Perspectives on research paradigms and validity: Tales from the Language Testing Research Colloquium

Brian K. Lynch, University of Melbourne
Liz Hamp-Lyons, Hong Kong Polytechnic University

Abstract

The educational measurement literature has been discussing new perspectives on reliability and validity for the past decade (Linn, Baker, & Dunbar, 1991; Messick, 1989, 1994; Moss, 1994, 1996; Wolf, Bixby, Glenn, & Gardener, 1991). In an attempt to describe the range of research paradigms and their associated perspectives on validity within the language testing community, we investigated the available abstracts of papers presented at the Language Testing Research Colloquium (LTRC) over the past sixteen years (Hamp-Lyons & Lynch 1998). In that study, we argued that we had found evidence that research in language testing has been dominated by a perspective on research and validity best characterized as "positivistic", with occasional evidence of alternative perspectives. Because of the limitations of our earlier approach, we conducted a follow-up study, in which (1) a selection of full papers from that dataset were collected and analyzed to compare our abstract-based impressions to the full paper versions, and (2) interviews with LTRC colleagues were conducted, on the basis of which we explored the ways language testing researchers locate themselves among possible research paradigms, and further, developed in-depth portraits of two of these colleagues.

The draft report of the follow-up study generated a wide range of reactions from both the research subjects and the journal reviewers. This paper focuses on the interviews we conducted, and also explores the nature of some of the feedback we have received as we have sought to find ways to report our interpretations of the problems and the findings to our language testing colleagues. It recognizes a fundamental difference between language testing conceived as measurement only and a conception that includes alternative assessment as a different "culture" (Wolf et al. 1991; Birenbaum 1996) underlain by a different research paradigm. It attempts to consider how the field of language testing might be able to more fully engage

Melbourne Papers in Language Testing 1999 Volume 8.1 pp. 57-93. The Language Testing Research Centre, The University of Melbourne.

with what McNamara (1999) has called a "revolution in epistemology."

1. Introduction

This study has grown out of our earlier work describing the abstracts of papers presented at the Language Testing Research Colloquium from its beginning in 1979 through the 1994 annual meeting (Hamp-Lyons & Lynch 1998), and out of a long, tortuous and ultimately (to us) fascinating set of interactions. The first group of interactions was with some of the language testers whose papers we had selected for our initial study, and whom we had approached to dialog in more detail with us about how they viewed their language testing work. Like our selection of complete papers from the abstracts in the initial study, although the selection of people to interview was somewhat arbitrarily based on our knowledge of them as people with identifiable roles/positions within the international language testing community, we attempted to select a group of individuals that would represent more-or-less equally the three research paradigms we were using as our initial framework (see below). In seeking to take their responses into account, we found ourselves on constantly shifting ground, for people are not static, and the field of language testing is certainly in a period of fast development. We found also great difficulty in separating out, clearly enough to convince the participants, our analysis of a specific paper/text of theirs written in a particular (and sometimes quite distant) historical period from their own perception of where they stand now in their intellectual development and relationship to the language testing community. We were forcefully reminded of the accuracy of the work of Becher (1981), who argued that academics define themselves and their intellectual contribution in large measure by the company they keep.

The second group of interactions was with the four reviews of a draft of the paper that we had submitted to *Language Testing*[1], which the editors had kindly provided. Our first reactions to the reviews were that some had gone beyond the bounds of professional courtesy in expressing strong objections to certain parts of our paper. Later we were able to appreciate the care some reviewers had taken to give

[1] A brief version of the paper originally sent to *Language Testing* was presented at the Language Testing Research Colloquium in 1996.

specific instances and explanations of what troubled them and why. Our third stage of reaction was to see how interesting the whole "story" of this project is, and to hope that some day we may be able to tell it in full as a case study in the sociology of academic research and publication. But the paper presented here is not that case study. Rather, it is a greatly reduced and more focussed revision of the initial submission. It also adds some reflection on the process of attempting to integrate our own understandings with a range of other researcher perspectives and reviewer reactions in order to strengthen our interpretation of the research paradigms that underlie current work in the field of language testing.

Although this study focuses on research paradigms, it has implications for discussions of test validity as well. In pursuing our line of inquiry, we have come to the realization that there is a fundamental difference between viewing "language testing" as a measurement-only enterprise, and viewing it as inclusive of forms of assessment that are non-quantitative and based on alternative research paradigms. That is, rather than seeing "alternative assessment" as being more properly labelled "alternatives in assessment", as espoused by Brown and Hudson (1998), we would argue for alternative assessment as a different "culture" (Wolf et al. 1991; Birenbaum 1996), with a different underlying research paradigm, and we would suggest that traditional validity frameworks, derived from positivist research paradigms (see below), will not be appropriate for judging alternative assessment practices. The notions of alternative assessment are, of course, primarily developed within the educational measurement literature in relation to ongoing debates about school reform (e.g., Terwilliger 1997, 1998; Newmann 1998; Wiggins 1998), but we find them useful for considering the impact of research paradigms on validity perspectives within language testing as well.

2. The background to this study

In a previous paper (Hamp-Lyons & Lynch 1998), we investigated the question of where the language testing research community stands in relation to alternative perspectives on validity. That study used the following framework (taken from Guba 1990) to investigate abstracts written for the Language Testing Research Colloquium (LTRC) between 1979 and 1994 in terms of the assumptions underlying the research.

Ontological level:	What do we think we can "know"? What is "reality"?
Epistemological level:	What is our relationship to the thing we are trying to "know"?
Methodological level:	How do we go about our pursuit of knowledge?

The framework above guided our categorization of the abstracts in the original study into one of three possibilities: a strongly positivist, strongly alternative,[2] or ambiguous/mixed approach to research.

Evidence of a positivist approach consisted, at the ontological level, of a realist perspective, or an indication that the researcher assumes that the object of our inquiry really exists, 'out there' in the world. An assumption that the object of inquiry is governed by immutable laws and mechanisms that are essentially independent of who, when, and how it is being examined, was also considered as evidence for this paradigm. We also decided to treat the modified, or postpositivist perspective (see Phillips 1990) which maintains that, although the object of our inquiry exists outside and independent of the human mind, it cannot be perceived with total accuracy by our observations, as within the positivist category.[3] Evidence of an alternative approach consisted, at the ontological level, of a belief that realities are multiple, dependent upon particular historical and cultural contexts as constructed in the minds of people (note that at the level of what there is to know, the relativist concept of "multiple realities" is

---

[2] "Alternative" will be used here to refer to the range of paradigms that differ fundamentally, along the dimensions we present here, from the positivist/postpositivist paradigm. This includes constructivist, feminist, criticalist, and interpretivist paradigms (see Guba 1994; Denzin & Lincoln 1994).

[3] This view has also been referred to as a critical realist ontology, and forms part of Cook and Campbell's (1979) validity typology, and is referred to by Messick (1989, p. 29-30) as the "constructivist-realist view." However, following Guba and Lincoln (1989) and Lincoln (1990), we use the term constructivist to refer to a paradigm which is distinctly alternative to positivist or postpositivist approaches.

different from the postpositivist concept of "multiple strategies" for gathering and analyzing data).

At the epistemological level, evidence for a positivist paradigm included an objectivist stance toward research. Such a stance demands that we remove our influence from the research setting, distance ourselves from the object of inquiry, in order to make an accurate correspondence between our observations and this reality. Because the modified objectivist perspective, while allowing that such objectivity is nearly impossible to achieve, retains the notion as an ideal to regulate research, it was also considered a positivist approach. In contrast, evidence for an alternative paradigm included a subjectivist stance toward research. This stance assumes that our attempts to "know" things are inherently and unavoidably subjective, that reality is dependent upon, rather than independent of, our inquiry, and that as a consequence, facts cannot be established as aspects of knowledge that are independent of values. As a consequence of the relativist ontology, it also assumes that there is no independent foundation from which to judge knowledge claims.

