EDITORIAL COMMITTEE:

Editor:
Anthony Liddicoat
Department of Linguistics
Australian National University
Canberra ACT 0200

Associate Editor:
Ursula Nixon
TESOL Centre
University of Canberra

Reviews Editor:
Timothy Shopen
Department of Linguistics
Australian National University

Editorial Board:
Edith Bavin
Camilla Bettoni
Christopher Candlin
Frances Christie
Michael Clyne
Susanne Döpke
Marietta Elliott
Michael Halliday
Linda Hand
Debra Hoven
Mariko Iwashita
Louise Jansen
Ian Malcolm
Mike McCausland
Tim McNamara
Helen Moore
Howard Nicholas
Anne Pauwels

The *Australian Review of Applied Linguistics* is the journal of the Applied Linguistics Association of Australia, and is published twice each year. Articles on any topic of relevance may be submitted to the editor at any time. Such articles should follow the conventions set out in the style sheet in this edition. Enquiries regarding advertising in the journal should be directed to the editor.

Printed by ANU Printing Services

APPLIED LINGUISTICS ASSOCIATION OF AUSTRALIA

The language testing cycle: from inception to washback

Series S Number 13

1996

Editor of Publications

Anthony Liddicoat
Department of Linguistics
Australian National University
Canberra ACT 0200

Editors of This Volume

Gillian Wigglesworth
Macquarie University
Catherine Elder
The University of Melbourne

ISSN 0817-9514

© ALAA

INVESTIGATING WASHBACK IN JAPANESE EFL CLASSROOMS: PROBLEMS OF METHODOLOGY

Yoshinori Watanabe
International Christian University

ABSTRACT

In Japan it has long been considered that university entrance exams have a negative impact on teaching and learning of English in schools. Recent research, however, suggests that the relationship between testing and teaching is not so simple and requires fine-grained ethnographic research in order to fully understand its complexities. This paper argues the case for the necessity of conducting empirical research before any conclusions about the presence or absence of washback can be drawn. An example of this type of research is offered and the processes involved in investigating washback in the Japanese context are described in such a way as to provide a model for future research in this area.

INTRODUCTION

The impact of exams, which has been known as 'washback' effects in the ESL/EFL literature, is a widely asserted but rarely attested phenomenon. Watanabe (in progress) has come up with more than 500 general claims concerning the impact of the Japanese university entrance exams on classroom practice, whilst finding only ten empirically conducted studies.

Although these studies provide directions which are helpful in formulating research questions, they do not provide the type of information necessary to innovate education through testing. First, in virtually all the studies surveyed, the notion of 'washback' was not fully conceptualised, nor was its scope clearly defined; therefore, it is often unclear what implications could be drawn for improving education through testing. For example, one of the findings reported by Ariyoshi & Senba (1983) was that third grade high school students took practice tests more frequently than second graders. This finding was used to support their argument that entrance exams were influencing high school education *negatively*, in that the lessons were geared increasingly towards

ARAL Series S Nº 13 (1996) 208-239

the coming exams as the grade became higher. However, it cannot be taken for granted that the taking of practice tests is intrinsically harmful. The authors needed to specify more clearly what they meant by 'negative', and to define the role of practice tests in their notion of washback. And before claiming negative effects they would need to examine how the practice tests were used by students and teachers.

Second, since most researchers did not fully analyse the exams which they considered to have an impact, it is not clear whether the observed phenomenon had really been caused by the exam. Fujita (1992) reported that students who were preparing for the university entrance exam tended to learn English by paying particular attention to complex grammatical structures. He speculated that this was because most of the questions employed in Japanese university entrance exams consist of those requiring of testees the ability to understand complex grammar. However, since no in-depth analysis of the exams was reported, his conclusion remains merely a surmise.

Thirdly, most of the studies have university students as subjects. Since these students have already taken the entrance exams, it is not always clear that what the researchers observe is attributable to the effect of the exams. Berwick and Ross, for example, demonstrated that most Japanese students lack motivation to learn English after entering university, and concluded that:

> the intensity of motivation to learn English hits a peak in the last year of high school... Once the university examinations are over, there is very little to sustain this kind of motivation, so the student appears in freshmen classrooms as a kind of timid, exam-worn survivor with no apparent academic purpose at the university
>
> (Berwick and Ross1989: 206)

However, because they dealt with university students, it is not clear from their research whether their students had in fact been motivated during their last year of high school.

Finally, most of the research collected data by indirect means, such as questionnaires and interviews. For example Saito et al. (1984) based their research into the effect of multiple-choice questions on medical education on the self-reports of teachers, and concluded that this type of question distorted teaching in various ways. For example teachers reported that "they tended to teach with an emphasis on the areas that would most likely be covered in the exam" (p. 38 [translation mine]) and that "students tended to study about details rather than main points," (p.43 [translation mine]). However, because their findings were based on questionnaire data rather than on direct observation, we are left in some doubt as to whether these findings accurately reflect either what was happening in the classroom or how students were preparing themselves for the exams.

In summary, then, the research to date has not been successful for the following reasons; first, the notion of the "effect" of the exam has not been clearly delineated; second, the target exams have not been fully analysed; third, data has been gathered from students after, rather than prior to the examinations; and fourth, conclusions have been drawn on the basis of indirect data collection methods, such as questionnaires and interviews.

While much of the investigation of washback is limited by lack of direct empirical support, there are exceptions. Wall and Alderson (1993), and Alderson and Hamp-Lyons (1994) used classroom-based studies to indicate that washback is not so straightforward a phenomenon as it is generally assumed to be. Wall and Alderson (1993) examined the impact of the Sri Lankan O-level English exam on teaching by observing classrooms before and after the implementation of a new exam and comparing the two sets of data. Results provided evidence for the presence of washback (both negative and positive) on lesson content but not on the method of teaching. The methodology used by teachers after the introduction of the new exam was essentially the same as that observed in the baseline study. They concluded that the exam is only one of the factors that "affect how innovations succeed or fail" (1993: 68). Alderson and Hamp-Lyons (1993), investigating the the washback effect of TOEFL, compared a total of eight non-TOEFL classes to eight TOEFL preparation classes. Results suggested a clear impact of the test on the "curriculum, text books, teachers, attitudes and the content of classes aimed at TOEFL". However, the impact on teaching methodology was much more complex, and in this area they concluded that "the differences between teachers are at least as important as the differences between TOEFL and non-TOEFL" (1993: handout).

To deal with this kind of complexity there is a clear need for different types of observational research in various contexts. The contexts in which tests are used differ greatly, and the type of exam and its potential influence also varies. Test purpose is also variable: some tests

are used to select students; others to place students in classes according to their proficiency levels. Thus, the result obtained in one research area may not be readily generalisable to another context. For example, in Japan, more than 1,000 different university entrance exams are carried out each year, so the results of research relating to a single examination like TOEFL, may not be applicable to the overall situation.

Washback is an extremely complex phenomenon. It can be conceptualised on different dimensions, involving a number of variables, such as teaching, learning, interaction between teachers and students, students' motivation, etc. and therefore needs to be examined from multiple perspectives. Classical experimental research is not appropriate because of the difficulty of controlling for all these variables.

