A critical evaluation of placement tests/skills grouping

Rory Britto
Kurume University

Reference Data:

Britto, R. (2006). A critical evaluation of placement tests/skills grouping.
In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2005 Conference Proceedings*. Tokyo: JALT.

I approach placement tests/skills grouping [PT/SG] from a critical perspective. Although they are commonplace, I question their purpose, design, and application. In discussing my research, I will show that they rely on flimsy theoretical support, go contrary to established educational principles, and rest on virtually nonexistent empirical evidence. It is hoped that this presentation will be of interest to those involved in testing and curriculum development, and issues of educational and language learning principles.

筆者は批判的な観点から習熟度別クラス編成(以下PT/SG)について検討する。PT/SGはめずらしいものではないが、それらの目的、構成、妥当性に疑問を感じる。この研究を説明するにあたって、PT/SGにはたいした理論上の根拠がなく、確立された教育原理に矛盾し、事実上存在しない経験上の証拠に依存しているということを示したい。本発表はテスティングやカリキュラムの発展と教育原理や言語学習の原理に関する事柄に携わっている人の興味・関心を引きつけるものになるだろう。

In this discussion of placement examinations and skills grouping (PT/SG), I approach the subject in terms of theory and practice. This division makes it seem as if there is a successive relationship between the two; that practice follows theory. However, this is likely due more to the linear nature of language than to any true relationship. The two interpenetrate so that such a division is, in essence, impossible, and at times the discussion will overlap. The basic relationship seems to be that theory has a foundational basis to the practices that we pursue, though the relationship extends in the other direction also. As Oller points out, "...successful practice is almost always founded in good theory and ...superior theory is almost always the one that works best in practice." (Oller 1983:x)

In addition, often when we speak of theory, the general impression seems to be that because of its academic nature, it is somehow removed from the practical, pragmatic world of teaching. One way to state this is that academic interest in theory often abstracts it from praxis. The way that Oller puts it (in

referencing Krashen, ibid) is to question whether there is a disillusionment among teachers in regards to theory. However, it seems clear to me that the choice we have in this regard is to base our practical educational decisions on theory or not.

In this paper, I argue that we should find grounding in theory for our educational decisions. In the present case, this refers to decisions about placement examinations and skills groupings (hereafter PT/SG). I take this position because of the alternative. That is to say, that if we don't base our decisions on theory, what we have are decisions that are arbitrary at best and whimsical at worst. That being the case, the strong version of the argument here is that we should abandon PT/SG of a certain type, or at least postpone them until certain conditions are met. These conditions will be discussed below.

Here, when we speak of theory as regards language learning, I find it useful to distinguish unified theories (sets of theories) from the theories (hypotheses) that treat of certain phenomena. Unified theories, as I use the term, are those that constitute schools of thought. For example, we have the Audio-Lingual school, the Grammar-Translation school, or Krashen's Monitor Model and so on. These theories focus on what language is, how it is learned, and the best way to teach it; our basic *approach* > *method* > *technique* paradigm.

On the other hand, we have phenomena specific theories or hypotheses; a good example being the matter at hand - placement examinations and their concomitant skills grouping. The distinction I am trying to make here is between macro-theory and micro-theory. In either case, whether a unified theory or a hypothesis, they must rest on the same foundations. That is, in order to be cohesive on the one hand and valid on the other, they must rely on evidence, verbal reasoning, and persuasive argument. The contention here is that support for a specific kind of PT/SG is deficient in all three.

What I am *not* saying here is that it is necessary to tie our educational decisions, through a chain of reasoning, to any particular unified theory. I do feel though, to repeat, that our educational decisions, especially those at the systemic level, should have an empirical basis.

Background

My investigation into PT/SG began when the university where I teach (Kurume University, Japan) instituted a placement program for students taking Oral English Level I, an English conversation class. I mention Kurume University only as a specific instance, with which I am familiar. It is not my purpose or intent to criticize this institution or any party responsible for decisions related to its placement program.

