C

A Discussion of Norm-Referenced and Criterion-Referenced Assessment, and the Application in the Computer Classroom of an International School

Introduction

As a relatively new teacher at an International School, I am concerned with how I assess, and whether I am assessing "correctly". As well, I am constantly wondering if the methods of assessment that I use in my classroom for the Primary Years Programme (PYP) computer classes and Middle Years Programme (MYP) Technology classes are the appropriate ones to be using. What follows is a discussion of what are seen as two categories of assessment, criterion-referenced and norm-referenced assessment. In it, I will outline what these terms mean, which ones I am using, why I am using them, and whether there is a need to change my approach, to adopt more of one than the other. More specifically, should I be using one assessment technique less than I am using now and the other more.

Definition of the Terms

By definition, the task of distinguishing between norm-referenced and criterion-referenced assessment appears to be a simple one, as evidenced by the various definitions and opinions which appear in the literature. However, as discussed below, the differences between the two are not as distinct as one might believe by simply reading the definitions.

When discussing norm-referencing, Frith and Macintosh (1984 p 6) state,

> "There is a tendency to assume that comparisons must of necessity be made between individuals. This is known as norm referencing."

Cunningham (1998 p 157) claims that,

> "...in most educational settings, obvious, easily defined standards are not available. It is under these circumstances that the use of comparisons with other students becomes necessary. Such comparisons are labelled norm-referenced assessment."

Futcher (1989 p 262) discusses,

> "Its [norm-referencing] theoretical foundation is one simple assumption: various skills which are measured are equally distributed throughout the population. In each skill, it is inevitable that some will be excellent, some will be poor, and the majority will be "roughly average"."

Gipps and Stobart (1993 p 32) claim that,

> "In norm-referenced assessments all the students' scores are put into a distribution table (or graph) and a certain percentage are assigned each grade (e.g. only 10 per cent will be awarded grade A, 20 per cent grade B and so on); or a cut-off point is

> chosen for passing, allowing a certain percentage to pass and
> the rest to fail"

While all of these references to norm-referencing have the common element of comparison for the purpose of somehow ranking individuals on a scale, there are some interesting differences to note. While all of the authors above discuss the comparison of individuals, Frith and Macintosh (1984) claim that norm-referencing is used specifically due to the general tendency of people to compare individuals. This is a different explanation than what Cunningham (1998) offers, when he says comparisons are necessary in education because there simply are no standards available. Perhaps Futcher (in Murphy and Moon, 1989) is more in agreement with Frith and Macintosh (1984), when he discusses a distribution within the population. It could be that his reasoning behind this distribution comes from his own natural tendency to compare individuals, and that, in general, skills are distributed amongst the people in a society. I feel that it can be understood from these definitions, that to norm-reference is to compare the performance of individuals, and to rank those performances on a comparative scale.

In education, then, the idea of norm-referencing is to compare the students being assessed to someone else in order to somehow rank them. The example presented by Gipps and Stobart (1993) is one way of specifically distributing the results of an assessment. However, to whom the students are compared can be varied. For example, it can be inferred from several of the quotes above that the students are being compared to other students within their own class. They are being compared to those who have been taught the same thing by the same teacher. However, as was discussed in the lectures of the Assessment unit, the students can also be compared to other students with whom they have had no contact. This is the case when the students are taking entrance examinations, when they are being compared to other students at the same grade level.

Criterion-referencing is quite different from norm-referencing, although, as will be discussed shortly, they are still somewhat interrelated. Gipps and Stobart (1997 p 46) state,

> "...criterion-referenced approaches specify what needs to be done in order to qualify, and the student who meets the requirements will pass – independent of whether others do..."

Futcher (1989 p 262) says the following,

> "..a criterion-referenced test is only interested in pass or fail: it divides students into two groups, those who can and those who cannot, those who meet the criterion and those who do not. It is not concerned with differences among the very good or very poor, or among those who just meet the criterion and those who just fail on the criterion. It divides the testees into two neat groups."

Glaser (1963 p 519) discusses,

> "Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance."

Finally, Gipps and Stobart (1993 p 32) state,

> "Criterion-referenced assessments, on the other hand, are designed to reflect whether or not a student can do a specific task, or range of tasks, rather than to measure how much better

> or worse his or her performance is in relation to that of other students."