At the methodological level, evidence for the positivist paradigm came from the use of controlled variables and manipulative designs of the empirical experimental tradition, including the modified position in which multiple strategies for gathering and analyzing data (including qualitative data) are used within an experimental or quasi-experimental framework. Evidence for the alternative paradigm, on the other hand, came from the use of a non-experimental, non-manipulative set of research procedures. These procedures involve the researcher forming interpretations, or constructions, from a close understanding of the data (observation notes, interview recordings, etc.) through cyclic processes (the potentially never-ending "hermeneutic circle") of data interrogation and analysis, interpreting these constructions and then refining and forming new constructions.

Our first study (Hamp-Lyons & Lynch 1998) had led us to multiple passes through the LTRC abstracts, including their categorization by paradigm, examination of the language used to discuss validity, and examination of whose "voices" were represented in the research, and we had concluded that the work presented at the LTRC was still mainly positivist. This was probably not surprising since the LTRC community is primarily concerned with measurement, with making quantitative assessment distinctions among persons. However, the

study did suggest that the language testing community's ways of discussing measurement and issues of validity had expanded over the years, with occasional examples of alternative perspectives. But the conclusions of that initial study were tentative, for several reasons. First, we had only examined abstracts, not complete research papers. Second, we were very aware of the degree to which the study had been carried out from a mostly positivist perspective. We had approached the abstracts with *a priori* categories, and we had attempted to reduce the complexity of the potential perspectives on validity into those categories. We had also analyzed the data and reported the findings as if we inhabited some detached, neutral position with regard to the perspectives on validity that we were investigating. Although the study had also pursued an analysis of themes, of ways of referring to validity, and of the 'voices' represented in the research, analysis types which were a departure from the strictly *a priori* paradigm categories, we decided that a more interpretivist or constructivist approach--that is, one that would represent the alternative paradigm(s)--to this investigation would be worthwhile. We knew of no work that had attempted something similar to this approach; we were interested in thinking about how to search for answers while using alternative paradigm strategies as much as we were interested in finding the (an) answer. Lacking a methodological road map, we planned a number of moves toward a more interpretive analysis of complete papers, retaining the *a priori* categories from our earlier study as a transition from the primarily positivist approach to one that used an alternative research paradigm.

3. Ways of looking closer

In the present study, we first identified a selection of full papers from each of our three categories--positivist, alternative, and ambiguous/mixed--of research paradigm orientation, aiming to include the broadest sampling of perspectives on validity possible but at the same time attempting to capture the 'flavour' of the LTRC core community, that is, those who make attendance at LTRC one of their chief professional activities of every year. We used the same basis for category assignment (detailed above) we had used in the initial study to categorize the full papers, as a check on the accuracy of our original judgements. We can think of this first analysis-type as a positivist way of looking at the data (see Appendix A for complete list of paper categorization). In our second "way of looking", we followed up our earlier exploration of "whose voices" (Hamp-Lyons & Lynch

1998) were represented in the research. We had argued in that study that one of the indicators of an alternative paradigm in research is that voices other than the researcher's are heard in the study. This inclusion of other voices acknowledges that there are other stakeholders in testing situations, and that these stakeholders 'know' things that have meaning for the study. We also felt that the examination of voices itself comes closer to an alternative approach to the data than to a positivist approach. We read the full papers to find out whether the promise in the abstract of various voices was followed through in the full text.

The third "way of looking" that we experimented with was a Toulmin analysis of the full papers in this dataset. This analysis (Toulmin 1977) looks at a text's logical structure to identify three key components of an argument: claim, data, and warrant. A claim is a statement about an entity or the relationships among entities; data are what provide the basis for making the claim; and the warrant is the authority to which the author appeals as justifying her or his right to make the claim or interpret the data in this way. A Toulmin analysis, we felt, would enable us to look more closely at fairly predictable textual elements and identify the kinds of assumptions and appeals researchers were making. By placing these assumptions and appeals into one or other of our categories, it would be possible first, to characterize the bases of arguments in the positivist tradition versus the interpretive tradition; and second, to attempt to disambiguate the papers in the "ambiguous" category by the same approach.

Our fourth "way of looking" was to select some individual researchers who had papers in one or more of our categories, to be interviewed concerning their perceptions and understandings of research paradigms and validity (see Appendix B for list of interviewees.) The purpose of the interviews was to obtain additional evidence for or against categorizing their selected papers by any particular research paradigm, as well as to gain a more in-depth sense of the individuals' own perspectives on those characterizations. Finally, we hoped that the interviews would generate for us some new ways of understanding and categorizing the LTRC community's perspectives on validity.

In exploring these four "ways of looking" at the LTRC community from the perspective of research paradigms, we moved back and forth within and between components of the data, the emerging

interpretations, and the various methods of analysis in an iterative fashion. For example, the Toulmin analyses led us to re-check some of our category characterizations; the voices analysis and the Toulmin analysis usually, but not always, provided complementary information; as we engaged in the interviews, sometimes our categorization of the paper(s) of the interviewee was altered because we were helped to see their work more clearly. The four "ways of looking" briefly described above were, we thought, four points along a possible continuum of research strategies that ranges from positivist to alternative. We hoped that these four taken together would both enable us to see our own community more clearly, and enable us to learn more about how to study paradigm issues from different paradigmatic perspectives.

Following on from the individual interviews, we felt the need to get still closer to the question of what characterizes work in language testing as positivist, alternative or something in between, and also to address in more depth the interesting question of whether language testing researchers see themselves as inhabiting one school of thought/research or another and, if they do, whether their view of themselves matches the view of their work emerging from our analyses. In order to do this, we decided to develop detailed portraits of two of the twelve researchers we had already interviewed, two firmly established as members of the LTRC and language testing communities. In essence, the two portraits represent a culmination of this iterative probing toward an understanding of the research paradigms that drive the language testing community. For reasons that may become clearer as the reader reads on, the components of our research study that we have chosen to focus on in this paper are the interviews and the in-depth portraits of two interviewees.

4. Focus on the interviews

Our interviews were actually conducted in two phases. In the first, we contacted the interviewees as we were developing the analyses presented above. We did not tell the individuals how we had categorized their papers, or even which papers we had selected, at that point. Instead we attempted to get some idea of how they would categorize themselves, using the descriptions of the research paradigms in this analysis, and how they would characterize their perspective on validity. In two cases, the first round interview was never concluded. For one of these, we were able to contact a

replacement person who interviewed with us during the second round only. (See Appendix C for interview guide questions.)

Many of those interviewed attended the presentation of our preliminary findings at LTRC 18, in Tampere, Finland, 1996. The reactions of some of those people showed us the need to obey the imperative of our chosen paradigm and seek more feedback and reflection on our initial interpretations and categorizations from all of the interviewees, and let them express their concerns and reservations directly in our paper. We returned to those interviewed with our interpretations of their papers and of what they had told us in the first round of interviews concerning research paradigms and validity. Depending on their stated preference, we sent them either the part of our draft paper that referred to their work and to our interview with them, or the whole first draft of the paper that resulted from these four "ways of looking", and invited them to engage in a further dialogue with us over e-mail. A number of them accepted, pointing out aspects that they saw as controversial in our treatment of their work and words.

In the following section we look at the interviews. First indicating how we characterized the paper chosen for inclusion in this study according to the paradigm categories we had identified (using the first three "ways of looking" described briefly above), we then quote from the interviews, attempting to portray each individual's perspective on validity, and perception of his/her research paradigm, as expressed in the first round of interviews. We explained to the participants, and stress here, that we did not claim, nor do we believe, that these categories necessarily reflect either the sum total of each individual's approach to research, or the individual's expressed view of their own perspectives on research paradigms and/or validity. Indeed, one thing above all else has become clear in our attempts to understand the LTRC community in terms of its relationship to research paradigms: most, if not all, researchers in our field resist being categorized or labeled as belonging to one paradigm or the other. A further complication was the fact that our methodology did not succeed in distinguishing between a 'snapshot' (or perhaps 'video clip' would in this case be a better metaphor) view and the whole body of each researcher's work, including the paradigmatic position, stable or changing, that lies behind it.

