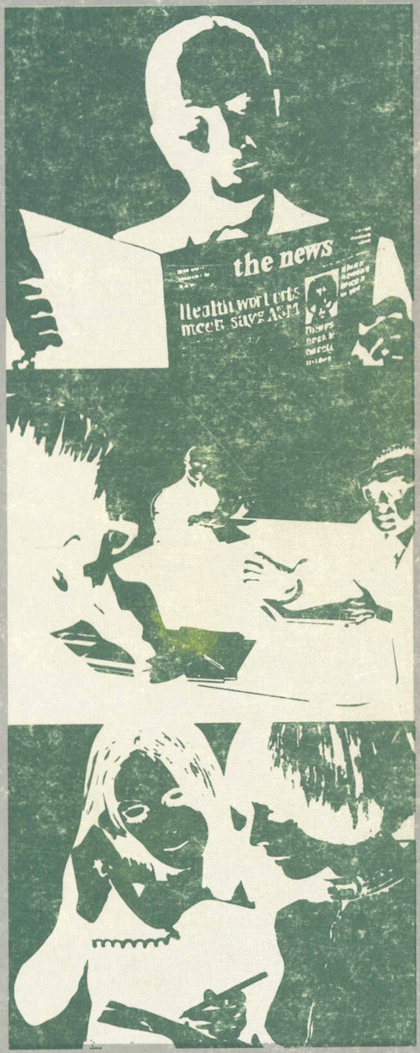
ELT documents

111 - Issues in Language Testing



The British Council

ELT documents
111 - Issues in Language Testing

Editors: J Charles Alderson
Arthur Hughes

The British Council
Central Information Service
English Language and Literature Division

The opinions expressed in this volume are those of the authors and do not necessarily reflect the opinion of the British Council.

ELT Documents is now including a correspondence section. Comments arising from articles in current issues will therefore be most welcome. Please address comments to ELSD, The British Council, 10 Spring Gardens, London SW1A 2BN.

ISBN 0 901618 51 9

CONTENTS

SECTION 3: **General Language Proficiency**

INTRODUCTION

This book arose from an occasion in October 1980 when seven applied linguists met in Lancaster to discuss what they felt were important problems in the assessment of learning a second or foreign language. This Symposium resulted, partly because of its informal nature and its deliberately small size, in an intense discussion in certain areas, a concentration which is rarely possible in conferences or large seminars. It was felt that the Symposium had been so useful that it was decided to make the discussion public, in order not only to let others know what had happened at Lancaster, but also to encourage and stimulate a much broader and hopefully even richer debate in the areas touched upon.

Testing has become an area of increased interest to language teachers and applied linguists in the last decade. Yet as Davies says (Davies 1979) testing has for many years firmly resisted attempts to bring it within the mainstream of applied linguistics. This is no doubt to some extent due to historical reasons, as both Davies and Morrow (this volume) suggest. In the era that Spolsky dubbed the 'psychometric-structuralist period' language testing was dominated by criteria for the establishment of educational measuring instruments developed within the tradition of psychometrics. As a result of this emphasis on the statistical analysis of language tests, a group developed, over the years, of specialists in language testing, 'Testing Experts', popularly believed to live in an arcane world of numbers and formulae. As most language teachers are from a non-numerate background (sometimes having deliberately fled 'figures') it is not surprising that they were reluctant to involve themselves in the mysteries of statistics. Consequently, an expertise developed in language testing and particularly proficiency testing, divorced from the concerns of the language classroom, and imbued with its own separate concerns and values which to outsiders were only partially comprehensible and apparently irrelevant. Despite the advent of Spolsky's third phase of language testing — the psycholinguistic-sociolinguistic phase (what Moller (this volume) calls the third and fourth phases — psycholinguistic-sociolinguistic and sociolinguistic-communicative phases) — 'testing' has not yet recovered from this image of being stubbornly irrelevant to or unconcerned with the language teacher, except for its embodiment in 'exams' which dominate many a syllabus (be it the Cambridge First Certificate or the TOEFL). Teachers who **have** felt they should be concerned with assessing what or whether learners have learned have found the jargon and argument—ation of 'Testing' forbidding and obscure.

problems of native speakers are similar to or different from those of non-native speakers. It is clear that the design and implementation of a new test instrument requires an enormous amount of research, development, effort and resources, which it is easy to underestimate. The same need for research would exist for any test, but particularly for a test that appears to be an ESP test, that claims to be innovative, to be an improvement over other tests and that deals with the future of people. We need to know whether the claims made in the *Specifications* document are substantiated by the evidence. Nevertheless, it was agreed that the *Specifications* document is important, despite its unsubstantiated claims because it highlights the central problem of ESP proficiency testing: matching the demands of test design with those of the people taking the test and with those of the sponsoring and receiving institutions.

SECTION 3

BASIC CONCERNS IN TEST VALIDATION[1]

Adrian S Palmer, English Language Institute, University of Utah, USA
and Lyle F Bachman, University of Illinois, USA

Introduction

Test validity is a complex issue, and to address its many facets in any degree of detail in the space available is a considerable challenge.[2] To make this possible at all, we have had to assume that the reader has some degree of familiarity with traditional views of validity. Consequently, we will review only briefly the basic types of validity. We then look in somewhat more detail into the nature of construct validity – the type of validity which we are currently investigating. Finally, we present some of the general results we have obtained in a recently completed construct validation study.

Types of Validity

Investigations of test validity are, in general, investigations into the extent to which a test measures what it is supposed to measure. This is however, a very general definition of validity, and it is useful to distinguish among several different types of validity. We will distinguish among four here.

Face validity

The first, and in our opinion the least important, type of validity is 'face validity'. Face validity is the appearance of validity – the extent to which a test looks like it measures what it is supposed to, but without any empirical evidence that it does. There is no statistical measure of face validity, and there is no generally accepted procedure for determining that a test does or does not demonstrate face validity.

[1] Prepared for presentation at the RELC Regional Seminar on the Evaluation and Measurement of Language Competence and Performance, Singapore, April 21-25, 1980.

[2] We would like to express our deepest appreciation to the participants in the 1979 and 1980 colloquia on the construct validation of oral tests, held at the TESOL national conventions. These individuals, too numerous to name here, have contributed to every phase of the research described in this paper – from the original expression of a need for such research to its design, implementation, and interpretaiton.

evidence that it does. There is no statistical measure of face validity, and there is no generally accepted procedure for determining that a test does or does not demonstrate face validity.

Content validity

The second, and a much more important, type of validity is 'content validity'. Content validity is the extent to which the selection of tasks one observes in a test-taking situation is representative of the larger set (universe) of tasks of which the test is assumed to be a sample. For example, if a test is designed to measure ability to speak a foreign language, yet requires the testee only to answer yes/no questions, one might doubt that this single task is representative of the sorts of tasks required in general conversation, which entails operations like greeting, leave-taking, questioning, explaining, describing, etc. The process of investigating content validity is basically a sampling process and requires a fairly complete description of the type of competence being tested.

Criterion-referenced validity

Another important but controversial type of validation is 'criterion-referenced validity'. Criterion-referenced validity is the extent to which a test predicts something that is considered important. For example, a test might predict success on a job, and, therefore, be very useful to an employer screening prospective employees.

It is important to note that in criterion-referenced validity, knowing exactly what a test measures is not crucial, so long as whatever is measured is a good predictor of the criterion behaviour. For example, a score on a translation test from a student's native language into English might be a very good predictor of how well a student would do in courses in an English-medium university – even though it might not be at all clear exactly what the translation test measures: the student's knowledge of English, his sensitivity to his native language, his ability to translate, his perseverance, or some combination of these or other abilities. One problem with criterion-referenced validity, then, is that a test can exhibit criterion-referenced validity without one's knowing what it measures.

Construct validity

The fourth type of validity is the relationship between a test and the psychological abilities it measures. This characteristic is called construct validity – the extent to which a test, or a set of tests, yield scores which behave in the ways one would predict they should if the researcher's theory

of what is in the mind of the subject is correct. For example, if it is claimed that a test measures 'knowledge of grammar', one should be able to demonstrate that one can measure knowledge of grammar (as a psychological property) to a certain extent independently of other purported psychological properties such as 'knowledge of vocabulary', 'knowledge of the writing system', 'ability to reason verbally', etc.

Construct validation in the language testing field, then, is a process of hypothesis formation and hypothesis testing that allows the investigator to slowly zero in on the nature of the competence of the language user. As more and more construct validation studies are completed, researchers can say with more and more conviction that the evidence tends to support one position, and not another one.

The MT-MM C-D Construct Validation Procedure

One powerful procedure for investigating construct validity is called by the rather forbidding name 'multitrait-multimethod convergent-discriminant construct validation.' First described by Campbell and Fiske (1959), this procedure requires gathering data that will let one assess two types of validity: convergent and discriminant.

Convergent validity

Convergent validity is evidence that if one wants to measure something or other (a specific trait), one can measure it in a number of different ways (that is, by using different methods of measurement) and still come up with more or less the same results. In other words, it is an indication of how well test scores **agree.**

Discriminant validity

Discriminant validity, on the other hand, is an indication of the extent to which test scores **differ.** Here, one looks for evidence that tests which are supposed to measure different abilities (referred to as 'traits' or 'constructs') actually do provide different information. For example, if a test of the trait 'mathematical ability' and another of the trait 'verbal ability' always gave the same results, that is, if they ordered the subjects taking the tests in exactly the same ways, there would be no evidence that the mathematical and verbal ability traits were actually distinct. Now, in order to assess discriminant validity, it is necessary that one measure several traits at one time. This necessity is the source of 'multitrait' element in the name of the construct validation procedure.

The effect of method

The multitrait-multimethod research model assumes that test scores reflect not only what it is that one is attempting to measure (the trait), but also the effect of the methods of measurement. In other words, a test consists of both trait **and** method components.

To enable one to assess the relative contribution of trait and method to test scores, two or more traits must be measured by a minimum of two distinct methods. This stipulation is the source of the 'multimethod' element in the name of the procedure.

Types of Construct Validation Studies

Because of their complexity, a relatively small number of construct validation studies of language tests have been carried out. Those that have been are basically of three types: principal-component analytic studies; correlational studies; and confirmatory factor analytic studies.

Principal-component analytic studies

Principal-component analytic studies constitute the majority of the construct validation studies to date. Principal component analysis is a technique for accounting for as much common variance as possible on a set of different tests using a minimum number of factors. As it has been used, this analytic technique has been widely criticised. A comprehensive review of the criticisms would go far beyond the limited scope of this paper, and, in any case, such reviews are available in Thorndike (1971), Vollmer and Sang (1980), and Werts, Linn, and Joreskog (1971).

One general problem is that principal component analysis cannot be used to examine any kind of structural model in which the elements in the model are correlated (as appears to be the case in models of language proficiency). The reason for this is that principal component analysis looks only at variance structure, not covariance structure. (The structure model which we will present later will specify the magnitude of the correlation between the elements in the model.)

Another general problem is that of commonalities – this is, the amount of variance the analysis attributes to something the various measures have in common. The reason this is a problem is that the common variance in a principal component analysis contains measurement error and method variance, which inflate the magnitude of the common variance.

In short, principal component analysis not only does not allow only to test the likelihood of specific structural models, but it also produces results which may be inherently biased toward finding a large general factor, no matter what data is analysed.

John Oller has summarised the evidence from many of the principal component construct validation studies in the appendix to his new book, *Language Tests at School* (1979). Oller considers the results of the studies in terms of three hypotheses. The first is the divisibility hypothesis, according to which language proficiency is divisible into a number of distinct components, such as knowledge of grammar, knowledge of vocabulary, speaking ability, reading ability, and so on. The second hypothesis is the unitary competence hypothesis, according to which language proficiency **cannot** be broken down into a number of sub-components which can be differentially measured. This hypothesis predicts, for example, that reading knowledge and speaking knowledge (as measured by tests of each) cannot, in fact, be distinguished. A third hypothesis expresses a position somewhere between the first two. Called 'the partial divisibility hypothesis'; it posits that a major portion of test variance is unique to specific tests. Oller concludes, after considering the data from a number of studies, that the second hypothesis, the unitary competence hypothesis, seems to be a better explanation of the data.

Multitrait-multimethod correlational studies of language tests

Three construct validation studies using the multitrait-multimethod convergent-discriminant design referred to previously have been conducted: Brutsch (1979), Clifford (1978), and Corrigan and Upshur (1978). These studies have attempted to assess the construct validity of tests of purportedly different language use skills (such as reading and writing, and speaking) and purportedly different aspects of language (grammar, vocabulary, etc).

The primary (but not the only) analytic technique used is the examination of the pattern of intercorrelations of test scores according to criteria set forth by Campbell and Fiske (1959). These Campbell-Fiske criteria will be stated and applied, for illustrative purposes, to data from the Bachman-Palmer study described later in this paper.

The results of the three studies cited above are, in general, inconclusive. We believe these inconclusive results to be due, in part, to problems with the tests (such as low reliabilities) and in part to limitations of the analytic techniques used to evaluate the data. For example, an examination of the intercorrelation of test scores through the Campbell-Fiske framework does not even allow us to postulate an underlying causal model, much less to examine the plausibility of the three models hypothesised by Oller.

Multitrait-multimethod confirmatory factor analytic studies

Multitrait-multimethod confirmatory factor analytic studies employ experimental designs which allow the separation of the effects of traits and method on test scores. In addition, they employ an analytic statistical technique, called confirmatory factor analysis. Described in detail by Joreskog (1969), confirmatory factory analysis allows one to make a statistical comparison between structural predictions of a model and the results obtained in an empirical study. For example, given two alternative models of language proficiency, such as the unitary competence model and a two-factor divisible competence model, and given a sufficiently rich set of data (specifically, an over-identified model as described in Alwin (1974), the researcher can compare the explanatory power of the two models by applying statistical tests of goodness of fit of each model to the data.

The Bachman-Palmer Study

Origins of the study

Now we would like to describe a construct validation study whose origins go back to the summer of 1978. During the Fifth Congress of the International Association of Applied Linguistics at Montreal, Peter Groot suggested that despite the general interest in oral testing, attempts to assess the construct validity of oral tests were few. As a result of this conversation, Groot and Adrian Palmer contacted a group of researchers in language testing and arranged a two-day colloquium on the construct validation of oral tests at the 1979 TESOL convention in Boston. At this colloquium, the participants discussed the current state of affairs in the validation of oral tests. The general feeling of the participants was that the construct 'communicative competence in speaking' had not been adequately defined and that the convergent and discriminant validity of tests purporting to measure communicative competence in speaking had not been established. As a consequence, the participants recommended that a construct validation project be instigated.

The authors of this paper (Lyle F. Bachman and Adrian S Palmer) agreed to carry out this study with the advice of the members of the colloquium. In this study, we investigated the hypothesis that two language use skills, speaking and reading, which differ both in direction (productive versus receptive) and in channel (aural versus visual) are psychologically distinct and can be measured independently.

Design

In the study, we used three different methods of testing (namely, interview, translation, and self ratings) to investigate two hypothesised traits (namely, 'communicative competence in speaking' and 'communicative competence in reading'). To test two traits by means of three methods requires a minimum of six tests. The tests we used are described briefly in Figure 1. The tests are described in the boxes, with the names of the traits listed down the left column and the names of the methods listed across the top.

We administered all six tests to a population of 75 speakers of English as a second language at the University of Illinois at Urbana. All were native speakers of Mandarin Chinese. The subjects included students at the University and their spouses. All six tests were administered individually, and total testing time for each subject was approximately two hours.

Results of correlational analysis

The intercorrelations of scores on tests used in this study are presented in Table 1. Of the six tests administered, four (the interview tests of speaking and reading and the translation tests of speaking and reading) were rated by two different examiners. For the purpose of our analysis, we have considered each examiner's ratings as a separate method (or a separate test). Thus, Int-1 on Table 1 stands for the interview as rated by interviewer number 1. Int-2 stands for the interview as rated by interviewer number 2, and so on. Considering the data in this way allowed us to set up a 10 x 10 matrix of inter-correlations.