Beginning in the 2000 academic year, we have administered placement examinations to all students registering for Oral English I. Although we do offer an Oral English II (with Oral I a prerequisite) and an Advanced English Speaking class (with Oral II a prerequisite) class, there are no placement exams for either one of these. The stated purpose and objective of this activity is to group students into classes by levels of like ability. The reason and goal for *that* being to facilitate teaching where the individual teacher can better focus on students' strengths and

weaknesses if those strengths and weaknesses are similar. A similar rationale, which will be discussed below, is that in a class of mixed abilities, were the teacher to focus on students with higher linguistic capabilities, the less successful learners would feel left behind, frustrated and consequently bored. In addition, were the teacher to focus his instruction on the less successful learners, those with higher abilities would be bored and frustrated studying material with which they are already familiar, and suffer from diminished motivation. In either case, it is claimed that this creates a disruptive presence in the classroom and the best way to alleviate that it is to homogenize the classes.

I should discuss the types of PT/SG involved. In ESL/EFL classes in general, there seems to be two: The *serial* type where students take an examination and are placed in different courses for which they receive credit for that level, and where the titles of the classes are different. It is also possible for students to advance sequentially through the levels. For example: English Conversation 101 [3 credits], English Conversation 102 [3 credits], etc. The divisions are referenced to a set criterion of scores.

The program in use at Kurume U. and under consideration here is of the second or *parallel* type: where students are grouped in separate classes for the same course according to their ability – high, medium and low. The title of the classes for which they register and the credits are the same (Oral English I [3 credits]). The divisions are referenced to relative scores.

When the PT/SG program was put into effect in the year 2000, I felt, as with the majority of teachers (Mosteller1996:811) that placement exams were probably a good thing, and had no real objection to having one in place. Ireson & Hallam (2001:107) note that 90% of language teachers favor homogeneous groupings over mixed groups. I imagine that I would have continued to accept PT/SG as normal in the course of events until I came across *The Book of Learning and Forgetting* by Frank Smith. There were two things in particular in that volume that caught my attention and stimulated my critical thinking about PT/SG. One was the statement that:

> What is still called "grouping students by age and ability" really means segregating them according to inexperience and inability, as if the aim were to make it impossible for students to help or learn from each other." (Smith 1998:47)

Smith's phrase "as if the aim...." in particular, caught my attention and caused me to question what the goal of separating students could really mean.

The other was a footnote concerning the findings of a study by Mosteller et. al. (1996), in which they found through a survey of the literature that 1) "contrary to widespread assumptions, few studies have been done in the efficacy of skills grouping" and 2) "there is no compelling evidence that it has a major impact, positive or negative, on learning." (Smith 1998:111)

Mosteller lists five findings from the 15 studies that met the criteria of providing data from experiments carried out in actual classrooms:

- The evidence for the effect of XYZ grouping is weak.
- Effects were found to be zero for high, average, and low achievers in studies of high methodological quality.

- Higher level students benefit a bit more from skills grouping, while lower level students benefit a bit more from whole class grouping.
- An intriguing observation in one study that whole class instruction was more effective, but skills grouped students spoke out more.
- More skilled students benefit a bit more from skills grouping while less skilled students benefit a bit more from whole class grouping.

Two paths opened to me; one was to question the groundings of skills grouping as a theory, and the other was to investigate what studies had been done in the field of Second Language Learning (SLL). Since the latter would naturally seem to include the former that is what will be discussed first.

Brief survey of the field

Following Mosteller's lead, I decided to conduct a survey of the field of SLL. This decision stemmed in part also from discussions at committee meetings dealing with the PT/SG program. I had brought up Mosteller's finding that there were few studies dealing with the subject, and of those none found a strong beneficial effect for skills grouping. One response was that Mosteller's findings were from general education and really didn't apply to Second Language Learning. I will address this point before continuing.

What seems to be said here is that the principles of language learning are somehow different from the principles upon which general education rest. Specifically, it was mentioned that other subject fields in general education are concerned with the transfer of knowledge, in contrast to language learning where the focus is on its communicative aspects. However, I feel that is only half of the story. In any field, there are two aspects to teaching/learning: knowledge and use. My understanding here is that *knowledge* includes the facts of any given field, and *use* refers to the cognitive aspects (i.e. skills of analysis, comprehension, interpretation, application, etc.) To say that the essence of language teaching/learning is communicative is to focus on the aspect of use only. The argument in addition, denies to other fields that the exercise of knowledge (use) is part of the educational process.

