THE TROUBLE WITH THE CURVE: AN ARGUMENT FOR THE ABOLISHMENT OF NORM-REFERENCED EVALUATION

By

GREGORY RAYMOND

University of Windsor, Ontario.

ABSTRACT

The norm-referenced evaluation system has been used to grade students, from elementary to post-secondary, for decades. However, the system itself is inherently flawed. Looking at the history of the norm-referenced system and its most famous tool, the Bell Curve, and taking examples from the author's own teaching experience, this paper examines the erroneous logic that makes the system, as a whole, invalid for grading at any level, particular for college or university. The paper goes on to propose alternatives to the norm-referenced system. Examining first a self-referenced system, and ultimately finding it, too, lacking, the paper turns to criterion-referenced evaluation. Criterion-referenced evaluation is shown to be not only the best choice for evaluation at a post-secondary level, but also proves to be the only viably fair system available to teachers in today's, grade-emphasised education system.

Keywords: Evaluation, Norm-Referenced, Self-Referenced, Criterion-Referenced, Bell Curve, Grading, English, Composition.

INTRODUCTION

Teachers have to give grades to students. It is a simple fact of post-secondary life. Students want to know how they are doing in a class, and schools want to know how a class is doing in its institution. Because there is so much vested in grades (for students, teachers, and schools), there is a huge drive to get grades "right." For similar reasons, grades also have to be defensible; if a student questions why he or she received a grade, a teacher must be able to explain. Finally, grades must be clear, even though, as Tara Fenwick and Jim Parsons (2000) explain in the book *The Art of Evaluation: A Handbook for Educators and Trainers*, "because learning involves more than just the physical, using only numbers is limiting" (p. 13); the system must somehow allow for clarity through fairly ambiguous and simple number or letter grades. This desire for correct, clear, defensible grades has lead to the creation of numerous different systems of evaluation, including the one that is in use at most Western post-secondary institutions now: normative-referenced evaluation. As a Composition teacher, however, the author has come to the conclusion that this particular system is inherently flawed. Normative-referenced evaluation (NRE) is useless when it comes to producing meaningful grades in Composition classes, and is unfair throughout the educational institution. While there are certain tools in the NRE arsenal that are particularly poignant examples of errors in the evaluative processes of post-secondary institutions, the entire system contains too many faulty assumptions and unfair distinctions and should be replaced.

The Reasons to Grade

Before we can look at the systems we use to grade, and why they do or do not work, we need to understand the purpose of grading. Essentially, before we can (or, at least, before we should) make decisions about the *how* of grading, we must understand the *why*. Why do we, as teachers, give grades? Here we must make a distinction – the difference between assessment, evaluation, and grading. These terms are often used interchangeably, and there is little agreement on exact definitions. For the readers purposes, however, the author shall use these definitions: Assessment is the explanation to a student of what that student did well, or is skilled at, and where he or she must improve. Grading is the ranking of a student based on a pre-existing scale of numbers (GPA), letters (A-F), or percentages. Evaluation is the combination of both, feedback and rank given together. Some teachers do all three of these throughout a class; some do only one or two.

Assessment, evaluation, and grading serve different purposes, and, as such, all have their own place in a post-secondary class.

Pure assessment is primarily a feedback tool. It gives a student an idea of what to work on. Evaluation, however, can serve multiple purposes. Fenwick and Parsons (2000) offer nine reasons for evaluation: (i) to compare performance to goals; (ii) to help learners make decisions about future actions; (iii) to monitor student's progress; (iv) to assess the teaching methods; (v) to revise the program; (vi) to provide information for the institution; (vii) to assess the learner's background knowledge; (viii) to determine learner satisfaction; and (ix) to develop self-assessment (p. 15-6). While these are all good purposes behind the assessment half of evaluation, how do grades work in this context? Do grades help students improve? Help them understand their mistakes? Warn them of the path they are treading in terms of academic achievement? Possibly the main uses of grades are to rank which students get which bursaries, and to satisfy the criteria set forth by our employers. But why do universities require a grading system? In Writing Relationships, Lad Tobin (1993) says that "grades are currently an integrated, even central, part not only of our academic institutions but also our entire society" (p. 60). It would be difficult to disagree; admission, scholarships, bursaries, and employment often hinge on students achieving and maintaining specific grades. But why is this the case? Why do institutions require students to achieve a certain rank or score to determine if they "learned" enough to get a credit for the course?