Ethnography, on the other hand may be a more suitable approach to understanding the complex nature of washback. For LeCompte and Preissle (1993: 3) ethnography is characterised by four research strategies. First, ethnography elicits phenomenological data that "represent the world view of the participants being investigated," and "participants' constructs are used to structure the research." Second, in this method, participant and nonparticipant observation is used "to acquire firsthand, sensory accounts of phenomena as they occur in real world settings." Third, ethnographers "seek to construct descriptions of total phenomena within their various contexts and to generate from these descriptions the complex interrelationship of causes and consequences that affect human behavior toward and belief about the phenomena." And fourth, the ethnographic researchers use "a variety of research techniques to amass their data."

The present paper is part of a larger study on the washback effects of the Japanese university entrance examinations (Watanabe, in progress). To demonstrate the value of the ethnographic approach it will focus largely on the research process. Provisional conclusions will nevertheless be presented at the end of the paper.

The paper is in several sections. The first section describes the Japanese educational context which is the setting in which the research was conducted. In the next section, the process of the research is described in several stages, the research questions are identified and the approaches employed to address them discussed. The final section summarises the provisional findings, and identifies the need for further empirical research using methods derived from ethnography.

UNIVERSITY ENTRANCE IN JAPAN

The Japanese university entrance examination system

In Japan the school year runs from April 1 to the end of March. The entrance examinations of most universities are administered in January and February. Each university department produces its own examination which is offered on the campus. As no information concerning the exam (such as scoring methods, weighting of each section, etc., not to mention the intended effect of the exam on teaching and learning) is made public, the most important sources of information for the examinees are the previous year's examinations.

There are three major types of university in Japan; i.e., national, local public, and private. According to the Ministry of Education (1994), out of a total of 534 four-year course universities, 98 (18.4%) were national, 46 (8.6%), local public, and 390 (73%), private. Entry to universities is highly competitive, since more than half of upper secondary level graduates hope to enter tertiary institutions. In 1993, for example, 59.6% of the total of 1,760,000 graduates took the exam; 67% were accepted into university. Thus, as elsewhere, demand outstrips supply. Further, most students who fail to enter their target university undertake a further year of study and take the entrance exam the following year. These so-called *ronin* students contribute to the increase in the number of applicants in the following year. In addition, universities are ranked and there is a widespread belief that entering higher rank universities guarantees better jobs after graduation. Thus university entry is highly competitive.

The teaching of English in Japan

With only a few exceptions (e.g., the natural science-related departments of some universities), virtually all Japanese universities require English or other foreign languages in their screening process.

Secondary schools are expected to follow the guidelines issued by the Japanese Ministry of Education. Various textbooks based on the guidelines are published by private publishers commissioned by the Ministry.

The role of English is gradually changing in Japan; English used to be merely a medium of gaining 'high culture' through the written language, but now, in reponse to the need for international communication, there is a greater emphasis on active language use involving exchange of information in both speech and writing. To meet these external demands, a new set of guidelines was officially announced in 1989, and implemented in 1994. The first entrance examinations based on the new curriculum will be held in 1997. The new guidelines are more specific than the preceding ones about the nature and range of oral communication skills (including daily conversation, note-taking, giving speeches, and debate), as well as reading and writing to be taught. Given this greater emphasis on communication in EFL teaching, the EFL examinations have been criticised for their lack of communicative content. The need for research into the impact of the exams is called for in such a context.

One of the purposes of this research then was to gather base-line data to allow eventual comparisons to be drawn between the current situation in Japanese EFL classrooms and that which will occur in 1997, once the new curriculum has been introduced.

THE RESEARCH PROCESS

The ideas and terms for the research process adopted here are drawn from Evertson and Green (1986). Figure 1 illustrates the process and the description which follows is based on the diagram.

Received wisdom

My initial interest in washback developed from my experience as a teacher, researcher, and student who had gone through the process of preparing for the university entrance examinations. During this period I subscribed to the prevailing view held in the Japanese society that exams had a negative impact on learning and teaching. It was believed that if the exam were to be improved, then innovation in teaching and learning would automatically follow.

Start questioning

This assumption was called into question when I conducted research into the relationship between language learning strategies and the screening method of a junior college (Watanabe, 1990). In this study, two types of college students were compared with respect to the range and types of language learning strategies they used. The first type of student had entered the college by taking the entrance examination; the second, by recommendation. The second type of student was completely exempted from the exam, and was accepted on the basis of high school records. It was predicted that the examination students would report using a narrower range of strategies, such as rote-memorisation of vocabulary items, translation, which would benefit their test-taking. The 'recommended' students, on the other hand, were expected to report using a wider range of more communicatively-oriented strategies (eg., learning vocabulary items by using them in actual contexts). The favourable high school records of these students were taken as an indication that they had done well in classrooms where the communicatively-oriented Ministry of Education guidelines had supposedly been implemented. Contrary to these predictions, the result indicated that the entrance examination seemed to have a positive effect on strategy use, in that the exam students reported employing a greater range of strategies than those accepted on the basis of school performance alone, and the type of strategies included communicatively-oriented strategies, such as learning English by using it in actual communicative situations. However, since the research was based on student self-reports rather than direct observation, the reliability of these findings was uncertain. Moreover, it was not clear what had actually been taught in their high school classes. Thus, this was the stage where I felt it necessary to conduct empirical research into washback.

Personal theory

Before designing the research, the initial theorising of washback was first expressed in the form of research questions. The questions raised at this stage were as follows;

- Does washback exist?
- What evidence enables us to say washback exists or does not exist?
- If washback exists, what is its nature (i.e., positive or negative)?
- If washback does not exist, why not?
- If washback exists, under what conditions?

