Investigaciones

Systematizing a qualitative evaluation experience. Toward a new concept of the learning process

Felipe Tirado Segura*
Alejandro Miranda**
y Ana Elena del Bosque***

Translator: Pablo Contreras Fresán
E-mail: deepcolearning@gmail.com

* Professor, UNAM-FES-Iztacala, Head of the Psychology & Education Program, SNI member.
Email: ftirado@unam.mx
** Professor, UNAM-FES-Iztacala, psicólogist dedicated to Education on line and culture process on the internet.
Email: correo@alejandromiranda.org
*** Associated Professor, UNAM-FES-Iztacala, Psychology Department.
Email: aneldelbosque@gmail.com

Approach: 08/09/10 • Acceptance: 08/07/11

REVISTA DE LA EDUCACIÓN SUPERIOR
ISSN: 0185-2760
Vol. XL (3), No. 159
Julio - Septiembre de 2011, pp. 9 - 28

Resumen

En este trabajo se reporta una investigación en la que se sistematiza un procedimiento de evaluación cualitativa a partir de la valoración hecha por los propios estudiantes de ensayos realizados por sus compañeros. El propósito es inducir el aprendizaje de competencias complejas, tales como la reflexión, el análisis crítico, la argumentación, la formación de juicios sustentados, a partir de la elaboración y evaluación de ensayos colectivos.

Palabras clave:

- Evaluación cualitativa
- Competencias complejas
- Aprendizaje

Abstract

This paper reports on the research of a systematized qualitative assessment evaluation process carried out by the students themselves on their peers' essays, as a way to induce the learning of complex skills, such as reflection, critical analysis, arguing and issuing supported judgments, through collective essay writing and evaluating.

Key words:

- Qualitative evaluation
- Complex competences
- Learning

Conceptual Framework

Schooling systems are highly structured and formalized, which makes them resistant to change and unable to cope with the continuous and rapid changes occurring today. The transformation is both necessary and urgent in order to solve problems and take advantage of new circumstances. One problem is the huge unmet demand for higher education therefore it is imperative to find ways to increase enrollment as well as to improve the quality of education. This is thought to be possible if we tap the many benefits now offered by digital technologies for delivering education and developing new ways to conceive the educational process.

The educational model does not necessarily have to be restricted to the classroom learning model, there is the opportunity of taking advantage of online connectivity and building mixed learning modalities. Mixed models, commonly referred to as blended learning or b-learning, combine distance and/or online learning with on-site or classroom learning. It is assumed that when these modalities are blended advantages can be taken from both. It is true that mixed models have coexisted with distance education for a long time, it was common that educational content and exercises delivered by mail also required students to attend regular, sporadic or voluntary, face to face tutoring sessions. Nowadays, however, the circumstances of distance education have profoundly changed because of the potential of computer assisted instruction, particularly when internet connections are available (Aretio García-Ruiz and Domínguez, 2007).

From the 1960s to the 80s, a whole field of experience designed to make use of computer-aided instruction (CAI) was developed. Under the teaching machine approach based on the 1950s behaviorism, ways to customize education were designed to give every student the opportunity to advance at their own pace, at the time and place that would best fit them.

Then in the 1990s, the speed of new microprocessors, the exponential expansion in memory capacity, higher quality multimedia resources (images, audio, video), hypertext and hypermedia links and networking via internet, had produced a revolution in the production and dissemination of information which was quickly used to improve educational strategies, raising concepts as diverse as Computer Based Education, Computer Based Training, Computer-Based Learning, E-learning, Teaching-Learning Environments, and so on.

In particular, communication via internet is seen as a boon for distance education systems, because it enables asynchronous communication, i.e. without having to interact simultaneously, in addition to abating the problem of distances, establishing communication in an expeditious manner and with different partners, which allows the development of management systems for social networks. These characteristics give rise to new educational approaches such as Computer Supported Collaborative Learning (Bartholomew, 2004; Stegmann, Weinberger and Fischer, 2007; Laurillard, 2009; Liguori

and Ritella, 2010, Mihaela and Stoicescu, 2011, Whittaker and Bonakdarian, 2011) which can be combined (mixed) with classroom education modalities. The experience referred to in this paper is supported under this model.