Convergent validity

The first hypothesis tested concerns convergent validity. The hypothesis states that correlations between scores on tests of the same trait which employ different methods (called validity indices) should be significant and positive. These validity indices are enclosed in the upper left and lower right triangles. All of these correlations are significant and positive, thus providing evidence of convergent validity for both the speaking and the reading tests.

Discriminant validity

The next two hypotheses tested concern discriminant validity. The first hypothesis is that correlations between different tests of the same trait (validity indices) should be higher than correlations between tests having neither trait nor method in common.

An example will illustrate exactly how this hypothesis is tested. Consider the validity indices in the left column of the upper left triangle (.88, .77, .76, and .51). These are correlations between test #1 (examiner #1's ratings on the oral interview test) with the scores on all other tests of speaking. We now wish to compare these with correlations between tests which share neither trait nor method. This set of correlations includes all the indices in the first column of the lower left hand box **except** the index inside the diagonal (.54).

For example, let us compare the .88 validity index with the four relevant indices in the column below it (.56, .58, .52, and .44). .88 is higher than all of these indices – providing evidence of discriminant validity. Note, however, that one of the validity indices in column 1 (.51) is lower than some of the indices in the column below it. If we work through all possible comparisons, we find that the first discriminant validity hypothesis is confirmed in 28 out of 40 cases for speaking, and 38 out of 40 cases for reading.

The second discriminant validity hypothesis is that correlations between tests of the same trait (the validity indices) should be higher than correlations between tests of **different** traits measured by the **same** method. Evidence for this type of validity is harder to obtain, since one has to find **low** correlations between tests which share the **same** method of testing. If the effect of the method is strong, it can exert a strong effect on pairs of test scores which share method.

To test this hypothesis, one compares the same set of validity indices used to test the previous hypothesis with the index within the diagonal in the lower left hand box. This criterion for discriminant validity is clearly met when we compare validity index .88 with the .54 index below it. It is clearly not met, however, when we compare the validity index .51 with the .54 index below it. Again, if one works through all possible comparisons, one finds that the second discriminant validity hypothesis is confirmed in 7 out of 10 cases for speaking and 4 out of 10 cases for reading.

The effect of method is particularly noticeable in tests using translation or self-rating methods. Of the indices in the diagonal in the lower left hand box, the intercorrelations between tests 3-5 which employ translation and self-rating methods (.64, .69, and .68) are clearly higher than those between tests 1 and 2 which do not (.54 and .46).

This completes an examination of the correlations using Campbell-Fiske criteria. We would like to emphasise once again that there are a number of problems associated with the Campbell-Fiske criteria (see Althauser, 1974) which lead us to favour confirmatory factor analysis, the results of which we turn to now.

Confirmatory factor analysis

Confirmatory factor analysis, as we have noted before, is a technique for statistically evaluating the goodness of fit of competitive causal models to a body of data. We have tested over ten models against our data, each involving different assumptions about trait-method interaction. The models with the best fit assume **no** trait-by-method interaction. In keeping with the limited goals of this paper, we present the results of the analysis of only two models. One model posits three method factors and **one** posits three method factors and **two** trait factors: competence in speaking and in reading (a version of the divisible competence hypothesis). The results of the confirmatory factor analysis are given in Table 2.

To test the hypothesis of distinct speaking and reading traits, we examined the difference between the chi squares of the unitary language factor model (50.722) and of the two trait model (34.980). The difference is significant at the p. < 001 level. Thus, we reject the hypothesis that a single language factor underlies the variables.

Having found that the model which best accounts for the data comprises two language traits (speaking and reading) and three methods (interview, translation, and self ratings), we examined the loading of each test on each of these five factors (as well as a uniqueness factor which includes specificity and measurement error components). Factor loads of the ten tests of the six factors are given in Table 3.

The high loading of the oral interview measures on the speaking factor (.819), compared to the relatively lower loading of the oral translation measures (.568) and the oral self-rating measure (.298), indicates that the oral interview method provides a better measure of speaking ability than do the translation and self-rating methods. An examination of the loadings of the interview, translation and self-rating measure on the reading factor leads us, by similar reasoning, to conclude that the translation measure (with a loading of .756 on the reading factor) provides the best measure of reading ability.

Loadings of the measures on the three methods factors (interview, translation, and self-rating) support these conclusions. Specifically, the oral tests load less heavily on the interview method factor (.459) than they do on the translation method factor (.729) and on the self-rating method factor (.734). This indicates that the effect of the method of testing on oral test scores is **least** for the interview method. In other words, assuming we are interested in maximising the effect of the **trait** (which we are trying to measure) and minimising the effect of **method** (which we are not trying to measure), we would choose the interview method to measure oral ability.

Looking at the effect of method on the reading test scores, we find that the translation method (which loads .611 on the reading tests) affects the reading test scores less than the self-rating method (.834) or the interview method (.972). We conclude, therefore, that of the three methods used in the study, the one which minimises the effect of test method on the reading test scores is the translation method.

The results of the confirmatory factor analysis can be presented in the form of a path diagram for the multitrait-multimethod model comprising two traits and three methods. This diagram is given in Figure 2. M_{1-3} and T_{1-2} are the three method and two trait factors which, confirmatory factor analysis indicates, best account for the scores on the measures (the X's). Double ended arrows indicate correlations. Single-ended arrows indicate factor loadings. Single-ended arrows from a number to a measure indicate the loading of that measure on a uniqueness factor – a factor which includes measure specific non-random variance as well as random error variance.

Summary

We feel that this study, and the two years of planning and discussion that preceded it, have yielded two important results: one methodological, the other empirical.

With respect to methodology, we feel that the application of confirmatory factor analysis to a body of data gathered in such a manner as to make it possible to identify and quantify the effects of trait and method on test scores allow us a far clearer picture of the nature of measured language proficiency than has been available using other types of analysis.

With respect to our empirical findings, we feel we have found strong evidence supporting Oller's divisible language competence model. In addition, we have evidence that of the three methods we used to evaluate proficiency in speaking and reading, the interview method provided the best measure of speaking and the translation method the best measure of reading. This should somewhat reassure the United States Foreign Service Institute, which has, up to now, had to rely primarily on faith and on face validity to justify their using these two methods in their testing programme.

Having obtained evidence for the distinctness of the speaking and reading traits, we are now in a position to examine further (1) the extent to which a common factor may or may not underly these distinct traits or (2) the composition of the individual traits.

Table 1
MTMM Correlation Matrix
All correlations sig at p < . 01, df = 74

		Speaking (A)				
		Int-1 (1)	Int-2 (2)	Trans-1 (3)	Trans-2 (4)	Self (5)
A	1	1.00				
	2	.88	1.00			
	3	.77	.72	1.00		
	4	.76	.72	.85	1.00	
	5	.51	.56	.46	.53	1.00
B	1	.54	.45	.62	.65	.58
	2	.56	.46	.64	.67	.60
	3	.58	.61	.64	.68	.46
	4	.52	.55	.62	.69	.49
	5	.44	.45	.47	.51	.68

		Reading (B)				
		Int-1 (1)	Int-2 (2)	Trans-1 (3)	Trans-2 (4)	Self (5)
B	1	1.00				
	2	.97	1.00			
	3	.65	.65	1.00		
	4	.65	.65	.94	1.00	
	5	.68	.68	.54	.54	1.00

Table 2

Comparison of chi squares for two models

	Rater factor loadings equal
	Model 1

1 Trait	χ^2 = 50.722
	df = 30
	p = .0104

	Model 2

2 Traits	χ^2 = 34.980
	df = 29
	p = .2052
	$r\lambda_{ti}\lambda_{tj}$ = .524

difference	$\chi_1^2-\chi_2^2$ = 15.742 df = 1 p $<$ .001

Figure 1

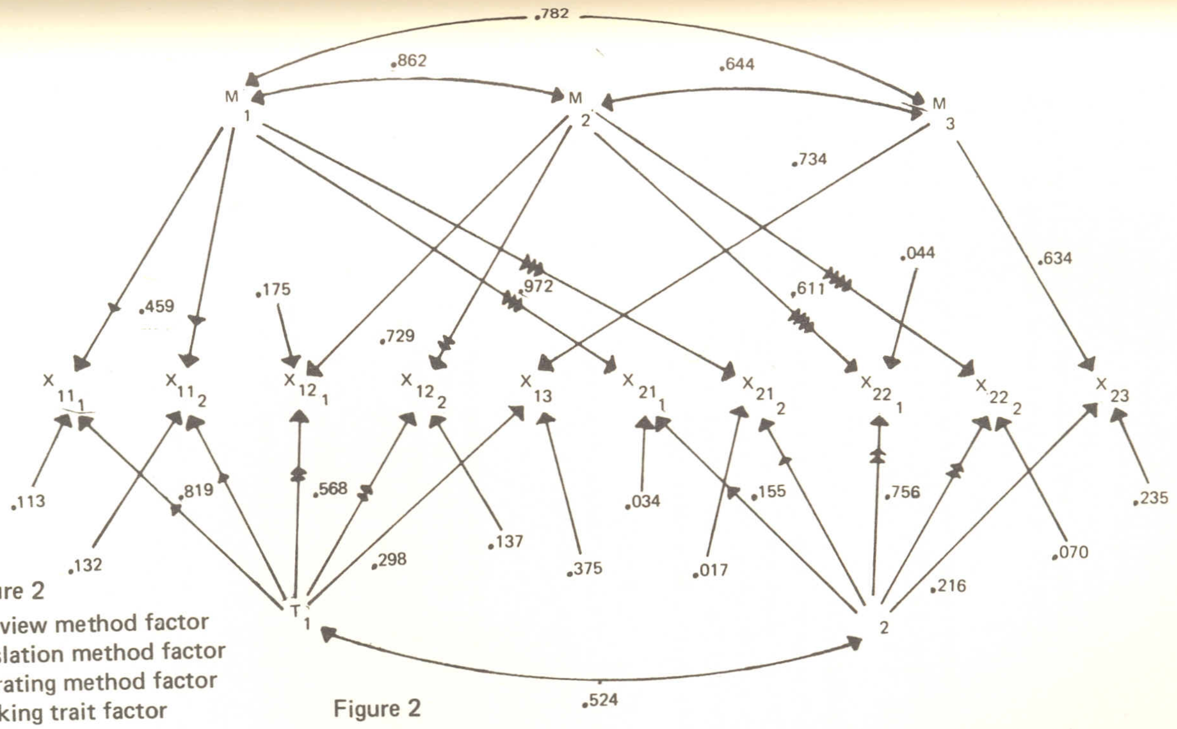
Multitrait-multimethod matrix for the Bachman-Palmer construct validation study

Methods Traits	Interview (1)	Translation (2)	Self-ratings (3)
Communicative competence in speaking (A)	For. Serv. Inst. (FSI) interview test of speaking	Translation test of speaking. Direct translation of dialogue from subject's native language into spoken English.	Self-ratings of speaking ability
Communicative competence in reading (B)	Interview test of reading. Subject is interviewed in his native language about contents of English reading passages.	Translation test of reading. Direct translation from English reading passages to subject's native language.	Self-ratings of reading ability

Table 3

Factor loadings (and standard errors) for measures

Measures	Speaking		Reading		Interview		Translation		Self-Rating		Uniqueness
Oral Interview 1	.819	(.082)	.000		.459	(.126)	.000		.000		.113
Oral Interview 2	.819	(.082)	.000		.459	(.126)	.000		.000		.132
Oral Translation 1	.568	(.091)	.000		.000		.729	(.098)	.000		.175
Oral Translation 2	.568	(.091)	.000		.000		.729	(.098)	.000		.137
Oral Self-rating	.298	(.097)	.000		.000		.000		.734	(.108)	.357
Reading Interview 1	.000		.155	(.140)	.972	(.085)	.000		.000		.034
Reading Interview 2	.000		.155	(.140)	.972	(.085)	.000		.000		.017
Reading Translation 1	.000		.756	(.097)	.000		.611	(.133)	.000		.044
Reading Translation 2	.000		.756	(.097)	.000		.611	(.133)	.000		.070
Reading Self-rating	.000		.216	(.113)	.000		.000		.834	(.104)	.235



Key for Figure 2

M_1 = interview method factor
M_2 = translation method factor
M_3 = self-rating method factor
T_1 = speaking trait factor
T_2 = reading trait factor
$X11_{1-2}$ = oral interview measures
$X12_{1-2}$ = oral translation measures
X_{13} = oral self-rating measure
$X21_{1-2}$ = reading interview measures
$X22_{1-2}$ = reading translation measures
X23 = reading self-rating measures

Figure 2

Path diagram for multitrait-multimethod model (2 traits, 3 methods)
χ^2 = 34.9804, df = 29, p = .2052

Double-ended arrows = correlations
Single-ended arrows between elements = trait and method factor loadings
Single-ended arrows from a number to a measure = uniqueness factor loadings

BIBLIOGRAPHY

ALHAUSER, ROBERT P
Inferring validity from the multitrait-multimethod matrix: another assessment. In: CONSTNER, H L, *ed.* Sociological Methodology 1973-1974.

ALWIN, DUNNE F
Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In: COSTNER, H L, *ed.* Sociological Methodology 1973-1974. San Francisco: Jossey-Bass, 1974.

BRUTSCH, SUSANNA M
Convergent/discriminant validation of prospective teacher proficiency in oral and written production of French by means of the MLA Cooperation Foreign Language Proficiency Tests. French Direct Proficiency Tests for Teachers (TOP and TWT) and self-ratings. PhD Dissertation. University of Minnesota, Minneapolis, Minnesota, 1979.

CAMPBELL, D T and FISKE, D W
Convergent and discriminant validation by the multitrait-multimethod matrix. In: Psychological Bulletin, *56,* 2, 1959.

CLIFFORD, REV T
Reliability and validity of language aspects contributing to oral proficiency of prospective teachers of German. In: CLARK, John L D, *ed.* Direct testing of speaking proficiency: theory and application. Princeton, New Jersey: Educational Testing Service, 1978.

CORRIGAN, ANNE and UPSHUR, JOHN A
Test method and linguistic factors in foreign language tests. (Paper presented at the 1978 TESOL Convention, Mexico City)

JORESKOG, K G
A general approach to confirmatory maximum likelihood factor analysis. In: Psychometrika, *34,* 183-202, 1969.

OLLER, JOHN W, Jr.
Language tests at school. London: Longman, 1979.

THORNDIKE, ROBERT L
Educational measurement. 2nd ed (Chapters 14 and 16). American Council on Education, 1971.

VOLLMER, HELMUT J, and SANG, F
On the psycholinguistic construct of an internalised expectancy grammer. (Paper presented at the 2nd Colloquium on the Construct Validation of Oral Tests, 1980 TESOL National Convention, San Francisco, California, 4-5 March, 1980)

WERTS, C E, LINN, R L, and JORESKOG, K G
Estimating the parameters of path models involving unmeasured variables. In BLALOCK, H M, Jr, *ed.* Causal models in the social sciences. Chicago: Aldine-Atherton, 1971.

WHY ARE WE INTERESTED IN GENERAL LANGUAGE PROFICIENCY?[1]

Helmut J Vollmer, University of Osnabrück, Germany

Introduction

I should like to start out by saying that language proficiency is what language proficiency tests measure. This circular statement is about all one can firmly say when asked to define the concept of proficiency to date. This is even more so when it comes to the construct of overall language proficiency, regardless of whether we want to refer to one's mother tongue or any second or foreign language. What exactly is this general language proficiency, does it really exist and what, then, is our particular interest in this construct either as test researchers, or as test developers or as users of test results? What models of general language proficiency (GLP) seem to be plausible, on what grounds and based on what theoretical insights? Is the concept of GLP related more to the competence level of a person, that is, to what the learner knows about a certain language (including knowledge about how to use it) or does it rather refer to the performance or skill level on which the learner actually demonstrates his/her knowledge in more or less meaningful communicative situations? If we consider proficiency to be a performance category we should then try and define it as some sort of attained level of mastery within a given language which can be observed and measured by a number of different methods. We would then immediately face the question in which way precisely GLP might differ from the sum or average of one's scores on any of the existing proficiency measures covering different aspects of language that one might be able to name (if not isolate). If we think, however, of GLP as an underlying ability to demonstrate one's knowledge of a language regardless of the nature of the task involved, the skill(s) implied, the measurement method used etc., we would then have to elaborate on the differences between the two terms 'overall proficiency' and 'competence' (if there are any at all) – no matter what theoretical framework for linguistic and/or communicative competence we may have in mind.