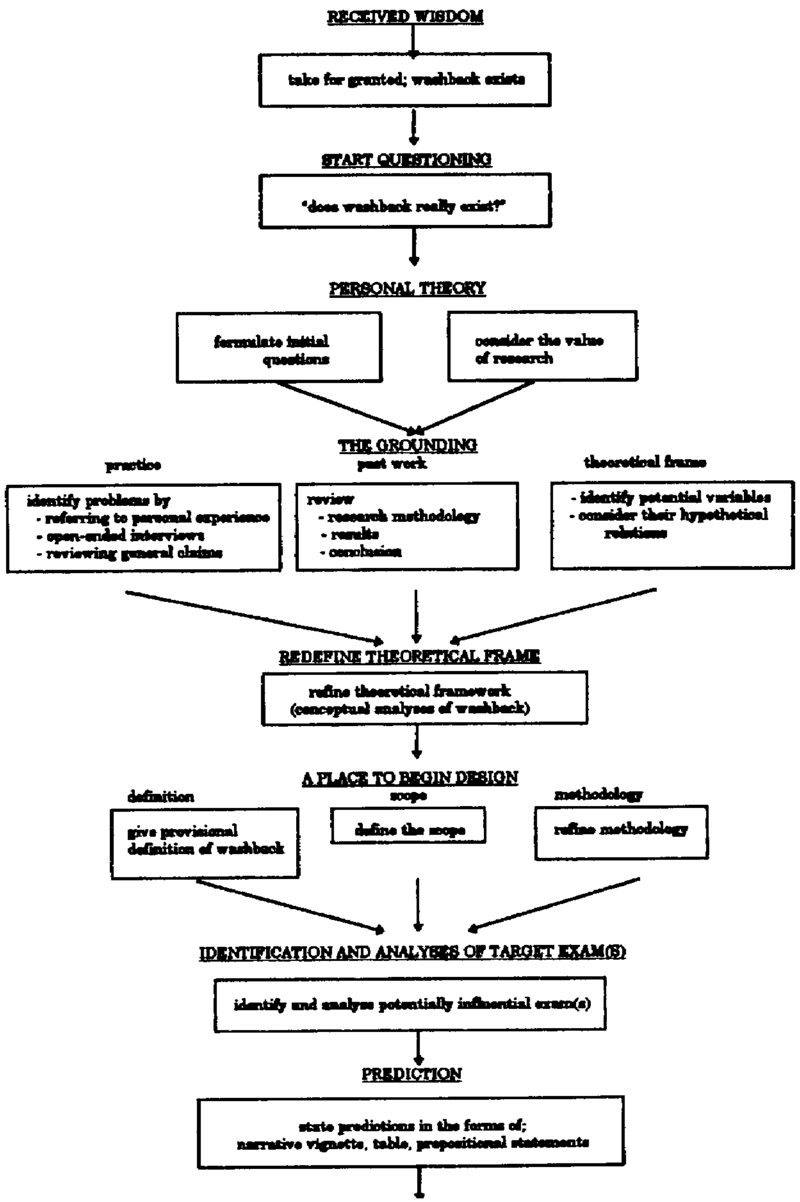

RECEIVED WISDOM
take for granted; washback exists
START QUESTIONING
"does washback really exist?"
PERSONAL THEORY
formulate initial questions
consider the value of research
THE GROUNDING
practice
past work
theoretical frame
identify problems by
- referring to personal experience
- open-ended interviews
- reviewing general claims
review
- research methodology
- results
- conclusion
- identify potential variables
- consider their hypothetical relations
REDEFINE THEORETICAL FRAME
refine theoretical framework (conceptual analyses of washback)
A PLACE TO BEGIN DESIGN
definition
scope
methodology
give provisional definition of washback
define the scope
refine methodology
IDENTIFICATION AND ANALYSES OF TARGET EXAM(S)
identify and analyse potentially influential exam(s)
PREDICTION
state predictions in the forms of;
narrative vignette, table, propositional statements

Figure 1: *Process of washback research*

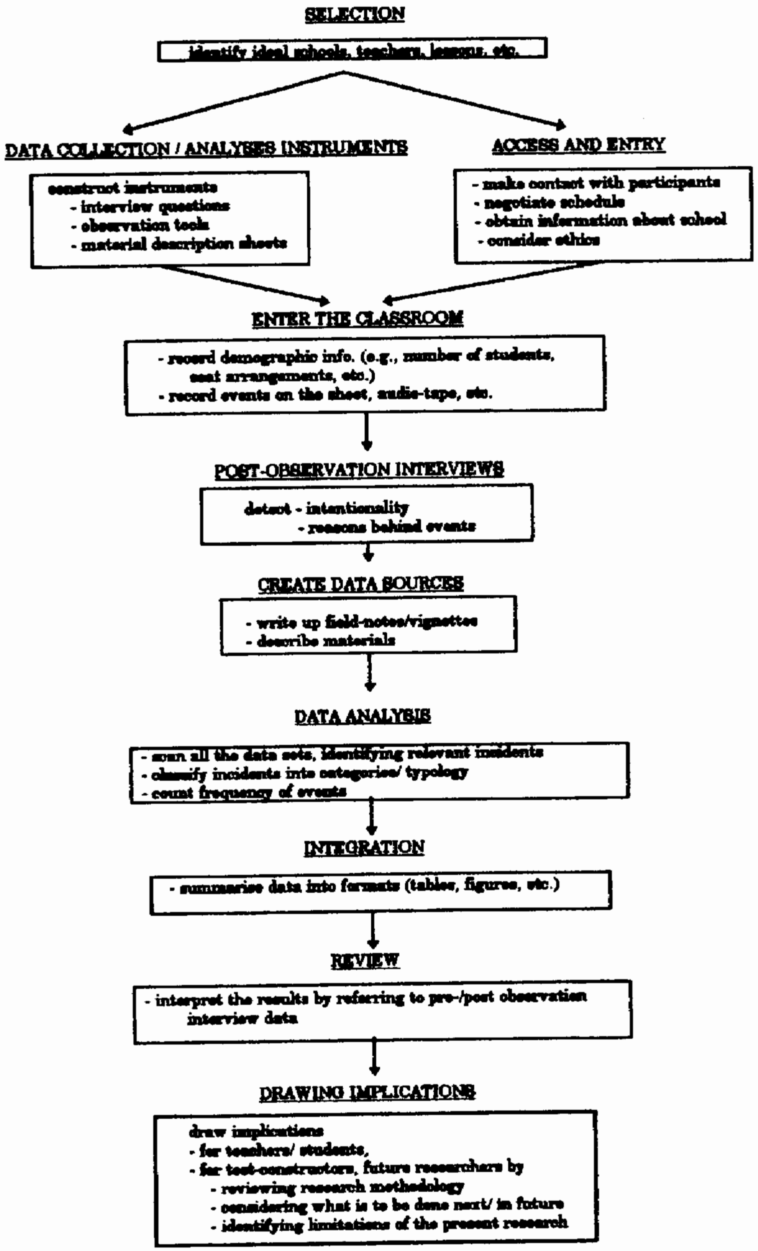

SELECTION
identify ideal schools, teachers, lessons, etc.
DATA COLLECTION / ANALYSES INSTRUMENTS
construct instruments
- interview questions
- observation tools
- material description sheets
ACCESS AND ENTRY
- make contact with participants
- negotiate schedule
- obtain information about school
- consider ethics
ENTER THE CLASSROOM
- record demographic info. (e.g., number of students, seat arrangements, etc.)
- record events on the sheet, audio-tape, etc.
POST-OBSERVATION INTERVIEWS
detect - intentionality
- reasons behind events
CREATE DATA SOURCES
- write up field-notes/vignettes
- describe materials
DATA ANALYSIS
- scan all the data sets, identifying relevant incidents
- classify incidents into categories/ typology
- count frequency of events
INTEGRATION
- summarise data into formats (tables, figures, etc.)
REVIEW
- interpret the results by referring to pre-/post observation interview data
DRAWING IMPLICATIONS
draw implications
- for teachers/ students,
- for test-constructors, future researchers by
- reviewing research methodology
- considering what is to be done next/ in future
- identifying limitations of the present research

Figure 1: Process of washback research (continued)

These five core questions remained unchanged throughout the research.

