CrossMark

Lang. Teach. (2017), **50.2**, 276–289 © Cambridge University Press 2017
doi:10.1017/S0261444816000471

First Person Singular

Forty years of doing second language testing, curriculum, and research: So what?

James Dean Brown University of Hawai'i at Mānoa

I started out as a budding English as a second language (ESL) teacher in 1976 at UCLA where I went through the M.A. TESL (Teaching English as a Second Language) and Applied Linguistics Ph.D. programs. Sadly, those two programs were 'disestablished' in 2014, which provides a stark lesson to any departments that think they are hot stuff that NOTHING LASTS FOREVER. Nonetheless, my training in those programs provided me with an excellent start down three professional paths: second language testing, curriculum development, and research methods.

Second language testing

One day, in my first TESL methods course, Russ Campbell was explaining some of the job opportunities that were available in the field. He covered teaching ESL/English as a foreign language (EFL), publishing materials, training teachers, and directing ESL/EFL programs. Just before the bell rang, he mentioned one other possibility—one that most language teachers weren't interested in because it required lots of math and statistics. Suddenly, I sat up straight and listened carefully. (As a French horn player for the first 24 years of my life, one whose dad played bassoon and brother played oboe, I was acutely aware that people who played odd and difficult instruments would always have a seat in the orchestra, and if they played reasonably well, they would be first chair.) What Russ said was that every program needed one person who was interested in language testing and statistics, and since I was good at math, I knew immediately that I had found my applied linguistics French horn.

Enter criterion-referenced testing

The next school term, I took a language testing course, and I continued taking testing and research statistics courses in the Educational Psychology (Ed Psych) department all through

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

the rest of my four years at UCLA. Fortunately for me, my first testing course in Ed Psych was taught by W. James Popham, who was one of the fathers of criterion-referenced testing (CRT is essentially good solid classroom testing that focuses on measuring what students have learned and ignores the normal distribution and related statistical concepts) (see Popham & Husek 1969). His influence is clear in a number of my articles and books (e.g., Brown 1996, 2005; Brown & Hudson 2002). At this point, I wanted to help people in our field understand the difference between the institutional-level norm-referenced tests (NRTs) that are done TO students for admissions and placement purposes and the classroom CRTs that teachers create FOR students (usually for diagnostic and achievement purposes) to promote learning by providing motivation and feedback.

In another stroke of luck, I took five courses with Richard Shavelson, who also greatly influenced my thinking on testing and research statistics. He had studied with Lee J. Cronbach, who first proposed generalizability theory (G-theory). G-theory allowed testers, for the first time, to study and statistically disentangle multiple sources of measurement error statistically and use that information to rationally design more reliable or dependable tests (see Cronbach et al. 1972). I use the two terms RELIABLE and DEPENDABLE on purpose because G-theory also usefully differentiates the RELIABILITY of NRTs from the DEPENDABILITY of CRTs. G-theory's practical capacity for analyzing and redesigning CRTs fit very nicely with Popham's theoretical CRT notions and led directly to my dissertation and a string of papers on using G-theory in a variety of different testing projects.

Unfortunately, everywhere I turn today I see people fundamentally—almost willfully—choosing to misunderstand the difference between NRTs and CRTs. As a result, worldwide, language-related tests are often poorly designed and their scores fundamentally misused. Sadly, almost all of my efforts to get administrators to do the right thing with NRTs have failed. But I stubbornly hold out hope for CRTs and keep trying to promote sound classroom assessment practices (e.g., Brown 2013a, 2016a) because I think that most language teachers fundamentally care about their students' well-being.

Even among language testers, I see the basic CRT notions twisted beyond all recognition: people regularly misunderstand the concept, mislabel CRTs, and analyze CRTs incorrectly. For example, a number of people have somehow decided that any test that involves a verbal description in a rubric (e.g., the American Council on the Teaching of Foreign Languages (ACTFL) and Common European Framework of Reference for Languages (CEFR) oral proficiency assessments) is a CRT. However, careful examination of such tests (in terms of their scope; the procedures raters use for making scoring decisions; the purposes of the decisions; the wording and levels of precision in the descriptors; and so forth) reveals that they are really just very imprecise NRTs of general language proficiency. In my view, such folks should go back to the original literature starting with Popham & Husek (1969) and Cronbach et al. (1972), and get it right.