Schooling systems are usually designed under an on-site learning paradigm, in which a group of students in a classroom is required to focus their attention on a presentation and arguments put forward by a lecturer, usually the teacher. Over time it has been observed that such a system tends to inhibit the divergence of ideas due to the natural handicaps it has for operating synchronously with diversity. A wide range of views cannot be addressed at the same time, consequently, this is not a paradigm that encourages students' creativity and critique. Most of the time arguments are held up by the teacher, therefore the model focuses on the teaching function and not on actions or learnings by the students.

In a learning process what the students do, not teachers, is most relevant (Biggs, 2003). The development of understandings in the appropriation of knowledge, as well as creative and critical thinking, have to come from each student and from their own reflections, hence an active student-centered model is required, which would allow them to build and develop their knowledge, beliefs, skills and attitudes, as is proposed in the constructivist model of education (Estes, 2004). It is therefore necessary to have a divergent model in which students have the opportunity to develop their own ideas and creativity, rather than focusing on the teaching function. The aim is to build models of divergent and distributed teaching among students themselves (Tirado, Bustos and Miranda, 2007).

The multiplicity of approaches that can arise when students have the opportunity to develop their own ideas, makes it impossible for a single teacher to attend to them all, hence a dynamic alternative may be the distribution of tasks and the integration of students into teams. These teams can be convened to develop essays in collaboration on the course curriculum topics, selected according to students' preferences, which opens the possibility of having a divergent model for each of the students to develop their creativity in their own arguments while analyzing them in collaboration with their partners, which makes for a distributed learning model based on co-construction (construction in collaboration with peers). The co-construction of essays can be regulated and induced by criteria established early in the course and defined as the criteria for both the work and the evaluation, to provide guidance on the parameters and regulatory activities of the course, as they guide the student's performance in addition to being the goals expected to be reached during the course and therefore to be evaluated on.

Writing essays is very favorable for promoting reflection, critical thinking and creativity, in that it involves expressing ideas in writing. Writing is the best instrument of thought, an ideal way to develop, define, express and communicate one's ideas (Olson, 1998). Writing is a resource for rethinking the writing, which allows text to be changed as often as desired, until it reaches the satisfaction of the person who writes, hence achieving their best results in terms of expression. When comparing oral expression of certain

ideas transcribed in verbatim with the same approach but expressed through writing done with some care, it is shown that there is great difference obtained through written expression.

The development of essays in collaboration operates through a process of distributed cognition, where cognitive processes are distributed around a collaborative environment, the group behavior united by common goals does not operate as the sum of isolated individuals, instead their actions take place within a sociocultural environment in which they are influenced and empowered by the group, given the interactions that occur between people and resources (aims, content, materials, devices, rules, tasks, ideas, arguments, counter-arguments) these are brought into actions (i.e., an essay), where learning and doing are inseparable ("learning by doing"). This is an exercise conducive to the promotion of ideas and opinion counterpoints, dialogic reflections and using arguments to support statements. This enables the production of more solid work as the product of the sum of ideas, a combination of efforts, can achieve better results (Bearison and Dorval, 2002). It is clear that two people think better than one and three better than two.

Collaboration also encourages respectful dialogue, making commitments and promoting shared responsibility, which shifts the focus of the evaluation, often based only on the academic content results, to another that considers collaborative construction as a learning process, as well as promoting substantive principles that contribute to building social and ethical practices of accountability and civic conviviality (Tirado, 2007).

Baker (2009) makes a number of remarks on the importance of formative assessment and accountability as processes that can improve instruction by taking measures that effect classroom teaching. Baker also notes that a system of accountability involves making people more responsible and may contribute to the reliability and validity of their assessments. Evaluation may serve different purposes, it must show "instructional sensitivity" for it is only reasonable to create opportunities for learning from the materials that would be measured in tests. Sensitive assessments are required to open the opportunity for students to perform meaningful high quality tasks, that may reflect their dominion over the complex content of a course, that would allow them to evaluate classroom practices, homework and project assignments.