Because a number of our interviewees raised concerns about our characterization of the part of their work included in our analysis, we conducted a second round of interviews in order to clarify our emerging interpretations and to allow the participants the opportunity to elaborate or challenge those preliminary understandings. This difference can be quite significant, especially if the paper we were using in our study was from some years earlier; most language testers (like most researchers in any field) are developing their understanding and position over time, and may not accept a characterization now that might have been accurate in the past. In the following sections, we have brought together our initial characterizations, quotes from the participants' first interviews, and in most cases quotes from the follow-up interviews. In keeping with the alternative paradigm we are exploring, we use the voices of the researchers themselves to uncover the complexities of any attempt to characterize individual members of the LTRC, or the LTRC as a discourse/research community. All of the interviews we conducted informed the conclusions of this study, and we quote only from those exchanges for which we received permission from the participants. In the interest of keeping the length of this article reasonable, we have selected six interviews (followed by the two interviews developed into portraits) to represent the range of perspectives discovered in the original set of twelve.

Fred Davidson

From the fact that one of us regularly collaborates with Davidson on research projects, we know that he combines a solid grounding and training in positivist methodology and research design with a concern for issues of educational reform and an openness to new perspectives on validity. We had categorized one paper of Davidson's as positivist, and another (co-authored) as ambiguous. The former paper (Davidson 1988) focussed on the use of factor analysis to determine the dimensionality of language tests. The methodology itself presumes a realist ontology and an objectivist epistemology, although there was an appeal to the use of other analyses in future research on the topic (e.g., "more extensive linguistic analyses... and retrospective data."--p. 67). However, in the latter, co-authored papers (Davidson et al. 1994; Lynch and Davidson 1994), although there were indications of the same realist, objectivist stance, there were also appeals to a more alternative paradigm approach--e.g., giving "...priority to teachers' knowledge and experienced over item

statistics when deciding on the value of test items." (Lynch & Davidson, p. 732)

In his interview, Davidson said:

> *[Scholarship is valid if] it is accepted within the present arc of development of the field to which it speaks. If such work is rejected, or if it is accepted after a long and protracted negotiation, then the validity is less sound... I perceive that one validates tests nowadays primarily by argument, much as a lawyer does (an old analogy). [email interview, 6/22/96]*

This notion of validation by argument is much like Moss's (1996) call for "enlarging the conversation" on validity. Rather than appealing to some neutral foundation for judging the validity of knowledge claims and test interpretations, Davidson is signaling the need for a negotiation across multiple perspectives.

Dan Douglas

We had initially placed Douglas' abstract (1989; co-written with R. Fagundes) in the ambiguous category, but were unable to verify this due to the unavailability of the complete paper. Nevertheless, we decided to pursue his perceptions of the research paradigm underlying his work via the interview.

Talking about his work in general, Douglas said: "The research paradigm that I've tried to work within since around 1983 or 1984 is 'grounded ethnography'--I want to understand communication from the point of view of the participants in communicative events." [e-mail interview 6/27/96]. In another email exchange, Douglas elaborated his sense of this paradigm and his critique of the traditional, positivist approach:

> *We assume, without thinking about it very much, that if we crank up the juggernaut of the scientific method and aim it at the research problem, the outcome will be a successful and valid response to the problem. Moreover, we subject our research to our peers in the field, and this also produces validity evidence - if our research gets published in reviewed journals, etc., then we assume it to be valid (notwithstanding subsequent critiques...), though perhaps this is more akin to 'internal reliability' measurement than to validity (pre-*

*Messick)... Anyway, that's the romantic response to your question (not a romantic view of validation, but a romantic response to your question, please note...). A more classical, analytical response might go something like this: I could validate my research in much the same way we validate tests: by providing evidence in such areas as fairness, cognitive complexity, consequences, content quality and coverage, meaningfulness, generalizability and cost effectiveness (Linn, Baker, & Dunbar 1992-sic). I *think* all of these could be applied to research as well as to performance testing, as L, B & D intended...As I say, I take a more romantic, touchy-feely approach to validation in research - if you follow the procedure, you'll do valid research - 'evidence' in the classical sense of providing validity evidence in testing, just isn't an issue." [e-mail interview 7/1/96]*

Although we were unable to follow up on these initial interviews, much of what Douglas had to say suggested a more alternative stance than we had initially ascribed to him. Despite the qualification of his response as "romantic", it suggests a questioning of the positivist notion that the "scientific juggernaut" gives us a foundation (realist) against which to judge all knowledge claims. The appeal to Linn, Baker and Dunbar (1991) may only provide a set of validity criteria that are essentially parallel to those of the positivist paradigm, but Douglas seems to be characterizing a paradigm for his research that accommodates evidence beyond that normally considered as appropriate within positivism.

Tim McNamara

We had categorized McNamara's paper (1994/1995) as positivist because of its focus on formal hypothesis testing within an essentially traditional, scientific approach to model building. However, we noted an appeal to other methodological approaches than those traditionally used in language test validation, specifically ethnomethodology and conversation analysis. This suggested that McNamara was looking in directions that might include our present characterization of the alternative paradigm(s)—specifically, that research which sees conversation as something other than an objective reality whose rules are independent of our attempts to construct an understanding of them.

In his first interview McNamara said:

> *I find a logical extension of Popper's position more powerful [than alternative paradigms such as Habermas's critical theory]...I'm skeptical of positions that involve a commitment to values... I'm allergic to irrationality... Societies are going to make decisions--we can either bail out or work and make a marginal difference between gross unfairness and the limits of human judgment...[as opposed to 'objectivity'] I believe in shallower and deeper understanding... I have a concept of 'profound insight' [does not equal 'the truth'] which is the result of having considered more positions and reconciling more ideas. [in person interview, 7/3/96]*

We judged these statements to be mostly indicative of a positivist perspective. However, in subsequent conversations it became obvious that McNamara felt that our interpretation did not capture the complexity of his approach to research or the fact that he was well acquainted with the alternative research paradigm, and had embraced some of its ideas. He also disagreed with some of our interpretations of other people's research perspectives, and observed in the discussion of our LTRC 18 (1996) paper presentation that our own research into the topic of perspectives on research paradigms and validity had remained more positivist (putting people and papers into categories) than alternative. This thought-provoking comment in fact led us to much of the reflection we have since done, some of which is represented in this published version, and we are deeply grateful to McNamara and our other 'critical friends' for their insights.

In subsequent interview exchanges with him, we discovered that McNamara's perspective on research and validity has evolved over the years since the paper we had examined. At present, he would consider himself to be using alternative perspectives, whilst retaining a healthy skepticism for any one paradigm as the answer to questions of validity. His thorough understanding of the potential role for alternative research paradigms was clearly communicated in a recent LTRC paper (McNamara 1999).

Bonny Norton

Like our investigation of Douglas, our initial understanding of Norton's abstract (Peirce (Norton) & Troy 1990) as representing an alternative research paradigm could not be checked due to the unavailability of that paper. A co-authored paper that she referred us

to (Peirce (Norton), Swain, & Hart 1993) on the same topic proved to be a more positivist approach, looking at 500 students instead of the original four-person case study and using quantitative data analysis only. However, Norton did suggest other more recent work that tended to substantiate our initial sense of an alternative paradigm underlying her research (Norton 1997).

Our characterization of Norton's work as alternative was borne out by our interviews with her.

> *I would have to say that the theory I am drawn to is theory that grapples with questions of social justice. I guess this is the vision that guides all my research. How can what we do make the world more just and compassionate? And to do this, we have to recognize that inequities exist in the first place, and that we have to be prepared to address questions of power. [email interview, 17/6/96]*

This alternative, critical theory position was borne out in later elements of Norton's exchange with us.