Overall English language proficiency (whatever that is)

More recently, I have been grappling with the issue of general English language proficiency (ELP). A number of tests (e.g., iBT TOEFL, IELTS, TOEIC) claim to be testing ELP, which

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

is largely understood to be the abilities of non-native speakers (NNSs) to understand and produce the spoken and written English of native speakers (NSs). From the beginning of my career, I have been uncomfortable with the notion of ELP, always referring to it as '*English language proficiency* (whatever that is)'. Largely because of conversations I had with Sandy McKay about *World Englishes*, *English as an International Language*, and *English as a lingua franca*, I have recently been able to clarify my thinking about ELP in terms of language testing (Brown 2014a, forthcoming; Brown with Nodoushan 2015), curriculum design (Brown 2012), and both topics (McKay & Brown 2015).

Unfortunately, the validity of ELP tests is generally tied closely to the NS model. Davies (2004) usefully characterized NSs in terms of their: childhood acquisition of English, intuitions about their idiolectal grammars, sensitivities to differences between their idiolectal grammars and standard English, abilities to produce spontaneous discourse, wide range of communicative competence, capacities for writing creatively, and abilities to interpret and translate into English (p. 435). Nonetheless, the notion of NS English remains an idealized, vague, ill-defined, and problematic concept from both language testing and curriculum development points of view. For example, given that verbal ability varies tremendously among NSs (as shown consistently on NS standardized tests), I must ask exactly what this NS target is. The verbal abilities of an average NS? The description provided by linguists in grammar books? Or, some idealization of an educated NS? And, which English are we talking about? British? North American? Australian? And, which dialect of that English are we talking about? And, who speaks this English? The Queen of England? George W. Bush? And, what about Kachru's (1986) notion of the Inner-, Outer-, and Expanding-Circle of Englishes? And, is this English the oral or written form, or both? The questions about the NS, the NS model, and ELP go on and on—leaving us with the idea of ELP (whatever that is), which is based on the so-called NS model, which is based on the notion of the NS, which is itself vaguely and poorly understood. (For a much fuller description of these issues, see Brown 2014a, forthcoming.)

This sort of thinking has led me to ask if the NS model of ELP is the only one that makes any sense for ESL/EFL testing and to explore other possible approaches to ELP. In Brown (forthcoming), I list and discuss a total of 14 approaches to ELP that can serve as a basis for ELP testing. Six are what I call top-down approaches that focus on variance caused by examinees interacting with the language for different types of English (including what I label traditional ELP, truth-in-advertising, inner-circle world Englishes, English as a lingua franca, global standard English, and functional approaches) and eight are bottom-up approaches (including what I label the effective communicator, scope of proficiency, scale of range, intelligibility, resourcefulness, symbolic competence, intercultural communication skills, and performative ability approaches).

In short, I had my doubts about the notion of ELP (whatever that is) from the very beginning, and the field has done little to help me resolve those doubts. Instead, it has led me to alternative approaches for testing ELP that may be more appropriate for large proportions of the students studying English in the world today. Notice that I am not arguing for any particular approach, but rather for the realization that approaches different from the NS model of ELP exist, ones that might be more appropriate and effective for testing ELP under many circumstances.

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

Second language curriculum development

My first exposure to real-world curriculum development started in 1980 when I was sent by UCLA to China to help develop an English for science and technology (EST) program. Though we had spent three months preparing at UCLA, when we arrived in Guangzhou, we really had no curriculum. However, we did have 225 students arriving in a month. So in a sense, curriculum development, and later program evaluation, just happened to us, and we had to react. We needed to design, set up, and implement one of the first EST programs in China called the Guangzhou English Language Center. Ultimately, I think we did a respectable job even though we were working under very difficult circumstances and with limited resources. However, we encountered a number of problems along the way (also see Brown 2009).