Education in language must, as in other fields, concern itself with the transfer of knowledge. That this knowledge is *at times* qualitatively different from say "factual knowledge" offers no mitigation to that fact. On the other hand, to say that other fields are only concerned with the transfer of knowledge is a mischaracterization. I feel that the root of this lies in the myth that mathematics or history for example, deal only with facts. However that is not the case, especially when we talk about the teaching or learning of these subjects. Education deals as much with how to treat information, as it does with the transfer of factual information itself. In fact, this is part of the critique of standardized testing and how it affects the classroom. Kohn (2000:29) notes that the quality of teaching suffers when teachers are 'forced' to teach the memorization of math facts and algorithms, rather than the understanding of concepts.

Anderson & Krathwohl et. al., focusing on general education, make the same kind of distinction as I have by

characterizing knowledge as *factual* or *conceptual*. They lend support to my point by sketching a trend away from a passive view of learning (which would be the 'transfer of knowledge' model), toward a "perspective that emphasizes what learners **know** (knowledge) and **how they think** (cognitive processes) about what they know as they actively engage in meaningful learning." (2001:38) (their emphasis and parenthesis)

The non-quantifiable aspects of analysis, comprehension, interpretation, and application are intrinsic to education in these other fields, and form their communicative dimensions. In turn, in second language education, we do necessarily engage in the transfer of knowledge basically in the same way done in other fields. We teach pronunciation, vocabulary and grammar structures that are the 'facts' of language instruction.

Another response to my presentation of Mosteller's finding was that in fact, there had been a discussion sometime in the 1980s and that the consensus was in favor of PT/SG.

I had to admit my ignorance of such a discussion. Although I felt that the burden of proof did not lie with me to find such a discussion and hence support for PT/SG, and since nonesuch was forthcoming, I did accept the unstated challenge. If I could find this discussion and support for the rationale of PT/SG, I could remove some of my doubts. Also, this would give me an opportunity to view and analyze the theoretical and empirical support that I had assumed existed. Thus my search began. The task was to do a library search. I began with *TESOL Quarterly* starting with the latest volume in the stacks – 2002. The original plan was to go through the table of contents of each volume, noting articles that had the word *placement* in the title. This was very time consuming and I began to wonder if there might not be a searchable database which could automate the process. Fortunately, *TESOL Quarterly* publishes a CD with such a database on it. The result of my search there was that there were a total of four articles dealing with placement examinations.

Other search results (manual and electronic) were; Modern Language Journal (2 articles since 1980), Language Testing (5 articles), English Language Teaching Journal (3 articles), Language Learning (none), and Applied Linguistics (none since 1980). While not complete, the result of my search of these six representative publications revealed 14 articles that had placement examinations in the title. In reviewing these articles, there seemed to be no discussion leading towards anything that could be characterized as a consensus on the effectiveness of placement examinations.

As a matter of fact, Fulcher referencing Wall, Clapham and Alderson, notes that,

> Although placement testing is probably one of the most widespread uses of tests within institutions, there is relatively little research literature relating to the reliability and validity of such measures. (Fulcher 1997: 113)

Important is what such research might yield in terms of validity. Research in construct validity would necessarily focus on the underlying theoretical construct of PT/SG, to which I now turn.