This stage involved consideration of the value of the research. Clearly research addressing these questions has important implications for Japanese society, since much energy, time and money are spent on the exam at individual, school, and national levels every year. Empirically-based research was seen as an essential investment.

The grounding

Evertson and Green (1986) divide this stage into three components; practice, past work, and theoretical frame.

Practice

The purpose of the practice component was to identify the existing problems with the exams. This information was derived from two main sources; firstly open-ended interviews with the people who had been involved in the entrance examination (including university students, high school teachers, those who were working at special preparatory school called *yobiko*), and secondly, investigation of a range of claims published in magazines, newspaper articles, TV broadcasts, etc. The claims were summarised on note-cards. Each item was then labelled with a tentative title. The source and page numbers were also referred to, and my comments were recorded in the margin. A sample card is shown below:

> [title] negative effects of objective type questions/ positive effects of subjective type questions
>
> ... Our university has recently decided to reform our entrance exam by replacing objective type questions with subjective type, because the objective test is likely to have extremely negative effects on students... (Nakamura, 1985: 11) [translation mine].
>
> [comments] What is referred to by 'subjective' 'objective'; unclear. No evidence is cited.

The note-cards were classified according to the titles, and the most frequently claimed problems were then summarised as sets of assumptions as follows:

Assumption 1: Aural/oral aspects of English are neglected in exam preparatory lessons.

Assumption 2: In exam preparatory lessons, a greater emphasis is placed on de-contextualised mechanical drills rather than communication exercises.

Assumption 3: In exam preparatory lessons, test-taking techniques are taught.

Assumption 4: Exam preparatory lessons are characterised by teaching based on grammar-translation, with detailed explanations about minute grammar points.

Assumption 5: The exam drives students to work harder.

Assumption 6: Different testing methods, such as objective/subjective, and direct/indirect, induce differential washback.

Assumption 7: If the target examination includes listening, then this skill is taught in the course.

Assumption 8: If the target examination requires translation, then the teaching is more likely to be based on grammar-translation.

These assumptions served as "base-line" assertions, around which the subsequent part of the research was organised.

Past work

In the component called past work, the following set of questions was addressed:

- has anything been done in the area?

If yes;

what has been done?
- what methodology has been used?
- what results/conclusions have been drawn?
- what has been lacking in past research?

- would there be any problems particular in the present context (i.e., the Japanese university entrance examinations)?

In addition to the problems identified through the literature review, summarised in the introduction of the paper, there was a problem particular to the context of Japanese educational system. In Japan, as already noted, there are more than 1,000 different types of exams administered every year. Therefore, it was anticipated that identification and selection of target exams for the research would be difficult. (This problem is dealt with in a later section on exam analysis.)

Theoretical frame

The theoretical frame was developed from Alderson and Wall (1993), specifically using their fifteen washback hypotheses. The variables expressed in these hypotheses were subsequently diagrammed to conceptualise washback as a phenomenon, as shown in Figure 2.

Redefine theoretical frame

The major task at this stage was to conceptualise washback as a phenomenon by examining all the information gathered at the previous stages. First, by reconsidering the set of assumptions summarised in the practice component (see above), washback was conceptualised on the three bipolar dimensions outlined below.

1) *Specificity.* This is best illustrated by example - there is an assumption that if a test has a listening component, then the teaching of listening will be emphasized in classes. This is placed on the highly specific end of the dimension, since this assumption refers specifically to tests which include a listening component. On the other hand, there is an assumption that entrance exams will drive students to work harder than otherwise. This type of assumption does not refer to any specific type of exam, but rather to the impact of the exams in general. Thus, this sort of claim is placed on the lower end of the specificity dimension.

2) *Observability.* This dimension ranges from washback on overt behaviours to that on covert internal behaviours. If we address, say, the assumption that the impact of the exam on students leads to changes in

their reading strategies, then this type of problem is placed on the less observable end of the scale. On the other hand, if the exam affects teaching and/or learning in a way that leads teachers and students to engage in mechanical drills rather than communication activities, then this type of issue is placed on the more observable end of the dimension.

3) *Intentionality.* Intentionality may be 'direct' or 'indirect.' Direct washback occurs when the teacher has a clear intention of preparing students for exams. Indirect washback is observed when the teacher does not have such an intention, but nevertheless uses materials similar to those often included in exams. For example, there was a case in the present research where the teacher was using past exam papers without any intention of preparing students for exams. This type of washback was deemed to be 'indirect' rather than 'direct,' and was treated separately.

These dimensions proved valuable in making decisions at subsequent stages of the research. For example, the specificity dimension is implicated in selecting types of target lessons and exams. That is, to investigate the more specific assumptions, it would be necessary to select courses which aimed at preparing students for a specific university exam (out of the hundreds of different types). On the other hand, to focus on the less specific assumptions in the research, it would be necessary to select those courses which aim to prepare students for examinations generally as well as regular courses without any preparatory purpose.

The observability dimension affected the construction of research tools, since investigation of internal learning processes demands different instruments from those required for observable phenomena.

Intentionality was investigated through interviews which explored the reason behind various incidents which had taken place in the classroom.

The relationship between the variables identified by Alderson and Wall (1993) were conceptualised in a diagrammatic form in Figure 2. Note that 'micro- and 'macro-contexts' have been added as variables in the present research, since it was deemed important to spell out these variables clearly, so that the meaning of the test was clear to other researchers. Without going into further details about the diagram (see Watanabe, in progress), suffice to say that it provided a useful starting point for planning the research, although this initial conceptualisation underwent modifications as the investigation progressed.

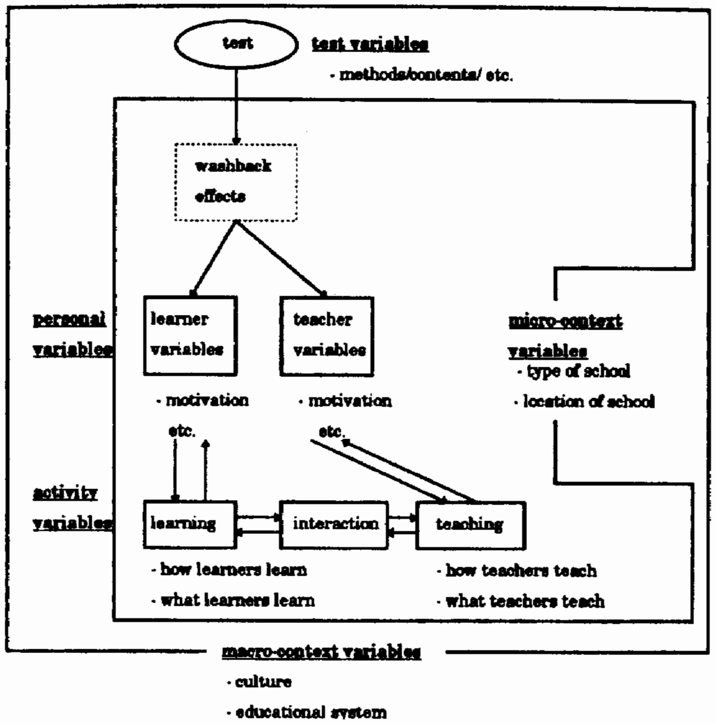


Figure 2: Relationship between variables in washback research

A place to begin research

After establishing an overall theoretical framework, the main observation research started. The task at this stage is divided into three components; formulating a working definition of the term "washback", defining the scope of the research, and determining methodology.