Getting people to do things

Teachers

We learned very quickly that language teachers tend to think of themselves as sovereign within their classrooms and that any extra work outside of the classroom is not their responsibility. Because we needed to create a curriculum from scratch, we desperately needed to get teachers to change their thinking about the relationship between teachers and curriculum development. We were able to solve this problem partly because Russ Campbell had wisely built curriculum development and research right into each teacher's job description and contract (i.e., 50% teaching and 50% research and curriculum development). We further clarified this division of labor by restricting each teacher's classroom workload to two to three hours per day and scheduling all of that in the mornings. Thus we were all free after lunch for what we described as class preparation, curriculum development, and research.

However, even with a structure conducive to doing curriculum and research, teachers vary in their training and beliefs, one result of which is that they do not always agree on the best ways forward. More specifically, teachers have belief systems in terms of the nature of language and language learning (sometimes called APPROACHES, see Richards & Rodgers 1982; Brown 1995) and how best to organize language teaching and materials (sometimes called SYLLABUSES, see Brown 1995, 2012, 2016b). We used afternoon meetings to provide the teachers with the information and time they needed to understand and discuss these issues. In terms of providing them with information, during much of the first year, we ran some of the meetings as workshops/seminars in curriculum development and research methods with different participants researching and presenting information as appropriate. We also used those meetings to organize our efforts, set deadlines, have teams report on their progress, and so forth—all of which was necessary to keep our curriculum development efforts moving forward. In addition, we were young and energetic and, more importantly, we didn't realize that what we were doing was impossible, so we just did it.

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

Students

Another problem we faced was getting the STUDENTS to cooperate. Like teachers, students have belief systems. In China at that time, most of our students had only experienced a form of grammar-translation method in their language learning. In contrast, the American teachers generally believed in communicative language teaching, functional syllabuses, skills acquisition, and what are today called task-based activities. While the students wanted us to tell them which paragraph to look at and then explain it in detail to them for the class hour, we preferred to demonstrate functions or skills and the language associated with them and then have the students practice all of that in pairs or groups. Initially, they did not like our approach at all. We were able to convince them to try our ways by gathering information in the form of needs analysis and by respecting them enough to patiently explain what and why we were doing what we were. For example, we showed them that their grammar and vocabulary recognition scores on our placement test were very high, and that their reading, writing, listening, and speaking scores were relatively low. Since most of them were scheduled to leave for English speaking countries in four months, we used a variety of strategies to gradually convince them that they would be better served when they arrived in those countries by being able to read, write, listen, and speak, than by knowing a great deal about English grammar and vocabulary. In addition to showing them their test scores, (a) we designed course tests (CRTs) that focused on their abilities to use the language rather than on usage, (b) we explained in all classes what they would need to be able to do when they landed in their new locations, and (c) we explained the importance of learning how to do those things as fluently as possible (sometimes at the expense of accuracy). In the end, based on attitude questionnaires that we regularly administered, we found that they were increasingly buying into what we were selling. Over the course of two years, we not only saw a great deal of change in their attitudes, but more importantly, we saw transformations in how they were practicing English outside of the class.

By gathering and using information from students, teachers, and administrators in the form of needs analysis, we were able to find a path down the middle that would make the curriculum largely acceptable to the teachers and students alike. We also learned that such information is often contradictory and confusing and that there is no such thing as TRUTH in these situations. Indeed, this is where I learned to use needs analysis information to create DEFENSIBLE CURRICULUM, which by definition involves a series of compromises between and among stakeholders that leads, through needs analysis, to student learning outcomes, assessment, materials, teaching, and program evaluation that are defensible in the sense that most stakeholders can live with them (for more on these ideas, see Brown 1995, 2012, 2016b).

Dealing with TENOR

Because of my experiences with English for specific purposes and with needs analysis, when I became the director of the English Language Institute at the University of Hawai‘i at

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

Mānoa, I naturally shifted its focus to English for academic purposes. My main motivation in doing so was to provide students and teachers with realistic and manageable targets. At the same time, while I was working summers and sabbaticals in Japan, I noticed that the EFL teaching in Japan was largely TENOR (teaching English for no obvious reason). I came to see TENOR as a destructive force because, in many teaching situations, it sets up unrealistic and unrealizable expectations. Typical students in Japan only study 1–4 hours of (non-intensive) English for 30 weeks per year spread across six years of intermediate and secondary school (with long summer breaks) and (in many cases) with low motivation, which only amounts to a maximum total of 720 hours (4 hours x 30 weeks x 6 years = 720) of EFL, not ESL, study. The students are nonetheless declared in the press to be failures because they cannot communicate with NSs who have had tens of thousands of hours of high motivation, intensive exposure to English, and concerted learning of English in school.