Even within the heading of criterion-referencing, it is possible to see gradations of assessment. The first quotes are discussing the difference between pass and fail, while the last discusses a gradation of tasks, or a range of tasks within the assessment. However, the generalities of criterion-referencing are clear. Contrary to norm-referencing, the important issue is whether the student can meet a given set of criteria. As Futcher (1989) and Gipps and Stobart (1993, 1997) clearly state, the issue is clearly not how students do compared to other students, but whether they meet the criteria. As well, I feel that while the others discuss specific tasks, Gipps and Stobart (1993) are also allowing for higher order accomplishments, as they discuss a range of, rather than single, tasks.

As will be discussed later in this paper, younger students in my classes are predominantly assessed along the pass/fail criterion (can he/she perform a particular task), while the older students, under the criteria of MYP Technology, have levels of achievement within the criteria. Upon reflection, I think that the reasons for this have to do with the level of learning of the classes. The younger grades are just beginning to learn specific concepts, and for this reason there are criteria for each task. For the older grades, however, these simple tasks are being incorporated within more complicated projects, which must allow for a wider range of assessment.

When comparing these two forms of assessment, it is clear that they both may have advantages, although I would argue that there can also be disadvantages. For example, in the case of norm-referencing, the achievements of a student are only known in terms of others, not whether they actually know what they should know. This consideration is discussed by Frith and Macintosh (1984), Nitko (1996) and Cunningham (1998).

Gipps and Stobart (1993, pp 32-33) also stated,

> "On norm-referenced tests there is no point in trying to get every pupil to achieve an average or above-average score since, by definition, these tests are designed to have half the population scoring above and half below the mean."

At the same time, several authors, including Nuttall and Goldstein (1984), and Gipps and Stobart (1997) discuss the idea that norm-referencing and criterion-referencing are more intertwined than originally stated.

As Futcher (1989, p 263) writes,

> "The question which springs to mind is how the criteria are set. Criterion-referenced tests have been over concerned with content validity since their inception, so that their content more accurately represents the ultimate performance they are meant to predict, and construct validity has been by-passed, so the criteria has been set in the following way: successful students in various academic courses or vocational fields have been tested, and their average scores used as the criteria for the new testees. In other words, norms for restricted population samples form the criteria. So far, therefore, criterion-referenced tests are just norm-referenced test in disguise…"

I would also argue, based on what I have seen, both in my school and within the MYP in general, that criteria are often built around what is felt, or understood, to be appropriate for that particular age group or class. Therefore, one is actually basing the criteria of criterion-referenced assessment on norms. These could be norms of previous years of classes, previous results of similar tests, etc. This concept is discussed by Gipps and Stobart (1997 p 46), who also give the following example,

> "A simple example is a swimming certificate. If a child swims the 20 metres then the certificate is hers, regardless of who else gets it. The norm-referenced equivalent would be a swimming trial from which the first three finishes qualify and the rest get nothing. It is important to recognise, however, that there are norm-referenced assumptions behind criterion-referenced measures. It would be foolish, for example, to set the criteria for the first music grade at a level which only accomplished musicians could attain, or for a driving test so that virtually nobody passed. The skill is to set criteria which are attainable by those for whom they are intended."

It is clear that the criteria for the swimming certificate are based on norms which have been established previously. It is very similar to the practice of teaching. Criteria for a specific grade level are based on the experience of previous students at that grade level, perhaps for that city or that country or for that organisation, and, therefore, the norms set by those other students. With the implementation of the MYP curriculum internationally, one would have a topic for discussion, as to whether it is appropriate to instil international norms on a local school environment.

Outline of my Practice

At the International School where I currently work, I am responsible for teaching and assessing the Computer classes for Preschool 3 to Grade 5, and teaching and assessing the

MYP Technology classes from Grade 6 (Year 1) to Grade 10 (Year 5). With the continuing growth of the school, this past autumn saw the introduction of grade 11 to our school, to whom Information Technology is now taught.

Assessment is done in a variety of ways. For the very youngest (Preschool 3, Preschool 4 and Kindergarten) students, I base much of my assessment on observation of their work. For the next levels (grade 1 to grade 3), I use a combination of observation and marking of work, which they have saved onto the computer. For the rest of the students, I use a combination of observation, marking of work and testing. Tests are usually a combination of short answer questions and multiple choice, which are often combined with a practical section that they perform on the computer.

For the MYP students (grade 6 to 10), most of the work in the classroom is based on the "Design Cycle" (International Baccalaureate Organization, 1998). Each part of this cycle has specific criteria, by which the work of the students is assessed. The marks from each part of the cycle are combined, to give an overall mark for that particular project.