In the present research, washback was defined as "the impact of the examinations on what would happen in the classroom" (Alderson and Wall, 1993: 115). At this stage, the research methodology was further refined. For example, it was decided to adopt a 'cross-sectional' rather than 'longitudinal' approach; that is, in order to examine washback, different groups were observed within the same time span, rather than the same group at different points in time. It was felt that the presence or absence of washback could be satisfactorily established by comparing and contrasting the classroom events relating to a range of different exams.

Identification and analysis of the target exams

As mentioned above, in the Japanese context, where more than 1,000 different types of exams are administered every year, it was not possible to analyse all the exams. Thus, reference was made to a compilation of past exam papers published by Obunsha, a private company which deals with exam-related materials. In the 1994 edition, a total of the 280 most popular university exams were compiled, including 101 from national/ local public university exams, and 179 from private university exams. From the analysis of the papers, the following question types were considered relevant to the general assumptions (see above) of the present research:

tests measuring aural/ oral aspects of English

- listening
- speaking (indirect, using written dialogue)
- paper test of pronunciation

reading tests

- translation (from English into Japanese)
- summary in Japanese
- gap-filling
- multiple-choice questions
- questions requiring a visual response

writing tests

- direct writing tests (e.g., essay, picture description, etc.)
- translation (from Japanese into English)

tests of grammar and word usage

vocabulary tests

On the basis of these categories, the number of universities employing each question type was counted and the following tendencies were observed. There were marked differences between the exam formats of national/local public university exams on the one hand and of private university exams on the other. National/local public exams comprised

73% of 'constructive response types' (requiring testees to produce answers in their own words); with the remaining 27% being 'selective response questions' (which require students to select an answer from options). Private university exams were made up of 91% of 'selective response questions' and only 9% of 'constructive response questions'. The constructive response type questions were generally either translations (from English into Japanese or vice versa) or passage summaries.

Common tendencies emerging from the analysis were first, that only 13% of the examinations included direct listening tests; second, only 10% included direct writing tests; third, only 5% employed reading questions requiring a visual response (e.g., picture, diagram, etc.); and fourth, approximately a quarter included de-contextualised grammar, word usage, and vocabulary questions.

Prediction

Predicted research outcomes were based on results of the test analysis. These ranged from general to specific on the basis of the specificity dimension established earlier. At the most general level, it was predicted that if washback existed, there would be differences in activities and materials used in regular lessons compared to exam preparatory lessons, where both were taught by the same teacher. More specifically, it was predicted that regular lessons would be taught based on the Ministry of Education textbooks, but that exam preparatory lessons would be characterised by:

- neglect of aural/oral aspects of English (Assumption 1)
- emphasis on de-contextualised mechanical drills (Assumption 2)
- emphasis on teaching of test-taking techniques (Assumption 3)
- heavy reliance on grammar-translation (Assumption 4) and
- greater student enthusiasm for learning English (Assumption 5).

Most specifically it was predicted that if washback did exist, there would be differences in courses designed for different target university exams. These differences were expected to manifest in the following areas:

- test method; i.e., objective vs. subjective, direct vs. indirect (Assumption 6)
- the skill area, listening in particular (Assumption 7), and
- grammar-translation (Assumption 8).

Selection[1]

Prediction 1 was to be investigated by observing high school classes. It was expected that teaching in regular lessons would reflect the communicative orientation expressed in the Ministry of Education guidelines, while the exam preparatory lessons would be oriented towards more formal aspects of English reflecting the exams. However, high schools were not appropriate for attesting the more specific predictions 2 and 3 since high school students vary widely in terms of their target exams, so the teacher must teach students aiming for various target exams at the same time. On the other hand, private extra-curricular institutions, (*yobiko* literally, 'preparatory schools') were ideal, since they offer a range of courses targeting specific university exams.

Schools were selected based on the following criteria: (i) in high schools, two teachers were selected who taught the same preparatory and regular courses; (ii) in the *yobiko*, two teachers were selected who taught two different target university exam courses; i.e., both taught courses aimed at the exam for national/ local public universities and the private university exams. Both teachers selected taught to the same two exams.

Access and entry

In identifying the schools, an attempt was made to select those with a reputation for successful student acceptance at prestigious universities, since it was assumed washback would be observable in such schools. However, various school background variables were also considered in the selection, such as educational methods, curricula, student characteristics, etc., since it was conceivable that if exams exercised sufficient power over education, then washback would override these background variables.

The following observations were made:

High School A			
Teacher A1	Regular	79 minutes	June 24, '94
	Exam preparation	79 minutes	June 27, '94
Teacher A2	Regular	57 minutes	June 25, '94
	Exam preparation	67 minutes	June 24, '94
High School B			
Teacher B1	Regular	98 minutes	June 29, '94
	Exam preparation	80 minutes	June 30, '94
High School C			
Teacher C1	Regular	47 minutes	Sept 29, '94
	Exam preparation	47 minutes	Oct 6, '94
Teacher C2	Regular	50 minutes	Nov 17, '94
	Exam preparation	48 minutes	Nov 28, '94
Yobiko			
Teacher Y1	A private university course	456 minutes	Aug 4 to 9, '94
	A national university course	451 minutes	Aug 4 to 9, '94
Teacher Y2	A private university course	508 minutes	Dec 25 to 29, '94
	A national university course	382 minutes	Jan 3 to 7, '94

Note that practical constraints meant that the selected samples were not ideal. First, in High School B it was not possible to achieve a perfect "pairing", with two different teachers teaching both regular and exam preparatory lessons, Teacher B1 was in fact the only teacher who taught both types of course. Also the length of observations possible in high schools was limited because of a conflict of schedules - the researcher could only carry out observations during the intersessions between terms at university. Unfortunately, high schools were also on holidays during these periods. In addition, high schools had various extra activities, such as mid-term exams, final exams, sport festivals, cultural festivals, etc. which made regular observation difficult. On the other hand, with the *yobiko* it was possible to observe whole intensive courses for specific target university exams, which were held during summer and winter holidays.

High School A & B were public schools located in rural areas. High School A had quite a high reputation for successful preparatory teaching in the area. High School B was reputed both for exam preparation and for their unique English teaching method - a type of immersion program where the entire lesson was taught in English. High School C was a private school located in a metropolitan area. This high school attracted numerous returning students, who had graduated from junior high schools in English-speaking countries. The *yobiko* was located also in a metropolitan area. This school attracted students from all over Japan, and had a high reputation for its emphasis on teaching English for entrance examinations.

Data collection and analysis instruments

While access to schools was being established, the data collection instruments were constructed. There were two major components.