For some students of English around the world who want to study at a university in an English speaking country, TENOR and the NS model MAY in fact be appropriate (though having walked through the Engineering Department at my university I would question even that). However, for the vast majority of students of English in the world, their goal is not to study in a university in an English speaking country, or even to study in a university at all. I really must ask why we insist on setting up this impossible dream of attaining the sainted status of NS and thereby guarantee that most students, even hardworking, conscientious students of goodwill will fail. Think about it. We expect everyone to reach a high level of ELP (whatever that is) in a short few years while only meeting a few times per week, and when they fail, we blame the _____ (fill in the blank with your favorite scapegoat: teachers, materials, students, curriculum planning, big government, unions, etc.).

Another problem with TENOR is that it typically holds that NNS students should learn to communicate with NSs of English and such communication should be assessed/judged in terms of how well they approximate the NS model. The reality for a majority of NNSs learning English is that they are likely to only use English to communicate with other NNSs (e.g., Malay with Tamil, Arab with Hungarian, Chinese speaking with Malay speaking Singaporeans). The implications of all of this are that students around the world are being taught an NS model of English: (a) that is ill-defined, (b) that they do not need, (c) that they may never use, (d) that they are very unlikely to ever achieve, and (e) that will lead them to a profound sense of failure. It seems clear to me that such learners would be much better served by being taught something relatively realistic, manageable, and learnable like English for specific purposes (see Brown 2016b), or locally defined English as an international language (EIL) based on their actual local needs to communicate in English (see Brown 2012, and McKay & Brown 2015).

Second language research methods

At UCLA, one of the most important lessons I learned came from watching young professors fail to publish enough and then get passed over for tenure and promotion and simply disappear—publish or perish was real. Because I wanted to be a successful professor, I started

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

publishing right away. I did this by writing publishable papers (that the field would want to read) instead of term papers (that only a professor would read) for all of my courses. As a result, when I graduated, I not only had a doctorate, but also seven statistical research reports and five other articles either published or in press—all of which set me up nicely for the job market.

Shifting research paradigms

Along the way, I was horrified to realize as a young researcher immersed in statistical research that most of the quantitative research in our field was poorly designed, analyzed, and reported. Because I was young and naïve, I made it a professional goal to try to remedy this problem by publishing straightforward books (e.g., Brown 1988, 2001) and articles (e.g., Brown 1991, 1992) on statistics and quantitative research design. I also volunteered to serve on what the *TESOL Quarterly* editor at the time called her 'statistical swat team.' Several of us young statistics buffs were tasked specifically with examining the quality of the statistics in the articles accepted by the journal. I thought this was a terrific idea, and I wholeheartedly supported it. Unfortunately, most of the light went out of my eyes because of one experience: I suggested crucial changes in the statistics for an article (i.e., reanalyzing multiple statistical tests that were uninterpretable with a standard omnibus ANOVA design and post hoc comparisons), but the first author, who was a so-called 'big name', simply disagreed and refused to reanalyze the data, and unfortunately, the editor caved in and published the study anyway. That experience very nearly cured my idealism. From that point on, I have wondered privately if there was much hope for our field ever producing consistently sound statistical research.

One day in 1991, while I was teaching in Brazil, a colleague (who favored qualitative research) was taking me around Rio de Janeiro for shopping. I remember talking about research and explaining why I thought qualitative research was pointless. She stopped me cold by asking me if I wouldn't be a stronger researcher if I could use both quantitative and qualitative research tools. I blithely continued my pursuit of quantitative research, but her question nibbled away at the back of my mind for several years.

Then one day, it dawned on me that I had been doing both quantitative AND qualitative research all along—especially in my curriculum work. Considering this idea more closely, I realized that I had always unconsciously included QUALITATIVE data in my curriculum work (e.g., meetings, classroom observations, interviews, analysis of textbooks, analysis of lecture discourse, open-ended questionnaire items) in gathering information from students, teachers, and administrators for the needs analyses, program evaluations, and other curriculum components. In retrospect, the numerical/quantitative parts of such work look more like afterthoughts (usually taking the form of Likert-item questionnaires). Instinctively, I would analyze even my questionnaire results not only numerically, but also qualitatively by including open-ended items in each section that would help me understand the numerical results. Thus I began begrudgingly to understand that I had been doing quantitative and qualitative research all along.