For example, in subsequent interview exchanges Norton noted that:

> *...'alternative assessment' may have less to do with the particular construction of a test, than with the way the test is interpreted and used. Tests that appear 'alternative' can still be interpreted and used in problematic ways, while 'traditional' tests may strive to be accountable to test takers. The challenge for testers, whether alternative or positivist, is to determine to what extent the unavoidable unequal relationship of power between test takers and test makers compromises the validity of a given test. [email interview, 6/30/98]*

There is no suggestion here that alternative assessment is in some way inherently more 'good' or 'just': traditional language testing research has also taken accountability as a presupposition. However, a characteristic of alternative paradigms in research and validity theory is that a test developer/testing body's perspective on accountability is met with and added to by the judgements and views of a wider range of stakeholders on what is 'good' and 'just.'

Carolyn Turner (second round interview only)

We had characterized Turner's paper (1989) as positivist because of its consistent reference to formal hypothesis testing within a traditional scientific methodology: "Having to specify the causal model(s) beforehand forced the researcher to explain how the measures were inter-related and selected for the specific purpose of operationalizing theoretical constructs." (p. 195)

However, in her interview Turner said:

> *At this point in time, I do not see my research situated in any particular paradigm. My original training may have been in the positivist tradition, but time and experiences have quickly removed me from any one fixed group of principles. My research is dictated by the context in which my study is situated... meaning the questions, participants, audience, etc. I believe there are truths (i.e., reality) out there, but we cannot observe them directly as such, due to all the specific factors always contributing to any one context. We can observe and document instances of truth, but they will not be complete or comprehensive... even when I venture into alternative/qualitative inquiry, certain aspects of that paradigm (i.e., "that paradigm" referring to the "classical/quantitative tradition") especially concerning procedure remain with me (e.g., the need for control of some of the context, organization and consistency in collecting data, etc.).*
>
> *...I would say for research findings to be valid, they should represent phenomena to which they refer in the most authentic manner possible (context)... I feel that researchers just have to be up front about the context of our research, and interpret it accordingly as we put forth the evidence for validity. The types of evidence we put forth (now that the view of validity has been expanded) once again are dictated by the context, questions, and audience in relation to the phenomena we are wanting to explain and understand.*
>
> *... I do not consider that there is a 'best' paradigm, nor that we have to subscribe to one, but that more realistically there are several paradigms and combinations of paradigms out there for us to tap into as we pursue our research in various contexts... We MUST address validity. However, the procedures we use need to be compatible with the type of research we are doing, and do need to be*

meaningful/appropriate within the community we are in. [email interview, 12/4/98, with clarification, 5/23/99]

This was the last interview to be conducted, initiated in June 1998 and completed in December 1998 (clarification in April 1999) and, thus, represents the most current thinking of any of the individuals in this study. Although the paper selected for our analysis remains a solid example of the positivist paradigm, it is clear that Turner's perspective on research and validity has evolved and that she makes use of an alternative paradigm, especially at the methodological level. Her ontological view (that "reality is out there" even if we can not observe it directly or completely) remains essentially positivist, but there is also the recognition that validity will be approached differently, will require different forms of evidence, depending on the research paradigm being used.

Caroline Clapham

We had classified Clapham's paper (1995) as ambiguous because it contained elements that seemed to include voices other than the traditional ones of testers and language experts--i.e., the test takers themselves. Although this particular study did focus on academic subject specialists, we saw it as an attempt to look at the object of research--ESP reading ability--as a reality to be examined from different perspectives, rather than existing independently of those interpretations. In other ways, however, the research seemed grounded in positivist notions of objectivity and statistical analysis as the key to valid inquiry.

In her interview Clapham said: "I was not aware of any research paradigm in which my research was situated, but I was aware that I was investigating an aspect of test validity, that I was basing my research on previous findings, and that I was, to some extent, using tried and tested methods of test design and analysis." [email interview, 26/6/96] She later added:

First and foremost my research methods are based firmly on the Popperian belief that knowledge is provisional. One can never <u>prove</u> anything for certain, but one can advance knowledge by <u>disproving</u> propositions. To be useful, therefore, a proposition or theory has to be verifiable. [email interview 7/2/96; original emphasis]

This emphasis on verifiability (there is a truth 'out there') indicates to us a fairly strong indication of positivist (as defined above) assumptions in her approach to research and validity.

In subsequent interview exchanges, we discovered that Clapham's perspective has changed substantially since the time of our initial investigations and the paper we selected. She commented that we had happened:

> *...(perhaps intentionally) to have chosen a moment when language testing research is in a state of flux, and I think that many erstwhile 'positivists' are becoming less certain of what they formerly saw as truths. There seems to be a general feeling in applied linguistics that experiments can be so riddled with unwanted variables that other methods of investigation are often more informative and it is probably the case that fewer researchers in the field are trying to copy the empirical methods used in the hard sciences. Worries are now expressed about the concept of reliability and the value of reliability indices, and qualitative methods of research are becoming more widespread. [email; 9/13/98].*

We feel that Clapham is here making a sound judgement about where language testing is going.

Like others we interviewed, Clapham was uncomfortable with being categorized into a particular paradigm. Even though we had tried to stress that the paradigm categorizations were meant to refer to particular instances of work--the papers we had selected--and not to the individual per se, after reading a draft of our paper Clapham warned:

> *...you should be careful about labeling researchers as belonging to any one paradigm. Especially as I get the distinct feeling in your paper, though I'm not sure if the feeling comes from you or from me, that you consider that those who indulge in quantitative research tend to be positivist, and that to be a positivist is BAD." [email; 9/13/98]*

We have given this comment much reflection, and discuss it later (also see discussion of Norton).

Going beyond interviews: Two "portraits" from our analysis

Our original plan had been to extend the interviews into full-fledged portraits of each of these researchers, a project in which we were defeated by factors of time—ours and our colleagues'. But we were able to develop two such portraits, and at the 1996 LTRC meeting in Finland, we presented two portraits which detailed papers and arrived at a characterization of dominant research paradigm and validity perspective for each (Lynch & Hamp-Lyons 1996). Here, we will provide a summary of those portraits, omitting the detail on individual papers. Our aim in doing this is to allow readers to judge for themselves whether such a richer, thicker "look" might have provided the more convincingly emic account that reviewers felt was needed; whether, as one reviewer said, "categorization of researchers in the way carried out by the authors" would have been appropriate if "an examination of a substantial body of work of the researchers" had been examined.

Elana Shohamy

Shohamy was one of the few people we identified as having papers within both the positivist and the alternative paradigm, and one of the first questions we wanted her to answer was whether she saw herself as being "in" one paradigm or the other. Our analysis of her papers (Shohamy 1984, 1993a, 1993b) suggested that she may have been moving from an identification with the positivist perspective to a closer affiliation with the alternative paradigm.

Although we saw clear evidence of research "paradigms" in Shohamy's work, in her interview she adamantly disavowed allegiance to any paradigm and, further, questioned the value of "paradigms" in general:

> *I don't see myself as situated in a specific paradigm and do not feel that any research or any researcher should be situated in 'a paradigm' I think that research (a procedure of getting to the truths) is too complex and important an issue to be categorized in a given perspective, and to be dictated by certain rules. In that way I am certainly in the view of Moss and other people who are talking about interpretive research, that is, gaining insight and understanding of phenomena, but unlike Moss I don't believe that gaining insight or interpretation requires anchoring oneself in a set or principles or*

> *paradigms, rather the truth can be arrived at through a variety of avenues, and variety of procedures, and I don't even like the term paradigm. [email interview, 6/21/96]*

Despite her refusal of a paradigm label, there is much in Shohamy's work and her reflections on that work that tends to articulate a truly alternative perspective on the validity of assessment and assessment research. For example, the "goals" of Shohamy's 1990 paper seem to include changing the "power dynamic" in the testing context she was investigating, or at least taking it into consideration as a part of the research. Also, in Shohamy's 1992 discussion of "a broader notion of construct validity", there is the notion that the responsibility of the tester does not end with acceptable reliability coefficients, as well as the Foucaultian notion of tests as ultimate form of social control.

