materials", but can only issue from careful consideration of what, in each case, *specific* learners need, what *specific* teachers can handle and what materials it is feasible or permitted to use in a *given* situation. Thus, one returns again to the "local context". Provided judgement and specificity are constantly borne in mind, subjectivity may be minimised, and evaluation will be not amount to a "black art". This is not to imply that materials evaluation is an exact science, because it involves human values. In dealings with human beings there are no "correct answers" and absolutes, only the best possible answers that can be produced in the circumstances. To expect of materials evaluation

that it should do more than to provide a *system* for the application of *consistent* human and pedagogical *judgement* is to view it in terms of the wrong paradigm.

Quantitative versus qualitative methods of evaluation

As remarked earlier, there are examples to be found of checklists which use a quantitative system of elicitation, i.e. arithmetical scoring, and others which elicit qualitative assessment, i.e. descriptive comments, and again, others which combine these systems. In the abstract, it is impossible to say which system is preferable.

Checklists to be responded to numerically can look relatively uncluttered, since they do not need to contain questions, but only "heading-type" criteria. Thus, for example, Tucker's checklist contains criteria such as "completeness of presentation" (in relation to pronunciation)—just three words as against at least one of Cunningsworth's questions running to over 50 words. However, one cannot supply a rating on any criterion in Tucker's list without reading his "talk-through" or "key", and it so happens in this instance that his explication of the criterion "completeness of presentation" runs to something like 150 words. Thus, one may have to decide on a trade-off here: supply a short checklist with a long and detailed key, or a more fully-specified list with a shorter key. And it is not totally true that only "quantitative checklists" are uncluttered: Matthews's list, for example, consists of only 19 criteria, such as "methodology", but invites descriptive or yes/no answers. Again, however, one has to read his key in order to interpret the criteria. If there is a real advantage in the quantitative approach, it is that checklists answered numerically are often quickly dealt with and facilitate comparison of the answers of different evaluators. The disadvantage is that if they contain no questions to be answered descriptively, they do not leave space for discerning reactions to materials on parameters which the checklist designer has not foreseen. They may also look misleadingly "objective".

By contrast, the advantages of the qualitative approach are that it gives evaluators the opportunity to express themselves freely and to describe reactions to materials on points overlooked by the checklist designer. The disadvantages are that evaluators may feel that the writing-task imposed is too much of a burden and that, where several evaluators are involved, it may be very difficult, if not impossible, to form a coherent whole out of descriptive answers phrased in different ways and dwelling on different aspects.

Furthermore, whether one uses quantitative or qualitative elicitation, there is no guarantee that more care will be put into a qualitative rather than a quantitative response. It may only take a second to put a "3" or a "4" against some criterion on a checklist inviting numerical scores, but one would hope that the score was dutifully considered. On the other hand, though it takes a little longer to write down, a response such as "The layout and presentation are excellent" does not necessarily mean that the respondent has thought about the matter carefully. It is perhaps also worth pointing out that questions to be answered with "yes" or "no" are not "qualitative" questions at all, but "quantitative" questions. The answers can only be counted, and one may as well ask that this or that feature be accorded a plus or a minus, a one or a zero.

Whether a checklist is quantitative or qualitative must again be related to the "local context" and the complexity of the evaluative task. However, the default assumption should

be that the more quickly evaluators can respond to a checklist, the more willingly they will work. Entering numbers or checking boxes is certainly less irksome than writing comments where any significant quantity of materials is concerned, and so, since it is also easier to compare quantitative evaluations, the quantitative approach has much to recommend it. However, the point about evaluators enjoying the freedom to comment on matters which have not occurred to the checklist designer is not to be forgotten. Therefore, even a basically quantitative checklist should supply a section for "further comments".

Finally, any checklist, quantitative or qualitative, to be used by several evaluators should be tested for inter-rater reliability. See, for example, Jones (1993: p. 48).

"Higher order" criteria or questions first

It would seem that the purpose of McDonough and Shaw's (1993) proposed two-phase approach to assessment through, first, a macro-evaluation and, second, a micro-evaluation, is to save time and effort. Thus the macro or external examination focuses upon aspects which can be determined quickly by looking at such things as the claims made by the publishers of a set of materials and their statements about the intended audience and the proficiency level aimed at, seeing whether a vocabulary list and index are included, establishing whether the layout and presentation are clear or cluttered, and so on. The macro-evaluation would lead, then, to the rejection of patently unsuitable materials and the identification of potentially appropriate ones, only the latter being investigated further in the micro-evaluation.