The answer to these questions lies in the differentiating of two terms: summative evaluation and formative evaluation. According to Fenwick and Parsons (2000), "summative evaluation occurs at the end of a unit or course of study. Its purpose is to summarize what the learner has accomplished and the growth that has taken place" (p. 27), whereas "formative evaluation occurs during the learning activities. Its purpose is to give feedback to learners about their progress or growth" (p. 28). We see, then, that it is summative evaluation that is required by post-secondary institutions; a single, holistic grade (whether it is a letter grade, percent, or grade point system) that reflects how successful a student was in a course. Using grades in this way allows for an external audit of how successful or unsuccessful a student was (in a quick, easy way, as opposed to having to read through notes to and on every student) and, by extension, how successful and unsuccessful a class is. Summative evaluation gives the grades that will follow a student through his or her academic career, lurking on transcripts and meddling with GPAs. For that reason, the purpose behind summative evaluation is somewhat out of the hands of individual teachers; it must be used to compare the student's "performance as it is, with performance as it should be" (Biggs & Tang, 2003, p. 164) and to express where they fall in that comparison.

Formative evaluation, on the other hand, is a more versatile tool that teachers can put to different uses. Feedback on essays, marks on small activities, and notes jotted in the margins of journals are all formative feedback that can help show students what they did well and in what areas they need to improve and give students an idea of where the currently stand in terms of grade. The grades in formative evaluation, then, serve purely as information for the student, an explanation (or, occasionally, a warning) of what they can expect in the future. The University of Windsor, where the author teaches, there is a policy that states that students must be given feedback constituting at least 20% of their final grade prior to the Voluntary Withdrawal deadline for the semester (Senate Policy: Policy F2, 2004). This is, on a very basic level, an enforcement of formative evaluation; it summarizes what the student has done *so far*, while still promising the chance to move up or down from that point by the final, summative grade for the class. Of course, the fact that this grade is often a summative grade of one major assignment or midterm test that cannot be revised or changed makes this policy less helpful to the students. It remains, however, an example of an institutionally enforced (almost) formative evaluation tool.

We see, then, that the why of grading depends on the situation. It would be nice to think that these two types of evaluation, formative and summative, are viewed equally by students. Realistically, however, because of the permanent nature of summative evaluation and how it

affects multiple aspects of a student's academic life, that is often the one that students are most concerned about. As Lad Tobin (1993) puts it, "grades remind us and our students that there is always a bottom line" (p. 59), and it is that bottom line that many students focus on. For that reason, we will focus here on how teachers create summative grades for their students.

The first step in deciding a grade for a piece of work is making a decision: is the work evaluated by looking at it in sections, or as a whole. Focusing, for example, on a Composition class assignment, should teachers give a pre-determined number of points for each aspect of an essay (grammar, format, thesis, supporting argument, tone) that is done "correctly," and then total those points up to determine grades? Or should teachers look at how well the essay works as a whole, and determine the grade it should receive based on that? In Teaching for Quality Learning at University, John Biggs and Catherine Tang (2003) suggest that the answer here depends on the purpose of the grading:

> *Analytic marking of essays or assignments is a common practice. The essay is reduced to independent components, such as content, style, referencing, argument, originality, format, and so on, each of which is rated on a separate scale. The final performance is then assessed as the sum of the separate ratings. This is very helpful as formative assessment; it gives students feedback on how well they are doing on each important aspect of the essay, but the value of the essay is how well it makes the case or addresses the question as a whole. The same applies to any task: the final performance, such as treating a patient or making a legal case, makes sense only when seen as a whole. (p. 183-4).*

Grading essays by sections, then, becomes useful in formative evaluation, allowing for process feedback on which areas need improvement and which areas are done well. For any form of summative assessment (our focus here), though, the grades need to be approached from a holistic viewpoint; the essay must function as a whole to pass. Biggs and Tang (2003) go on to suggest that some critics argue that this holistic method is faulty because it is too subjective, but they respond by pointing out "awarding marks is a matter of judgment too, a series of mini-judgements, each one small enough to be handled without a qualm" (p. 184). Analytic or holistic, any marking a teacher does is going to contain some measure of subjectivity. The decision between the holistic or component-based evaluation, then, has to be made based on which makes more sense. For assignments that require a whole product (as opposed to tests, with multiple, individual questions), it only makes sense to grade the assignment as a whole.