Question after question arises once we investigate more deeply into the concept of GLP which is not clear at all as yet. In my paper I would like to share with you some of the problems which my colleague Dr Sang and I came across in studying the structure of what we thought to be 'linguistic competence' in German learners of English as a foreign language. The research project that I am referring to is based at the Max-Planck-Institut für Bildungsforschung in Berlin. One of the main objectives of this project is to study the theoretical claims and empirical evidence put forward in support of either the 'unitary competence' hypothesis or the 'divisible competence' hypothesis and to further contribute to this controversy by presenting our own research findings and careful interpretation of them (cf. Sang/Vollmer 1978). The basic question here is – as you might know - whether or not all performances in a second/foreign language can be traced back to a single underlying factor, the so-called 'General Language Proficiency Factor' (GLPF) and whether or not it seems theoretically plausible and valid to interpret the appearance of such a factor as an indication of the existence of a unitary cognitive ability at work. If so, the wide-spread belief in relatively distinguishable, more or less autonomous dimensions of linguistic competence and their realisation on the performance level (the 'four skills') which most foreign language teaching (and testing) nowadays is still based upon would have to be questioned, if not overthrown. The research situation concerning this problem, which implies one of the central issues of language testing theory, is quite controversial. Basically speaking there are two different lines of research which operate to some extent apart from one another (without really relating their arguments and tentative findings to each other). Let me now turn to a brief outline of these two positions with respect to their definition of proficiency in theoretical and operational terms.

Conflicting views of language proficiency

1 The divisible competence hypothesis

The first of the two research branches referred to has concentrated on attempting to identify those areas/dimensions of linguistic achievement which could be interpreted along the lines of meaningful learning objectives and which were able to structure the learning and teaching process of a second/foreign language in a plausible way. Theoretically this approach is based on the more or less implicit assumption that there is (most likely) no such thing as a single unitary language ability but (more likely) a number of specific linguistic – and non-linguistic – competencies or areas of competence underlying language behaviour. It is hoped that these competencies can be identified and related to each other more distinctly and systematically as our knowledge advances, and that they can be further broken down some day into sub-competencies, eg into components or aspects contributing to the successful operation of a certain area of competence. It must be added, however, that within the last twenty years there has never been a strong version of this claim. Rather a great number of researchers seem to have

[1] I would like to thank my friend and colleague Fritz Sang for his valuable comments on an earlier draft of this paper. Of course, I assume full responsibility for the views expressed here as well as for all kinds of errors.

adopted this general outlook (possibly for lack of a convincing theoretical alternative) and used a multidimensional 'model' as their unquestioned starting point. Accordingly, they have devoted time and effort only in identifying and naming those competencies that could be plausibly related to the different skills or aspects of language behaviour on the performance level. In addition, investigation into the structure of foreign language aptitude seemed to support the view that the acquisition of another language other than one's native tongue was dependent on at least three different language-specific factors within the learner (besides non-linguistic variables like motivation etc.).

This approach has been labelled (somewhat unjustly as I believe) the 'discrete-point approach', although in reality (at least for its most outspoken proponents) it has always been a mixture of 'discrete-point' tests and some 'global' testing (see, for example, the matrices given in Valette (1967) or Harris (1969); for a discussion of this 'disjunctive fallacy' as a whole cf Farhady (1979).

Certainly sets of items that test the control of specific elements of the second language (phonemes, intonation patterns, vocabulary or structural items, and the like) are discrete-point tests, as most multiple-choice items are discrete-point items. But the testing of the so-called 'integrated skills' like reading or listening with comprehension questions based on a longer reading or listening test do in my understanding very much focus on global aspects of the language independent of the item format used. Tasks like these require the integration of different elements of knowledge in order to understand and interpret language in context. Even if a longer reading or listening passage is scored on the basis of specific elements implied (in a manner that parallels discrete-point items) I would still consider it to be a global measure more than anything else. As concerns language proficiency it was normally thought of as being best approached by a whole battery of language tests (instead of only one or just a few). Each of the tests was supposed to aim at a unique aspect of knowing a language and/or handling it on different levels.

As early as 1961 J B Carroll worked out a rationale for describing and measuring language proficiency along the multidimensional lines outlined above. Carroll pointed out that the validity of a proficiency test does not only depend on whether a representative sample of the English language had been covered. It is more important yet, according to Carroll, that the success of the testee in coping with future language situations, future learning situations as well as certain forseeable social situations in real life can be adequately predicted with some degree of certainty on the basis of the test results. Therefore one has to select and combine those dimensions of test performance which are relevant to future tasks and situations. In other words, the proficiency of a learner (his degree of mastery of the foreign language) cannot be judged or measured in abstract terms. A test of proficiency, according to Carroll, has always to be validated externally against the criterion of 'having sufficient English to operate in given situations' (Carroll 1972:315). Carroll goes on to specify ten relevant dimensions of test performance which include those elementary aspects of knowledge and the four integrated skills: listening comprehension, reading comprehension, speaking and written composition. These dimensions are to be combined in a specific manner each time. They should be given different weighting according to their relative importance depending on the purpose of the testing and based on the findings of future job or task analysis, that is, on the results of the externally validated description of qualifications needed.

As far as I can see the term 'overall proficiency' or 'GLP' was never used (and maybe has no place) within this theoretical framework. As long as the purpose of proficiency testing is to determine whether a learner's language ability corresponds to specified language requirements it makes more sense to speak of a learner's 'specific proficiency' in relation to the content area defined and the criteria used. For example, in placement tests we want to know whether a student is proficient enough to enter this or that course, or we want to find out whether a learner is to able to read professional literature in another language with a specific level (such as 80 or 90 per cent) of accuracy, etc. The Foreign Service Institute of the United States has developed a number of proficiency tests that are meant to indicate to what degree a person can function in the foreign language. Again, the reported language ability of a candidate is defined by a predetermined set of functional categories: having sufficient German, Spanish, Russian etc. to carry out an informal conversation, to chair a meeting, to explain a statement of policy, to do this or that . . .

In all of these cases nobody would dare to make a judgement on a person's overall foreign language proficiency, but only on a limited, yet seemingly well-defined aspect of language proficiency based on the tests used. The crucial question, of course, is that of validity: do the tests really measure what they purport to measure, what language tasks, what content areas, what communicative situations etc. are being sampled, how are the levels of correctness and appropriateness being defined and identified, how justified are the predictions made as to a person's functioning in that language? The very problems of sampling and prediction suggest that we always include some judgement of a learner's 'transfer ability' (if he or she is able to act with language in this or that test situation, he or she will probably be similarly successful in situations not included in the test or not forseeable at all). In other words, a certain understanding of a person's generalised state of knowledge or ability to use this knowledge – however vague – seems to be implied in any proficiency concept. It is exactly here where the second of the two research branches starts.

2 The notion of 'overall proficiency'

In the late sixties it was Spolsky who asked: What does it mean to know a language or how do you get someone to perform his competence (as contradictory as this formulation sounds). He argues that 'knowledge of a language' was more than having command over a certain amount of vocabulary or mastering its isolated elements. It was **knowing the rules** of a language, as he put it.

> Knowing a language is a matter of having mastered these (as yet incompletely specified) rules; the ability to handle new sentences is evidence of knowing the rules that are needed to generate them (Spolsky 1973: 173).

Spolsky thus reminds us of 'two vital truths about language, the fact that language is redundant, and the fact that it is creative' (1973: 167). To him knowledge of a language, being a matter of knowledge of rules, is the same as 'underlying linguistic competence'. This operates in all the different kinds of performances, be they active or passive (the latter being an equally creative process on the part of the learner).

Almost everyone would agree with Spolsky so far. It is worth noting that he only speaks of an 'underlying linguistic competence', not of a 'unitary competence'. In another context he considers knowledge of rules to be the 'principal factor' (1973: 174) in the understanding as well as in the production of messages (not the one and only factor explaining all sorts of language behaviour). This distinction which I try to make here is quite important. It becomes clearer, I think, when we follow Spolsky's suggestion that we could find out about 'knowledge of a language' equally well when testing passive or active skills:

> This last does not of course mean that an individual's performance as a speaker is the same as his performance as a listener; such a claim would clearly be ridiculous, for it would be tantamount to saying that anyone who could read a Shakespeare play could also write it. All that it does claim is that the same linguistic competence, the same knowledge of rules, underlies both kinds of performance.
>
> (Spolsky 1973: 174).

I take this quotation to be a clear indication of the shift of focus from the differences between the skills (and how they might relate to underlying competencies) to what they might have in common by way of a shared basic competence stretching out into all the skills. But in trying to explain the ability to read (and understand!) a Shakespeare play or to write one we will have to take other competencies (constructs) into account – besides and on top of 'knowledge of rules'. If our focus of interest is concentrated on the assumed central linguistic competence (or that portion which may be common to the operation in all the skills) the additional cognitive forces (those which are **not** common to all the skills) do not disappear – they are simply out of focus (for the time being).

My interpretation of the concept of an 'underlying linguistic competence', which does not imply it to be necessarily unitary, is somewhat dimmed again by Spolsky's introduction of another term, that of 'overall proficiency' (1973: 175).

> some way to get beyond the limitation of testing a sample of surface features, and seek rather to tap underlying linguistic competence
>
> (Spolsky 1973: 175).

This sentence can easily be misunderstood in that it suggests that competence of a foreign language learner can be **tested directly** (or at least more directly) rather than measured through any of its manifestations of the performance level known so far – which is not possible! What Spolsky refers to is the development of 'competence-oriented' tests (others say 'integrative' tests) as valid indicators of learners' success in handling actual performance, calling for normal language functioning based on the principles of redundancy and creativity.

The sentence quoted above could very well nourish a second misunderstanding by suggesting that linguistic competence can be measured by a (singular!) test of overall proficiency. Moreover, the term 'overall' does not only imply 'basic', but also 'comprehensive', as if **all** the possible aspects of a person's language behaviour (and the ability structure governing his or her performance) could be grasped exhaustively in one proficiency measure. This view, though, is not shared by the author quoted. When asked at the 1974 Washington Language Testing Symposium for a clear definition of overall proficiency, Spolsky answered:

> It should be obvious by now that I can't say that precisely, or I would have. It's an idea that I'm still playing with. It has to correlate with the sum of various kinds of things in some way, because it should underlie

> any specific abilities. In other words, I have the notion that ability to operate in a language includes a good, solid central portion (which I'll call overall proficiency) plus a number of specific areas based on experience and which will turn out to be either the skill or certain sociolinguistic situations
>
> (Jones/Spolsky 1975: 69).

Taking this uncertainty as it is, other authors like John W Oller had picked up the notion of overall proficiency and had experimented in the meantime with a number of measures in foreign language testing aimed at tapping the postulated GLP, namely with different form of the Cloze test and dictation.

3 The unitary competence hypothesis

Oller and others believe that there are good reasons for assuming that linguistic competence is not only the principal factor underlying all language skills, but that this competence is unitary (cf for example Oller 1976, Oller/Hinofotis 1976). In his theoretical work Oller tries to convince us that this (assumed) unitary competence is more than just a construct, that it 'really exists'. In addition, he asserts that all processes of comprehending and producing utterances, of understanding and conveying meaning (in whatever mode by whatever medium) are governed by this one indivisible intellectual force – in L1 as well as in any L2. In terms of psycholinguistic modelling Oller has offered an interpretation of this assumed force (or basic human ability) as an 'internalised expectancy grammar' at work (cf Oller 1974; 1978). This concept can be based partly on research done in cognitive psychology, especially as to perceptual processes in general (not restricted to language perception). On the other hand one has to be rather careful in adopting or applying results or non-language-specific insights from cognitive psychology to a theory of language processing. Neisser himself, one of the authorities in that field, turns out to be much more cautious in 1976 than in his basic work published in 1967 (for further discussion of the plausibility of Oller's psycholinguistic construct 'expectancy grammer' see Vollmer/Sang 1979).

As to the comparison of language reception and language production as psychological processes, their structural equation does not seem justified at the moment or it seems a bit overhasty at least. Though the results of psycholinguistic research to date indeed suggest some commonalities between the encoding and the decoding system, production and comprehension can probably not be seen as mirror images. Many attempts have been made to account for their unique characteristics by postulating different underlying processes. The role played by syntax is a case in point here. To our present knowledge, the syntactic level seems to be much more important for the process of planning and producing an utterance than for perceiving and decoding it, whereas in the latter case the semantic level seems to be predominant. Generally speaking, the differences between knowing how to analyse input and knowing how to construct output apparently outweigh the correspondences between these two processes. Evidence continues to come in from many sources that language as comprehension and language as production are so profoundly different that any attempt to describe language 'non-directionally', or neutrally with respect to its interpretive and expressive functions, will be highly controversial, if not fail. I am not ready, however, to claim that there are basically two distinguishable competences, one associated with understanding language, one with producing meaningful utterances (although this might be so). This 'two competences hypothesis' may be considered to replace the construct of one indivisible linguistic competence – or else all the competences named could be looked upon as hierarchically ordered, pertaining to different levels, each having its own scope, not excluding one another (theoretically). I know that a position like mine would need further explication to be better understood and needs, above all, further research to back it up and make it more plausible. Unfortunately, I cannot go into it any deeper in this paper (for discussion of some aspects, however, cf Fodor, Bever, Garrett 1974; Straight 1976; Vollmer/ Sang forthcoming).

My main point here is that almost anything one can say about language processing, especially about speech production, is still very speculative, even by the standards current in psycholinguistics. There are a vast number of uncertainties and many open research questions to be solved before any one of the theoretical models can hope to reflect psychological reality (a claim that Oller makes). One of the major problems with the writing of Oller, then, is the speediness with which (suitable) pieces of research from other disciplines are incorporated into his theoretical framework – and the firmness with which certain positions are taken forcing the reader to follow (and believe!) the expert – as if no doubt were possible. From a theoretical point of view the notion of a general language proficiency as the manifestation of an underlying unitary competence interpreted along the lines of an expectancy grammar is still very vague and not convincing at all (as we shall see in more detail in the next section). So is the empirical evidence for both the unitary and the divisible competence hypothesis (as I shall point out later).

General language proficiency defined

In this part of my paper I would like to develop some of the critical points concerning the notion of GLP and the testing of proficiency in a more systematic way. I have organised my thoughts under three different headings:

Proficiency and competence
General language proficiency and cognition
The dynamics of general language proficiency

1 Proficiency and competence

Let us reconsider once more whether proficiency, especially the concept of a GLP, pertains to the performance level and thus to overt language behaviour, or whether it represents a construct on the competence level reflecting our understanding of how we think that different uses of a language have been integrated internally within a learner. One dictionary which I looked up defines proficiency as an 'advanced state of attainment in some knowledge, art, or skill.' Such a definition is useful though it must be elaborated upon, especially since both the knowledge and the skill level could be meant if someone is said to be proficient.

When we turn to Carroll's (1968: 57) suggested chart of linguistic performance abilities (all based on assumed underlying competences) it becomes evident that according to this author the term 'proficiency' relates neither to actual (and measurable) performances not to the competence level in the sense of knowledge of a language. The 'proficiencies' or aspects of proficiency seem to form a level of their own – somewhere in between performance and competence (in the Chomskyan sense of the terms). Carroll (1968) speaks of linguistic performance abilities. Their relative degree of development decides what a person's language proficiency looks like, what it is made up of, which of his or her performance abilities contributes to what extent to the overall picture (expressed by a total score) of mastery of a second language. In discussing Carroll's earlier work on *Fundamental Considerations in Testing for English Language Proficiency of Foreign Students* (1972) I have already pointed out that in testing proficiency we are not only interested in an examinee's actual strengths or weaknesses, in particular fields of linguistic knowledge or lack of it. What we are mainly concerned about is how this knowledge is put to use, how bits and pieces of this knowledge are being integrated on different levels of performance, how language is handled with more or less facility in terms of the total communicative effect of an utterance. Another important aspect in this context is, of course, a person's ability to get by even in situations where knowledge is deficient, where uncertainties as to the appropriateness of an utterance according to social conventions exist or psychological restrictions in interaction have to be dealt with. To these latter aspects more attention has been paid ever since the broader concept of communicative competence (made up of linguistic competence plus something else which we have yet better to define) has been introduced (cf the recent work of Canale/Swain 1979, especially their concept of strategic competence which comprises something equivalent to Carroll's linguistic performance abilities plus a set of less language-bound social-interactional abilities).[2]

Carroll's view of foreign language proficiency focusing on the narrower construct of linguistic competence can probably best be summarised as an accumulated index of a person's (predictable) mastery of and functioning in L2. This index is inferred from actual measurements on different levels of performance, which are taken to be manifestations of covert linguistic performance abilities which in turn are all thought to be based on underlying competences.