While the general idea is excellent, it should be stressed that the criteria included in the macro-evaluation should be those relevant in the local context, i.e. not necessarily McDonough and Shaw's. Another way of looking at the matter is to say that a checklist should proceed from "higher order" questions to those which are of "lower order" in this context, even though these might be more technical than some higher up the list. For example, the McDonough and Shaw list does not include the criterion "price", which, however, appears in some form in the lists of Tucker, Williams, Matthews and Sheldon, but always towards the end, presumably because it is an unexciting consideration. Yet it could well be that an institution with a restricted budget has to impose a non-negotiable limit on the price to be paid for textbooks. If so, "price", dull as it is, becomes a "higher order" concern, and should go near the beginning of the checklist, so that no evaluator's time is wasted by working through the whole list when the textbook under scrutiny cannot be purchased anyway. It can be envisaged that a really user-friendly checklist might not only arrange criteria in descending order of priority, but might also provide "exit points" at various places for materials not fulfilling expectations on crucial parameters. The argument against this is the line often taken in the literature that "no textbook is perfect", the implication of which seems to be that teachers must accommodate to the best compromise available, and that therefore a total profile of any set of materials examined is necessary in order to see what the shape of the compromise might be. The counter-argument and the proposal here is that it is a question of isolating those aspects over which there can be no compromise, and putting them high up on the list. In other words, the macro-evaluation stage should contain all the hurdles which really do have to be cleared.

Including learners in the evaluation process

Should earlier remarks on the difficulty of classroom trials have seemed to dismiss learners from the evaluation process, any such impression must now be corrected. Since materials are "used on" learners, these should have a voice, and their reactions should be taken to heart. To what extent they can be directly involved in the selection of new materials is, obviously, problematic, and will depend upon factors such as age, the feasibility of presenting them with genuine choices and the liberality or otherwise of educational authorities. However, there may in many contexts be at least two measures which could be adopted, subject to the proviso that learners have had some experience of being the recipients of language teaching as well as of being people that materials are "used on". One is to administer the sort of questions which Breen and Candlin (1987) suggest, e.g. "What particular subject-matters (topics, themes, etc) interest you? What would you like to find out more about through the new language?". The other is that, at whatever point a decision has been taken to change materials, i.e. to buy different ones, they be asked to fill in a questionnaire (or, if more suitable, give oral feedback) on the materials employed to date. This will not implicate them in the selection of the new ones (in some educational contexts, unthinkable) but could lead to the determining of motivating and demotivating features, to be borne in mind in subsequent selection. Teachers may, of course, also enquire informally in the course of classes what students think of current materials, but the dangers of this, if the consequences are not thought through, hardly require spelling out.

Universals in materials evaluation

Sheldon (1988: p. 246) says, surprisingly, in view of his caution, regarding his own checklist, that "consumers would emphasize other factors that relate specifically...to their own unique situations" (1988: p. 242):"...we need to discover whether a *de facto* evaluative consensus exists at all, and whether there is any foundation upon which universal criteria could be erected".

The first part of the statement is a red herring. If no other message has emerged from this paper, it is to be hoped that the one which has impact is: "base materials evaluation on the local context". If we are selecting textbooks for teaching German to third-year pupils at a comprehensive school somewhere in England, we do not *need* to know whether our criteria accord with those of evaluators choosing German textbooks for 14 year-olds somewhere in the U.S.A. or Mongolia or Brazil, neither do evaluators in those countries *need* to know whether their criteria correspond with those of the evaluators in England or in each other's countries before they can pursue the task. That it would be *interesting* to know whether an "evaluative consensus" exists, and that it would be *informative* to compare notes, are different propositions.

Much the same could be said for the second part of the statement. We do not *need* to discover "whether there is any foundation upon which universal criteria could be erected". It is of academic concern to pursue this problem, but in the meantime, there is no reason to suppose that materials evaluation cannot be carried out optimally within the bounds of what is known and possible today. This is not to advocate complacency, but materials evaluation, like teaching, is an activity not to be suspended until "all is known". It must proceed, whether fully informed by research or not. In any case, it is already predictable that the only true universal criteria which can figure in any checklist are theoretically trivial, e.g. the cost, availability and durability of materials. More complex

criteria, such as *cultural bias*, will obviously have different exponents, depending upon the setting in which given materials might be employed.