Assuming, then, that summative evaluation after the activity and/or semester should determine how successfully a student performed and that summative grading on activities in Composition class should use holistic grading, we can see the importance of a clear and fair manner of evaluation or grading. After all, if the techniques used to determine grades are not fair and consistent, then how can they accurately reflect performance? If they do not accurately reflect performance, how can we fairly use them as the academic currency they have become? The drive for "accurate" and "fair" methods of evaluation has lead many institutions to focus on norm-referenced grading. The most (in) famous tool (amongst students, at least) in the norm-referenced system is one often reviled by students – even if they do not, in fact, know exactly what it means: The Bell Curve.

Purpose: Examining the Problematic Nature of Norm-Referenced Evaluation

The History of the Bell Curve

The bell curve (more formally a "normal curve" or "Gaussian curve") is a graphic representation of the Gaussian Function. In its original (mathematical) use, it represented a probability model which holds that "most phenomena occur around a middle point, while few occur at either the high or low extreme ends" (Fendler & Muzzaffar, 2008, p. 63). For example, if you recorded the results of flipping fifty coins one million times and graphed the frequency of those results, the graph would, theoretically, be a Gaussian curve with the most common results resting around the middle (a roughly equal distribution of heads and tails) while results towards either extreme (all heads or

all tails) would be less common.

The bell curve's presence in education is largely due to Francis Galton, a nineteenth century scientist. Galton created a measurement model which "postulated the existence of stable traits that, in a randomized (and large) population, were distributed along a bell curve" (Potter & Baker, 2011, p. 11). In other words, the graph depicting probability of phenomena could be used to demonstrate the level of a given skill (for example, mathematical aptitude) in a large, random population; some would excel at math, most would be average, and some would be poor. In The Truth About Testing, W. James Popham (2001) explains how this idea was adopted by the U.S. Army during World War I, when they instituted a test for new recruits called the Army Alpha text. This test forced recruits to demonstrate skills in multiple areas, such as spatial awareness, mathematics, and vocabulary; the results were compared to those of previous test-takers, a norm group. This comparison allowed the new recruits to be ranked along a normal curve to see their own standings. "Put briefly, the Alpha was intended to permit comparisons among test-takers' relative intellectual abilities (as defined by the test's items)[…] The Alpha, therefore, was a predictor test and, as such, was definitely an *aptitude* test" (Popham, 2001, p. 41) and, as an aptitude test, this form of testing was met with very good results. As such, "the Army Alpha's assessment strategy became the template for almost all of this nation's subsequent standardized testing, irrespective of whether that testing was supposed to serve as an aptitude assessment or an achievement assessment" (Popham, 2001, p. 42). Once the slide into education began, it became commonly accepted as fact. Lunn Fendler and Irfan Muzzafar (2008) reflect on this:

> *It has been made reasonable to assume that in any educational task, a few people will excel, most will be satisfactory or average, and a few will fail. Because so many people believe that the bell curve represents the way things are in nature, the idea of a normal distribution has been naturalize in education and, to some extent, in U.S. society at large. (p. 64).*

The school system, and the post-secondary system in particular due to the important nature of their summativegrades, began to rely heavily on the bell curve.