Currently, as a young teacher, the construction of criteria for the lower grades has been based on a combination of a variety of curricula from other schools, for students of the same age level. This has been combined with my personal assessment of what my students have been able to accomplish in those same skill areas. As my experience increases, I am better equipped to make judgements regarding whether a particular skill is more suitable for one grade instead of another. Also, it must be taken into account that some classes, in general, are more adept at acquiring skills, and the criteria must be adjusted accordingly.

As I have now discovered, I am setting my criteria based on the norms set by a variety of other schools, as well as on the achievements of the previous students of a particular grade level in my own school. In a sense, I am trying to see whether my current students of a particular grade are more, equally, or less successful than previous students in my class, or than other same-age students in other schools. Part of the motivation behind this is to ensure that certain standards are being met, both in the learning, and the teaching, of the subject. Certainly, if a particular student or class is not as successful as his/her/their predecessors, I would want to analise my own input to be sure that I am doing the best for my students. At the same time, it is like a set of checks and balances to be sure that the criteria that has been set is appropriate. As discussed above, it might be necessary to adjust the criteria slightly dependent on how the individuals, and the class as a whole, are performing.

Criterion-Referenced and Norm-Referenced Assessment in my Classroom

Upon reflection, I realise that all of the assessment that I do in my classroom is criterion-referenced assessment. When I observe the students, I am watching to see if each student is able to perform skills that have been taught. When I mark assignments, I am assessing whether they were able perform the tasks that were assigned to them, based on certain criteria. As well, when I set a test, it is with the express desire to see if the concepts that have been taught in class have been learned. All of these tasks are carried out to see if the student is able to perform, in comparison to the criteria which has been set for that particular task. For MYP Technology projects, I am ensuring that projects meet the requirements set out in the criteria, outlined by the International Baccalaureate Organization, which has been adapted slightly for each grade level that I teach.

The criteria for different forms of assessment for PYP students can be implicit, in that there is not always a written list of criteria for the students to achieve during a particular activity. This is most often the case with observation. When I am observing their work, it is usually the case that I have given them a specific task, and I am observing to see whether they are able to carry out that task. For example, if I were to say "create a new folder on the desktop" then I am observing whether they are able to do this or if they are having any troubles performing this skill. In this case it would be implicit that I am observing to see whether they were able to follow this instruction. In general, I find that I am very aware of how I am assessing them. I take some care to ensure that I am assessing all students consistently according to the criteria. I feel, because my criteria occasionally have a range of expectations for each task, it is possible to appropriately assess the performance of each student.

In the case where PYP students are performing a more complicated task, for example, that they were asked to draw something in a drawing programme, a range of criteria might be used. The assessment would then be more detailed, to include not only the specific tasks of using particular tools, but also how the use of those tools is applied to the final product, and how that product has been created.

For marked work, there is usually a set of instructions which accompany the activity. In most cases, the instructions are both written on the board and repeated orally. In this case, the instructions are explicit, as the marks would be based on whether the instructions were carried out or not.

In the case of a test situation, the criteria are more implicit, as there have not been, up to now, an exact list of what criteria will be assessed on a particular test. It is usually the case that a list of topics for a test will be written on the board. While this list is quite specific (eg. changing the size of pictures in a photo editing application, or adding/deleting pages in a desktop publishing application), it is not explicitly stated exactly which activity will be on the test, and it is implied that you should know how to do all of them.

With the MYP Computer Technology, the criteria for the "Design Cycle" (International Baccalaureate Organization,1998) are very explicit. Each student is very aware of what needs to be done, in order to receive a given mark. However, part of the frustration of the MYP, on the Computer Technology side of things, is the fact that, while the students are working on activities based on this cycle, there is nothing to say what level of skill these same students should achieve in the use of, for example, word processors, spreadsheets, databases, or design. If a student in Year 4 of the programme is doing a Computer Technology project, and is designing a database, then there is a good chance he or she will develop certain skills in this area. However, as all students are allowed to design their own projects within the given framework of the assignment, it is possible that one student becomes very adept as creating databases and another becomes adept at spreadsheets. There are set criteria for assessing how they follow the "Design Cycle" (International Baccalaureate Organization, 1998) throughout this project, but no criteria for assessing whether or not they have achieved a certain skill level. For a student who will continue his or her studies within the MYP, this may not be an issue, but for those continuing their studies in another programme this may prove to create complications. I find this to be one of the drawbacks of the MYP programme – it doesn't seem to take into account the mastery of certain skills at certain levels, even as a general standard. It is for this reason that tests that I set are always criterion-referenced - I am interested to learn if each student can do a certain set of tasks, that is to say, meet a certain set of criteria. I use it as a tool to gauge whether the ideas and skills presented in my class have been learned, and more importantly, understood.