If there is a need for a consensus and the discovery of universals, this need does not relate to criteria, but to *procedures*. Assuming that the recommendation that all evaluations be based upon the local context is valid, one must expect criteria to vary and even to be idiosyncratic. Agreed, systematic procedures, however, would ensure that evaluators were "speaking the same language", even if the topics and themes "spoken about" varied.

CONCLUSION

An attempt has been made in the above to demystify materials evaluation by:

(a) presenting a picture of materials evaluation as a "total" process, but

(b) delineating which parts of this process are likely to be feasible, especially in the POST-PUBLICATION PHASE, and

(c) recommending how these feasible parts may be realised,

(d) starting with the assumption that the local context will determine the criteria for evaluation and that

(e) searching for universal criteria does not, in the meantime, need to hold materials evaluation up

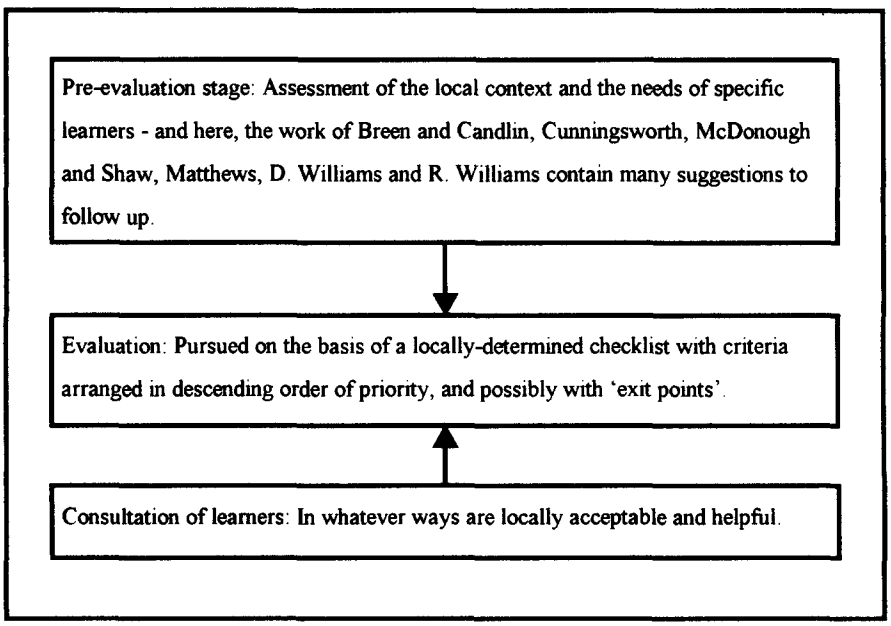

Fig. 2. A model for evaluation in the POST-PUBLICATION PHASE.

(f) providing it is pursued rationally, consistently and systematically.

A model for evaluation in the POST-PUBLICATION PHASE is shown in Fig. 2.

Pace Sheldon, materials evaluation *is*, precisely as Hutchinson and Waters (1987: p. 97) state, a "matching process: matching needs to available solutions". How publishers improve the "available solutions" is up to them. In free market economies, they do, of course, have licence to proffer what they like. But those in charge of selecting materials should be aware that it is entirely within their own power to weigh published offerings in the balance, and to make informed choices between them. There are no "black arts" where rationality, consistency and systematicity are present.

NOTES

[1]Which is perhaps why British University Teachers in the "career grades" are called Lecturers or Readers, meaning the same thing, i.e. people who read "the book" and relay the readings to the students. Professors, by contrast, may pronounce *ex cathedra* because their wisdom goes beyond the one or two books available...

[2]If Tucker appears first in this and subsequent allusions, it is because he is to be recognised as the pioneer in the field. Though his criteria are outmoded, his system is still inspirational.

[3]Here in the sense of a bludgeon for dispatching freshly-landed fish. Herrings are, of course, trawled and a priest is not used, but *red* herrings deserve to be knocked on the head.

[4]One could argue, of course, that many "decision stages" would be possible before this one—the publishers could reject the initial proposal, or change their minds about the project before the materials designer had finished work on the manuscript, or reject the manuscript when completed. Provision for manifold eventualities is not included here, in the interests of keeping the model relatively straightforward.

[5]The point about the free market economy is made because in some educational systems, it would not be accepted that materials might introduce learners to values not approved by the state, e.g. by depicting an "alien" form of political organisation. Conversely, the state might consider it to be in the national interest to positively inculcate certain values through the medium of education. In either case, the state is likely to commission materials, and there is no place for evaluators who assess the materials between production and application in the classroom, except, perhaps, on strictly-defined parameters.