Where Normative-Referenced Evaluation Fails

This expectation that a class worth of students' academic results would naturally be bell curved, however, was not a logical one. First, one of the largest problems with assuming a normal curve in the classroom is that the normal curve is taken from a large group; most classes simply do not have enough students for a normal curve to establish itself. While there are some classes, especially mandatory, low-level ones, which can have hundreds of students, the majority do not, and a number of different studies have found that the bell curve simply does not hold true in groups of 60 or fewer students (Fenwick & Parsons, 2000, p. 124). Additionally, returning to Galton's proposed model, this notion of a normal curve in human skill level is found in a *randomized* group. However, as Biggs and Tang (2003) put it, "the ability of our students is not likely to be normally distributed because our students are not randomly selected" (p. 171); to be in such a class in the first place, a student must be intelligent (after all, most schools require a certain grade average for admission) and would be more likely have some pre-existing knowledge of the subject than a randomly selected person from outside of academia. Finally, the notion that students' grades should fit on the bell curve is flawed due to the time of when grading takes place. Consider the Army Alpha; it was given to *new* recruits to assess their abilities untrained. In academia, however, grades (especially the ever-emphasized summative grade) come at the end of an activity or learning period, *after* the teaching process. This means that a teacher is no longer evaluating a randomized assortment of people; rather, the evaluation is of a group of people who have all been given the knowledge necessary. If the point of teaching is to improve the knowledge base of the students, then one must assume that, after the teaching has taken place, the students should all rank as above-average in the area that was taught. They should know more than the majority of randomized people who have not necessarily been taught the subject. In A Primer on Authentic Assessment, Michael K. Potter and Nick Baker (2011) phrase this idea quite succinctly: "if the goal of teaching is to help students learn, then effective teaching narrows or

eliminates a 'spread' of grades" (p. 14). For these three reasons, it is erroneous to assume that classroom grades will naturally spread themselves along a bell curve, especially for end-of-term, summative grades.

While there are numerous reasons the assumption that grades should naturally spread along the bell curve is incorrect, the assumption itself is not the major problem with this form of normative evaluation. The real problem begins when teachers adapt their grades to fit the curve, regardless of whether they naturally do or not. For some time, a common belief among many universities was that the letter grade C was average "in that C grades were given to students who neither excelled nor failed relative to their peers. C was the 'norm'" (Potter & Baker, 2011, o. 10). This holds true at the University of Windsor, where the very first rule of the Faculty of Arts & Social Sciences Grading Policy states "Instructors in large enrolment lower level classes should grade so that the 'average' grade, or expected performance of the average student, is within the 'C' range" (Grading Policy – Faculty of Arts & Social Sciences, 2001). If C should be average (the highest point in the bell curve), then, the logic goes, the majority of people should get a C and the grades are worked to fit this notion regardless of the quality of work the class provides (Potter & Baker, 2011, p. 11). If fifty percent of the class receives a score of eighty-five out of one hundred, than that eighty-five percent grade is "curved down" to be worth a C. Similarly, if the majority of the class receives only twenty percent, than that twenty is "curved up" to be a C. The problem is that this invalidates the performance of the students, because, even if every student in the class does exceptionally well or exceptionally poorly, the teacher is limiting the number of students who can receive an A or an F. The entire concept of grading along the bell curve makes grades arbitrary and malleable from one situation to the next, which removes any possible belief of the "fairness" of grading.

If the bell curve is such an inappropriate tool for grading, then why is it used at all? Biggs and Tang (2003) suggest that "grading on the curve also appeals to administrators, because it conveys the impression that standards over all departments are 'right', not too slack, not too stringent" (p. 174). If too many students score too highly in a class, the class must have little value because it is easy, known colloquially as a "bird course." On the other hand, if too many students fail, the class still has little value, because the teacher is clearly failing to deliver information to the student. The bell curve, then, creates the illusion of perfection; since the majority pass, the teacher is doing his or her job, but since only a few are ranked very high, the teacher is not being too easy. This "goldilocks effect" of needing things to be balanced just right, however, results in the need to adjust grades to reflect those "just right" conditions, rather than to reflect the actual work being produced by a group of students.

The bell curve clearly has no place in university grading. However, it is merely one (albeit one very infamous) tool in the much larger system of normative-referenced evaluation. NRE is any method of grading which "compares one learner's performance to others in the same group and is governed by the belief that a 'normal' standard of particular skills, understandings, or attitudes will emerge for a particular group" (Fenwick & Parsons, 2000, p. 40). While the bell curve is the term that most students are familiar with from NRE, it is not the only technique. Any time one grade is taken primarily through comparison to another student's work, it is a form of NRE, and is therefore haunted by the same problems that follow the bell curve, namely that it focuses more on the most common result and how the students' work compares to it, than on the quality of the work itself. Royce Sadler (2009) explains that "although rarely stated explicitly, the rationale behind grading by proportions is the classic market approach to regulating value when there are no stable, independent reference points" (p. 816). The fact that there are always a limited number of A grades to be given out (regardless of how many students may complete A-quality work) raises the perceived value of that A. At the same time, however, this sacrifices the integrity of it.