Not without reason Canales and Gilio (2008) have pointed out that evaluation in Mexico has focused on the quantitative properties of the educational system –factors associated with performance, such as percentage of failing grades, hits on the PISA tests, student teacher ratios, the number of full-time faculty, the number of teachers with graduate degrees, the number of publications, etc.– and has neglected the assessment of the teaching-learning process itself, therefore this type of evaluation cannot result in an improvement of such process.

The concept of evaluation has changed over the centuries, from evaluation focused on the act of judging, which was gradually replaced by more technical measuring notions out of which the psychometric view gained

prominence, continuing to be used to this day, supplemented by other, rather more qualitative oriented categories for approaching the evaluation of the training process. Thus various comparisons have been built, some focused on population distribution and others on achievement criteria.

Now we intend to take another step toward an evaluation based on the execution of constructed responses, where students demonstrate their competences on the tasks they are expected to execute in the contexts where they are required to perform them, hence this approach has been referred to as authentic evaluation. This new vision seeks that those performing the evaluation process stop conceiving it as independent from teaching and learning, emphasizing that evaluations are conducted as part of the educational process (Ahumada, 2005; Diaz Barriga, 2006, Castillo and Cabrerizo, 2007).

One substantive function of evaluation is to learn about how the educational process is being conducted, in order to improve it. Hence Ahumada (2005) suggests designing teaching and evaluation in terms of the dominion of processes, in a personalized and differentiated way, giving greater importance to the evaluation's diagnostic function than to its administrative function, considering its productive nature and making use of multiple techniques and assessment tools to allow for an authentic evaluation as an alternative, evaluating so that the implementation of competence-based programs, where students start tackling tasks, from the beginning, with the intention of solving real problems, therefore making learning more functional, applied and significant. In this way the focus is on the learning processes, on the student's doings and not on those of the teacher, thereby turning it into a formative evaluation rather than a summative one.

An evaluation of this type should provide information to students about their learning process, so that relevant corrections to it can be made. It is important that students take responsibility for their own learning and use this evaluation as a means to enable them to achieve the proposed knowledge and competences, so student participation is central to this evaluation modality, thus emphasizing the potential of self-evaluation and co-evaluation.

Evaluation is usually the domain of teachers, who are also responsible for translating it into a score for the accreditation of the course, but this process can be socialized, incorporating students to contribute to it. Co-evaluation involves several benefits such as: students taking a more active participation in it, making the evaluation process horizontal for students to learn to evaluate issues they have previously studied and prepared, which promotes critical and responsible thinking requiring them to refine and support their judgments, on top of being a review of what was previously studied. This, in turn, favors meta-cognitive self-regulated learning, for it requires students to become aware of their rights and wrongs, and as Baker (2009) has indicated reliability and validity elements of the evaluation may also increase.

Castañeda (2006) notes that in a community of learning and practice where goals, tasks and responsibilities are shared, co-evaluation represents an important tool to support students becoming familiar with criteria, values and learning goals as well as to develop skills of discussing and arguing.

It allows to focus on how to carry out the tasks assigned for learning throughout the program, and is no longer just a way to assign a final score at the end of the cycle.

Many teachers resist co-evaluation on the assumption that it is risky. They assume that judgments by students may not be grounded, lack substance, present biases due to sympathies or antipathies, or agree on benefits for mutual convenience. For co-evaluation to be effective, it is important to request that students expose arguments and point to empirical evidence to support their claims. This implies that learners have to reflect on what they do and how they do it, as well as to assess the results of their work, it requires to be clear that learners are not being evaluated, but rather the activities, work and products performed.