[6]There is indeed an argument about fairness here. It is easy to say that materials designers and publishers are in the business to make a living, if not a profit, and there are certainly instances of unscrupulous textbook publishing going back 300 years or so. But, on the other hand, there are teachers in the business to make a living (profit, unfortunately, not often entering the question) who do not regard it as incumbent upon themselves to furnish their own materials and who have come to expect others—in the most general sense, publishers—to offer them such requisites in the manner of tools with which to carry out their trade. Thus, the line of argumentation here is in no way intended to imply that publishers simply flood the market with anything they think vaguely saleable, and that evaluators and teachers must automatically be suspicious of all offerings to the point of ingratitude. The line is a more relativist one: It so happens that today there are vast offerings on the market. Among them, there are bad ones, and these should be identified and rejected. On the other hand, among them also are many representing careful design and quality, but not all of which—perhaps relatively few—will be appropriate for any particular population of learners, so that here also identification and rejection, but also selection, must be carried out.

[7]It is possible to argue, quite rightly, that teachers constantly evaluate materials in retrospect, once they are stuck with them over a length of time. However, in the present context, materials evaluation is intended to refer to a process occurring before rather than after the final adoption of a set of materials.

[8]The distinction being made here is that some checklists (e.g. Tucker's) contain "heading-type" criteria, such as "functional load" whereas others (e.g. Cunningsworth's) contain questions, such as: "Does the material have variety and pace?". Both criteria and questions are meant to elicit information on the topics they introduce, and if criteria and questions are not disambiguated anywhere in the present paper, "criteria" should be understood as the eponym, standing for both "heading-type" criteria and questions.

[9]The saint's suffering is depicted "realistically" enough. As for the "action" he is having his innards drawn out with a windlass. (He was a Christian martyr who became the patron saint of Mediterranean sailors, but the manner of his martyrdom may have been a "back-projection", since his attribute as a patron saint was a windlass or capstan—I am grateful to John Nash for this information.) The work hangs in the Vatican.

REFERENCES

ALLWRIGHT, R. L. (1981) What do we want teaching materials for? *ELTJ* **36**(1), 5–17.

BREEN, M. and CANDLIN, C. (1987) Which materials? A consumer's and designer's guide. *ELT Docs* **126**, 13–28.

CUNNINGSWORTH, A. (1984) *Evaluating and Selecting EFL Teaching Materials*. London: Heinemann.

GOODMAN, P. and TAKAHASHI, S. (1987) The ESL textbooks explosion: a publisher profile. *TESOL Newsletter* **4**, 49–51.

HUTCHINSON, T. and WATERS, A. (1987) *English for Specific Purposes: A Learning-Centred Approach*. Cambridge: Cambridge University Press.

JONES, F. R. (1993) Beyond the fringe: a framework for assessing teach-yourself materials for *ab initio* English-speaking learners. *System* **21** 453–469.

KELLY, L. G. (1969) *25 Centuries of Language Teaching*. Rowley, MA–Newbury House Publishers.

LOW, G. (1989). Appropriate design: the internal organisation of course units. In Johnson, R. K. (ed.), *The Second Language Curriculum*. Cambridge: Cambridge University Press.

MCDONOUGH, J. and SHAW, C. (1993) (pp. 136–154) *Materials and Methods in ELT: a Teacher's Guide*. Oxford: Blackwell.

MATTHEWS, A. (1985) Choosing the best available textbook. In Matthews, A., Spratt, M. and Dangerfield, L. (eds): *At the Chalkface*, pp. 202–206. Edward Arnold.

MESSIH DAOUD, A. (1977) Evaluating an English language textbook. *Workpapers in Teaching English as a Second Language* **XI**, June, 113–117.

O'NEILL, R. (1982) Why use textbooks? *ELTJ* **36**(2), 104–111.

SHELDON, L. (1988) Evaluating ELT textbooks and materials. *ELTJ* **42**(4), 237–246.

TUCKER, C. A. (1975) Evaluating beginning coursebooks. *English Language Teaching Forum* **XIII**(3/4), 355–361.

WILLIAMS, D. (1983) Developing criteria for textbook evaluation. *ELTJ* **37**(3), 251–255.

WILLIAMS, R. (1981) A procedure for ESP textbook analysis and evaluation on teacher education courses. *The ESP Journal* **1**, 155–162.