Another area that supports the notion of Shohamy's alternative perspective is the "voices" that are represented in the research. While in her 1983 paper, only test developers and test researchers are represented, in the 1990 and 1992 papers, we "hear" teachers and other stakeholders such as school principals and government bureaucrats. If these voices, and also those of students and test takers, are not actually "heard" in this research, they are at least considered and their opinions (at least teachers and bureaucrats) are reported and reflected upon, including their representation in the news media. A good example of Shohamy's consideration of other voices can be found in the 1992 paper (as published in 1993):

> *On the basis of interviews with ten teachers, five being teachers whose classes had failed, the most obvious impact was found to concern emotional involvement and stress... They spoke endlessly about the test--indeed, they were happy to have the opportunity to do so; they expressed anger and frustration, and were very critical of the test. (Shohamy 1993: 13)*

On the other hand, there remains a strong positivist flavor to some of her comments on the proper conduct of research, including an explicit concern for "science" and "empiricism":

> *every avenue is 'kosher' providing the research follows disciplined inquiry principles, or as we like to call them scientific principle, that it is not just based on beliefs, authority, prejudices, hunches and intuitions, but based on empirical procedures of observations,*

> *experimentations, validity principles, systematic examination of questions, collecting evidence, externalizing and opening data to a system of check and balances, reliability, etc. [email interview, 6/21/96]*

This concern for "disciplined inquiry", of course, does not distinguish it as positivist--alternative research aims to be "disciplined" as well. However, the appeal to the "scientific principle" as opposed to mere "beliefs" might indicate an underlying assumption that there is some sort of "objective" truth that can be captured with an adherence to the one proper method--the scientific method. This would indicate a fundamentally positivist (or modified positivist) position on the nature of what we are trying to know, the ontological aspect of research paradigms, with implications for how we define the appropriate relationship between ourselves as researchers and that which we are trying to know (the epistemological aspect). Does Shohamy espouse the notion of an objective "truth"? This she tends to deny:

> *Thus, I view science as interpretive, contextual, dynamic and fluid, socially and culturally constructed and represented by various means and forms -numerical, verbal, qualitative and quantitative. It continues to seek the truth , but it is realized that there is no ONE truth but multiple truths, multiple knowledges, and many avenues to get there. [email; 6/21/96]*

It may be, from Shohamy's perspective, that "knowledge" (things as they "really" are) can be discovered with the proper disciplined inquiry (an objectivist epistemology), but that the nature of what we're trying to know has to be seen as having multiple manifestations in that external reality. This suggests an interesting variant to Guba's analysis of paradigm types and characteristics. He sees critical theorists, Marxists, and feminists, for example, as being positivist at the ontological level (the nature of reality), but constructivist/relativist at the epistemological level (relationship of knower to the known). That is, there is "the truth", as an entity that exists external to our attempts to know it, but our attempts are necessarily value-laden and relative to particular social, political and historical contexts. Shohamy's expressed perspective suggests a constructivist/relativist ontology, but a positivist epistemology--that there are multiple truths, but that our way of approaching those

truths requires us to dissociate ourselves from values, to be "objective."

Shohamy summed up her perspective on validity in her follow up email interview [7/23/96]:

> *I think that I am in language testing to represent the victims, I am there to protect these who get penalized by tests. The fact that I do it using scientific methods is the only way I know how to find out things. I, too, don't buy the notion that science is the only way to know, yet, given my training and my affiliation (a univ.!) this is how I do my work.*

Lyle Bachman

Bachman was not only one of the founders of the LTRC, he has been one of its most influential members. From the earliest days of LTRC, Bachman's name has been associated with statistical sophistication and the use of rigorous analytic methods to model and test language behavior. It is perhaps surprising, then, to realize that of the three Bachman papers in this data set (Bachman, Lynch, Mason 1995; Bachman et al. 1988, Anderson et al. 1991), only one was originally placed in the "positivist" paradigm and two were placed in the "ambiguous" paradigm. The fact that the positivist paper is the most recent of these three could suggest that Bachman is moving more toward a positivist paradigm. However, this is clearly an artifact of our selection process. It would have been possible to choose papers involving Bachman that would have most likely all been categorized as positivist (for example, papers presented at LTRCs with Palmer in 1981, 1982, 1983, and 1988). We chose the ambiguous papers, instead, in order to help us clarify Bachman's perspective, and our analysis revealed a complex picture.

When we interviewed Bachman, we were interested to learn whether he viewed himself as a positivist in line with our perception of his reputation, or whether his self-view was different than that: if he believes himself to be changing, in what ways and in what direction? The picture of himself that Bachman painted in response to our first question--"In what ways do you see your research as being situated in a particular research paradigm, and how would you label or describe that paradigm?"--was of someone solidly in the "ambiguous"

category; that is, across both the positivist and the alternative paradigms.

> *I see my own research as fitting squarely within the construct validation (CV) paradigm, with all that entails--logical analysis and theorizing, building descriptive or explanatory frameworks to guide research, collecting information and providing descriptions or explanations. I also think that this requires both positivistic/quantitative and hermeneutic/qualitative approaches to research. What CV entails for me, is building and investigating a theory of language assessment procedures and the performance of individuals (designers/developers, administrators, raters, takers, users) within these procedures. So the model I'm envisioning will need to address not only the more traditional issues of reliability and validity, and the effects of assessment procedures on performance, but also issues of how and in what ways the process of design and development, and the individuals who are involved in this, contribute to the performance of individuals on the assessments, and to the way the assessments are used and perceived. I thus see everything from the first realization of a need for assessment, or the first idea for an assessment task, to the logical evaluation of the potential usefulness of actual assessment tasks, to the collection of evidence, to the logical consideration of consequences and ethical values, to be fair game for construct validation. [e-mail interview 6/24/96]*

These comments are quite reminiscent of those by Shohamy we have quoted already [6/21/96 interview].

Although the tenor of Bachman's reply to our first question is one of balance across paradigms, much of his research seems to operate within a dominant positivist paradigm, and the positivist flavor came through more strongly in his reply to our second question--"How do you define and determine the validity of your research?"

> *I think I'd consider my research in terms of usefulness... I'm interested not only in conducting research that may increase our knowledge and understanding about what happens in the language assessment process... but also in doing research that addresses practical, day-to-day problems in language assessment. ... At the "pure," psycholinguistic/cognitive research end, we need to be concerned with reliability, and internal and external validity (these apply, by the way, to both the quantitative and qualitative paradigms, albeit in slightly different operationalizations). What we often*

consider only implicitly are the qualities of authenticity and interactiveness (usually lumped together under some vague notion of validity), impact (ethical values that are implicit in the research, consequences, both positive and negative, for the field) and practicality. I think it would be interesting to start to formulate a set of questions, or considerations for applied linguistics research based on these qualities. At the "applied/action" end, we often ignore reliability and validity almost entirely, typically give only lip service to authenticity and interactions, and focus on impact and practicality." [e-mail interview 6/25/96]

Bachman seems to be saying here, 'if it's useful it's valid,' yet we know him well enough to know there is more to it than that. His answer to the final question--"How does your understanding of research paradigms influence your approach to validity?"--opens this up more.

... people like Cronbach and Messick were saying essentially what I had felt the need to find--an approach to validation that was clearly theory-driven, and which could accommodate a wide range of evidential approaches. What really hooked me, I suppose, was the analogy with theory falsification, which I still firmly believe. Thus, in a nutshell, I see my approach to validation as being directly related to my understanding of research paradigms. If we can consider abilities, cognitive styles, affective and personality characteristics, etc. to be, on at least one level, constructs, then I suppose any research paradigm that investigates the nature of these, the relationships among them, and how they affect behavior or performance, can be considered an instance of construct validation. Since I'm convinced that many of these factors cannot be investigated quantitatively, then I feel we need to utilize any and all research paradigms that are appropriate to the particular question.' [e-mail interview 6/27/96]

We seem to see in Bachman's thinking an interesting dichotomy between his theorizing and his approaches to what he has called theory falsification. In his theorizing, he embraces different paradigms and is open to the presence of many voices; in his approach to research, in what he uses as data and his perspective on how to establish reliability and validity, he seems to be far more likely to privilege certain kinds of evidence. Of the three papers we looked at in this analysis, for example, only the LTRC 1990 paper contained any voices other than those of test developers and testing researchers (and we believe those voices were introduced into that paper by Andrew Cohen). In contrast, several others in our study of complete

papers, such as Shohamy, Alderson, Norton, and Cohen, included voices other than testing researchers and test developers. Ultimately, while theoretically "open" to the alternative paradigm, Bachman seems methodologically and, perhaps, epistemologically grounded in the positivist perspective.