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

I think it had taken me such a long time to realize this basic truth because I had two strong biases: one against the idea of curriculum development being research and the other against the idea of qualitative methods being research. However, once I had redefined research in my mind as 'any systematic and principled inquiry' (a definition that I eventually published in Brown 2004: 477), I understood that many things that I was doing, like curriculum development, were actually research, whether I understood that or not. I suspect now, that I was simply embarrassed by my reliance on qualitative data, and rationalized my qualitative tendencies by explaining and defending the results quantitatively. Such cognitive dissonance typically leads me to a turning point.

That turning point came when I co-taught an introductory research course with Kathy Davis that was designed to be 50/50 qualitative and quantitative methods. In that course, I learned that qualitative research can be systematic, as well as principled. I even came to accept my qualitative tendencies, but with an eye to making my qualitative research dependable, credible, confirmable, and transferable (after Davis 1992, 1995).

I continued searching for strategies to combine qualitative and quantitative research methods that would make me a stronger researcher overall. In the process, I came across a book by Newman & Benz (1998), who argued for combining qualitative and quantitative methods into a single 'qual/quant' continuum. I began to publicly advocate balancing quantitative and qualitative research methods (e.g., Brown 2001, 2004). I also continued to look for ways that qualitative and quantitative research methods could CROSS-VALIDATE each other in ways that would make both stronger. Inevitably, I stumbled into a new trend called MIXED METHODS RESEARCH, in which cross-validation is only one way that qualitative and quantitative research methods can re-enforce each other by examining how the two methods converge, diverge, elaborate, clarify, exemplify, and interact with each other (all of which I explain in Brown 2014b).

In short, over nearly 40 years in applied linguistics, I have seen the notion of research grow in quality and quantity, and I myself have grown to accept and use quantitative, qualitative, and mixed methods research methodologies. Naturally, there are times when one or the other research methodology is best used alone with a particular set of issues or research questions. Thus, I have no problem arguing for more effective use of very advanced statistical techniques (e.g., Brown 2015), while also advocating for doing qualitative research methods well (e.g., Brown 2001) and using mixed methods research when that is appropriate (e.g., Brown 2014b). All of these realizations have led me to be more tolerant of the variety of research methods used in our field (see e.g., the variety of methods in Brown & Coombe 2015). I am not saying that I think our field has stopped producing shoddy research. In fact, it is still challenging to find exemplary quantitative, qualitative, and mixed methods studies in our field to cite in my writing and teaching. However, I think I can now judge the research in our field with a relatively open mind, whatever the research methodology, in terms of the degree to which it is systematic, principled, convincing, interesting, and useful.

Thus, I have little patience with quantitative researchers who denigrate qualitative interpretive methods given that: (a) quantitative research methods are so often applied badly in our field, (b) qualitative methods have a number of strengths not shared by quantitative methods, (c) qualitative research methods fit the skill sets of many language teachers better than numerical analyses do, and (d) qualitative research can be just as systematic and

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

principled as any other research methodology. At the same time, I see absolutely no reason for qualitative researchers to argue that quantitative research is useless—especially given that quantitative methods are by far the dominant form of research in the world today as well as in our field (in terms of numbers of publications, grants, citations, or however it is measured). In addition, when I hear people referring to 'today's postmodern world', I have to ask myself what world those people live in and if they are delusional or just western-centric. After all, the vast majority of the research done in today's world has nothing to do with the western notion of postmodernism.

Shifting research processes

While writing Brown (2013b), I came to a number of other realizations about research. In that paper, I looked back on one thread of my language testing research that included more than a dozen studies on cloze testing. The paper addressed a number of issues, but most importantly, I considered how I had learned from my mistakes and how I had learned more than I had set out to by looking at all of this research at one time. What I noticed was that, beginning with Brown (1980), the anomalies and mistakes in each of my studies brought up issues and questions that naturally led me to the next study, which in the long haul led me to increasingly better understandings of how cloze tests work.