The first comprised sets of pre-observation interview questions. Questions focused on information about the teacher's background, for example, education, age, major field of study. Questions about the teacher's perception of the entrance exam were avoided since such questions might affect his/her teaching, which in turn might pollute observations.

The second component consisted of lesson description sheets (in addition to pencils, audio-recorder, stop-watch). The observation sheets were designed to record the following types of information: time sequence, materials used, teacher activities/speech, student activities/speech, what was written on the chalkboard, observer's comments and the questions which came to the observer during the observation. These questions were asked of the teacher at post-observation interviews.

Enter the classroom

Non-participant observation was employed in the research: that is, the observer's involvement was kept to a minimum and the observer made

every attempt to avoid disturbing the lesson. When the teacher entered the room, stop-watch and audio-recording were started at the same time.

The initial purpose of the observation was to develop 'low inference' categories - clear enough to be specified in behavioural terms. An attempt was made to record new events, felt to be relevant to the research predictions, in as precise and concrete a manner as possible.

Each lesson was recorded on audio tape, to permit subsequent frequency counts based on the categories developed.

Post-observation interviews

After observing the lessons, post-observation interviews were conducted with the teachers concerned. The interview had two major purposes: to gather data about the teacher's perception of the impact of the exams, and to examine the intentions which lay behind the various 'events' occurring in the classroom. These data served as a source for interpreting the results.

Create data sources

Immediately following the observation and interviews, while memory was still fresh, the lesson description sheets were reviewed, and ambiguous parts clarified by listening to the audio tape. In addition, the interview data was summarised in propositional formats. Subsequently, the descriptions of the classroom events were summarised in the form of 'narrative vignettes', ("a more elaborated, literally polished version of the account found in the field-notes" Erickson, 1986: 150). The vignettes were designed to illustrate the overall picture of the classrooms and to assist with remembering the overall flow of classroom discourse at the stage of the frequency analysis.

At this stage, the classroom materials, which consisted of the textbook and supplementary exercises, were analysed according to the question type categories identified at the stage of the exam analysis (see above).

Data analysis

Both materials and classroom events were analysed. The latter are focussed upon here as these were the core of the research process. The purpose of the data analysis was two-fold: first to identify categories relevant to the predictions, and systematise them into a coding scheme for counting the frequency of incidents; and second, to code incidents which belonged to each category by listening to the audio tape.

Where possible, the data analysis was conducted immediately after the observation. The initial stage of the analysis involved repeated scannings of the lesson description sheets in order to become familiar with the data. Next, various incidents deemed relevant to the research predictions were circled. The identified incidents were word-processed in propositional form, and labels were attached indicating the nature of the incident, the teacher involved and the context in which the event occurred. For example:

23/6/1994

1 HA/TAe teacher reads aloud a paragraph.

t's turn/ read aloud/ a paragraph/ aural-oral/ English

24/6/1994

2 HA/TAr teacher calls on students to translate a sentence.

t's turn/ interact/ elicit/ Jpn/ a sentence

24/6/1994

3 HA/TAr student translates English into Japanese.

s's turn/ translate / a sentence/ Jpn —> Engl.

HA = high school A
TA = Teacher A
e = exam preparatory lesson
r = regular lesson

Incidents were dated and numbered for future reference. After identifying and labeling all the incidents, they were classified into categories. To systematise the categories, reference was made to the COLT (Communicative Orientation of Language Teaching) scheme

(Allen, Froehlich and Spada 1984) because of its flexibility and comprehensiveness for capturing a wide range of communicative features of language teaching. This system was referred to in two ways; first, some category labels of the system were adapted to the categories derived from the analysis of the field-notes; and second, the format of the system was employed to systematise the separate categories into a larger system and render it workable for actual data coding.

All the derived categories were then systematised into a typology as follows:

- classroom organisation patterns
  - teacher-fronted
  - pair-work
  - group-work
  - students' presentation
- interaction
  - turn-taker
    - teacher or student/s
  - language medium
    - English or Japanese
  - types of information and exchange patterns
    - requesting pseudo or genuine information
    - giving predictable or unpredictable information
    - reacting to form or message
- activities
  - choral
  - read aloud
  - paraphrase
  - translation from English into Japanese
  - translation from Japanese into English
  - summary in Japanese
  - summary in English

topics

examination
language form or use

use of metalanguage

incidents of laughter

The last category 'incidents of laughter' had not been predicted, but the category was incorporated, since it was found during observations that laughter was an important indicator of classroom atmosphere (cf., Alderson and Hamp-Lyons, 1994).

On the basis of the typology, a new coding scheme, COEPREC (Communicative Orientation for Exam Preparatory Classes), was developed. A sample copy is provided in Appendix A. Like COLT, the system is divided into two sheets, A and B. Sheet A was prepared to measure 'classroom organisation patterns' and 'the use of materials', which were appropriately measured by referring to the length of time, while Sheet B, various categories, which were to be measured by 'frequency'. The audio-recording was played again to code classroom events on COEPREC, and the length of time and the frequency were subsequently computed. Because the focus of this report is on the qualitative aspects of the research, we will not elaborate here on the coding and the data analysis procedures. Interested readers are referred to Watanabe (in progress).

Integration

The results of the coding were subsequently summarised into tables, called 'Washback Grid.' A sample table, which contains the information concerning teaching methodology of high schools, is provided in Appendix B. Note that when summarising data into tables, some COEPREC categories were combined with others. For example, in order to examine the frequency of using English aurally/orally, all the utterances made in English for interaction and activity were combined into one larger category "aural/oral." By examining closely the whole table, three forms of washback were observed in each category; i) washback does not exist, ii) washback exists with some teachers but not with others, and iii) washback exists with all teachers. For example, the table in Appendix indicates that the examinations influenced some teachers but not other teachers in the area of 'types of information exchanged'. It had been predicted from Assumption 2 (i.e., in exam preparatory lessons, a greater emphasis is placed on de-contexualised mechanical drills than communicative exercises) that 'pseudo/predictable' information, (i.e., the type of information that is known between the two persons, thus, indicating that the mechanical exercises were conducted) would be exchanged more frequently in exam preparatory than in regular lessons. However, in results obtained thus far, this type of assumption was not completely supported; washback was clearly present only in the lessons of Teacher C, whereas in the lessons of other teachers evidence of washback was slight.

Review

After the data had been summarised, the results were interpreted by referring to the pre-/post-observation interview data. This task was important in answering the basic question posed earlier: whether washback does or does not exist. From time to time during the process of interpreting the classroom events against the interview data, several possible interpretations emerged. When this occurred each of the "rival" explanations was noted.

For example, listening was taught in some of the high school exam preparatory lessons, but not in others. One of the reasons emerging from interviews was that the teachers who included listening in their class believed that the exam preparation could provide students with a good opportunity for developing communication skills. On the other hand, those who did not include listening claimed that there were too few support-materials available, and that they did not know how to teach listening skills effectively.