5. What the interview process taught us

Analyzing the interview data provided perhaps the richest and most difficult challenge of all that we attempted. Conducting the interviews in two phases allowed us to check our emerging interpretations of which research paradigms and approaches to validity were present in the work of those being interviewed. This lengthened the process of the research considerably, but because it allowed us to incorporate a fuller sense of the individuals participating in the study, and make certain they had the opportunity to challenge our interpretations and offer their own, we were able to respond to some of the criticisms we had received in our preliminary presentation of the research at the LTRC 1996.

This dialogue with our study participants was an important aspect of attempting to move toward an alternative paradigm, but was time-consuming and occasionally uncomfortable (for us and our participants). For some of our participants, it also proved to be an unsuccessful methodological approach. Faced with the reactions of our participating colleagues, some of whom felt personally attacked by our descriptions, we ourselves felt that the methodology had not done what we had hoped: it had not carried an authenticity and validity that would enable our colleagues to accept its results, even if they were not especially pleased with them. But confronting that problem has been an important lesson for us. We have learned the assailability, the inherent weakness, of an alternative paradigm compared to the power of the dominant paradigm. Our efforts to construct meaning from the data inevitably meant having those interpretations laden with our own values and influenced by the alternative paradigm we were consciously using. We attempted to keep those influences transparent by building a thick, detailed account of our processes and findings; however, we found ourselves facing four reviews from *Language Testing* that contained such unhelpful comments as "... a somewhat sickly continuous inspection of one's own navel—Accept."

With the encouragement of the editors of *Language Testing* and many of our research participants, we emerged from the resulting depressions not only older and wiser, but in some ways heartened by the rethinking that the reviews, harsh though they were, engendered. In particular, we found the advice of one reviewer that we should "consider the reaction of reviewers such as this to their paper as also representative of the LTRC community, and, importantly, as gatekeepers, and should include references to our comments in their fully alternative account" very congruent with what we were trying to accomplish. The reviewers found the interviews valuable, but criticized our imposition of our perceptions on the interviewees: we have tried in this version to respond to that criticism without retreating from our original aims. They felt we needed to clarify our own position "within the inner circle": we acknowledge that some might put us there. They felt we needed to clarify that the researchers we chose to study were "not representative of language testing in general," a criticism we find difficult to judge, given the enormous variety of activity under the name of "language testing."

We have persisted with this vastly different version of our paper because the effort it has entailed to respect the responses of our participating colleagues (responses which within the alternative paradigm must by definition be treated as having claims to validity), take account of the views expressed by the *Language Testing* reviewers, and yet create a text which would meet our own expectations, has led us to lengthy, painful but ultimately educational reflections on our own research processes, values and assumptions.

6. Conclusion

We began by thinking that a deeper understanding of how a fairly small group of language testers (Language Testing Research Colloquium participants) was developing its views of validity and of approaches to language testing research could inform thinking about what 'counts' as research in language testing. We have come to see that the trouble we have had with this paper and its audiences is itself part of the paradigm question. We are therefore hesitant to make any claims or conclusions about the LTRC or about the researchers represented here. LTRC members do not see, or wish to see, their work as being situated in a particular research paradigm. Many resist the sense of being boxed into one way of doing things [Clapham, Davidson, McNamara, Norton]; others find the notion of paradigm

unclear [Clapham], unhelpful [Shohamy] or simplistic [Alderson]. Some have recognized the power of a "dominant paradigm" and argue for alternatives [Cohen], while others will argue against an exclusion of the dominant paradigm whilst remaining open to alternatives [McNamara].

The dominance of the positivist paradigm in what we 'saw', compared to the general resistance among our study participants to being labelled as "positivist" because they 'saw' themselves in other ways, may well be an artifact of using primarily published papers (this issue was raised by Alderson in a personal communication, and commented on by one of the reviewers). That is, what gets published, especially in refereed journals, may be an overly conservative indicator of existing ideas and perspectives concerning approaches to research and validity.

However, the existence of a thread, running through the interviews, in which many of the people interviewed expressed an explicit belief in falsifiability, can be seen as an example of the continuing dominance of the positivist paradigm (at both the epistemological and methodological levels).

We did see the emergence of some emic (if you will) labels within language testing for research paradigms and characterizations of validity. For example, Bachman and others [Chalhoub-Deville, Turner] explicitly linked their perspective on validity with a "construct validation paradigm." Several of those interviewed described "validity" as being what is accepted by people in the field [Davidson, Cohen, Douglas], or as the result of "convincing arguments" [Davidson, Norton]. Others acknowledge "guiding principles" that may or may not correspond to existing statements concerning research paradigms: "advancing knowledge" [Clapham, McNamara], "deep understanding" [McNamara], "fairness" [McNamara, Douglas], "reform" [Davidson], a concern for "social justice" [Norton] and a concern for "the victims of tests" [Shohamy].

It is instructive that our colleagues resist equally being labelled in <u>any</u> paradigm. We agree with Clapham that we have captured the LTRC at a crossroads in its thinking about the issues and processes of language testing. Indeed, we have found ourselves at the same crossroads. Part of the difficulty we experienced in this study may be the fact that the central activity of language testing is measurement,

an inherently positivist enterprise. If we believe that measurement (the quantification of constructs), as traditionally conceived, is the only or best way to assess all language-related activities and concepts, then an alternative research paradigm is not going to be useful or appropriate. Measurement (and testing) adopt the positivist view that language (and language use) exists independently from our attempts to understand it; that it is an objective entity that can be measured, if somewhat imprecisely at times, with the proper tools and procedures. If, on the other hand, we adopt the alternative research paradigm perspective, language becomes viewed as something that is created and exists in the act of our using, inquiring and interpreting it, not as an independent, objective entity waiting to be discovered and measured. And it is here that we see the connection between alternative research paradigms and the notion of alternative assessment. Alternative research paradigm signals a different set of assumptions about the nature of language, not the use of qualitative methodology to investigate the characteristics of measurement and testing (the procedures by which we gather systematic data for measurement). Alternative assessment is intended as an extension of language testing beyond measurement (beyond testing) to include other approaches to understanding and evaluating language ability and use, not as "alternatives in assessment" (Brown & Hudson 1998). This does not mean ascribing automatic validity to such research and assessment, nor does it imply that one paradigm is "good" and the other "bad", but it does require the recognition that traditional validity frameworks will not do justice to this work. We think this recognition should be included in a response to McNamara's (1999) call for a consideration of the "revolution in epistemology" occurring outside the field of language testing.

We hope our attempt to look inside language testing in alternative ways, and to reveal ourselves struggling to find valid ways to do that, will provide an opportunity for some language testing researchers to think about their beliefs and research processes in ways that are often not facilitated by the pressure of project deadlines, budgets, etc. We hope too that debate and self-reflection about what we do and how we do it will continue to be a thread within the language testing community.

7. References

Anderson, N., Perkins, K., Cohen, A., & Bachman, L. 1991. Construct Validation of a Reading Comprehension Test: Combining Sources of Data. *Language Testing, 6*(1), 41-67. [LTRC 12: same title]

Bachman, L. F., Lynch, B. K., & Mason, M. 1995. Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257. [LTRC 15: "Investigating Variability in Tasks and Rater Judgments in a Performance Test of Foreign Language Ability".]

Bachman, L. F., Kunnan, A., Vanniarajan, S., & Lynch, B. K. 1988. Ability and Task Analysis as a Basis for Examining Content and Construct Comparability in Two ESL Proficiency Test Batteries. *Language Testing, 5*(2),128-159. [LTRC 10: same title.]