The role of novice researchers

More broadly speaking, writing Brown (2013b) taught me a number of things about research processes. First, we should stop thinking about research studies as one-shot affairs that are finished and done. Unfortunately, a good deal of the published research in our field today is made up of rewritten M.A. theses and Ph.D. dissertations. This means that much of our published research is being conducted by researchers who are doing their first, last, and only study. One consequence is that, because first-time researchers lack the experience or clout to get large sample sizes, much of our published research is based on small numbers of participants and lacks statistical power. A second consequence is that, since first-time researchers often have the naïve view that research must include, by definition, a treatment and control group and perhaps a pretest and posttest with the purpose of comparing two teaching methods, they are almost immediately in over their heads. I have repeatedly found that such studies are extraordinarily difficult to carry out because controlling all necessary variables is nearly impossible. The biggest problem is that learning a language takes an enormous amount of time, effort, and motivation to be successful, and unfortunately, for these comparative methods studies, a 30-minute or even 90-minute treatment typically has very little short-term impact on any student's overall language learning, much less long-term impact. Another problem is that the treatment and control group design only buys them (as Brian Lynch put it at a long ago conference) a demonstration that their treatment IS BETTER THAN NOTHING. Since such studies also open a Pandora's box of design problems (teacher

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

effect, equality of the groups at the beginning, practice effects, and so forth), they are very difficult to plan and carry out, especially for novice researchers.

I did one such study with a colleague when I was a novice researcher (Brown & Hilferty 1987) that compared a treatment group which was taught 50 ten-minute lessons on REDUCED FORMS (aka, CONNECTED SPEECH) versus a placebo group who practiced minimal pairs drills. We compared the two groups for effects on three types of listening comprehension scores. That study taught me how very difficult such comparative methods studies are to conduct, control, and publish. I vowed then and there to never do a treatment and control group study again—a vow that I have steadfastly kept.

But I did learn a valuable lesson from that study when Ann Hilferty and I went one time to talk to Jim Popham (of CRT fame). We brought up all the problems we were having with our treatment and control group study (Brown & Hilferty 1987). He just laughed and talked about how difficult such studies are to carry out. He advised us to ignore the control group and focus on the effectiveness of our treatment by studying our reduced forms teaching and materials carefully and in depth in terms of whether or not they worked, how well they worked, why they worked, how the students felt about them, how they could be improved, and so forth. As with many other things, I think Popham was 100% correct with his solution to our problem.

I am not saying that people should not do M.A. theses and Ph.D. dissertations. However, I would suggest that such novice researchers might benefit more from doing much-needed replication studies or working with more experienced researchers within very clearly defined research agendas. At the moment, researchers generally avoid doing replication studies because many think that replication studies are difficult to publish. Hence, strategies should be developed to provide incentives to novice researchers to conduct replication studies from which they (and the field) could learn a great deal. For example, the journals in our field could turn this trend around by setting up special sections for papers that are replication studies, or establishing websites specifically dedicated to publishing replication studies.

The role of senior researchers

I eventually concluded that our field should rely more heavily on professional researchers. As a field, we need to work out strategies that encourage experienced researchers to keep doing empirical data-based quantitative, qualitative, or mixed methods research throughout their professional lives. The problem that slowly creeps up on senior professors is that they encounter (a) OTHER DEMANDS on their time that gradually pull them away from empirical research and (b) increasing ROADBLOCKS to doing empirical studies.

The OTHER DEMANDS take many forms like: doing the endless bureaucratic paperwork; taking on administrative duties and committee work; sitting on local, national, and international governing boards; holding elected office in professional organizations; reviewing articles for journals; becoming journal editors; pursuing and obtaining grants or other funding; traveling to do invited keynote or plenary speeches and other invited lectures/workshops; writing invited think-papers (like this reflection); and so forth. Is it any wonder that senior faculty find themselves without the time to do empirical research?