Another example is the use of the grammar-translation method. It has generally been argued that the teaching methods of pre-college EFL in Japan are based on grammar-translation because the university entrance examinations require translation and grammatical analyses. However, it seemed that teacher orientation was a more important factor in determining what happened in the classroom than the content of the exam. One teacher used grammar-translation irrespective of the target exam, while the other used this method when he was explaining grammatically complex sentences, and when he was teaching for the target exam which often tested translation and complex grammatical structures. Thus, the latter teacher seemed to be influenced by the exam, while the former did not. The differences seem to be partly attributable to the teachers' educational backgrounds. The teacher who used grammar-translation had majored in theoretical linguistics at post-graduate level, while the other teacher who translated texts only occasionally, had gone through a teacher training course where communicative teaching was emphasised. Another possible source of the difference seemed to come from the teachers' past experience of learning English; for example one teacher reported that he felt it very difficult to teach English by other methods than the one by which he was taught.

At this stage, in addition to interpreting the results, the whole research process was also reviewed by examining what was lacking and what could be done in future research. For instance, as mentioned earlier, this was a cross-sectional study, so it left unanswered the question of whether teaching strategies would change as the exam period approached. Such an issue can only be addressed through longitudinal research, in which the same class is observed at different periods over time.

Drawing implications

The research findings thus far suggest that the incidents which were observed in each lesson did not greatly differ according to the purpose of the lessons (i.e., regular or exam preparatory) of high schools or the target exams of the *yobiko* lessons (i.e., national/local public or private). Differences between teachers and schools were great. The results suggest that changes to exams will not automatically lead to changes in teaching; but rather, teacher factors seem to play a greater role in deciding what happens in the classroom. It may be that in-service/pre-service teacher training should incorporate teaching methods which show how exams can be used to develop communicative ability. As the research progresses, however, it is expected that other implications will be drawn for other audiences, including test constructors, the Ministry of Education, etc.

CONCLUSION

In this paper, an approach to washback research which draws on ethnographic methods has been described. Despite the initial intention of demystifying the research procedure, the description may still seem to be highly complex. Such complexity is necessary and desirable because the phenomenon of washback itself is complex, and we are still at a preliminary stage of the research.

I would argue that an approach informed by the principles of ethnographic research is an appropriate way to deepen our understanding of the nature of washback. In support of this argument I now return briefly to the four research strategies identified by LeCompte and Preissle (1993) which were referred to earlier in this paper.

1) Ethnography elicits phenomenological data that represent the world view of the participants being investigated and participants' constructs are used to structure the research.

In ethnographic research, a researcher is only one of a range of participants, so his/her perception represents only view amongst many. Because tests are used in a particular context for a specific purpose, it is important before embarking on the research to identify the problems which participants themselves see as important. Otherwise, the research may neglect problems considered crucial by participants in favour of those which may be deemed irrelevant and extraneous. Consequently, the research results are likely to be sterile, having few implications for those most affected by the problems. It is for this reason that I have attempted to base my research not only on the findings of previous research, and on my own experience as an EFL teacher but also on views of washback gathered from a range of different sources, including other teachers, students and the media.

2) Ethnography employs participant and non-participant observation to acquire firsthand, sensory accounts of phenomena as they occur in real world settings.

There are two key phrases in this statement; "firsthand, sensory accounts of phenomena" and "real world settings." First, if the washback research were not to gather firsthand data, we would be obliged to take what teachers and students say about how they feel about the effects of examinations at face value, even though their perceptions may not reflect what they are actually doing. As Hopkins correctly points out

> there is often incongruence between a teacher's publicly declared philosophy or beliefs about education and how he or she behaves in the classroom.
>
> (Hopkins, 1985: 48)

Second, ethnographic research stresses gathering natural data. If for example it were to be found that a test administered as part of a controlled experiment had no impact on teaching and learning, this result could not be generalized to actual exam preparatory courses since tests administered under experimental conditions would be likely to be perceived by participants as having no educational consequences. Tests always play a certain role in a specific context, so washback research needs to take these contextual factors into account.

3) In ethnographic research, the researchers seek to construct descriptions of total phenomena within their various contexts and to generate from these descriptions the complex interrelationship of causes and consequences that affect human behaviour toward and belief about phenomena.

As pointed out by Alderson & Wall (1993), the exam may be only one of the factors that affect how innovations succeed or fail, and this study also seems to support this assertion. In other words, numerous variables other than the exam, such as the teacher's beliefs, his/her familiarity with the teaching method, etc., seem to be involved in determining what happens in the classroom. This type of insight could not have been gained without an attempt to describe the total phenomena of the classroom, including the teacher's perceptions about his/her teaching.

The small portions of the on-going research reported in this paper are sufficient to indicate that the washback effects are highly complex phenomena - a fact which is somewhat at odds with popular opinion. Educational innovations do not seem to spring automatically from innovations to examinations. Rohlen, an anthropologist, once observed five Japanese high schools, and

found that there was a large amount of variation among them. He states:

> Neither Japanese society nor its high schools are monolithic, a point too regularly ignored in foreign treatments of Japan.
>
> (Rohlen, 1983: 43)

His final comment is also suggestive for washback research;

> It is easy to suggest cause and effect relations between schooling and the lamentable aspects of society, and it is equally difficult to disprove them. Sweeping interpretations containing some answers appeal to the popular mind and are useful in campaigns to reform education.
>
> (Rohlen, 1983: 327)

The present research likewise warns us against basing educational innovation solely on public opinion.

4) Ethnographic researchers use a variety of research techniques to amass their data.

Direct observation is important, but not the only method appropriate for washback research. Rather, various research methods, including self-report, document analysis and interview, should be considered to complement each other. Without the interviews, for example, it would not have been possible to detect the reasons (or intentions) behind teachers' behaviour in the classroom. The problem is which method should be employed at which point, and it is hoped that this paper will help researchers to make an informed decision in conducting their own research.

Before establishing a theory of washback, we clearly need more empirical data gathered from different contexts. Results of such research will be of great assistance to all those involved in testing, when they have to make decisions about the construction and use of tests. Pilliner (1973) once said: "the most important requirement of a good exam is that it should be educationally beneficial" but, regrettably, this has thus far been a neglected aspect in the cycle of testing.

NOTES

1 'Selection' is the qualitative equivalent to the term 'sampling' used in the positivistic tradition. According to LeCompte and Preissle (1993), "selection requires only that the researcher delineate precisely the relevant population or phenomenon for investigating, using criteria based on theoretical or conceptual considerations, personal curiosity, empirical characteristics, or some other consideration" (p. 57).

REFERENCES

Alderson, J. C. and D. Wall (1993) Does washback exist? *Applied Linguistics* 14,2:115-129.

Alderson, J. C. and L. Hamp-Lyons (1994) TOEFL preparation and the communicative classroom. Paper presented at TESOL conference, Baltimore, March.

Allen, P., M. Froehlich, and N. Spada (1984) The communicative orientation of language teaching: and observation scheme. In J. Handscombe, R.A. Orem, and B. P.Tayler (eds) *On TESOL '83*. Washington D.C: TESOL.