Becher, A. 1981. Toward a definition of disciplinary cultures. *Studies in Higher Education, 6*(2), 109-122.

Birenbaum, M. 1996. Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F.J.R.C. Dochy (Eds.) *Alternatives in assessment of achievements, learning processes, and prior knowledge,* (pp. 3-29). Dodrecht, Netherlands: Kluwer Academic Publishers Group.

Brown, J. D., & Hudson, T. D. 1998. The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675.

Clapham, C. 1995. What Makes an ESP Reading Test Appropriate for Its Candidates? In: A. Cumming & R. Berwick (Eds.), *Validation in Language Testing,* pp. 171-193. Clevedon, Avon: Multilingual Matters. [LTRC 14 - same title; full study also published: 1996: The Development of IELTS: a Study of the Effect of Background Knowledge on Reading Comprehension. Cambridge: Cambridge University Press.]

Davidson, F. 1988. An Exploratory Modeling Survey of the Trait Structures of Existing Communicative Language Test Datasets.

Paper presented at LTRC 10 (also unpublished PhD dissertation, UCLA).

Davidson, F., Lynch, B., Cho, D., & Larson, S. 1994. Criterion-Referenced Language Test Development (CRLTD): An Overview. Paper presented at LTRC 16 (a modified version also published: Lynch, B. K. & Davidson, F. 1994. Criterion-referenced language test development: Linking curricula, teachers, and tests. *TESOL Quarterly, 28*(4), 727-743).

Denzin, N. K., & Lincoln, Y. S. 1994. Introduction: Entering the field of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research*, pp. 1-17. Thousand Oaks, CA: Sage.

Deville, C. W., & Chalhoub-Deville, M. 1993. Modified Scoring, Traditional Item Analysis, and Sato's Caution Index Used to Investigate the Reading Recall Protocol. *Language Testing, 10*(2), 117-132. [LTRC 15 - same title]

Fagundes, R. & Douglas, D. 1989. Strategic Competence and the SPEAK Test: An Exploration of Construct Validity. Paper presented at LTRC 11. [paper unavailable]

Guba, E. G. 1990. The Alternative Paradigm Dialog. In E. G. Guba (Ed.) *The paradigm dialog*, pp. 17-27. Newbury Park, CA: Sage.

Guba, E. G., & Lincoln, Y. S. 1989. *Fourth Generation Evaluation.* Newbury Park, CA: Sage.

Guba, E. G., & Lincoln, Y. S. 1994. Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research*, pp. 105-117. Thousand Oaks, CA: Sage.

Hamp-Lyons, L. & Lynch, B. K. 1998. Perspectives on Validity: A Historical Analysis of Language Testing Conference Abstracts. In A. J. Kunnan (Ed.), *Validation in language assessment*, pp.253-276. Mahwah, NJ: Lawrence Erlbaum.

Linn, R. L., Baker, E. L. & Dunbar, S. B. 1991. Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher, 20*(8), 5-21.

Lincoln, Y. S. (1990). The making of a constructivist: A remembrance of transformations past. In E. G. Guba (Ed), *The paradigm dialogue*, pp. 67-87. Newbury Park, CA: Sage.

Lynch, B. K., & Hamp-Lyons, L. 1996. Research paradigms and perspectives on validity. Paper presented at the 18th Annual Language Testing Research Colloquium, Tampere, Finland.

Lynch, B.K. & Davidson, F. (1994). Criterion-referenced language test development: Linking Curricula, Teachers, and Tests. *TESOL Quarterly, 28*(4), 727-743.

Messick, S. 1989. Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. 1994. The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher,* 23(2), 13-23.

McNamara, T. F. 1999. Validity in language testing: The challenge of Sam Messick's legacy. Messick Memorial Lecture, 21st Annual Language Testing Research Colloquium, Tsukuba University, Japan, 28 July 1999.

McNamara, T. F. 1994. Models of Performance in Second Language Performance Tests. Paper presented at LTRC 16 (also published: 1995. "Modelling performance: Opening Pandora's box". *Applied Linguistics, 16*(2), 159-179).

Moss, P. A. 1994. Can There Be Validity without Reliability? *Educational Researcher,* 23(2), 5-12.

Moss, P. A. 1996. Enlarging the dialogue in educational measurement: voices from interpretive research traditions. *Educational Researcher, 25* (1), 20-28.

Newmann, F. 1998. An exchange of views on "Semantics, psychometrics, and assessment reform: A close look at 'authentic' assessments". *Educational Researcher, 27*(6), 19-20.

Norton, B. 1997. Accountability in Language Assessment. In C. Clapham and D. Corson, *Language Testing and Assessment, Volume 7, Encyclopedia of Language and Education*, pp. 313 - 322. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Peirce, B. N. & Troy, K. 1990. What the Autonomous Language Learner Can Teach Us About Assessment. Paper presented at LTRC 12. [paper unavailable]

Peirce, B. N., Swain, M., & Hart, D. 1993. Self-assessment, French immersion, and locus of control. *Applied Linguistics, 14*(1), 25-42.

Phillips, D. C. 1990. Postpositivistic science: Myths and realities. In E. G. Guba (Ed.), *The paradigm dialog*, pp. 31-45. Newbury Park, CA: Sage.

Shohamy, E. 1984. Comparison of Six Methods for Testing Reading Comprehension in EFL. *Language Testing, 1*(2), 147-170. [LTRC 5 - same title]

Shohamy, E. 1993a. A Collaborative/Diagnostic Feedback Model for Testing Foreign Languages. In D. Douglas & C. Chapelle (Eds.) *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium*, pp. 185-202. Alexandria, VA: TESOL. [LTRC 12 - A diagnostic/feedback model for assessing achievements and proficiency in foreign languages]

Shohamy, E. 1993b. The Power of Tests: The Impact of Language Tests on Teaching and Learning. *NFLC Occasional Papers*, June 1993, 1-19. [LTRC 14 - The power of a test: A study on the effect of a reading comprehension test on language learning]

Spaan, M. 1993. The Effect of Prompt in Essay Examinations (in D. Douglas & C. Chapelle (Eds.) *A new decade of language testing research: Selected papers from the 1990 Language Testing Research*

Colloquium, pp. 98-122. Alexandria, VA: TESOL. [LTRC 12 - The effect of the prompt in essay examinations]

Terwilliger, J. 1997. Semantics, psychometrics, and assessment reform: A close look at 'authentic' assessments. *Educational Researcher, 26*(8), 24-27.

Terwilliger, J. 1998. Rejoinder: Response to Wiggins and Newmann. *Educational Researcher, 27*(6), 22-23.

Toulmin, S. 1977. *The uses of argument*. Oxford: Oxford University Press.

Turner, C. E. 1989. The Underlying Factor Structure of L2 Cloze Test Performance in Francophone, University-Level Students: Causal Modeling as an Approach to Construct Validation. *Language Testing, 6*(2),172-197. [LTRC 11 - same title]

Wiggins, G. 1998. An exchange of views on "Semantics, psychometrics, and assessment reform: A close look at 'authentic' assessments". *Educational Researcher, 27*(6), 20-22.

Wolf, D., Bixby, J., Glenn, J. & Gardener, H. 1991. To Use Their Minds Well: Investigating New Forms of Student Assessment. *Review of Research in Education, 17*, 31-74.

Appendix A. Paradigm Categorizations

Positivist

1. Shohamy, E. LTRC 5 1983. Comparison of Six Methods for Testing Reading Comprehension in EFL.

2. Hamp-Lyons, L., Henning, G., & De Mauro, G. LTRC 10 1988. Construction Validation of Communicative Writing Profiles.

3. Lynch, B. K., Davidson, F., & Henning, G. LTRC 10 1988. Person Dimensionality in Language Test Validation.

4. Davidson, F. LTRC 10 1988. An Exploratory Modeling Survey of the Trait Structures of Existing Communicative Language Test Datasets.

5. Turner, C. E. LTRC 11 1989. The Underlying Factor Structure of L2 Cloze Test Performance in Francophone, University-Level Students: Causal Modeling as an Approach to Construct Validation.