Whether it is through bell curving or through some other technique, NRE is an inappropriate method of evaluation for university studies. It is based on the unsustainable assumptions discussed above. It is a system in which the grades are arbitrary, and "education quality and student learning [becomes] irrelevant" (Potter & Baker, 2011, p. 14).

Perhaps most harmful of all to the students, however, is that NRE creates the belief among students that "it matters where they stand in relation to their peers" (Potter & Baker, 2011, p. 12); how well a student does is not only compared to, but is directly based on, how well every other student did. As Sadler (2009) says, "knowing relative standings may be important for some purposes, but rank ordering should follow from, not lead, the determination of grades" (p. 809); a student should be evaluated on the quality of his or her own work and given a grade that reflects that, not a grade that reflects the quality of work as it compares to all of the other students.

Recommended Alternatives to Normative-Referenced Evaluation

If we accept that normative-referenced evaluation is not appropriate for university classrooms, how should we grade? Consider the Composition class at the University of Windsor, where the final assignment is a portfolio of writing samples. There is no answer key for this assignment, no checklist for the points made, or multiple choice Scantron to rely on. How, then, can a teacher fairly decide between a C submission and a B submission? I am reminded of my high school English classes, and how often teachers suggested that "it is impossible to get one hundred percent right in writing." Surely, if grades are to be meaningful as cross-discipline currency, it should be just as possible to get perfect in an English portfolio as it is in a math exam. The desire for a fair and effective grading system leads to two options: self-referenced evaluation and criterion-referenced evaluation.

Implications of Self-Referenced Evaluation: Benefits and Detriments

Self-referenced evaluation tries to grade students based on how they have improved; it "compares what the learner understands or can do today to what he or she understood or could do in the past" (Fenwick & Parsons, 2000, p. 41). This, the theory suggests, allows any student to get a good grade if they work hard enough, and avoids the trap of norm-referenced evaluation by putting no limits on how many students can get any specific grade. One of the simplest ways to do this would be a straight comparison between work early in the semester and work late in the semester, or to hold a "pretest" and "post-test" to determine if the students have gained knowledge or skills as a result of the course. Popham (2001) points out, however, that "the traditional pretest/post-test design doesn't work – at least, it doesn't work if you're trying to determine whether teachers have been effective" (p. 129). This is because it is impossible for a teacher to make the two tests equal in difficulty. If the tests are different, than one is inevitably going to be more difficult than the other. If they are the same, however, then you encounter the problem that Biggs and Tang (2003) call "backwash," where "students learn what they *think* they will be tested on" (p. 169). If they have already encountered the material in test form, they will naturally be extra vigilant when learning that specific material; in effect, the identical post-test will be evaluating how well the students can remember the questions from the pretest – which questions they answered correctly and which ones they did not – and not how effective the course was in relating *all* of the material to them. Potter and Baker (2011) suggest a second major problem with self-referenced evaluation: that it "would make the grade a matter of rewarding improvement instead of achievement" (p. 14). The problem here is a matter of fairness – how can a teacher fairly evaluate students based on improvement when students may be coming into a course with different levels of ability? Imagine a Composition class in which there is a first year English as a second-language (E.S.L.) student and a student who received straight A's in her high school English classrooms. Realistically, the likelihood is that the student with a solid background in English will produce superior pieces of writing throughout the course; however, it is also quite possible that the E.S.L. student will improve quite drastically, while the English student remains at the same level through the course. Would we then give the E.S.L. student a better grade than the English student, because the E.S.L. student had improved by a greater degree than the English student? Students do not all come into a classroom with the same level of skill or knowledge, so a self-referenced evaluation system cannot be used to grade fairly across a number of students.