While the criteria sometimes make it difficult to assess specific skills (eg. computer skills) within a project of MYP Technology, the criteria are general enough to be able to assess the wide variety of projects which are created by my students. While I hesitate to agree whole-heartedly with Eisner (1985), that criteria/objectives need to be less rigid for some subjects, I feel that the criteria of MYP Technology can, as Eisner (1985 p 33) demands,

> "yield behaviours and products which are unpredictable. The end achieved ought to be something of a surprise to both teacher and pupil."

A Technology class of fifteen students, given the same outline, has shown me that it is very possible to create fifteen very different projects/responses which can all be assessed against the same criteria.

As with the younger students discussed above, it is important to apply the criteria consistently when assessing each project. Currently, I am the only teacher who gives MYP Technology instruction. As the school grows , there is the likelihood that there may be two such teachers. In that case, t would be especially important, not only to ensure that I am assessing consistently within a given group of students, but that all the teachers assessing the subject were assessing consistently in comparison to each other.

Based on the very definition of these two types of assessment, it would be impossible to get the same results when applying them to any given situation. For example, assume that a criterion-referenced test were to be given, and all the students received a grade above 70%. By the standards, this might be looked at as a good result, because all of the students "passed" the test. However, under the definition of norm-referenced assessment, the students would have to be ranked, with perhaps a passing grade given only to those who received a grade that was above the average. Therefore, whereas the criteria-based results would give the whole class a passing grade, the norm-referenced results would have had half of the class receiving a failing grade.

However, as has been mentioned earlier in this paper, and will be discussed shortly, the two forms of assessment are not always mutually exclusive. The criteria for the criterion-referenced assessment may have been based on previous students, other schools, city-wide expectations etc. which were the norms upon which the criteria was based.

As has been established, my assessment in the classroom is almost exclusively criterion-referenced. I am not able to think of a single instance where I would rank any student in my class higher or lower than any other student. It is interesting to note that I have not had any parents approach me to ask how their son or daughter is doing in my class, compared to the other students. The only situation that I could can recall is that of a student asking me how he had done on a test "compared to the others". In this case, he was asking me for a result based on norm-referencing. However, since the test was not constructed based on norm-referencing criterion, he was not able to obtain the same information that he have had the assessment been constructed based on norm-referencing.

There have been many examples in the literature of the use of norm-referenced assessment. Some of them include IQ tests (Black, 1998), standardised reading tests, national curriculum key stage tests, and GCE examinations (Gipps and Stobart (1997), in other words, the achievement of one student compared to the other students in the class.

In my classroom, I am implementing the use of various criteria, be it criteria laid out by the IB organisation, or criteria which has been collected and collated to form an assessment grid. While these may, in part, rest on norms for a particular age group, there is currently no reason for me to rank the students in any given class. It is for this reason that I feel that norm-referencing assessment does not currently have a place in my classroom. I do, however, think that there may be a time when there will be exceptions to this statement. For example, while grade 11 students of this past year were all studying the same computer subject with me, there may be a time when they have a choice between three different options. In that case, it may be necessary to rank the students, to give preference when they choose between those three options.

Are These Two Approaches Always Mutually Exclusive?

In my situation, I feel that criterion-based referencing is the more appropriate form of assessment. It gives a clear indication of how a student is doing relative to a given set of criteria or objectives. The important issue to keep in mind, is to have appropriate criteria, which can be used to asses the objectives set out for a particular task.

As a teacher, it is up to me to set the various criteria for these classes. When I arrived at the school, while there were general outlines for teaching, there was not a specific criterion-based curriculum for the teaching of Computers or Technology. In fact, while the MYP has criteria based assessment for 5 specific areas of work, there is no set criteria for the actual computer/technological skills which a student in a particular year should achieve.

It has been necessary for me to develop the curricular criteria for the various classes from scratch. This has meant searching the internet for various school curricula to see what students are learning at different stages. This would be a good example of where the norm-referencing comes into play. In these cases, I have been looking for a norm-referenced foundation on which to build my criterion-referenced curricula.

As well, it seems to me that if you create an assessment which is criterion-referenced, it is still possible to rank the outcomes of that test from highest to lowest. However, because criterion-referenced testing is designed to assess what a student knows, it is possible that a wide spread in this ranking can be seen as a reflection of how much has been learned in that particular unit, or how effectively or ineffectively something has been taught.

In the more senior grades, while I am not required to do norm-referenced assessment, I am, perhaps sub-consciously, doing just that. I am aware that, in general, there are some students who do better in my class than others. I am able to anticipate quite accurately how a student will do on a test, and am surprised when a student does not perform as I expect on a test. It is important, as was discussed earlier, to apply the criteria consistently and uniformly for all students.