The hypothesis: Some representative examples

Gaffney & Mason

> A common problem for the EFL teacher is the class that is too heterogeneous in ability levels for all students to be taught according to their needs. The instructor understandably must concentrate on teaching the majority of students in a class. Those too weak to keep up will become frustrated and may give up while those much better than the average are frequently bored; either sub-group may then become a disruptive presence in the classroom. (Gaffeney & Mason 1983:97)

In the above statement, there are certain assertions that may be either *a priori* or *a posteriori*. If they are *a priori*, they must be necessarily true in all possible classrooms. However, this extreme claim is impossible to falsify. If they are *a posteriori*, they must be empirically true. However, Gaffeney & Mason do not provide solid evidence on which these assertions are based. Part of the evidential support would be operational definitions of key words and the terms used (i.e. what do "common," "too heterogeneous," and so forth mean?). "Understandably," "much better," "frequently," and "disruptive" are imprecise terms. They seem to be clear, but that is because of their anecdotal familiarity. (David Griffiths, personal correspondence)

Looking at the substance of what Gaffeney and Mason assert, we can ask what it means to 'teach to the majority' of our students. One way I like to put it in my talks and presentations on the subject is to accept the statement at face value. Putting this concept into practice with a class of 30 students for example, let us assume we have 5 students at the top end and 5 students at the bottom end of the ability scale. To teach to the majority would mean to teach to the middle twenty students in that class, effectively ignoring 10 students or a full third of the class. I doubt that this is the intent of the proponents of PT/SG. Actually their program is intended to address such a problem. However, as I point out later, it may not be possible to homogenize our students with such instruments as we have available. Also, there seems to be very little evidence that there is any real beneficial effect, regardless of the assumptions made.

Problems of assumption can also be found with other statements of the purpose of grouping and placement.

Ilyin

> Students of heterogeneous background and ability should be placed at the proper level of ESL instruction if they are to learn or stick with the courses at all. (Ilyin 1970:xx)

The main focus of my critique here is what is meant by 'should'. One wonders through what processes the author arrived at that conclusion [actually a proposition]; however there is no discussion of this point in the article. Here too, there is no presentation of evidence in support of such assertions.

Brown

In a discussion of the benefits of norm-referenced tests, Brown notes that they can be used to homogenize groups of students according to aptitude, proficiency, abilities or all three. He claims that once this is done,

> …teachers can carefully tailor their classroom activities, exercises, homework, and so forth to the needs of a clearly defined group of students. Any teacher who has ever had to teach students with a wide range of aptitudes or abilities will easily understand the value of this benefit…. (Brown, J.D. 1995:41)

Brown seems to be making the argument that there is some correlation between the needs of students and the so-called level that is measurable by placement instruments. The subtle argument here is that it is indeed possible to group students in a clearly defined manner; doing that reveals certain needs that we are then able to address; and further that doing so is more effective than not. We are provided with no supporting research literature attesting to the veracity of these claims.

Such tailoring of activities in combination with the type of PT/SG we are discussing raises issues of parity that will be discussed further on.

Biggs and Moore

In a critique of skills grouping, Biggs and Moore (1993:164) lay out two arguments for skills grouping, or what they call streaming [italics theirs]:

1. *Classes of homogeneous ability are easier to teach.* Students are more likely to be working at the same pace and so the teacher can adjust the pace and level of instruction to suit the maximum number of students. (Biggs and Moore 1993:164)

Interesting in view of our previous discussion on general education vs. EFL, is that this argument from 'general education' echoes almost exactly two arguments from Second Language Learning: 1) Brown's argument that we can adjust the content to the level of students, and 2) Gaffeney and Mason's argument that we must teach the majority (maximum number) of students.

2. *The students are more comfortable.* The bright are less likely to be bored, the dull to feel lost. In particular, the less bright student is protected from the humiliating knowledge of how far ahead the brighter students really are.

Again it seems that the principles of 'general education' parallel closely those of EFL, or at least Gaffeney and Mason who mention the frustration of the 'weak' and the boredom of 'those much better'.

Biggs and Moore's critique the first argument in two ways, first by noting that it is "plausible on the assumption that only expository whole class methods are used". (ibid pg. 165) Again, the argument above that EFL/ESL classes are different because they don't rely on expository methods (*sic*) runs into trouble. If classes are easier to teach when homogenized only in an expository setting which SLA presumably doesn't employ, then homogenization isn't really available to us. Further, Biggs and Moore seem to imply that methods other than expository are available to, and desirable in 'general education'.