Let us find out now how the notion of GLP relates to the performance and competence level. It was Spolsky in 1975 who stated clearly that overall proficiency could not be considered identical with linguistic competence.

> It's something that presumably has what Alan Davies would call construct validity. In other words, it depends on a theoretical notion of knowledge of a language and the assumption that while this knowledge at a certain level can be divided up into various kinds of skills, there is something underlying the various skills which is obviously not the same as competence. You have to allow, of course, for gross differences. For example, if somebody is deaf he won't be good at reading or writing, and if somebody has never been exposed to speech of a certain variety he won't be good at handling that. And after allowing for those gross, very specific differences of experience, whatever is left is overall proficiency (Jones/Spolsky 1975: 67).

[2] Canale/Swain (1979) postulate three different dimensions of communicative competence in their theoretical framework: grammatical competence, sociolinguistic competence, and strategic competence. After having reviewed all the relevant literature it appears very unlikely to these authors that communicative competence could be reduced to only one global language proficiency dimension.

The model of Canale/Swain, however, is not yet based on any empirical investigation, as far as I know.

Apparently the basic idea is that a speaker of a second language acquires not only certain skills but at the same time builds up a general understanding of that language (knowledge of the rules). In other words, it is suggested that all speakers develop and have a general proficiency simply by being exposed to a language. This GLP may be acquired by different sense modalities, but once it is there it can then be employed in any of the skill areas — even in those not particularly trained. It can also be applied to a vast number of future situations - even to those which are not foreseeable. 'Theoretically, at least, two people could know very different parts of a language and, having a fairly small part in common, still know how to get by. That's where overall proficiency becomes important' (Jones/Spolsky 1975: 69). It almost looks as if GLP stays with a person once it has been formed. On the other hand it seems to be a cognitive potential consisting of language-specific knowledge (sets of rule systems) being stored which is thought to be the software of a generalised ability to operate in that language. Spolsky gives the following (construed) example:

> Someone is exposed to the traditional method of learning a language, that is, a grammar-translation approach at school, and then goes to live in the country for two months. At the beginning of the two months that person would test out completely at O or something on any kind of oral test. But he already has this overall proficiency that is just waiting for new experiences
>
> (Jones/Spolsky 1975: 70).

Although many questions remain unanswered it should be pointed out in summarising that for researchers like Spolsky and even more so for Oller the notion of GLP has become a psychological construct, something non-observable any more. It has thus moved in its theoretical meaning towards the competence level, with a clear connotation of an unfolding cognitive ability to operate in a language.

2 General language proficiency and cognition

In my opinion when we are chasing after GLP what we really want to get at is the centre of what might be called the **general cognitive apparatus** of a person. Whether theoretically justified or not we hope to be able to form a quick and somewhat overall picture of a learner's generalised ability to make use of the instrument of a foreign language more or less successfully in all possible situations. We are not concerned about the previous training of a testee or any curriculum programme in particular (in this respect proficiency tests differ from achievement testing). On the contrary we are looking for a more or less sound basis in order to make predictions about a person's future behaviour. In measuring GLP it is hoped to find an indicator of how adaptable a person is or might be, how well he or she will act or function within a social system including language use (and non-verbal means of interaction). The language side of communication is thought to be highly dependent on the development of what might be termed the **language processing mechanisms** in general. In terms of information theory the GLP factor is considered by its proponents to represent something like the central core of human cognition, a person's executive programme governing all subroutines and their coordination: linguistic, pragmatic etc. The fundamental problem involved here is, of course, that it cannot be determined with any degree of certainty what human cognition is made up of, how it functions, what cognitive abilities are implied in language learning and language use, whether an individual's performance in different languages (eg L1 and L2 or different L2) is governed by the same underlying cognitive factor or factors. As interesting as Oller's proposal of an analogy between perception and production of language is, as stimulating as the idea of language production as 'a kind of synthesis-by-analysis' (Oller 1978: 45) and the construct of an expectancy grammar as a whole may be — all of these thoughts are highly speculative and just a bit too sloppy for real life decisions to be based upon them. Neisser, for example, after having suggested in 1967 that speech is perceived by 'analysis-by-synthesis' no longer believes that this can be literally true: 'The listener's active constructions must be more open and less specific, so that they are rarely disconfirmed' (Neisser 1976: 32). Cognitive operations in language production are even less understood. Generally speaking, human cognition seems to be a much broader capacity than its language - specific manifestation may make us believe.

I do not say, however, that language proficiency doesn't have anything to do with cognitive abilities and processes. It certainly does! There is hardly any doubt that language proficiency (in L1 as well as in L2) strongly relates to IQ and to different aspects of academic achievement. The decisive question is whether or not this is only one dimension of language proficiency (related to general cognitive and academic skills) or whether or not language proficiency is basically defined by the central core and can thus be called 'global'. Spolsky in 1975 stated that the ability to operate in a language only '**includes** a good, solid central portion' (Jones/Spolsky 1975: 69; emphasis by H J V). In a recently published article Cummins distinguishes between 'a convincing weak form and a less convincing strong form of Oller's arguments' (1979: 198; cf his proposed terms 'cognitive/academic language ability' (CALP) and 'basic interpersonal communicative skills' (BICS)). Cummins tries to prove that everybody acquires BICS in a first language and that CALP and BICS are also independent of one another in L2. I find his contribution quite useful — it is another piece of evidence against the unitary competence hypothesis (the strong form of Oller's claim).

Speaking of language proficiency as a generalised cognitive ability sounds very much as if we were thinking of it as a fixed state or the end product of development, if not even as a personality trait. This is especially true in its German translation where GLP could mean 'Stand der Sprachbeherrschung', but as much 'Allgemeine Sprachfähigkeit' closely associated (at least connotatively) with terms like 'Begabung' or 'Intelligenz' (in the sense of a quality that is either innate or, after having been acquired somehow, is considered to be rather stable). In this particular context we cannot take seriously enough a criticism developed by the sociological school of the Symbolic Interactionism against the traditional trait concept and picked up by interactional psychology during the past few years. This criticism goes like this, that the unit of analysis in the behavioural sciences cannot be the structure of human capabilities (the assumed stable 'traits') but will have to be the interrelationship between task situation and persons involved. This understanding of human behaviour goes well together with what has been the outcome so far of second language acquisition research along the lines of the **interlanguage hypothesis.** According to this theory language is acquired (in a natural setting) or learned (through formal instruction) in terms of a creative construction process which is more or less open-ended in the direction towards a native speaker's competence (target language). Proficiency then is interpreted as a dynamic construct, as the relative degree or level of competence a person has reached by the time of measurement. This competence, though, cannot be developed *ad infinitum,* as some researchers believe. Much discussion has been devoted therefore to the concept of **fossilisation** during the past years. This phenomenon, however, seems to be dependent on so many variables (cf Selinker/Lamandella 1978) that, for the time being, we have to assume that (almost) everyone can further develop his/her linguistic code and thus language proficiency under favourable circumstances – either by being trained or by finding the language system at hand not functional any more for one's social and personal needs (for discussion of this point cf Sampson 1978). On the other hand, linguistic knowledge apparently can just as easily be forgotten (or 'decomposed') in the process of not wanting or not having to use it.

By and large it seems justified to consider foreign language acquisition and language use as a dynamic process. Testing language proficiency means making a cut at a given point in time in order to form a more or less rough idea of a person's state of advancement. In view of the dynamics of language development it will indeed be advisable to test proficiency not only once, but (wherever possible) time and again. We should use different versions of the same measurement instrument as well as different methods altogether (one of the reasons being to control the effect of measurement method). In spite of all the statistical problems involved here I am quite convinced that each single measurement will add to the information we might already have of a person's language ability. It is very unlikely, indeed, that a single type of test will reflect any full assessment of the facets of language command (this human faculty which is very intricate and complex). In this respect I strongly agree with Ingram (1978) and Farhady (1979), and with Davies (1978), for that matter.

Some empirical considerations

Despite the fact that I have almost run out of space I would now like to add a few comments on the empirical side of the problem under consideration. To put it very crudely, neither one of the two opposing hypotheses about the structure of language ability has strong empirical evidence in its favour. In our research project at the Max-Planck-Institut für Bildungsforschung in Berlin we were able to show that the factor analytic studies done in support of the multidimensional model of language competence are by no means convincing as to the methods used and their interpretation of the statistical results. They do not offer a clear picture at all; one could say they tend to discourage the strong version of the divisible competence hypothesis (that each of the four skills is based upon a separate underlying factor with little or no interrelation at all)[3]. Yet many questions remain open (for example, as to number and nature of tests included etc.). On the whole the results cannot be simply ignored or done away with as being irrelevant (cf Vollmer/Sang forthcoming).

Likewise the empirical evidence presented in favour of the unitary competence hypothesis, when being re-evaluated, turns out to be by far not as strong and clear-cut as had been asserted by its proponents. For example, in some of the factor analyses presented there is indeed only one factor (within the limits of a sufficient eigenvalue) which can justly be taken as a general factor (cf Oller 1976). In other studies, however, a number of

[3] This is supported by Hosley and Meredith (1979) in a recent study on the structure of the construct of English proficiency, as measured by the TOEFL. According to their data the divisible competence hypothesis could be rejected. Instead of adopting the unitary competence hypothesis, however, they suggest a 'hierarchical skills theory' for consideration, which seems to be 'compatible with, but not derivable from, the present data' (1979: 217).

factors showed up (cf Oller/Hinofotis 1976; Scholz *et al* 1977), and I can't quite understand why the unitary competence hypothesis should be the best explanation fitting these results (for further details see Sang/Vollmer 1978; Vollmer/Sang forthcoming).

As to our own research results a strong first factor (being the only one worthy of interpretation) emerged. But to our understanding the appearance of this factor could not be easily interpreted as an indication of the existence of a global language proficiency in the sense of the strong form of this argument (cf Sang/Vollmer 1978 and forthcoming).

I am afraid I'll have to go into factor analysis and testing theory just a bit more to make my point clearer. To cut things short, it is most important to make a clear distinction between what was later labelled the 'principal component model' on the one side and factor analysis in the narrower sense on the other side ('principal factor model'). In the latter model, according to Spearman's theory, factors represent either that portion of the variables under study which they have in common with other variables (so-called **common factors**) or that portion which they share with no others (so-called **unique factors**). In addition to a single general common factor which all tests included in his analysis would load high on, Spearman expected to see a number of unique factors on each of which only one of his tests had a substantial loading and the remaining tests a load of zero. Assuming that it is possible to concentrate the entire common variance of the tests on the general factor the residual correlations between the tests would then have to go to zero. Now up to this point researchers like Oller and Spearman are in agreement, at least in terms of their language. However their arguments begin to diverge when it becomes a matter of solving what is known as a problem of commonalities, ie determining the percentage of common variance. Here we run into a basic difference between the principal component model and the principal factor model.

Simply speaking the principal component model (the method used by Oller and by ourselves) doesn't even allow for the appearance of unique factors on top of the common factor expected. On the contrary, the principal component model produces one or more factors where each of the factors comprises common as well as unique variance in an indistinguishable manner. What we want, of course, is to have all the common variance concentrated on one factor whereas the others then only carry specificity. This is exactly what the principal factor model has been developed for and this is why it is superior to the other model.

But the problem with factor analysis is tougher yet when seen from a basic theoretical point of view. All classical forms of factor analysis including the ones mentioned so far are mostly used as **explorative methods,** that is to say, they work even without any piece of foregoing theory. All of these statistical procedures produce factors under almost any circumstances. We will never be able to select the meaningful factors from those that are pure artefacts. In other words, the structural hypothesis of a unitary factor, being the simplest under conditions given, has always quite a good chance of being confirmed, even if it does not represent at all any adequate description of the relationships among the several linguistic skills. Therefore we have to look for newer types of the so-called **'confirmatory' factor analysis** that allow a statistical comparison between theory-guided **structural predictions** and test results on the other hand. What we need is to expand our theoretical knowledge to a point enabling us to make precise **structural predictions** which are sound and reasonable. What we suggest in the end is the application of alternative research strategies: drawing the attention away from factor analysis as a seemingly convenient tool which doesn't help very much to solve the problems posed. Those alternative research strategies would mainly have to focus on language processing theory. They would have to throw light on those internal processes which determine a certain language behavior — preferably on experimental grounds. Here, of course, we touch on the question that many researchers are concerned with nowadays: it is the question of **construct validity** of those tests commonly used as indicators of general language competence[4]. We will never really understand what the correlations between tests of different skills mean, what they signify, and why some are higher than others — unless we better understand and are able to model more precisely the cognitive potentials and task specific operations on which performance in the various language tests depends. Only when our

[4] In this context it is interesting that the correlation between a Cloze test developed at Southern Illinois University and meant to be a measure of overall language proficiency on the one hand and an FSI-type oral interview on the other hand was no higher than .60, as reported in Jones (1977: 257; cf also Hinofotis 1980, where this correlation, however, is not mentioned any more at all). This moderate correlation with a speaking test suggests, I think, that at least speaking proficiency cannot be adequately predicted by a test of overall proficiency — or at least not as well predicted as the other skills. If that is true I cannot understand how anyone can conclude that the Cloze test could replace more complicated and more costly ESL testing procedures without substantial loss of information. I personally consider it to be really a substantial loss of information if we build our judgement of a person's general language proficiency on some measure which does not adequately represent his or her speaking ability. For me it would not be a valid measure of general language ability then.

theoretical knowledge increases, when we have moved further ahead towards construct validation, only then might factor analysis prove again to be useful under certain (restrictive) conditions.

Conclusion

After this excursion into methodological perspectives let me summarise my thoughts and then come back to the question: with all the dilemmas indicated, why are we (still) interested in 'General Language Proficiency'? I think the answer to this question has several aspects to it:

> Proficiency testing has a **distinct social function** which has to do with 'future needs or control purposes', as Davies (1977: 46) so aptly put it. These social implications of proficiency measurement, their so-called predictive value, can be severe for any individual involved. For this very reason it is all the more important that we really understand what our judgement and prediction is based upon, that the construct is as valid as are the tests designed to assess it. If there is any doubt or any considerable degree of uncertainty as to what proficiency (in theoretical terms) really is or what proficiency tests really measure it would be irresponsible, in my opinion, to continue to use the construct (as if it were well defined)[5] or to administer any proficiency measure (as if we were sure about its validity) and use the test scores to make more or less irreversible decisions. There seems to be a certain tendency of many an administration to nourish the belief in the necessity and in the validity of proficiency testing. I would seriously question, however, many a case in which proficiency tests are being exploited for placement purposes or, even worse, career decisions to be based upon them. We should not take it for granted that proficiency testing is done worldwide: each single situation in which such cutting decisions on the basis of proficiency scores are said to be necessary should be questioned and the procedures applied should be publicly called for justification over and over again. We as a society on the whole simply cannot afford to classify people and divide them up (allotting educational and professional chances of different kinds) as long as the question of construct validity of the instruments used are not clarified somewhat further. This is especially true with the concept and measures of GLP[6].