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

Table 1 Challenges and suggestions for meeting them in three areas of applied linguistics

Three Paths	Challenges	Suggestions
Second Language Testing	Criterion-referenced testing done poorly	Keep pushing administrators to understand and apply NRT/CRT differences in their decision making, but focus on helping teachers to do sound classroom assessment that benefits students
	Many language testers incorrectly label and analyze CRTs	Such folks should read the foundational works by Popham & Husek (1969) and Cronbach et al. (1972), and get it right by recognizing the clear differences between and uses of NRTs and CRTs
	Overall English language proficiency (whatever that is)	Abandon the NS standard for ELP in most cases and explore instead some or all of the 13 alternative approaches to ELP listed in Brown (forthcoming)
Second Language Curriculum Development	Getting teachers to do things	Build curriculum development and research into teachers' idea of professionalism and job descriptions, and provide adequate training in curriculum development and research skills
	Getting students to do things	Respect students, teachers, and administrators enough to gather needs analysis information and then use that information to create, organize, and explain your new DEFENSIBLE CURRICULUM to all stakeholders
	Dealing with TENOR	Teach global English only to that minority of students who actually need it; teach more manageable and realistic English for special purposes or locally defined EIL to the rest of the students, as appropriate
Second Language Research	Shifting research paradigms	Remain open and flexible to advances in quantitative, qualitative, and mixed methods research and adopt the benefits of new approaches that fit their training and personalities
	Novice researchers doing one-shot treatment/control studies	Avoid doing treatment-control group studies and comparative methods studies; develop policies to encourage novice researchers to do replications or work in teams with more experienced researchers
	Senior researchers distracted and blocked from doing research	Develop policies to encourage senior researchers to continue doing empirical research throughout their careers with collaborative grants, commissioned studies, organized collaborations, etc.

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

The ROAD BLOCKS to doing empirical research also take many forms: relentless human-subjects requirements that make us apply for EXEMPTION from human-subjects requirements (how is that for a circular waste of time) so as to protect our universities from lawsuits; human-subjects requirements that make it impossible to do sound empirical research (like the need to explain the purpose of a study to the participants before it is conducted); self-righteous 'peer' reviewers who know the 'truth' because they just finished graduate school; increasingly strenuous permissions requirements (e.g., requiring permissions for every single quotation) by some of the publishers (who are scared to death of being sued for quoting *Harry Potter*); the ever declining budgets at universities around the world; the anti-intellectual attitudes in society generally; and so forth.

I, for one, find the prospect of doing empirical research exhausting. However, as this paper illustrates, my interest in empirical research persists, and I do continue to do primary research, but like everyone else, my time is often absorbed by other activities that are both easier and often MORE PRESTIGIOUS.

I'm not sure what the solutions are to these problems. Perhaps organizations like Teachers of English to Speakers of other Languages ACTFL, International Language Testing Association, and American Association of Applied Linguistics could set up policies for research (with grants, commissioned studies, organized collaborations, and other incentives) that would encourage senior researchers to keep doing empirical research.

In the final analysis, the field and all of the researchers in it, whether novice or experienced, should begin thinking of research as a procession of studies over a period of years. Along with that view, it would be important to always: be aware of the research of others; accept the premise that we learn from our mistakes; expect the unexpected; be alert for anomalies and new questions that arise; take a longer view of our work; and accept that progress in any field requires a process of consensus building over time.

Conclusion

Table 1 shows a summary of the key issues and challenges that I have discussed in reflecting on my time in second language testing, curriculum development, and research. In the process, I have tried to outline where I think the future lies in my areas of applied linguistics. Will the field follow all of my suggestions? I doubt it. However, I trust that others are grappling with some of these same issues, and I hope that my musings here will prove helpful as they develop their own views and move forward in their careers.

References

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal* 64.3, 311–317.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