Secondly, they point out that grouping with measurements of general abilities does "not reduce variability of *particular* classes, so that in practice there is considerable overlap in performance levels between classes." (ibid

165) (italics theirs). Ted Power (2003) makes the same general observation, noting that, "[even] though a great many language teaching institutions use placement tests to group students, the classes that result are often sadly heterogeneous". (also see Gillis-Furutaka below)

The critique of the second argument mentions a study by Corno & Snow that shows it to be "true only of the students in the top streams". For the rest of the students, they cite other studies that not only show that "students feel less comfortable when streamed, but perform worse". (*ibid* pg. 165) We are able to apply their discussion to the idea that teachers must teach to the majority, when they assert,

> "[teachers] adjust *upwards* to cope with mixed ability classes, *downwards* to cope with low ability classes, so that low ability students are exposed to better teaching and better role models when in mixed ability classes, while the good students are better taught and achieve well, anyway. (Biggs and Moore 1993:166) [italics theirs]

To return to the point of whether we must teach to the majority of the class as if they were isolated from the rest, Corno and Snow (1986) themselves cite studies that show that lower ability students *perform better in mixed-ability settings* that use a small-group approach. This was in a discussion on adapting teaching to individual differences, in which they also found that "higher achievers benefited from this form of instruction." (in Wittrock 1986:263) Accordingly, it seems to be an issue less of the level make-up of classes than methods used.

Brown also addresses the 'must teach to the majority' argument in his treatment of individualizing instruction in the Second Language classroom by noting that:

> "[each] student in a classroom has needs and abilities that are unique." Usually, the most salient individual difference that you observe is a range of proficiency levels across your class and, even more specifically, differences among students in their speaking, listening, writing, and reading abilities. Small groups can help students with varying abilities to accomplish separate goals. The teacher can recognize and capitalize upon other individual differences (age, cultural heritage, field of study, cognitive style, to name a few) by careful selection of small groups and by administering different tasks to different groups. (H.D. Brown 1994:174)

In committee discussion, one justification that the placement examination was a good test was that the results after pre-test item analysis and subsequent revision yielded a well balanced bell-curve. This was to show that indeed there was a variety of ability levels. One thought experiment that I like to apply here is to take the bell-curve and apply the principles of PT/SG. As a hypothetical, say we cleave the curve at the upper 30% percentile and the lower 30% percentile, then we have three groups.

If we again apply a placement test [perhaps even the same test] to any resultant division, it is my intuition that we will produce another, smaller well-shaped bell-curve. With repeated applications of the test and divisions, the size of the curves will get smaller, but the shape will remain the same with each group producing smaller and smaller bell-curves until we come to a single student in the final group. One response to that has been that the groups that are produced

will become more alike with each application. This may or may not be true with the only definition of ability being certain test scores. But the point is that the groups will still remain of mixed ability as revealed by the bell-curves.

Comfort

For the students to feel more comfortable (hence less 'disruptive') under the PT/SG hypothesis, not only would it be necessary for levels to be homogenized and content to be adjusted accordingly, but the students would have to be aware of the levels of the other students in the class. Further, it would have to make a difference to them.

In an article evaluating a new streaming program, Gillis-Furutaka (2002) discusses student response to the grouping program at her university. After quite a rigorous screening process, using three different tests (the Oxford University Press Quick Placement Test (QPT), the Comprehensive English Language Test (CELT), and their university entrance exam scores) and then reclassifying according to mid-year performance assessments, Gillis-Furutaka expresses surprise that despite such filtering, up to 40% of the students thought that members of their class were at a different level. Across Skills and Content courses respectively, 23.5% - 25% felt that their classmates were of a higher level and 12% - 15% thought their classmates were of a lower level.

In addition, she found that after the mid-year reclassification (16% of the students were moved to a higher or lower class), among the students who were not moved, 90%+ found little or no change in class atmosphere.

Further, when asked explicitly what was important to them in terms of class make-up, her students seemed to indicate that being with students of the same level of ability was most important to 18% of the class, while being with students with whom they can work well was most important to 34.9%.