I do not propose, of course, to do away with proficiency testing altogether, as some authors do. Proficiency measures are badly needed, for selection as well as for research purposes, and should therefore continue to be developed and improved (in terms of higher validity). However, as test designers, testers or testees alike we have to **bear in mind all the limiting conditions** that are yet connected with the concept of proficiency, especially again when it comes to overall proficiency. These uncertainties will have to show in the way we are interpreting test results as well as in the carefulness with which we are formulating conclusions or suggest decisions.

In any case, it seems necessary to use more than one type of test (or several sub-tests) in trying to assess so complex (and dubious) a thing as communicative competence (or language proficiency, if you like) – if it were only to make sure that we don't arrive at too narrow a view of a person's foreign language abilities and that our judgements are not unsound or made too quickly.

In addition, any proficiency measurement should not be done at a single point in time alone but should definitely be repeated under varying circumstances because in all probability each single measurement will add to our information and understanding of a person's language ability and language use. This is so even if we use different versions of the same measurement instrument.(The methodological problems involved here are left aside on purpose).

It has been argued in the past that this suggested procedure (more than one type of test plus repeated measurement) is uneconomical in a double sense: it is too expensive and too time-consuming. This objection is true, but the answer is: we have no other choice, things being as they are. Considerations

[5] In a very recently published article reviewing a number of proficiency tests used in the United States, Dieterich *et al* (1979) speak of 'the nebulous character of language proficiency' (1979: 547) in their conclusion.

[6] A somewhat different attitude is expressed by Valette (1975) when she summarises her view on the need of prognostic instruments (especially referring to aptitude testing): 'Within the American educational framework, prognostic tests have but one legitimate use: to predict success in particular cases where an agency (governmental, industrial, etc.) needs to train a small number of personnel in a foreign language. Under such conditions, budget constraints and time factors demand that every effort be made to find the most suitable 'risks', that is, those candidates with the greatest chance of completing the course. Under such conditions, the fact that other equally suited candidates might be excluded due to the imperfections of the prognostic instrument is not a matter of concern since only a limited number of trainees are required in the first place' (1975: 10f.).

of practicality and/or money involved should – at least for a moment – be definitely kept apart from the question of validity (cf Stevenson, in press, for discussion of this point).

The notion of a GLPF and its strong form of interpretation by Oller and others that this dimension is unitary, representing the central core (in an absolute sense) of all that is meant by proficiency in a language (cf Cummins 1979), seems to be very seductive indeed. Apparently, it especially attracts the attention of administrative minds. To the, I believe, GLP is something like a handy label suggesting that the bothersome problems of evaluating people as to their language ability and making predictions as to their future behaviour and success could be solved easily now and effectively with a single instrument (which, admittedly, might need some more refinement, as the Cloze, for example). This is probably one of the main reasons why more and more people (teachers, administrators, managers of personnel) have become interested in GLP and how to assess it quickly (cf the rising awareness and demand for developing 'a' cloze test for English as a Foreign Language in the Federal Republic of Germany).

This perception of GLP implies – at least in my understanding – a good portion of wishful thinking. At the same time it has a strong affinity to the mentality of social engineering inasmuch as personal responsibility for evaluating other people, for making social judgements and decisions (all of them human acts than can be questioned, discussed and can potentially be revised, depending on the power structure) is hoped to be totally replaceable by 'objective' measures. It would be much easier to demonstrate (with the help of 'unquestionable' data) that certain people belong in certain categories, that the responsibility for any social (and socio-economic) consequences lies on the side of the testees themselves.

Fortunately, or unfortunately, things are not as clear-cut (yet). The notion of a GLP is in no way convincing so far, neither theoretically nor from an empirical point of view. The factor analytical data presented cannot be taken as strong evidence to support the unitary competence assumption. As a structural hypothesis it is much too general: a strong GLPF will always show up in statistical analysis explaining more or less of the common variance in a large number of L2 language measures. But the percentage of variance explained differs from study to study and, on the whole, is not high enough to be satisfied with (assuming, by the way, that all the tests included are reliable and valid). After having extracted the first factor (interpreted as GLPF) we cannot be sure at all that the remaining variance is nothing but unique and error variance. On the contrary, some factor analytic studies have indicated that there might be several common factors. As small as this counter evidence may be, it can definitely not be neglected.

Therefore, our search for more than one common factor underlying language performance will have to go on.

Wherever 'global language proficiency' in terms of Oller's strong claim is asserted to be measured we should be very sceptical. It may even be appropriate for some of us being test researchers to inform and back up examinees who begin to question the validity of language proficiency measures in general and demand explanation and justification of what they went through, why, and how exactly judgements were found and decisions arrived at. This may be a nightmare to many a tester and may certainly complicate the testing business – but that's what it is anyway: complicated and highly explosive in its social implications.

Editor's note: This is a revised version of a paper first presented at the First International Language Testing Symposium, held at the Bundessprachenamt, Hürth, 29 - 31 July 1979. The proceedings have since been published in KLEIN-BRALEY, C and STEVENSON, D K, eds. *Practice and problems in language testing, 1* (Bern and Frankfurt am Main: Lang, 1981).

BIBLIOGRAPHY

CANALE, M and SWAIN, M
Theoretical bases of communicative approaches to second language teaching and testing. Toronto: Ontario Institute for Studies in Education, 1979.

CARROLL, J B
Fundamental considerations in testing for English Language proficiency of foreign students. In: Testing the English proficiency of foreign students. Washington, DC: Center for Applied Linguistics, 1961. *Reprinted in:* ALLEN, H B, and CAMPBELL, R N, *eds.* Teaching English as a second language: a book of readings. 2nd ed. New York: McGraw-Hill, 1972, pp 313-321.

CARROLL, J B
The psychology of language testing. In: DAVIES, A *ed.* Language testing symposium: a psycholinguistic approach. London: Oxford University Press, 1968, pp 46-69.

CUMMINS, J
Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. In: Working papers on bilingualism, *19,* 197-205, 1979.

DAVIES, A
The construction of language tests. In: ALLEN, J P B, and DAVIES, A, *eds.* Testing and experimental methods. (The Edinburgh Course in Applied Linguistics, Vol 4) Oxford University Press, 1977, pp 38-104.

DAVIES, A
Language testing. In: Language teaching and linguistics abstracts, *11,* 145-159 and 215-231, 1978.

DIETERICH, T G, FREEMAN, C and CRANDALL, J A
A linguistic analysis of some English proficiency tests. In: TESOL Quarterly, *13,* 535-550, 1979.

FARHADY, H
The disjunctive fallacy between discrete-point and integrative tests. In: TESOL Quarterly, *13*, 347-357, 1979.

FODOR, J A, BEVER, T G, and GARRETT, M F
The psychology of language: an introduction to psycholinguistics and generative grammar. New York: McGraw-Hill, 1974.

HARRIS, D P
Testing English as a second language. New York: McGraw-Hill, 1969.

HINOFOTIS, F B
Cloze as an alternative method of ESL placement and proficiency Testing. (Paper presented at the annual meeting of the Linguistic Society of America, Philadelphia, 1976) *In:* OLLER, J W, Jr, and PERKINS, K *eds.* Research in language testing. Rowley, Massachusetts: Newbury House, 1980, pp 121 - 128.

HOSLEY D, and MEREDITH, K
Inter- and intra-test correlates of the TOEFL. In: TESOL Quarterly, *13,* 209-217, 1979.

INGRAM, E
The psycholinguistic basis. In: SPOLSKY, B, *ed.* Approaches to language testing. (Advances in language testing series 2) Arlington, Virginia: Center for Applied Linguistics, 1978, pp 1-14.

JONES, R L
Testing: a vital connection. In: PHILLIPS, J K, *ed.* The language connection: from the classroom to the world. (ACTFL foreign language education series 9) Skokie, Illinois: National Textbook Company, 1977, pp 237-365.

JONES, R L and SPOLSKY, B, eds
Testing language proficiency. Arlington, Va.: Center for Applied Linguistics, 1975.

NEISSER, U
Cognitive psychology. New York: Appleton-Century-Crofts, 1967.

NEISSER, U
Cognition and reality: principles and implications of cognitive psychology. San Francisco: Freeman, 1976.

OLLER, J W, Jr
Expectancy for successive elements: key ingredient to language use. In: Foreign language annals, *7,* 443-452, 1974.

OLLER, J W, Jr
Evidence for a general language proficiency factor: an expectancy grammar. In: Die Neueren Sprachen, *75,* 165-174, 1976.

OLLER, J W, Jr
Pragmatics and language testing. In: SPOLSKY, B, *ed.* Approaches to language testing. (Advances in language testing series 2) Arlington, Virginia: Center for Applied Linguistics, 1978, pp 39-57.

OLLER, J W, Jr, and HINOFOTIS, F B
Two mutually exclusive hypotheses about second language ability: factor-analytic studies of a variety of language tests. (Paper presented at the annual meeting of the Linguistic Society of America, Philadelphia, 1976) *Published with a different subtitle* ('Indivisible or partially divisible competence') *in:* OLLER, J W, Jr, and PERKINS, K, *eds.* Research in language testing. Rowley, Massachusetts: Newbury House, 1980, pp. 13-23.

OLLER, J W, Jr and PERKINS, K
eds. Research in language testing. Rowley, Massachusetts: Newbury House, 1980.

SAMPSON, G P
A model of second language learning. In: Canadian modern language review, *34,* 442-454, 1978.

SANG, F and VOLLMER, H J
Allgemeine Sprachfähigkeit und Fremdsprachenerwerb: zur Struktur von Leistungsdimensionen und linguistischer Kompetenz des Fremdsprachenlerners. Berlin: Max-Planck-Institut für Bildungsforschung, 1978.

SANG, F and VOLLER, H J
(forthcoming) *Modelle linguistischer Kompetenz und ihre empirische Fundierung. In:* GROTJAHN, R and HOPKINS, E, *eds.* Empirical research on language teaching and language acquisition. (Quantitative linguistic series) Bochum: Studienverlag Dr Brockmeyer.

SCHOLZ, G and others
Is language ability divisible or unitary? A factor analysis on twenty-two English proficiency tests. (Paper presented at the 11th Annual TESOL Convention, Miami, Florida, 1977) *Published in:* OLLER, J W, Jr, and PERKINS, K, *eds.* Research in language testing. Rowley, Massachusetts: Newbury House, 1980, pp 24-33.

SELINKER, L and LAMANDELLA, J T
Fossilisation in interlanguage learning. In: BLATCHFORD, C H and SCHACHTER, J, *eds.* On TESOL '78: EFL policies, programmes, practices. Washington, DC: Teachers of English to speakers of other languages. 1978, pp 240-249.

SPOLSKY, B
What does it mean to know a language? or how do you get someone to perform his competence? In: OLLER, J W, Jr, and RICHARDS, J C, *eds.* Focus on the learner: pragmatic perspectives for the language teacher. Rowley, Massachusetts: Newbury House, 1973, pp 164-176.

STEVENSON, D K (in press)
Beyond faith and face validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests. (Paper delivered at the Colloquium on the Validation of Oral Proficiency Tests at the 13th Annual TESOL Convention, Boston, 1979) *In:* PALMER, A S, and GROOT, P J M, *eds.* The validation of oral proficiency tests. Washington, DC: Teachers of English to speakers of other languages.

STRAIGHT, H S
Comprehension versus production in linguistic theory. In: Foundations of language, *14,* 525-540, 1976.

VALETTE, R M
Modern language testing: a handbook. New York: Harcourt Brace Jovanovich, 1967.

VALETTE, R M
Directions in foreign language testing. New York: New York: Modern Language Association, 1975.

VOLLMER, H J and SANG, F
Zum psycholinguistischen Konstrukt einer internalisierten Erwartungsgrammatik. Trier: Linguistic Agency, University of Trier (LAUT), Series B, No 46, 1979.

VOLLMER, H J and SANG, F (forthcoming)
Competing hypotheses about second language ability: a plea for caution. Submitted for publication in: Applied psycholinguistics.

REACTION TO THE PALMER & BACHMAN AND THE VOLLMER PAPERS (1)

Arthur Hughes, University of Reading

My immediate reaction to the these two papers was the strong feeling that I needed to put them aside and work out for myself just what was involved in claims made about unitary competence and general language proficiency. I must confess that I have not always been sure what was being claimed, with what justification, and at times I have even wondered whether anything very interesting was being said at all. This paper is an account of the thinking that has helped me reduce my uncertainty. It also makes suggestions for the improvement of research in the area. I shall make points and present arguments as briefly as possible, allowing the discussion to provide whatever expansion is necessary.

We say that someone is proficient at something when he can perform certain tasks to what we regard as an acceptable standard.[1] So it is with language proficiency, though our notion of adequacy is as likely to be norm-referenced as it is criterion-referenced. The question arises immediately: what are the relevant tasks for the assessment of language proficiency? The answer is that we are at liberty to choose whatever language-based tasks we like: solving anagrams, finding rhymes, judging the grammaticality or acceptability of sentences, making translations, or even doing cloze tests. It seems to me, however, that certain tasks — such as reading a book with understanding, writing a letter, or holding a conversation — are more central to our interests. The uses of the four skills in performing more or less natural language functions is, after all, what most modern language teaching supposedly has as its objective. It follows from this that the study of the performance of such tasks is an essential part of research into language proficiency, just as it is against such performance that proficiency tests must be validated. It is not good enough to administer the writing ability section of TOEFL, which once correlated at around .7 with some writing task but which is essentially a test of the grammar appropriate to formal English, and then claim that you have measured writing ability.

[1] Vollmer worries whether proficiency is a matter of competence or performance. The distinction is unnecessary. We might use it if we thought it helped us think more clearly about the matter in hand, but I believe that, at least if it is the Chomskyan distinction(s) we have in mind, it does quite the opposite.

The 'natural' language tasks that most concern us can be classified according to:

1 which skill or skills (of the four) are involved
2 subject matter
3 style
4 regional variety
5 function

and, doubtless(6) something you have thought of that I have omitted. There are thus in principle a great many proficiencies. We ought not to need experiments like that of Bachman and Palmer to tell us that, even when measures are norm-referenced, individuals do not show equal ability in each of them. How then is the unitary competence hypothesis (UCH) to be defended? Presumably by pointing out:—

1 The individual will have had unequal exposure to or practice in the different skills, styles etc. You are unlikely to speak English very well if you have never heard it, however much you have read it. Nor will you understand British English as well as American English when you have previously only heard the latter. The unitary competence hypothesis must assume, then, equal exposure and practice. (Even though proficiency tests are thought of as looking forwards rather than backwards, if they were genuinely intended to predict longer-term future (rather than tomorrow's) performance, they would have to take into account previous exposure and practice (or some measure of aptitude)). Research must control for, or at least take into account, exposure and practice.

2 Non-linguistic factors will inhibit performance in some tasks but not others. These may be of two kinds:—

a emotional or cultural; like shyness or excessive respect. For the tester, provided the inhibition is more or less permanent and not just a product of the testing situation, inferior performance may be considered to give a true picture of the subject's ability. For the researcher interested in the UCH, on the other hand, it can be argued that differences attributable to such factors are of no significance, and must be controlled for.

b some physical defect; poor eyesight, cleft palate. Because the eyes and mouth are at the periphery of the language processing system(s), they may be dismissed as non-linguistic and irrelevant to the UCH. The temptation will be, however, to regard as irrelevant anything which results in a difference in performance between skills; for example, differences between visual and auditory memory which might well contribute to differences between reading and listening ability. If this line is pursued far enough (something which the competence-performance distinction encourages), anything which language processes do not have in common will be excluded from consideration, and the UCH will inevitably be substantiated.

If we ignore differences in performance attributable to (1), (2, a), and those parts of (2, b) that we think reasonable, would we expect performance in the different skills to be equivalent (with norm-referencing)? Our expectations might depend on what we know about (1) language processing, and (2) such learner variables as (a) aptitude and (b) motivation.

1 The first thing to say about language processing is that we know very little about it (see Butterworth (1980) for recent confessions of psycholinguists). One can, however, indulge in conjecture. It would be strange, I suppose, if there were (at least) four completely independent processing systems. To the degree that we monitor our speech, speaking and listening processes are presumably related. And when one skill is firmly established before the development of another is begun, a parasitic relationship between them seems inevitable (for example, my subject Blanca (Hughes 1979) had still read no English after 6 months learning of the language through conversation with me; yet she was able to read immediately I presented her with a book). At the same time, psycholinguists would seem to be moving towards the view that each process is sufficiently different to necessitate quite separate study, something reflected in the nature of books recently published in the field.