6. Spaan, M. LTRC 12 1990. The Effect of the Prompt in Essay Examinations.

7. Cohen, A. D. LTRC 12 1990. The Role of Instructions in Testing Summarizing Ability.

8. Bachman, L. F., Lynch, B. K., & Mason, M. LTRC 15 1993. Investigating Variability in Tasks and Rater Judgments in a Performance Test of Foreign Language Ability.

9. Deville, C. W., & Chalhoub-Deville, M. LTRC 15 1993. Modified Scoring, Traditional Item Analysis, and Sato's Caution Index Used to Investigate the Reading Recall Protocol.

10. McNamara, T. F. LTRC 16 1994. Models of Performance in Second Language Performance Tests.

Alternative

1. Madsen, H. S. LTRC 6 1984. Retrospective Student Evaluation of Testing.

2. Dickinson, L. & Haughton, G. LTRC 10 1988. Collaborative Assessment by Master's Candidates in Tutor-Based System.

3. Cohen, A. D. LTRC 11 1989. The Taking And Rating of Summary Tasks.

4. Peirce, B. N. & Troy, K. LTRC 12 1990. What the Autonomous Language Learner Can Teach Us About Assessment.

5. Shohamy, E. LTRC 12 1990. A Diagnostic/Feedback Model for Assessing Achievements and Proficiency in Foreign Languages.

6. Shohamy, E. LTRC 14 1992. The Power of a Test: A Study on the Effect of a Reading Comprehension Test on Language Learning.

7. Wall, D., & Alderson, J. C. LTRC 14 1992. Examining Washback: The Sri Lankan Impact Study.

8. Hamp-Lyons, L. LTRC 15 1993. Applying Ethical Standards to Portfolio Assessment in ESL.

9. Gordon, C. & Hanauer, D. LTRC 15 1993. Test Answers as Indicators of Mental Model Construction.

10. Milanovic, M., Saville, N., & Hong, S. S. LTRC 15 1993. A Study of the Decision-Making Behaviour of Composition.

Ambiguous

1. Clark, J. L. D. LTRC 8 1986. A Study of the Comparability of Speaking Proficiency Interview Ratings Across Three Government Language Training Agencies.

2. Alderson, J. C. LTRC 10 1988. New Procedures for Validating Proficiency Tests of ESP? Theory and Practice.

3. Bachman, L. F., Kunnan, A., Vanniarajan, S., & Lynch, B. K. LTRC 10 1988. Ability and Task Analysis as a Basis for Examining Content and Construct Comparability in Two ESL Proficiency Test Batteries.

4. Fagundes, R. & Douglas, D. LTRC 11 1989. Strategic Competence and the SPEAK Test: An Exploration of Construct Validity.

5. Anderson, N., Perkins, K., Cohen, A., & Bachman, L. LTRC 12 1990. Construct Validation of a Reading Comprehension Test: Combining Sources of Data.

6. Hale, G. A. & Courtney, R. G. LTRC 12 1990. Note Taking and TOEFL Listening Comprehension.

7. Brown, A. LTRC 14 1992. The Role of Test-Taker Feedback in the Test Development Process.

8. Clapham, C. LTRC 14 1992. What Makes an ESP Reading Test Appropriate for Its Candidates?

9. Luoma, S. LTRC 15 1993. Validating the Certificates of Foreign Language Proficiency: The Usefulness of Qualitative Validation Techniques

10. Davidson, F., Lynch, B. K., Cho, D., & Larson, S. LTRC 16 1994. Criterion-Referenced Language Test Development (CRLTD): An Overview.

Published Versions of Papers (other than those in reference list)

Alderson, J. C. 1988. New Procedures for Validating Proficiency Tests of ESP? Theory and Practice. *Language Testing, 15*(2), 220-232. [LTRC 10 - same title]

Brown, A. 1993. The Role of Test-Taker Feedback in the Test Development Process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*(3), 277-304. [LTRC 14 - The role of test-taker feedback in the test development process]

Clark, J. L. D. 1987. A Study of the Comparability of Speaking Proficiency Interview Ratings Across Three Government Language Training Agencies. In K. M. Bailey, T. L. Dale & R. T. Clifford (Eds.) *Language testing research: Selected papers from the 1986 Colloquium.* Monterey, CA: Defense Language Institute. [LTRC 8 - same title]

Cohen, A. D. 1993. The Role of Instructions in Testing Summarizing Ability. In D. Douglas & C. Chapelle (Eds.) *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium,* pp. 132-161. Alexandria, VA: TESOL. [LTRC 12 - same title]

Cohen, A. D. 1994. English for Academic Purposes in Brazil: The use of Summary Tasks. In C. Hill & K. Parry (Eds.) *From testing to assessment: English as an international language,* pp. 174-204. London: Longman. [LTRC 11 - The taking and rating of summary tasks]

Cook, T.D., & Campbell, D. T. 1979. Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin Company.

Dickinson, L. & Haughton, G. 1988. Collaborative Assessment by Master's Candidates in Tutor-Based System. *Language Testing, 5*(2), 233-246. [LTRC 10 - same title]

Gordon, C. & Hanauer, D. 1995. The interaction between task and meaning construction in ESL reading comprehension tests, *TESOL Quarterly, 29*(2), 299-324. [LTRC 15 - Test answers as indicators of mental model construction]

Hamp-Lyons, L. 1993. Applying Ethical Standards to Portfolio Assessment in ESL. Paper presented at LTRC 15.

Hamp-Lyons, L., Henning, G., & De Mauro, G. 1988. Construction Validation of Communicative Writing Profiles. Paper presented at LTRC 10.

Hale, G. A. & Courtney, R. G. 1994. The Effects of Note-Taking on Listening Comprehension in the Test of English as a Foreign Language. *Language Testing, 11*(1), 29-48. [LTRC 12 - Note taking and TOEFL listening comprehension]

Luoma, S. 1993. Validating the Certificates of Foreign Language Proficiency: The Usefulness of Qualitative Validation Techniques. Paper presented at LTRC 15.

Lynch, B. K., Davidson, F., & Henning, G. 1988. Person Dimensionality in Language Test Validation. Paper presented at LTRC 10 (also published: 1988. Person dimensionality in language test validation. *Language Testing, 5*(2), 206-219).

Milanovic, M., Saville, N., & Hong, S. S. 1993. A Study of the Decision-Making Behaviour of Composition. Paper presented at LTRC 15. [paper unavailable]

Murray, N. & Madsen, H. S. 1984. Retrospective Evaluation of Testing. *Selected papers from: Deseret Language and Linguistics Society 10th Annual Symposium.* Provo, UT: BYU. [LTRC 6 - Madsen, H. - Retrospective student evaluation of testing]

Wall, D., & Alderson, J. C. 1993. Examining Washback: The Sri Lankan Impact Study. *Language Testing, 10*(1), 41-69. [LTRC 14 - same title]

Appendix B. Interviews

(positivist)	(alternative)	(ambiguous)
1. McNamara, Tim	1. Cohen, Andrew	1. Davidson, Fred
2. Turner, Carolyn [second round only]	2. Shohamy, Elana	2. Alderson, Charles
3. Chalhoub-Deville, Micheline	3. Norton, Bonny	3. Clapham, Caroline
4. Bachman, Lyle	4. [Wall, Dianne not completed]	4. Douglas, Dan

*** Categorizations of interviewees refer to our initial impressions of the single paper examined in the first stage of this study, NOT to the people and their work in general.**

Appendix C. Interview Guide (first stage)

1. In what ways do you see your research as being situated in a particular research paradigm, and how would you label or describe that paradigm? [By "paradigm", we intend: a set of guiding principles and assumptions about the nature of what we are researching (i.e., the nature of "reality"), our relationship to what we are researching (e.g., "objective", "subjective"), and the procedures we use to carry out that research.]

2. How do you define and determine the validity of your research?

3. How do you define and determine the validity of tests and other forms of assessment?

4. How does your understanding of research paradigms influence your approach to validity?