Implications of Criterion-Referenced Evaluation: A Possible Solution

What teachers need, then, is a method of evaluation in which the grades can be held up for scrutiny and still be found fair and accurate, and teachers should be able to explain what would qualify as an A level of work before any work is produced by students. Fenwick and Parsons (2000) suggest the answer is criterion-referenced evaluation (CRE), where a teacher "compares a learner's performance to an absolute, external standard or criterion" (p. 39). In other words, CRE requires a teacher to provide criteria for an assignment in advance, and to give those criteria to the students. It should detail exactly what is needed at each grade level, and all grading should be made by comparing the work completed with that list of criteria. Biggs and Tang (2003) put the idea for CRE forward in these relatively simple terms: "Say what you want students to be able to do, teach them to do it and then see if they can, in fact, do it" (p. 177). The major advantage that a criterion-referenced system has over a normative-referenced system is that, as grades are decided based on a comparison of work quality to a pre-set, external rubric, evaluation allows teachers to identify how well a student performed and grade independently of how any other student scored. This allows grades to be more concrete, and less variable from classroom situation to classroom situation. Potter and Baker (2011), who call CRE "standards-referenced grading," explain the difference between this system and the norm-referenced system:

> *In essence, the difference between norm-referenced and standards-referenced grading parallels the difference between arbitrariness and subjectivity. Norm-referencing forces us to divorce grades from actual achievement, which means our grading decisions are necessarily arbitrary and indefensible. Standards-referencing, on the other hand, involves subjective judgments that can be defended by reference to the criteria specified and the standards from which they are derived. (p. 18).*

Criteria-referenced or Standards-referenced grading is a more constant, defensible, and fair system than any normative-referenced system.

There are some critics who suggest that the criterion-referenced system leads to problems of non-specificity and fairness; without reference to and comparison with other work, how can a teacher clearly and fairly defend why one student received a B and another received an A? The answer begins with the standards or criteria the teacher is using to grade. Potter and Baker (2011) write that "criteria should be as clear and explicit as possible about what their relevant standards require" (p. 17), and it is this explicit criteria that allows for clear grading explanations. But then, the opposition to CRE suggest, is it not a subjective decision that determines who gets what grade? In short: yes. CRE does require some amount of subjective decision-making: "yes, *this* paragraph does what the criteria ask, and no, *this* one does not." This kind of subjectivity, however, is not a bad aspect of criteria-referenced grading. At least, it is an unavoidable aspect that is no worse in CRE than in any other system. Consider Lad Tobin's (1993) succinct point: "Assessment is never objective or clean; it is never easily and painlessly resolved" (p. 58). The simple fact is that grading is, at its core, a subjective, opinionated act. In CRE, however, the subjective decisions can be defended by reference back to the assigned criteria that were given to the students with the assignment. It is subjective, yes, but it is not unfair or indefensible.

Even with these defences of CRE, many universities still favour normative-referenced systems for their policies; however, that does not mean there are no examples of functional CRE systems in universities. Biggs and Tang (2003) explain that "despite the prevailing norm-referenced cast of mind at undergraduate level, the sheer logic of criterion-referenced assessment is generally seen in assessing theses and dissertations" (p. 180). Due to the individual work and unique subject matter of theses and dissertations, when a Master's or Doctorate level student submits such a paper, there is no work to compare it against. Normative-referenced assessment becomes impossible. Similarly, because the student only submits one such paper (rather than numerous, smaller papers), self-referenced assessment is equally impossible. Therefore, the only option is to compare the student's work to an external set of pre-determined properties, in other words, by using CRE. If such a system is not only functional, but necessary, at a graduate level, then it clearly works and should be equally successful at an undergraduate level.

Implications of the Abolishment of Grading: The Future?