Brown, J. D. (1991). Statistics as a foreign language – Part 1: What to look for in reading statistical language studies. *TESOL Quarterly* 25.4, 569–586.
Brown, J. D. (1992). Statistics as a foreign language – Part 2: More things to look for in reading statistical language studies. *TESOL Quarterly* 26.4, 629–664.
Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. New York: Heinle & Heinle.
Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.
Brown, J. D. (2004). Research methods for applied linguistics: Scope, characteristics, and standards. In A. Davies & C. Elder (eds.), *The handbook of applied linguistics*. Oxford: Blackwell, 476–500.
Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (new edition). New York: McGraw-Hill.
Brown, J. D. (2009). Language curriculum development: Mistakes were made, problems faced, and lessons learned. *Second Language Studies* 28.1, 85–105. Retrieved 5 December 2014, from www.hawaii.edu/sls/sls/wp-content/uploads/3-Brown.pdf.
Brown, J. D. (2012). EIL curriculum development. In L. Alsagoff, S. McKay, G. W. Hu & W. A. Renandya (eds.), *Principles and practices for teaching English as an international language*. London: Routledge, 47–167.
Brown, J. D. (ed.) (2013a). *New ways of classroom assessment, revised*. Alexandria, VA: Teachers of English to Speakers of Other Languages.
Brown, J. D. (2013b). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies* 7.1, 1–32.
Brown, J. D. (2014a). The future of world Englishes in language testing. *Language Assessment Quarterly* 11.1, 5–26.
Brown, J. D. (2014b). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University.
Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research? In L. Plonsky (ed.), *Advancing quantitative methods in second language research*. New York: Routledge, 9–20.
Brown, J. D. (forthcoming). World Englishes and international standardized English proficiency tests. In C. Nelson, B. B. Kachru & Z. G. Proshina (eds.), *Handbook of World Englishes* (2nd edn.). Malden, MA: Wiley-Blackwell.
Brown, J. D. (2016a). Assessments in ELT: Teacher options and making pedagogically sound choices. In W. A. Renandya & H. P. Widodo (eds.), *English language teaching today: Linking theory and practice*. New York: Springer International.
Brown, J. D. (2016b). *Introducing needs analysis and English for specific purposes*. New York: Routledge.
Brown, J. D. with M. S. Nodoushan (2015). Language testing: The state of the art. *International Journal of Language Studies* 9.4, 133–143.
Brown, J. D. & C. Coombe (eds.). (2015). *Cambridge guide to research in language teaching and learning*. Cambridge: Cambridge University Press.
Brown, J. D. & A. Hilferty (1987). The effectiveness of teaching reduced forms for listening comprehension. *RELC Journal* 17.2, 59–70.
Brown, J. D. & T. Hudson (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
Cronbach, L. J., G. C. Gleser, H. Nanda & N. Rajaratnam (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
Davies, A. (2004). The native speaker in applied linguistics. In A. Davies & C. Elder (eds.), *The handbook of applied linguistics*. Malden, MA: Wiley-Blackwell, 431–450.
Davis, K. A. (1992). Validity and reliability in qualitative research on second language acquisition and teaching: Another researcher comments . . . *TESOL Quarterly* 26, 605–608.
Davis, K. A. (1995). Qualitative theory and methods in applied linguistics research. *TESOL Quarterly* 29, 427–453.
Kachru, B. B. (1986). The power and politics of English. *World Englishes* 5, 121–140.
McKay, S. & J. D. Brown (2015). *Teaching and assessing locally defined EIL*. New York: Routledge.
Newman, I. & C. R. Benz (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum*. Carbondale, IL: Southern Illinois University.
Popham, W. J. & T. R. Husek (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement* 6, 1–9.

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press

Richards, J. C. & T. Rogers (1982). Method: Approach, design, and procedure. *TESOL Quarterly* 16.1, 153–168.

James Dean Brown ('JD') is currently Professor of Second Language Studies at the University of Hawai'i at Mānoa. He has spoken and taught in many places ranging from Brazil to Venezuela. He has published numerous articles and books on language testing, curriculum design, research methods, and connected speech. His most recent books are: *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (2012 from NFLRC); *New ways in teaching connected speech* (2012 from TESOL); *New ways of classroom assessment, revised* (2013 from TESOL); *Practical assessment tools for college Japanese* (2013 with K. Kondo-Brown & Tominaga from NFLRC); *Mixed methods research for TESOL* (2014 from Edinburgh University Press); *Cambridge guide to research in language teaching and learning* (2015 with C. Coombe from Cambridge University Press); *Teaching and assessing EIL in local contexts around the world* (2015 with S. L. McKay from Routledge); *Developing courses in languages for specific purposes* (2015 with J. Trace & T. Hudson from NFLRC); *Introducing needs analysis and English for specific purposes* (in press 2016 from Routledge); and two others that are currently in the works.

https://doi.org/10.1017/S0261444816000471 Published online by Cambridge University Press