Levels

That there remains "considerable overlap in performance levels between classes" or whether classes are truly rendered homogeneous by PT/SG is questioned elsewhere. Touching on both points that the discussion of placement examinations was settled in the 1980's and on the present point, Stern states, "[how] to group students into classes that make educational sense is a much debated issue in language teaching," explaining that,

> [determining] proficiency criteria for such levels is problematic, and… even within levels there is liable to be considerable heterogeneity with respect to different aspects of proficiency, not to mention differences in aptitude motivation and other individual factors". (Stern 1992:350)

Bachman also indicates that the issue remains unsettled when he points out,

> Test designers and experts in the field disagree about what language tests measure, and neither the designers nor the experts have a clear sense of the levels of ability measured by their tests. (in Brown & Gonzo, 1995:419)

All the more is the case with the type of PT/SG I have been focusing on. The level of one's language ability shouldn't really be a function of how well one does in relation to others on the same examination. If we do indeed talk about levels of language, it makes more sense were the level in relation to achievement on a criterion scale, rather than a reference to the norm.

This brings us to the focus on the levels themselves. Two questions that are related that have a direct bearing on the pragmatic application of PT/SG are, "How many levels are there?", and "Where do we make the divisions?"

In a section on in-house measures of proficiency in designing experiments, Thomas (1994) lays out three arbitrary methods and three less arbitrary methods (her terms) of separating test respondents. These were gathered from differing studies. Of the three arbitrary methods, what she calls the "crudest technique" is a) to divide them into two groups: those who passed the test and those who failed it. A more conventional method is b) to divide them into equal groups of three (high, mid, and low). The third method is c) to divide the experimental group by boundaries that have already been set by institutes for other purposes.

The three less arbitrary methods include 1) setting boundaries by more than one standard deviation below the mean, and one standard deviation above the mean, yielding three levels with the range of the mid-level two standard deviations; 2) dividing subjects at "clear gaps" in the range of scores, and; 3) dividing subjects into three levels according to results of two listening tests, then combining the bottom two levels for comparison with the higher level.

The types of division that we have used at Kurume U. are type a) and c) of the arbitrary methods and type 2) of the less arbitrary methods. This will be discussed below. As can be expected, the divisions were not quite what one interested in dividing students by like ability should be, in any division.

Ability

Another question we should ask is, "What are we measuring?" It is the claim of those who support PT/SG that a given score reflects 'language ability', and that this translates into classroom performance. Bachman addresses this point by saying:

> …we now know that a language test score cannot be interpreted simplistically as an indicator of the particular language ability we want to measure [in this case, general language ability]; it is also affected to some extent by the characteristics and content of the test tasks, the characteristics of the test taker, and the strategies the test taker employs in attempting to complete the test task. What makes the interpretation of test scores particularly difficult is that these factors interact with each other. (in Brown & Gonzo, 1995:421) (brackets mine)

There are many more factors that go into the mix, personal as well as social. One personal factor that we may be measuring instead of raw language ability is anxiety. Bradshaw (1990:15) finds that "…an excessive degree of anxiety can have debilitative effects on the performance of some test-takers." She points out that tests that emphasize evaluation (as placement tests do) increase anxiety in "those

who are most concerned about their achievement." She cites research by Alpert and Haber (1960) that suggests the effects of anxiety "will vary according to aspects of the testing situation or of the individual test-taker." She indicates that research done by Wine (1971) found debilitating test effects "in those who were most highly test-anxious...," with Madsen (1982) finding "similar debilitating effects on the performance of adult EFL students, as shown by an examination of test correlations."

Further, according to Bradshaw, the test itself and such things as "timing, clarity of instructions and familiarity with test type," may also be responsible for variation in scores. In her study on C-test, one of the implications she draws in pointing out "the possibility of debilitative effects on test scores" is that "it may be the difficulty of test technique… which causes adverse reactions, rather than the difficulty caused by low English proficiency." (Bradshaw 1990:26)

An important example of contributing social factors comes from the findings of Skehan that

> "…there is an even correlation between the rates of syntactic acquisition in a first language. Moreover, he found a greater correlation between second language aptitude and social class and parental education. These two elements were found mixed in with vocabulary development in a factor termed family background. Not only does family background correlate with second language aptitude, it also correlates quite highly with foreign language achievement. (Gass & Selinker 1994:249)