In the end, part of our understanding of the nature of language proficiency will be in terms of language processes. But the end is not near, and I agree with Vollmer that we should be sceptical of explanations that make use of concepts like 'synthesis by analysis'.

2 a Even if relatively independent processes are involved in the four skills, equivalent performance in each could result from differences in language aptitude. Aptitude for one skill might be a perfect predictor of aptitude for the other three. Evidence (perhaps not particularly reliable) from Pimsleur et al and Gardner and Lambert, however, would point to this not being the case.

Obviously language processes and aptitude are related; the aptitude we are talking about is for developing these processes. Nevertheless, similarity of processes and aptitude are logically distinct explanations of apparent unitary competence. Oller has spoken about both, without, as far as I know, presenting them as alternatives. I would suggest that a full

understanding of language proficiency will be facilitated by (and may depend on) further research into aptitude.

b Similarities and differences in performance between skills might be due to different degrees and types of motivation for the possession of these skills. Gardner and Lambert's (1972) work would suggest that this is a line of research worth pursuing, using more sophisticated measuring techniques.

I said earlier that we must get subjects to perform a variety of 'genuine' language tasks. What I want to add now is that we should measure performance on each task according to as many criteria as possible. It is essential, I think, that judgements should be independent; judges A B C would rate according to one criterion and no other. Where a single judge is required to provide ratings according to a number of criteria, a high correlation between them seems inevitable (eg FSI interview in Oller and Hinofotis, which results in a separate factor).

While on the subject of the conduct of experiments in this field, I want to suggest that the subjects be of as limited a range of ability within which it is possible to discriminate reliably. Too wide a range will bias the results in favour of the UCH, obscuring interesting and important differences. It is no surprise that when the ability range in Oller and Hinofotis's (1980) experiment was reduced, another factor emerged.

I have talked so far about the four skills and varieties. The other dimension of proficiency along which separable components have been thought to lie is the linguistic: grammar, semantics (or vocabulary), phonology/graphology. However plausible such components may seem, it must be remembered that levels are for the convenience of linguistic description and theory, and while some correspondence with units or stages of processing seem plausible, a one-to-one relationship is by no means inevitable. What is more, within linguistics there is not always agreement on the number and nature of the levels appropriate to descriptions of particular languages.

Even when there is agreement, it is clear that levels are not discrete, that they interact eg phonology and syntax in English negative contraction. In language learning there are similar interactions, for example Rodgers' (1969) finding that success with which items of Russian vocabulary were learned depended largely on the ease with which their phonetic shape allowed them to be anglicised. And the difficulties error analysts experience (or should experience) in assigning errors to levels are well known.

In the light of what has been said in the previous paragraph it would not be surprising if it proved impossible to separate out these linguistic components when evaluating performance on various linguistic tasks and to establish them as factors underlying performance on all the tasks. But I do think it is worth trying, provided that 'real' language tasks are involved and that the supplementary tests meant to measure control of the separate components are 'pure' (I am thinking of vocabulary items in the grammar sections of the English Language Battery (Edinburgh) and the English Placement Test (Michigan). Judgments of performance might be supplemented by linguistic analysis (eg types of structure used/misused in written and spoken output).

It should be clear, I think, from what has gone before, that I regard most of the research that has been done in this area as deficient in one respect or another. Few conclusions can be safely drawn. Interesting work **can** be done, but it must be more carefully controlled, using more or less homogeneous groups performing 'real' language tasks (amongst others). Whatever is done, I fear that it will be a long time before progress in the study of language processing will be sufficient to improve significantly the quality of language proficiency tests; and factorial studies of language performance are unlikely to provide more than clues as to the nature of language processes. What promises to bring more immediate benefits, at least to tests, is the continued investigation of the relationships holding between performance on a variety of language tasks. Whatever the fate of the UCH, if it continues to stimulate worthwhile research it will have done that much good.

BIBLIOGRAPHY

BUTTERWORTH, B
ed. *Language production* (Volume 1, Speech and talk) London: Academic Press, 1980.

GARDNER, R C and LAMBERT, W E
Attitudes and motivation in second-language learning. Rowley, Massachusetts: Newbury House, 1972.

HUGHES, A
Aspects of a Spanish adult's acquisition of English. In: Interlanguage studies bulletin (Utrecht), 4 1, 1979.

OLLER, J W Jr, and HINOFOTIS, F B
Two mutually exclusive hypotheses about second language ability: indivisible or partially divisible competence. In: OLLER, J W, Jr and PERKINS, K. Research in language testing. Rowley, Massachusetts: Newbury House, 1980.

OLLER, J W Jr, and PERKINS, K
Research in language testing. Rowley: Massachusetts: Newbury House, 1980.

PIMSLEUR, P, SUNDLAND, D M and McINTYRE, R D
Underachievement in foreign language learning. In: IRAL, 2 2, 1964.

RODGERS, T S
On measuring vocabulary difficulty: an analysis of item variables in learning Russian-English vocabulary pairs. In: IRAL, 7, 327-343, 1969.

REACTION TO THE PALMER & BACHMAN AND THE VOLLMER PAPERS (2)

Alan Davies, University of Edinburgh

The general versus specific (or unitary versus divisible) competence debate is a classic of psychology and no doubt of philosophy too. It has, like all great disputes, traces of the grand and unsolvable binary themes, of nature versus nurture and realism versus nominalism and perhaps good versus evil. My own view is that the structure of competence or skills or ability is partly a practical issue and partly a mathematical choice. (Of course it is also a philosophical question but in this instance that seems to me not amenable to proof). From a practical point of view it is important whether one views language (or any other 'ability' or 'skill' or 'competence') as a whole or as an array of parts – the implications for syllabus, for testing and even for varying learner activities are obvious, as are the criteria for judging eventual success. The mathematical issue may be less obvious but it is a well-known chestnut of applied statistics, viz that in Factor Analysis 'solutions' there are (at least) two ways of presenting the results, either as a superordinate with several (possible) subordinates (Type A) or as a set of equal partners (Type B). The Principal Components method of Factor Analysis will typically produce a Type A solution, the Rotation method a Type B. The great exponents of Factor Analysis have indeed represented the Type A (Spearman's general factor, involving a hierarchy) and the Type B (Thurstone's group factors) solutions. But there is no way of preferring one method (and therefore solution) to the other, short of appeal to one's view of the universe or of arithmetic elegance.

My position, then, on the issue of General Language Proficiency (GLP) is that it is essentially a non-issue theoretically. At the same time the practical implications are important.

I will now consider some of the arguments in the two papers and then make some procedural suggestions. In both papers the authors refer to J W Oller whose work has renewed interest in the question of language test validity (essentially the question of what should one test and therefore of the structure of abilities – one or more factors) through empirical rather than speculative research. It must be said that the discussion over Oller's work, which he has fostered, has been on a slightly separate issue. Oller's data show that his integrative or as he says pragmatic tests eg dictation, cloze, represent total EFL proficiency better than any other single test or indeed combination of tests. Whether this is so or not it is **not** an argument for the unitary factor view since, as Oller would agree, both dictation and cloze are so integrative

that they contain most or all language abilities. Now, if you construct a test that already contains everything you cannot then argue that it contains everything. So, as I see it, the 'best test' data and arguments of Oller are not necessarily of relevance in the GLP debate. I will not, therefore, deal directly with the Oller results, and turn first to Palmer and Bachman.

Here the two authors present results of a two-year study in construct validation carried out by them but monitored and encouraged by participants at two TESOL colloquia on oral tests in 1979 and 1980. The general method employed was that of the multitrait-multimethod model. In this case the design allowed for three Methods and two Traits, the methods being: interview, translation and self-ratings, and the traits: communicative competence in reading and communicative competence in speaking. Apart from the methodological interest of the project the aim was to investigate the validity of the construct: communicative competence in speaking. Palmer and Bachman assembled six tests (3 methods X 2 traits) which they administered to an N of 75 non-native speakers at the University of Illinois. The results are presented and discussed in the light of two methods of analysis, a correlational analysis and a confirmatory factor analysis. (Note that the third method they mention, principal component analysis, is not used, apparently for reasons of bias).

I do not always see eye to eye with Palmer and Bachman about the nature of correlation. For example, they say . . 'if a test of the trait "mathematical ability" and another of the trait "verbal ability" always gave the same results, that is if they ordered the subjects taking the tests in exactly the same ways, there would be no evidence that the mathematical and verbal ability traits were actually distinct'. While that is true, so is the converse, there would be no evidence that they were **not** distinct. Correlations are indicators of shared variance not or equivalent identity. Again, the kind of argument used about correlation sizes makes me uneasy. Here is a typical example: 'The effect of method is particularly noticeable in tests using translation or self-rating methods. Of the indices, in the diagonal in the lower left-hand box the intercorrelations between tests 3 – 5 which employ translation and self-rating methods (.64, .69, and .68) are clearly higher than those between tests 1 and 2 which do not (.54 and .46)'. Apart from the lack of mention of reliability of rs here and of any tests of significance between rs what is missing is the recognition that eg the first two rs mentioned may represent different segments of variance space ($.64^2 = .41$ and $.69^2 = .48$). Now it is difficult enough to compare repeated rs of X on Y since until they reach .7 they may occupy quite different variance space, but it is even more difficult with rs between quite different pairs. To say that the r of X on Y is bigger than the r of A on B is not necessarily very instructive.

I find in Palmer and Bachman another problem, that of distinguising clearly between Method and Trait. (I understand that this was an issue at the second colloquium in 1980). Palmer and Bachman select Communicative Competence in reading as a trait and translation as a method. But it could be maintained that translation is a trait and reading a method. Or better that they are both combinations of method and trait. No method it seems to me can ever be entirely free of the trait it seeks to realise. Interview, like translation, again seems to me as much trait as method. And so on. Only very technical 'methods' (like multiple-choice questioning) may be trait-free and I am not sure even about these. Next the arguments against Principal Component Factor Analysis. I don't understand these, either that PrinComp can be used only for variance structure and not covariance structure, or that 'principal component analysis cannot be used to examine any kind of structural model in which the elements in the model are correlated . . .' Why not? Surely most factor analysis studies deal with correlated elements.

Notice that in terms of correlations (Table 1) both for reading and speaking it is self-rating that is the 'best' method in that it shares least with the other methods. What that argument indicates is the slight absurdity of comparing correlation sizes.

In spite of my animadversions it seems to me that Palmer and Bachman do demonstrate their hypothesis, *viz* that **according to their analysis** the two traits, reading and speaking, differ when method is controlled.

The issues raised by Palmer and Bachman are arithmetical ones, they have to do with differing interpretations of factor analysis. Vollmer presents us with a different kind of argument. In the first place, Vollmer is deliberately offering a critique of the GLP position. In the second place, he advances his arguments from a theoretical standpoint. (He tells us that he has supporting data but does not in this paper present them.) What, he asks, can GLP be a description of? Is it about competence or about performance? Is it a metaphor, a way of talking about language ability; is it a construct (like competence) which enables us to idealise language itself? (He recognises that this implies some static nonvarying view of language ability.) Or is it an argument about language skills (ie performance)? If that then it would be possible to combine in one's view of language ability both GLP (= competence) and the divisible view (ie performance), though of course empirically we might find that in performance too the GLP position could be maintained.

Vollmer points out that the divisible competence hypothesis has been assumed by 'a great number of researchers' who have adopted this view for want of a convincing theoretical alternative. So while there is no (or little) experimental evidence for this view there is a lot of experience and, as

Vollmer indicates, many of our assumptions belong to this divisible competence position. At the same time there has always been the related assumption of 'transfer ability', *viz* that there is the likelihood of performance on one test being substantially correlated with performance on another.

Vollmer then shows how the concept of 'overall proficiency' has inevitably merged into the second major hypothesis, that of a unitary competence. Vollmer's position is that this unitary competence view, widely promoted by eg Oller in his discussion of an 'internalised expectancy grammar' is not justified and that it flies in the face of substantial evidence in favour of two competencies (at least), those related to comprehension and production.

Vollmer characterises the central idea of GLP as a psychological construct, identified in some way with the 'general cognitive apparatus' of a person. The trouble with this view, Vollmer suggests, is that it fails to incorporate the necessary dynamic of L2 proficiency, a dynamic which is captured by eg the interlanguage hypothesis.

From his own empirical studies Vollmer claims that neither the unitary nor the divisible hypothesis has strong grounds for acceptance. Vollmer points to the differing but equivalent solutions provided by Factor Analysis and concludes that for all known practical (especially selectional) situations a GLP construct is dangerous and altogether too simple. What is needed, Vollmer suggests, is more theoretical work on construct validity. Which takes us back full circle to Palmer and Bachman who, it will be remembered, start from a construct validity hypothesis and then carry out empirical research to test that hypothesis. So far so good.

Finally, I want to make a suggestion as to how we might approach the GLP issue by indicating the different kinds of argument involved. First, there is the **philosophical argument:** this may be what is meant by construct validity if it allows for testing. Thus the argument that GLP applies to both L1 and L2 seems to me interesting and testable. The argument that speaking and reading in an L2 are/are not combined in a GLP is, as Palmer and Bachman show, testable. Second there is the **competence-performance** argument. Since this is **either** a philosophical **or** a practical issue (ie we are testing one or the other) this merges into one of the other arguments. Third, there is the **practical** argument which is well considered by Vollmer and which says in view of our lack of clarity it is best to gather as much evidence as possible from a wide variety of tests; this has special force in diagnostic testing. Fourth, there is the **factor analysis** argument, though this does seem to produce conflicting results. Important arguments that are not much discussed are those dealing with **language variation** (Vollmer mentions this through time – his dynamic – but what of inter-speaker variation: whose GLP are we talking about?), with **predictive validity** and with the **'one best test'** idea, integrative, communicative or pragmatic.

REPORT OF THE DISCUSSION ON GENERAL LANGUAGE PROFICIENCY

J Charles Alderson, University of Lancaster

Preamble

The debate about the unitary competence hypothesis revolves around the issue: is there one underlying proficiency in a second language, or are there several proficiencies, which are separately measurable and teachable? Is Reading separate from Writing or Listening? Is knowledge of vocabulary distinct and separable from knowledge of grammar? If there are no differences between these 'knowledges' or 'skills' (or at least demonstrable differences), the practical consequences are quite considerable, namely that one only needs to measure 'proficiency', and the test used is relatively unimportant. The pedagogic consequence seems to be that one need not worry about teaching, say, writing and speaking explicitly, since teaching reading alone will affect general language proficiency and, 'automatically', writing will improve. Of course, the existence of a general language proficiency factor could also be used to justify the opposite of concentrating on teaching reading in order to improve reading ability: in order to improve reading it might be argued, one can validly teach the writing, speaking and listening skills at the same time, since they all relate to General Language Proficiency.

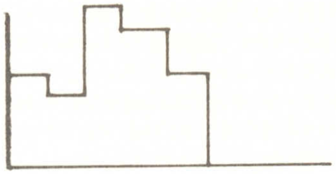
It was apparent during the discussion of this topic that there was a problem of level of abstraction or generalisation in the identification or acceptance of the existence of one general language proficiency factor: since all humans have language, in one sense at least there can only be one general language proficiency factor underlying Language. General Language Proficiency (GLP) from such a perspective is what language is. The more interesting discussion starts when one looks at less abstract levels: at such levels it is self-evident that people have different skills: some people can speak and understand when spoken to, but cannot write or read acceptable English. Other people have merely a reading knowledge of a language, and are unable to write it, speak it or understand the spoken form of it. However, these differences would be more apparent than real if there is a common underlying competence.