Ariyoshi, H., and K. Senba (1983) Daigaku nyushi junbi kyoiku ni kansuru kenkyu (A study on preparatory teaching for the university entrance examinations). *Fukuoka Kyoiku Daigaku Kiyo*, 33: 1-21.

Berwick, R. and S. Ross (1989) Motivation after matriculation: are Japanese learners of English still alive after exam hell? *JALT Journal*, 11: 193-210.

Erickson, F. (1986) Qualitative methods in research on teaching. In Wittrock, M.C. (ed.)*The handbook of research in teaching*, 3rd ed. New York, Macmillan.

Evertson, C. and J. Green (1986) Observations as inquiry and method. In Wittrock, M.C. (ed.) (1986) *The handbook of research in teaching*, 3rd ed. New York, Macmillan.

Fujita, T. (1992) Readiness model of optimal input: a comparison between Japanese high school and non-high school students of English. *Sophia Linguistica*, 31: 122-143.

Hopkins, D. (1985) *A teacher's guide to classroom research*. Milton Keynes, Open University Press.

LeCompte, M. D. and J. Preissle (1993) *Ethnography and qualitative design in educational research*. San Diego, Academic Press.

Ministry of Education (1994) *Monbu Toke Yoran* (Statistical abstract of education, science and culture). Tokyo, The Ministry of Education.

Obunsha (1994) *Zenkoku daigaku nyushi mondai seikai* (Compilation of past exam papers). Tokyo, Obunsha.

Pilliner, A. (1973) Assessment - principles and practice with special reference to education in Pakistan. Unpublished manuscript, The British Council.

Rohlen, T. P. (1983) *Japan's high schools*. Berkeley,University of Berkeley Press.

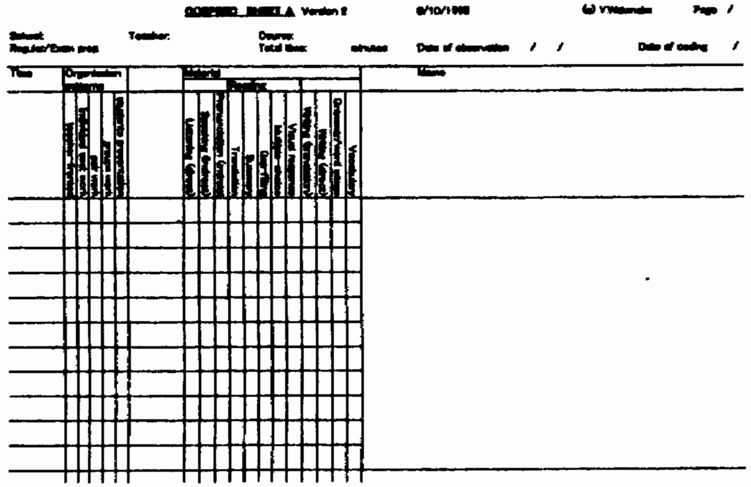
Saito, T., S. Arita, and I., Nasu (1984) Taki-sentaku tesuto ga igaku kyoiku ni oyobosu eikyo (Effects of multiple-choice questions on medical education). *Nihon igaku kyoiku shinko zaidan kenkyu jose ni yoru kenkyu hokoku sho.*

Wall, D. and J.C.Alderson (1993) Examining washback: the Sri Lankan impact study. *Language Testing*, 10, 1: 41-69.

Watanabe, Y. (1990) External variables affecting language learning strategies of Japanese EFL learners. MA dissertation, Lancaster University. (ERIC Document Reproduction Service No. ED 334 822).

Watanabe, Y. (in progress) Washback effects of Japanese university entrance examinations - classroom-based research (a tentative title). Ph.D. thesis. Lancaster University.

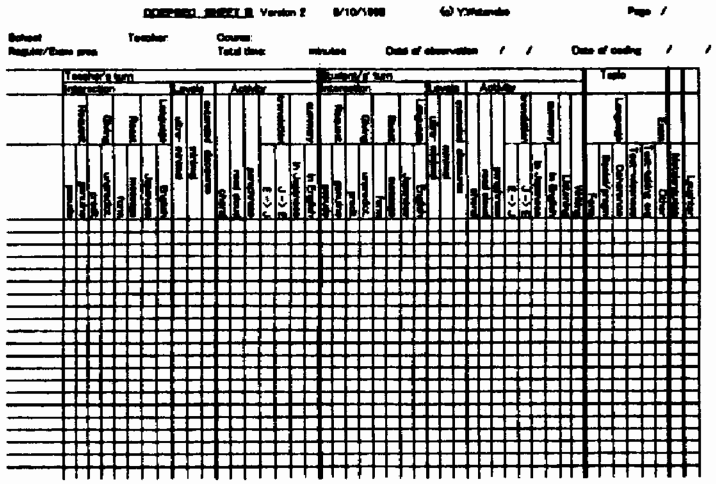
APPENDIX A: CODING SCHEME (COPREC)



COPREC_SHEET A Version 2 9/10/1998 (c) Y.Watanabe Page /

School: Teacher: Course:
Regular/Exam prep Total time: minutes Date of observation / / Date of coding /

Time
Organisation patterns
Material
Realia
Vocabulary
Memo



COPREC_SHEET B Version 2 9/10/1998 (c) Y.Watanabe Page /

School Teacher Course:
Regular/Exam prep Total time: minutes Date of observation / / Date of coding / /

Teacher's turn
Interaction
Levels
Activity
Request
Giving
Read
Language
translation
summary
Student's turn
Interaction
Levels
Activity
Request
Giving
Read
Language
translation
summary
Topic
Language
Exam
Other

APPENDIX 2: WASHBACK GRID

METHODS OF TEACHING											
Teachers	Aural/oral			Grammar-translation				Medium		Organis. Patterns	Type of information
				Translat.			Meta-lg.	Eng	Jpn		pseudo/ predict	genuine/ unpredict
	U	M	E	U	M	E						
Prediction	R>E			R<E			R<E	R>E	R<E	R > E	R<E	R>E
TA		■								■		
TB		■	x	■		x	■			■	▨	
TC			■	■	■		x			■	■	■
TE		■	x	■	■	x					▨	▨
TF		■	x	■	■	x	x			■	▨	▨
NOTES										Similar procedure		

Key:

■ [illegible]	▨ [illegible]	No/little indication of the presence of washback	x = frequency very small in both regular and exam prep

Aural/oral: Interaction/activities which were conducted in English;
U = Ultraminimal/ M = Minimal/ E = Extended discourse.
Grammar-translation
Translat. = Translation;
U = Ultraminimal/ M = Minimal/ E = Extended discourse.
Metalg. = Metalanguage
Medium: Medium of instruction: interaction done in English
Eng = English; Jpn = Japanese
Organis. Patterns: Organisation patterns
Type of information: Types of information exchanged
pseudo/predict = pseudo/predictable information
genuine/unpredict = genuine/unpredictable information
R = regular lessons; E = exam preparatory lessons
R>E ; the frequency larger in regular than in exam preparatory lessons
R<E ; the frequency smaller in regular than in exam preparatory lessons
R=E ; the frequency similar between in regular and exam preparatory lessons