It just may be that in our attempts to measure foreign language ability, and divide students accordingly, that we may be fostering and perpetrating the social divisions that contributed to the scores that they have produced. Perhaps, we are not so much measuring 'language ability' as a personal quality, but the quality of the education systems from which the students emerge. Studies cited by Ireson and Hallam (2001:18) seem to "demonstrate that middle class children occupied the top stream while the working class children populated the middle and lower stream." It is highly possible that this is also the case in foreign language learning where affluent parents have the wherewithal to provide their children with tutors, supply them with extra study materials, and send them abroad on homestay programs.

Actually, without a clear definition of what like ability means and how to determine the optimal score differential, there is no real way of telling if these divisions are of like ability or not.

The argument has been put forward that they are 'more alike'. The most that can be said is that scores on a test are closer than otherwise. Another thought experiment is that on a 70 point exam, it is possible that two students scoring the same 35 points could have missed every item that their counterpart got right. It would be difficult to say they are of like ability. It is possible that a different student answered only 35 questions, getting them all right, while still another answered all 70 questions getting half right. It is possible that student A answered questions in different sections of the test than student B. Yet their scores taken as simple indicators of ability would classify them as alike. There are too many possibilities that cannot be accounted for by the theory on which PT/SG rests.

Practice

I will outline briefly here how Kurume U. implemented PT/SG for two years: 2004 and 2005. The data for previous years is nonexistent and/or incomplete. Previous to 2004 the method for dividing the classes was the "clear gap" method. What this resulted in was a large mid-level cohort and two smaller groups on either end. In 2004 for reasons that I am unable to discern, the classes were divided into equal thirds. Part of the problem, as I perceive it, is that these divisions were not made according to equal thirds of the score continuum. This would have aligned better with the goal of grouping them by like ability. Instead, they were divided into equal thirds of the number of students taking the exam. The results, as to be expected, were inconsistent with the stated purpose. In other words, the groups were not of like ability as determined by scores on an examination. The differential between the top scorer and the lower scorer of each group were quite different, with the widest gap in the upper and lower groups.

The scores went from nine points on the lower end to fifty-five points on the higher end of a 60-point exam. The range of scores is shown in Table 1.

Table 1. Range of scores in 2004

2004	Range	Gap
High Group	36 ~ 55	19
Mid Group	28 ~ 35	7
Low Group	9 ~ 27	18

Partly in view of these inconsistencies, in 2005 the method of division was changed. In this instance, the number of students in the mid-group was expanded and the numbers in the upper and lower groups were decreased. Out of 27 classes available 8 were allotted to the higher level, 11 to the mid-level and 8 to the lower level. Still the divisions were made according to numbers of test takers, not test scores. This resulted in the following table.

Table 2. Range of scores in 2005

2005	Range	Gap
High Group	37 ~ 54	17
Mid Group	28 ~ 36	8
Low Group	2 ~ 27	25

Even in the absence of a definition of 'like ability', it seems obvious that the goal of grouping students along those lines has not been realized.

Parity

Having divided the students into the above groupings the question of parity arises. As originally stated, one of the purposes of separating students in such manner is to be able to adjust materials more closely to the needs of the students. In making such adjustment, what we in effect are doing is changing the content of the course. Adjustments are necessarily made in terms of quantity and quality. So that, if we provide higher groups with more (and more difficult) vocabulary; if we provide them with a wider range of grammatical structures; if

we assign them more (and more challenging) tasks, what we are doing is giving them a qualitatively different education from those in the lower groups. Yet, because of the nature of parallel placement we can make no adjustments for that in our grading. An 'A' grade given to a student in the higher group is equivalent to an 'A' grade in a lower group although the educational demands are different, This happens in spite of the fact the credits given are the same.

Even if it were possible to divide students by like ability in the manner that we have chosen, the question remains, does this affect anything?