When someone is proficient, we mean that s/he can perform a task, any task, to a required level (the criterion). If an individual is given several tasks to perform and he performs them differently, does this mean that he has differing abilities (proficiencies)? Would one, in any case, expect performances in the different skill areas of a second language to be equivalent? It was suggested in the debate that given the vast variety of sociolinguistic settings in which it is possible to perform in a second language, one would surely anticipate a variety of different proficiencies underlying performances on various tasks. Common sense suggests that we do not **need** empirical research of the Bachman and Palmer or Oller kind to prove the obvious. Curiously, however, research results, particularly those based on principal component and factor analyses do not always show the different proficiencies one expects.

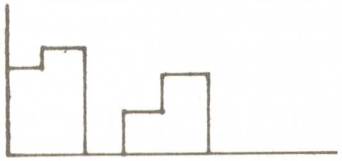
One view was that the reason for this might be the method of analysis: the statistical method one chooses conditions the results one gets. Principal Component Analysis is intended to simplify data, to reduce it to one best solution if possible, and is therefore likely to result in one factor emerging from the intercorrelations of, for example, a battery of apparently different language tests. Maximum likelihood factor analysis on the other hand, looks for as many factors underlying the data as possible, and tests emerging factors for significant contributions to variance. That the number of factors one discovers may be an artefact of one's statistical model is suggested also by a study by Oller and Hinofotis (1980). A given set of data was analysed according to the principal component method, and one factor was revealed. When the data was factor analysed using a Varimax rotation procedure, two factors emerged, one relating to FSI interviews, and the other to cloze and multiple-choice test performance. In order to disprove the General Language Proficiency hypothesis, clearly the method of analysis one selects should not be the one which **favours** the hypothesis. It was generally agreed that in research of this kind one should always use maximum likelihood factor analysis.

It was suggested that another reason for the failure of several factors to emerge from the data may be the lack of homogeneity of the groups studied. Studies have suggested that when a relatively homogeneous subgroup of a population is studied, more factors emerge than when one looks at the heterogeneous population. It was not, however, entirely clear in the discussion why this should be. The claim is that to allow interesting differences among individuals to emerge, one needs to study a relatively homogeneous group, ie to allow the maximum possibility for several factors in language proficiency. But this does not seem to be reconcilable with the consideration of differences in performance: if the sample studied included people with deformities, for example, or gross differences in length of exposure to the second language, that would increase the likelihood of more than one factor emerging. Thus increased heterogeneity should lead to a multi-factor solution. Should one exclude from one's sample such extreme differences in performance? If so, when do differences in performance cease to be 'extreme'? When the results they yield prove our hypothesis? This, it was felt, was unsatisfactory.

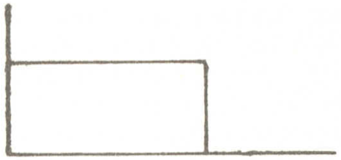
The argument that heterogenous groups lead to a (false) unifactorial solution was illustrated with an example. Assume that Tom, Dick and Harry form a heterogeneous group with respect to language proficiency, such that Tom is much better than Dick, who in turn is much better than Harry. If one attempted to see whether there are separate abilities in, say, writing and listening, with such a group, one would find that Tom was better than Dick in both writing and listening, and Dick was better than Harry in both abilities also. Thus only one factor **could** emerge from such a group, given the constant rank order of subjects. Yet interesting differences might exist between Tom's writing and listening abilities, and simply be obscured. The important question, is: if we are interested in differences within one individual, why do we group individuals together at all? Doing so is quite likely to obscure differences, unless the differences within individuals are the same for all or most individuals (which may not be likely). To put the argument graphically, it might be found from test data that individual T. has a profile of various abilities (test scores) that looks like:



Whereas individual H. has a profile that looks like:



That is, both appear to have interesting differences in abilities. However, if one groups the scores, the result is:



One might argue that one wants data from groups rather than individuals for reasons of realiability. One might also be interested in comparability: Harry's writing may be worse than his listening **and** that is true for everybody else also, which is an interesting fact. Similarly, of course, it is also interesting if Harry's writing is worse than his listening and such is **not** the case for anybody else.

Considerable discussion took place over the reasons for the interest in General Language Proficiency. Why is one at all interested in differences between writing and listening? What does one want to know? Do we want to know if different individuals merely find writing more difficult than listening? Or do we want to know whether there is a relationship between writing and listening? If there is a relationship, what are the consequences? Does the fact that one can perhaps predict listening ability from writing ability help us to produce better tests? Does it help us to understand language proficiency better? It was suggested that there are obvious teaching implications of such relationships, of the existence of one general language proficiency factor, or of several closely related proficiencies. There are also practical consequences for language testing of the existence of such relationships, or, indeed, of the existence of only one language proficiency. If there is only one general language proficiency, it may be argued that any language test that taps this proficiency will suffice. This, essentially, is the 'one best test' argument, put forward in relation to integrative test like cloze test and dictation tests. The argument is that there is such a high correlation between, say, a cloze test and a variety of other types of tests that for all practical purposes it does not matter whether one uses a variety of other tests, or the cloze test alone. The question raised was whether testers (and, more crucially, testees and teachers) are happy to test just one 'skill' knowing that it will tap the one general language proficiency factor to the same extent and with the same efficiency as any other test? Are we content to ignore writing and speaking in our tests (because they are difficult to test or impractical) since we 'know' we will measure the ability underlying Writing and Speaking with more practicable tests of listening and reading? It is commonly argued in favour of 'indirect' measurement of writing in, for example, the TOEFL that such measures correlate highly with more direct measures.

It was generally agreed that the only reasonable view to take on the 'One Best Test' argument must be that of Vollmer, namely that, regardless of the correlations, and quite apart from any consideration of the lack of face validity of the One Best Test, we must give testees a fair chance by giving them a **variety** of language tests, simply because one might be wrong: there might be no Best Test, or it might not be the one we chose to give, or there might not be **one** general language proficiency factor, there may be several. The one Best Test argument derives from the Unitary Competence Hypothesis, (since if competence is unitary, any test will measure it). It may be that the Unitary Competence hypothesis is too simple. The fact is that it is too dangerous at a practical level because it implies that it does not matter

which test one uses, and that is unacceptable when decisions are being taken, on the basis of test data, which affect people's lives. There may be arguments for using a smaller number or a narrower range of language tests when it is not important that our measurements be accurate. In other words, our purpose in testing, and the consequences of making wrong decisions about people, will affect **what** testing we do, and how many tests we give, regardless of the posited existence of one general language proficiency factor. A related issue, already raised in the discussion of communicative language testing in connection with extrapolation, was also mentioned at this point. If one needs to predict how well someone will perform in a seminar discussion, does one have to measure **that** performance? If a cloze test will predict the performance, should we not be content with the cloze test? The problem with such an argument, apart from the ethical arguments already mentioned, is that it presupposes that we can actually measure seminar performance. This is precisely what communicative language testing attempts to do, but so far there has been little convincing success.

A doubt was expressed as to whether it is possible to say anything meaningful about the existence of one general language proficiency factor until we have investigated specific competences. It was claimed that in the UK typically we do not attempt to test general language proficiency by producing general language tests, but that there is a tradition of testing for specific purposes, and in particular for academic study purposes. The theoretical notion behind ESP testing is that there is no General Language Proficiency. There would appear to be a conflict between the proponents of the Unitary Competence Hypothesis and ESP testers, who would argue that one must identify the real, sociologically defined activities and skills, and measure them. Presumably, however, it **is** possible to investigate GLP without prior investigation of specific competences, at least in the sense of an ESP test, and that would be by looking at what people can do in the most context-free situation. R A Kelly, for example, in his tests for the Department of Education in Canberra, Australia, has attempted to limit content and context by taking real texts and substituting the lexical items with other lexical items and nonsense words, in an attempt to get at a general language proficiency. One's view of whether it is necessary to go to such lengths depends in part upon whether one believes that context **determines** meaning (a strong hypothesis which would lead one to reject work like Kelly's) or whether context merely **conditions** meaning (the weak hypothesis). One may wonder also whether it is **possible** to remove context from text, if one includes within context the knowledge and experience that the testee/reader/interlocutor brings to the text or communication. The search for non-specialised texts for Reading Comprehension tests (texts comprehensible to the educated lay person, for example) looks like an attempt to find neutral content. However, does not this very search presuppose the existence of a general language proficiency factor? We choose 'neutral' content for our general tests in order not to bias the test in favour of one type of knowledge or experience, yet we predict from performance on such a test to performance within a specific context – assuming thereby that there must be a general language factor underlying both performances. If it is unacceptable to assume that there is a general language proficiency factor, then are we not driven to ESP-type tests? Of course the reverse argument also holds: that if there is a general language proficiency factor, then text **content** does not matter, and one can predict from performance on one text to performance on any other text. If this is an unreasonable inference from the Unitary Competence Hypothesis, then presumably the hypothesis would appear to be untenable. It was not clear whether the alternative to the UCH meant positing the existence of different proficiencies or abilities for different types of texts, or one proficiency for familiar texts, and another one for unfamiliar texts, or one for general texts and one for specific texts. In addition, the relationship was unclear between the notion of a 'core language' – basic structures, key vocabulary, and so on – and the Unitary Competence Hypothesis. It was felt that the Unitary Competence Hypothesis necessarily required that there be a core language in which to be proficient.

A question was raised of the implications for test profiles, of the sort done by ELTS, of the existence of a general language proficiency factor. If there were only one general language proficiency, then profiles across different tests would not be necessary, or, conceivably, possible. It was not clear to what extent a general language proficiency factor would allow for differences in profiles, that is differences among individuals across a series of tests, regardless of the differing test difficulties.

It was agreed that if there were one general language proficiency **across** languages there would be important educational consequences. For example, it may be true that one's reading ability in a second language is affected (or determined) by one's reading ability in the first language – certainly the general language proficiency theory would suggest this. There would appear to be value in research which examined the relationship between reading abilities in the first and second languages, and, if possible, related that to levels of proficiency in the second language. The results of Alderson, Bastien and Madrazo (1977), Clark (1979) and Alderson (1980) would suggest that it is not impossible to investigate the Unitary Competence Hypothesis **across** languages, by hypothesising the existence of two factors, one in the first language and one in the second.

General agreement was reached that research was needed into a number of areas related to but not necessarily derived from the UCH. Research into the relationship between abilities in first and second language would, as suggested, be of practical interest. There is a need to develop tests of criterion behaviours and then to relate a series of relatively indirect tests to those criteria to determine their best predictors and the interrelationships among tests. Such information should not only help us to understand better the nature of language proficiency, and language performance, but also enable us to improve our language tests, and possibly facilitate the understanding of general principles to make tests more efficient. The method of such research would be to assemble a battery of potential predictors – all selected or constructed according to a theory of language processing and production – to relate them to each other and to create and develop a data bank of test relations and performances. Clearly, it would be as important to know which tests were unrelated to others, or which were unsuccessful predictors, as it would be to know which tests were related or predictive.

Similarly it is necessary to research the relationships between performance on a test with one type of subject content, and that on a similar test with different subject content. Such research would throw light on the relationship between general language proficiency and specific language proficiencies, and between general language tests and ESP-type tests. A suitable vehicle for such research might be the new ELTS tests, with similar reading tests in the six areas of Physical, Medical, Life and Social Sciences, Technology and General Academic Studies.

The importance of the Unitary Competence Hypothesis was felt to be at least as much its capacity to generate controversy and research as its inherent truth, or lack of it.

BIBLIOGRAPHY

ALDERSON, J C
L_2 Reading: A Reading Problem or a Language Problem?
Paper presented at TESOL Conference, San Francisco, USA, 1980.

ALDERSON, J C BASTIEN, S and MADRAZO, A M
A comparison of Reading Comprehension in English and Spanish.
Research and Development Unit Report Number 9, UID, CELE, UNAM, Mexico: mimeo. (1977).

CLARKE, M
Reading in Spanish and English: Evidence from adult ESL students.
Language Learning, Vol. 29, No. 1, pp 121-150. (1979).

DAVIES, A
Language Testing: Survey Article. Linguistics and Language Teaching Abstracts, Vol. 11, nos. 3/4, 1978.

DAVIES, A, MOLLER, A and ADAMSON, D
The English Proficiency of Foreign Students in Higher Education (Non-University) in Scotland. A report to the Scottish Education Department, Edinburgh.

EGGLESTON, J F, GALTON, M and JONES, M E
Science teaching observation schedule. (Science Council research Studies) Macmillan Education, 1975.

KELLY, R
On the construct validation of comprehension tests: an exercise in applied linguistics. (PhD) University of Queensland, 1978.

MURPHY, D F and CANDLIN, C N
Engineering Lecture Discourse and Listening Comprehension. Practical Papers in English Language Education, Vol. 2, pp. 1-79. (1979).

OLLER, J W, Jr and HINOFOTIS, F B
Two mutually exclusive hypotheses about second language ability: indivisible or partially divisible competence. In: OLLER, J W, Jr and PERKINS, K. Research in language testing. Rowley, Mass.: Newbury House, 1980.

ISSUE OR NON-ISSUE: GENERAL LANGUAGE PROFICIENCY REVISITED

Helmut J Vollmer, University of Osnabrück

At the time when I wrote my paper 'Why are we interested in 'General Language Proficiency'?' to be presented at an International Language Testing Symposium in Germany, I intended to clarify some of the basic claims and questions associated with that concept. I knew that Bernard Spolsky was going to take part in that symposium having been invited as the main guest speaker. I considered Spolsky to be one of the major proponents of that concept and was all the more surprised to find out that he was not any more, but was rather critical of the testing business altogether (cf Spolsky 1981).

In the meantime much discussion has been going on internationally and, equally important, substantial research results have been presented ever since. In addition, the methodological aspects of the issue have been elaborated upon and new theoretical advances have been proposed with the effect that no firm answer seems to be possible to the question 'Is there really only one single factor of language ability?' in the near future. I would not go as far as to assert that the issue of General Language Proficiency (GLP) 'is essentially a non-issue theoretically' (Davies,this volume), but it certainly has changed its quality and forcefulness as a research question: the discussion has led away from the macro-level of mathematical data reduction procedures to a psychologically more informed and better motivated information-processing view of language performance and to the intensive analysis of tasks and individual differences on a micro-level of behaviour.

In order to illustrate somewhat further what I mean by these general remarks I have divided my response into three sections:

1 Empirical findings
2 Methodological and theoretical advances
3 Final remarks.

Empirical findings

Much progress has been made during the past two or three years in proving that the structure of foreign language ability is not likely to be one-dimensional. As far as I can see no one seriously claims the Unitary Competence Hypothesis (UCH) to hold true empirically any more. On the contrary, even John Oller at a Testing Symposium held last summer (1980) at the University of New Mexico in Albuquerque, publicly admitted that he might have gone too far in postulating only one unitary, indivisible underlying ability all the different language activities a learner can engage in. It will be interesting to find out how far Oller has come to a substantial revision of his position in a new book edited by him on **Issues in Language Testing Research** to which the present author also has contributed (cf Vollmer/Sang forthcoming).

In my own research (done jointly with F Sang at the Max-Planck-Institut für Bildungsforschung in Berlin) I tried to demonstrate that the empirical evidence presented in a great number of studies does not really enable one to decide in favour of one or the other theoretical positions. On the basis of an extensive analysis and re-analysis of all the relevant factor analytic studies up to 1979 (including Gardner/Lambert 1965, Lofgren 1969, Caroll 1975, Oller 1976, Steltmann 1978, Sang/Vollmer 1978, Hosley/Meredith 1979, Oller/Hinofotis 1980, Scholz et al. 1980 among others) it was shown in particular that the strong versions of both hypotheses (unitary versus divisible competence) can hardly be justified and would clearly have to be rejected on the basis of the data available. It should be remembered here that according to the multidimensional model of foreign language ability several independent factors were expected, one for each single cell in the component-by-skill-matrix. The strong version of the UCH on the other hand asserted that only one GLP factor should appear, explaining the whole amount of common variance among all sorts of language performances. There seems to be no substantial support for either one of these two extremes (cf Vollmer 1980; Vollmer/Sang forthcoming).