There is one final alternative to the normative-referenced grading that permeates the post-secondary institution, and it receives varying amounts of positive and negative feedback: the abolishment of grading altogether. Rather than assign what some instructors feel to be too arbitrary a rank (either percentage or letter grade), teachers can award students who complete the course with something as simple as a "credit" or "satisfactory" mark. The advantage to non-graded courses, the logic goes, is that "the learners are more receptive to feedback. Because they are not worried about losing marks and figure that if they work hard they'll get credit, they often relax and focus on improvement" (Fenwick & Parsons, 2000, p. 128). Non-graded courses would make giving feedback easier for the teacher as well. The feedback that you give to your students are suggestions, not evaluation. Fenwick and Parsons (2000) point out that "adults have difficulty being evaluated" (p. 25) because, by the time they are in university, most adults "consider themselves competent, self-reliant, and self directing [but] are once again in a learning situation" (p. 25). By removing the grading from courses, adults may no longer feel that their belief in their abilities is in conflict with what the teacher thinks of them. The author has certainly encountered this problem of a student's opinions of their own ability differing from his (and it was, the author believe, compounded by the fact that he was not too much older than the student in question and so did not have the mantle of parent-like authority that some teachers have). The author particularly remember a student who questioned, not the feedback that he gave her, but the grade it amounted to. Further discussion, though, revealed that she understood and agreed with his comments, including the comments highlighting unclear language or grammatical mistakes. When the author asked what it was she did not understand about her grade, her response was that she "was always an excellent writer in high school," and so clearly she believed should still be graded as such in university. The author believes that the comments, removed from the grade, would have benefited this student in a way that they did not when the grade was given, because she could look no further than the grade.

Unfortunately, the abolishment of grading is currently not a real option for most courses. As Potter and Baker (2011) put it, "while [the abolishment of grading entirely] might be a desirable goal, it isn't feasible at this point because many scholarship, graduate school, and employment decisions are made on the basis of grades" (p. 34). There is far too much interest vested in grades as academic currency, proof of aptitude, or intelligence for them to be removed from the post-secondary system. For now, at least, grades are a necessity.

Conclusion

There are many ways and many reasons to evaluate students. While formative assessment is helpful during the course of teaching, it is summative grading that will follow the students and affect their academic careers. Similarly, while analytic evaluation of parts may work for formative assessment, it is holistic grading, in Composition at least, that must be used to reach the summative grade on work. Whatever the reason for grading, if there is any hope for grades to be fair, accurate, and defensible, the grading must be done in a criteria-referenced system. Clear, detailed criteria must be given to students when the assignment is first assigned, and it is in reference to those criteria that the teacher must determine how successful or unsuccessful a given piece or work is. As Royce Sadler (2009) says, "students deserve to have their work graded strictly according to its quality, without their responses on the same or similar tasks being compared with those of other students in their group, and without regard to the students' individual histories of previous achievement" (p. 809). While grading may be an unavoidable necessity of the post-secondary institution, the teachers, can work to make it fair. The first step towards that is the abolishment of a normative-referenced system of grading and its most infamous tool, the bell curve.

References

[1]. Biggs, J., & Tang, C. (2003). *Teaching for quality learning at university*. Maidenhead, England:McGraw-Hill, 2003.

[2]. Fendler, L., & Muzaffar, I. (2008). The history of the bell curve: Sorting and the idea of normal. *Educational Theory, 58*(1), 63-82.

[3]. Fenwick, T., & Parsons, J. (2000). *The art of evaluation: A handbook for educators and trainers*. Toronto, ON: Thompson Educational Publishing.

[4]. Grading Policy – Faculty of Arts & Social Sciences. (2001). *University of Windsor*. Retrieved from http://web4.Uwindsor.ca/units/fass/fassTop.nsf/831fc2c71873e46285256d6e006c367a/828d04255c1494d08525778b00515d18/$FILE/GRADING%20POLICY.pdf

[5]. Policy F2: Provision of meaningful feedback to students on their in-course performance prior to the voluntary withdrawal deadline. (2004). *University of Windsor*. Retrieved from http: //web4.Uwindsor.Ca/units/senate/main.nsf/SubCategoryFlyOut/2C5F23B2BDB061FE8 525791100568767

[6]. Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

[7]. Potter, M. K. & Baker, N. (2011). *A primer on authentic assessment*. Windsor, ON: University of Windsor, Center for Teaching and Learning.

[8]. Sadler, R. (2009). Grade integrity and the representation for academic achievement. *Studies in Higher Education, 34*(7), 807-826.

[9]. Tobin, L. (1993). *Writing relationships: What really happens in the composition class*. Portsmouth, NH: Boynton/Cook Heinemann.

ABOUT THE AUTHOR

Gregory Raymond is a Graduate Student in the M.A. English: Language and Literature program at the University of Windsor in Ontario, Canada, where he previously studied both English and Drama. He is currently researching English as performative language in theatre, and the unique applications of the oral tradition of English. He teaches Composition to non-English major undergraduates.