Conclusion

Not only is definition of ability and the criteria for the divisions vague, the goals of PT/SG are never clearly stated. To be able to divide students by like ability in order to adjust materials accordingly is a proximate goal. The purpose of that should be to effect some change in either the classes as a whole or in the students as individuals. The nature of those changes is never clearly stated. Are the students supposed to learn better? Are they able to learn faster? Are the classes supposed to be more fun? What exactly, is supposed to happen? Further, in order to say that a goal has been met, some kind of assessment must take place. In other words, there must be some measure of success for a program such as PT/SG. As with any program we put in place, we expect it to be successful. How do we measure any of these aspects in order to say that we have been successful?

PT/SG is a practice that relies on a hypothesis that is untested. It is my conviction that it is untested precisely because it is untestable. Its terms are as ill-defined as its purpose and criteria for success. There is virtually no support in the research literature for its continued use. It is my recommendation that parallel PT/SG be abandoned until certain criteria are met and the questions that have been raised above answered. The time spent in implementing such a program could be better put to use in developing methods that involve all the students and motivate them to achieve what they can to the best of their individual abilities.

Rory Britto is an Associate Professor at Kurume University. He holds an MAESL from the University of Hawai'i. His main research interest is educational psychology.

Acknowledgments

I would like to thank Dr. David Griffiths, formally of York University, Canada, and Nicholas Warren, Associate Professor at Fukuoka Women's Junior College for their invaluable input in developing some of the ideas contained in this paper.

References

Anderson, L. W. & Krathwol, D. R. [eds.] (2001). A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives. New York: Longman..

Bachman, L.F. (1994). What Does Language Testing Have to Offer? In Brown, H.D. & Gonzo, S (eds.), *Readings on Second Language Acquisition* (pp. 415 – 447) New Jersey: Prentice-Hall.

Biggs, J. & Moore, P. (1993). *Process of Learning*. Australia: Pearson Higher Education.

Bradshaw, J. (1990). Test-Takers' Reactions to a Placement Test. *Language Testing 7*(1) 13-30.

Brown, H.D. & Gonzo, S. (1995). *Readings on Second Language Acquisiton*. New Jersey: Prentice-Hall.

Brown, J.D. (1995). Developing Norm-referenced Language Tests for Program-Level Decision Making. In J.D. Brown and Yamshita, S.O. *LanguageTesting in Japan* (pp. 40-47), Tokyo: Japan Applied Materials.

Corno, L. & Snow, R. E., (1986). Adapting Teaching to Individual Differences among Learners. In Wittrock, M. L., *Handbook of Research on Teaching* (pp. 605-629) , New York: Macmillan.

Fulcher, G. (1997). An English Language Placement Test: Issues in reliability and validity. *Language Testing 14*(2). 113-139

Gaffney, J. & Mason, V. (1983). Rationalizing Placement and Promotion Decisions in a Major ELT Program. *TESOL Quarterly Vol. 17*(1) 97-108

Gass, S. M. & Selinker, L. (1994). *Second Language Acquisition: an introductory course.* New Jersey: Lawrence Erlbaum Associates.

Gillis-Furutaka, Amanda and Sakurai, N. (2002). *Curriculum Change and Streaming in the department of English at Kyoto Sangyo University*. [Online] Available: <www.jalt.org/pansig/2003/HTML/2002.htm>

Ilyin, D. (1970) Structure Placement Tests for Adults in ESL Programs in California. *TESOL Quarterly 4*(4) 323-330

Ireson, J. & Hallam, S. 2001). *Ability Grouping in Education*. London: Sage Pubs.

Mosteller et.al., (1996). Sustained Inquiry in Education: lessons from skill grouping and class size. *Harvard Educational Review. 66*(4) 797-828.

Oller, J.W. (1983). *Issues in Language Testing Research*. Rowley, Mass: Newbury House.

Power, T. (2003). [Online] Available: <www.btinternet.com/~ted.power/esl0804.html>

Smith, F. (1998). *The Book of Learning and Forgetting*. Amsterdam: Teachers College Press.

Stern, H.H. (1993) *Issues and Options in Second Language Learning*. London: Oxford University Press.

Thomas, M. (1994). Assesment of L2 Proficiency in Second Language Acquisition Research. *Language Learning 44*(2) 307-336.