However, we are reminded by Frith and Macintosh (1984 p 7) that there can be a third method of assessment:

> "This polarisation between norm referencing and criterion referencing, as the latter form is called, is, as Rowntree (1977) reminds us, rather misleading in that it is too narrow. He points out that we can assess the performance of individuals by comparison, with some predetermined criterion (criterion referencing), or a norm established by colleagues (norm referencing), but it is also possible to judge them against their own previous performance."

This third example of referencing, commonly known as "ipsative-referencing" or "self-referencing" (University of Western Australia, 1998), is something that I also use in my classroom. When students do their first projects based on the "Design Cycle" (International Baccalaureate Organization, 1998), they are not expected to receive near perfect marks. However, as students better understand the cycle and its application, they are better able to address the various aspects of the criteria in their work. It would stand to reason, then, that if a student worked diligently, that his or her mark would tend to improve as her or she worked on more projects. This is something that I have witnessed in my classes. One class, where two students received final project grades of 1 and 3, respectively, at the beginning of the year, are now receiving marks of 3 and 6, respectively. This shows an improvement, not compared to the rest of the class, but relative to their own work, over time. There is a very important observation to be made. While the individual might be on the "low end of the scale" in terms of the rest of the class (norm-referencing), he or she is making improvements in respect to his or her own previous achievement.

Conclusions

In general, norm-referenced assessment and criterion-referenced assessment are two very distinct forms of assessing students. While it has been discussed that criterion-referenced assessment may have part of its foundation in the realm of norm-referencing, the results of these two forms of assessment are interpreted in very distinct ways.

Norm-referenced assessment is specifically set up to rank the students performing a particular task. This ranking is divided into categories based on the mean result of the group undergoing the assessment, or some other norming group, with part of that group inevitably falling into the "failing" or "below average" category.

Criterion-referenced assessment is assessment which gives specific criterion, or goals, that are to be achieved. In this case, if all those who are assessed reach these goals, then they are showing a certain level of performance. In criterion-referenced assessment, it is possible for all of the students in a particular group to achieve a grade of "pass".

At my school, as a teacher of Computers/Technology, I am now more aware that, while my criteria are perhaps based in certain norms, the assessment that I am doing with my students is criterion-referenced. At this stage of the school's development, in my classroom, there has been no need for any student to be labelled "the best", and certainly not "the worst". There may be a need for this in the future, however, until then, let the student be assessed according to specific objectives. If there needs to be a comparison, let the student be compared to how they themselves have worked in the past. What matters most is that progress is assessed consistently, on an individual basis, according to clearly stated criteria.

References

Black P (1998) Testing: Friend or Foe? Theory and Practice of Assessment and Testing, London, The Falmer Press

Cunningham G K (1998) Assessment in the Classroom: Constructing and Interpreting Tests, London, The Falmer Press

Eisner E W (1985) The Art of Educational Evaluation: A Personal View, London, Falmer Press

Frith D S and Macintosh H G (1984) A Teacher's Guide to Assessment, Leckhampton, Stanley Thorne Publishers Ltd.

Futcher G (1989) Measurement or Assessment: A Fundamental Dichotomy and its Educational Implications in Murphy P and Moon B (Ed.) Developments in Learning and Assessment, London, Hodder & Stoughton Educational

Gipps C and Stobart G (1993) Assessment: A Teacher's Guide to the Issues, 2nd Edition, London, Hodder and Stoughton

Gipps C and Stobart G (1997) Assessment: A Teacher's Guide to the Issues, 3rd Edition, London, Hodder and Stoughton

Glaser R (1963) Instructional Technology and the Measurement of Learning Outcomes: Some Questions American Psychologist, No. 18, pp 519-521

International Baccalaureate Organization (1998), Middle Years Programme Technology, Book One, International Baccalaureate Organization (IBO)

Nitko A J (1996) Educational Assessment of Students, 2nd Edition, Englewood Cliffs, Prentice-Hall Inc.

Nuttall D L and Goldstein (1984) Profiles and Graded Tests: The Technical Issues in Effective Assessment and the Improvement of Education – A Tribute to Desmond Nuttal, Murphy R and Broadfood P (Ed.) (1995) London, The Falmer Press

The University of Western Australia (1998) Some Key Concepts in Assessment Issues of Teaching and Learning Vol 4 No 5
(located at http://www.csd.uwa.edu.au/newsletter/issue0598/assessment.html)