It cannot be denied, however, that in a number of studies – either in its original form or even after the reanalysis – only one single strong factor remained (cf. Oller 1976, Sang/Vollmer 1978, Steltmann 1978, or as it were, Carroll 1975) explaining anything between 76% and 55% of the common variance. In all of these cases the appearance of a strong first factor could not be easily interpreted in terms of the UCH but was rather to be labelled as a 'pseudo-general factor' for several reasons. The number and range of variables investigated was lacking in all of the studies mentioned. In other words, important aspects of language behaviour (above all: the productive side of handling a language) were not included for consideration at all (or not sufficiently, at least). Moreover, the relatively small number of variables, highly correlated with each other, also meant that there was hardly any chance in factor analysis to divide those variables up in more or less homogeneous groups indicating dimensions of language proficiency. The probability for a one-factor-solution was rather high from the very beginning, without proving very much about the structure of the variables under investigation as it is (or might be) in reality. We would only consider it to be a clear piece of evidence for the assumption of one-dimensionability, therefore, when a one-factor-solution showed up even if a large number and a broad variety of tests were included in the analysis. Yet in a case like this the probability of the appearance of more than one factor will rise again: whenever twelve or even more language variables were included (Carroll 1958, Pimsleur et al. 1962, Gardner/ Lambert 1965, Lofgren 1969, Bonheim/Kreifelts et al. 1979, Scholz et al. 1980) statistical analysis led to at least three different factors; but again, none of these structures found can be interpreted materially in terms of the strong form of the divisible competence hypothesis.

In order to arrive at a sound judgement on the dimensionality of language ability, one would have to include a variety of tests to measure productive performances, namely to assess writing and speaking skills on a somewhat discoursal and communicative level. As a guideline, one should perhaps measure the four integrated skills by at least three different methods/instruments each (combining various approaches and formats). In addition, we will have to take the necessary precautions to ensure that our samples are more or less homogeneous as to the range of previous experience and exposure, because heterogeneity of a population might very well lead to the appearance of an artificially strong first factor (without having a substantial meaning in terms of a structural hypothesis). In this connection it is also very important to make comparative studies between second language acquisition in a natural versus formal setting. It might be especially interesting here and worthwhile to find out how far the distinction between 'creative competence' and 'reproductive competence' (Felix 1977) seems to hold empirically.

As to our own (deficient) data (cf. Sang/Vollmer 1978, 1980) I (nevertheless) undertook a follow-up study trying to test alternative explanations for the appearance of such unexpectedly high inter-correlations between our six variables (Pronunciation, Spelling, Vocabulary, Grammar, Reading and Listening Comprehension). The rationale behind this procedure was finding out whether or not the correlations could have been 'produced' at least in part or even substantially by the influence of a third, intervening variable like motivation of the learner. Four groups of variables were investigated in relation to the proficiency measures:

school setting, curriculum, teaching methods
(complex versus simple skill approach)
aspects of learner motivation (including achievement motivation and attitude towards school, preference of subjects and interest in learning the foreign language)
intelligence
achievement in the first language (German as mother tongue).

The results of this study cannot be reported here in any detail: the main findings, however, may be mentioned (cf. Vollmer 1980):

a The first three groups of variables hardly show any effect on the correlations between the language tests: neither is the amount of common variance explained by the first factor reduced in any significant way nor is the factorial structure itself affected at all.

b As soon as the influence of the first language (measured by an array of 11 different tests including knowledge of words and grammatical forms as well as discourse analysis) on the performance in the second language was controlled statistically, the average correlation between any two subtests in L2 went down from .45 to .28. At some time two factors emerged instead of one: the first one being interpretable as a dimension of complex skills in understanding (38.8%), the other one being associated with simple and basic knowledge in L2 and especially with 'Pronunciation' (16.9%). These results basically indicate that the linguistic and cognitive ability or abilities already built up in acquiring German as L1 heavily influence (but do *not* determine!) the learning of L2, namely the proficiency profile in English as a foreign language. Once the influence of L1 is controlled for, we are left with two factors to be interpreted as specific competencies having genuinely to do with L2. And these competencies seem to be more or less independent of one another (on the basis of a varimax rotated factor solution). A General Language Proficiency across L1 and L2 does not seem to exist.

Neither the quality of the tests used nor the data itself allow any stronger argument to develop at the moment. To be more explicit: The test results of L1 and of L2 have yet to be factor-analysed together. In addition, the two factors gained should not be mistaken as psychologically real abilities, but should be taken as a convenience, as handy constructs so far. What I simply wanted to demonstrate is the direction in which future research might advance. In this regard, I strongly agree with what was said in the Symposium discussion.

In view of all the empirical data we have so far, I have come to the conclusion that in all probability there is no such thing as a 'Unitary Competence' across languages. It also does not make sense to postulate the existence of two separate factors in the human mind, one GLP factor in the first language and one in the second. At least for the second language the UCH can hardly be upheld in the light of all the counter-evidence mentioned earlier.

Consequently, it might be worthwhile to consider other versions of unidimensional and multidimensional models which are less strong, which would have

the advantage of being more plausible (at least from the data side) and thus being more acceptable (even as competing hypotheses).

The development of weaker forms of theoretical assumptions (as introduced by Cummins (1979), for example, in the case of Oller's UCH) seems promising and helpful in the attempt to give our research efforts a more productive focus.

It will also be necessary in the future to test another alternative: hierarchical models of foreign language abilities. These models might possibly offer a better explanation for the different sets of data given and might describe the structure of foreign language competence more adequately.

Methodological and theoretical advances

It might very well be that one of the greatest shortcomings in the analysis of linguistic and communicative competence is to be seen in the inappropriateness of the procedures applied in finding out about its structure. Too much concentration on factor analytic models is a case in point here. Only recently Upshur and Homburg (forthcoming) have demonstrated quite convincingly that the UCH is but one possible causal model among others and that this one-factor model is not at all the best fit to the data they investigated. In particular, they show how the choice of method is determined by the underlying theoretical model which is assumed to hold true.

In our own work (cf. Sang/Vollmer 1980; Vollmer/Sang forthcoming) we have been able to prove that the principal component data reduction procedures 'produce' quite different structural results than would be the case in aiming at principal factor solutions. This difference in method can lead to strikingly divergent interpretations of one and the same set of data as exemplified by the reanalysis of Oller/Hinofotis (1980) or Scholz et al. (1980): in each particular case it meant interpreting the data alternatively either **for** or **against** the UCH, depending on the analytical producedure chosen and the number of factors generated thereby. I therefore cannot agree with Davies, who considers this choice to be merely a matter of 'arithmetic elegance'. It definitely is not — at least not in the way it is sometimes handled in the literature.

But again this argument must not be exploited too far. Both types of factor analysis are generally used in a purely exploratory way and thus are good only for helping to generate hypotheses, but not for testing them. Accordingly, there have been two main objections as to the use of factor analysis: first, it produces a structure under almost any circumstances; secondly, it does not offer any criteria for determining whether the structure found is only a chance product or indeed a replicable representation of the domain under investigation. It is not at all clear, therefore, what the factors thus produced really mean, with no theory backing up any kind of interpretation. The only chance of reducing the risk that any factorial structure found is (mis)interpreted too quickly as a reflection of reality is to describe a structural hypothesis as detailed as possible **before** the analysis is begun. This chance, of course, is further narrowed down when only one factor is expected. But, independent of this expectation, the chosen type of method tends to maximise the variance among different language performances on the first factor anyway (this being true for the principal component as well as for the principal factor analysis). This is why the possibility of method-induced artefactual results cannot be ruled out in the case of a single-factor solution just as much as in the case of multiple-factor solution within classical factor analysis. In other words, the assumption of some sort of GLP factor being the simplest model under conditions given has always a fairly good chance of being verified – even if the model may not be an adequate representation of the relationship between the variables involved. These objections, however, do not hold good any more in the light of newer forms of the so-called 'confirmatory' factor analysis (used by Palmer/Bachman) which allow a statistical comparison between the predicted model and the results actually achieved. The same is true for path analysis as it is advocated by Upshur/Homburg (forthcoming).

Possibly we have been too much preoccupied with the assumption of factorial casuality, with the interpretation of factors as underlying abilities or 'latent traits' within an individual. This interpretation has been seriously questioned. Consequently, a critical reassessment of classical test theory as well as psychometric theory is under way. We are having to ask ourselves what our tests really measure.

In order to understand better this change of focus (or change of paradigm, as it were), it might help to look beyond the narrow borderlines of the language testing business. As early as 1977, R J Sternberg presented his componential approach to human intelligence. Sternberg elaborates on the severe intrinsic limitations and even misuses of factor analysis being the main tool within a differential approach to the human mind. He reminds us of an important distinction between two types of theory, the psychological theory and the mathematical one. 'The psychological theory states how the mind functions. The mathematical theory states the way a set of empirical data should look' (Sternberg 1977: 29f.). As to the limitations of factor analysis they are summarised as follows:

> 'First, the machinery of factor analysis rather than the investigator formulates and thus has control over the mathematical theory, resulting in a reduced ability of the investigator to perform theory-comparison operations. Second, solutions are indeterminate because of the infinite number of possible axis rotations. Third, factor analysis is done over items, and hence cannot be used as a means for discovering or explicating the processes that are used in solving individual items. Fourth, intelligence and abilities exist within the individual, but factor analysis is between individuals (except in rare cases)'
>
> (Sternberg 1977:36).

Sternberg continues by analysing the information-processing approach in its advantages and limitations on the other hand. He finds that it suffers from none of the intrinsic limitations of the factor-analytic method. Yet none of the three methodologies (computer simulation, subtraction method, additive-factor method) within the information-processing approach seems to '(1) provide a means for studying systematically correlates of individual differences in performance; (2) provide a common language across tasks and investigators; or (3) prevent overvaluation of task-specific components' (Sternberg 1977: 63).

The author therefore wants to synthesise an approach 'that would capitalise upon the strength of each approach, and thereby share the weaknesses of neither' (1977: 65). In his componential analysis of human intelligence the fundamental unit is the **component.**

> 'A component is an elementary information process that operates upon internal representations of objects or symbols . . . The component may translate a sensory input into a conceptual representation, transform one conceptual representation into another, or translate a conceptual representation into a motor output. Componential Analysis is therefore akin to information-processing analysis in that its elementary unit is a process rather than a construct representing a static source of individual differences' (Sternberg 1977: 65).

Factors then cannot be interpreted as 'latent traits', but rather as mathematical representations of 'reference abilities' defined as 'constellations of components that in combination form stable patterns of individual differences across tasks' (Sternberg 1977: 78). In this theoretical and methodological context a **general** component would be one which is 'mandatory in all tasks of a specified kind', whereas a **group** component is optional. 'It is used in only a subset (or group) of the tasks being considered' (1977: 319).

The implications of this componential approach for the study of foreign language proficiency/ability have only just begun to be investigated. In this situation we can only hope to progress somewhat further by combined theoretical efforts: What does it mean to know a language and to act in it from an information-processing point of view? What are the cognitive processes in understanding and producing meaningful utterances? For only when we develop a clearer picture of the individual strategies, processes, operations etc. as well as of the task-specific demands that are involved in foreign language learning and testing shall we be able to devise valid foreign language tests in the future.

Personally, I have been working on a componential model of foreign language ability during recent weeks. In comparing comprehension with production processes I am trying to show that their structural-procedural equality does not seem to be justified. Although the results of psycholinguistic research to date indeed suggest that the encoding and the decoding system in the human mind have much in common, production and comprehension as macro-activities can probably not be seen as mirror images. Attempts have been made to account for their unique characteristics by postulating specific as well as general underlying processes. As to the former, the role of inferencing as opposed to planning procedures is a case in point here. Generally speaking, the differences between knowing how to analyse input and knowing how to construct output apparently outweigh the correspondences between these two acts of discourse (for more details cf. my paper presented at the Lund AILA Congress, Vollmer 1981).

Final remarks

In my opinion, the concept of GLP, as defined by Oller, for example, has largely served its purpose. It has stimulated controversial debate and a lot of research activities and thereby provoked international communication of some width. Yet within the narrower boundaries of the problem as it was originally posed many an argument and some hard data have been put forward against the assumption of only one internal general factor of language proficiency. I cannot finish, therefore, without contradicting Davies in what he labelled the 'philosophical question' behind the GLP controversy: to him this question (general versus specific or unitary versus divisible competence) does not seem 'amenable to proof'; to me **it does**: the UCH can more or less be rejected on theoretical as well as empirical grounds (though more evidence is needed to strengthen the proof).

At the same time, the rejection of the UCH does not necessarily mean support for the multidimensional model of language ability in its strong form and traditional definition. As I tried to indicate above, the whole problem of (foreign) language proficiency (and of GLP, consequently) will have to be redefined in theoretical and methodological terms. This is not to say, of course,

that we do not need our philosophical strength and intuitive virtues any more. On the contrary, these should be applied to more promising issues on a less abstract level of investigation, eg what goes on in the learner solving a specific task, interacting with others in varying socio-linguistic contexts, trying to understand a specified topic, item, or text, reacting to it, organising his/her own thoughts and expressing these more or less adequately etc. It could very well be that in all these activities a number of processes are at work that are 'general' and 'differential' at the same time in the sense that in combination they form more or less stable patterns of individual differences across tasks and over time. Other processes may be more task-specific. A single component may enter into many different constellations of tasks and individual solutions thereof.

This **Componential view of language proficiency** needs to be elaborated upon before we can better judge its value. Certainly it is not the only view possible. A totally different approach to defining and measuring foreign language proficiency, namely **functional testing** (with its highly contextualised items, areas, and levels of performance and its notion of specific competencies for specific purposes) should likewise be followed through – for practical reasons just as much as for theoretical considerations. Both lines of development seem to be equally necessary.

BIBLIOGRAPHY

CUMMINS, J
Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. Working Papers on Bilingualism 19, 1979, 197 – 205.

OLLER, J W (ed)
Issues in language testing research. Rowley, Mass.: Newbury House (forthcoming).

OLLER, J W and HINOFOTIS, F B
Two mutually exclusive hypotheses about second language ability: factor analytic studies of a variety of language tests. In: Oller, J W and Perkins, K (eds.): Research in language testing. Rowley, Mass.: Newbury House 1980, 13 – 23.

SANG, F and VOLLMER, H J
Allgemeine Sprachfahigkeit und Fremdsprachenerwerb. Zur Struktur von Leistungsdimensionen und linguistischer Kompetenz des Fremdsprachenlerners. Berlin: Max-Planck-Institut fur Bildungsforschung 1978.

SANG, F and VOLLMER, H J
Modelle linguistischer Kompetenz und ihre empirische Fundierung. In: Grotjahn, R and Hopkins, E (eds.): Empirical research on language teaching and language acquisition. Bochum: Brockmeyer 1980, 1 – 84 (Quantitative Linguistics vol. 6).

SCHOLZ, G. et al.
Is language ability divisible or unitary? A factor analysis of twenty-two English proficiency tests. In: Oller, J W and Perkins, K (eds.): Research in language testing. Rowley, Mass.: Newbury House 1980, 24 – 33.

SPOLSKY, B
Some ethical questions about language testing. In: Klein-Braley, C and Stevenson, D K (eds.): Practice and problems in language testing I. Proceedings of the First International Language Testing Symposium, held at the Bundessprachenamt, Hurth July 29 – 31, 1979. Frankfurt, Bern: Peter Lang Verlag 1981, 5–21.

STERNBERG, R J
Intelligence, information processing, and analogical reasoning. Hillsdale, N J: Erlbaum 1977.

UPSHUR, J A and HOMBURG T J
Some relations among language tests at successive ability levels. In: Oller, J W (ed.): Issues in language testing research. Rowley, Mass.: Newbury House (forthcoming).

VOLLMER, H J
Spracherwerb und Sprachbeherrschung: Untersuchungen zur Struktur von Fremdsprachenfahigkeit. Osnabruck 1980. (To appear in Tubingen: Gunter Narr Verlag 1981).

VOLLMER, H J
Receptive versus productive competence? – Models, findings, and psycholinguistic considerations in L2 testing. In: Proceedings I: Sections and Workshops of the 6th AILA Congress. Lund, Sweden 1981.

VOLLMER, H J and SANG, F
Competing hypotheses about second language ability: a plea for caution. In: Oller, J W (ed.): Issues in language testing research. Rowley, Mass.: Newbury House (forthcoming).