To do the co-evaluation, learners must have the relevant criteria, these must be explicit and put forward for their consideration from the outset of the course. To accomplish, the teacher must first plan and write these criteria through headings, which are assessment scale guidelines for establishing the proficiency levels related to the performance a learner must show in relation to the process and specific products. Such headings represent a formative evaluation as they enable determining the quality of implementation based on an exercise of critical reflection in evaluating tasks that do not have right or wrong answers, but instead encourage an assessment within a wide range of qualifying criteria that go from performance ranging from incipient, that would be typical of a novice, to the level of an expert who has full dominion, that are then referred within a range of assessment criteria (very bad-bad-regular-good-very good), which then would be translated to an ordinal numerical scale (1-2-3-4-5) which allows for mathematical operativity.

The criteria answer questions such as what features characterize the performance of a specialist or an expert, and what are the characteristics that distinguish between excellent, good, average and poor performance. It is important to review the criteria with students before carrying out the task or activity to be assessed by this instrument, so that they learn in advance the criteria by which the performance will be evaluated; thus reducing the subjectivity in the co-evaluation exercise (Diaz Barriga and Hernandez, 2000; Quesada, 2003, Ahumada, 2005; Diaz Barriga, 2006).

Procedure

The development of the experience upon which this work is based is supported by the considerations referred to above that depart from three approaches. The first is that the evaluation can not only account for, but also induces the educational process. The second is that this process can be induced through the development followed by the co-evaluation of collective essays, which represent complex tasks, as they are the written expression of ideas that require critical and creative contributions expressed through well-supported arguments. The third is that students are able to

perform the co-evaluation in a responsible, sustainable, reliable and valid manner.

Evaluating the quality of an essay is a complex task, for it is inevitably based on subjective appreciation judgments, as it aims to assess the quality of a complex body of ideas. However, it is possible to make a less subjective evaluation if it is narrowed by previously defined clear criteria (headings) and is based on peer reviewed judgments expressed and debated collectively, which could not occur when only one teacher is doing the evaluation. The peer review neutralizes, if not at least mitigates, some of the subjectivity if it adheres to standards and expresses the foundations that underlie the judgments, which can be further evaluated (meta-evaluation = evaluation of the evaluation).

In our view it is assumed that co-evaluation by peers offers three great advantages, the first is that judgments come from similar and shared circumstances (peers), the second, is that by being collective it brings a plurality of judgments and the third, and most important, is that it impacts the students' learning process. It is well known that what is evaluated becomes important, so attention is paid and students seek to perform the expected actions in order to obtain a good mark. Hence, evaluation largely determines the work of students, in the same way an evaluation strategy can be designed and used to induce learning, letting students know in advance the criteria and procedures to be used. Therefore it is desirable to provide students with the regulatory criteria to be used in the headings, so that from the beginning of the course they try to comply with these precepts in the process of drafting their essays. For example, our students were told that the essay would be evaluated according to the organization, clarity, sufficiency, bibliography, and so on.

Evaluations have been limited to measuring learning achievement, not to induce learning, whereas the two functions can be compatible and complementary. The ability of students to participate in the evaluation process (co-evaluation) as an educational resource, is being wasted, because, as already indicated, it is believed that if students are in charge of evaluating they will be partial and unjust, or they will give high scores to each other in an act of complicity, or conversely, they will make each other fail in retaliation for rivalries. This study shows that all of this is not true, or does not have to be so, given that systematic procedures can be developed for responsible and substantiated co-evaluations to be performed by students.

To transfer the function of evaluating to students constitutes a formative principle of educational relevance, as it entrusts students' capabilities to make substantiated judgments and in a responsible manner, promoting their ability to exercise a fair assessment. In making the evaluation horizontal it becomes transparent, dimming the teacher's power to establish marks in utter discretion, which is also relevant as it contributes to the student's civic education.

In our procedure, the teams for doing collaborative essays were conformed based on choosing a course topic that was found to be of interest by

the students. All topics were presented in the classroom seminars throughout the semester. For each topic there were at least three teams, so the first team could evaluate the second, the second the third and the third the first, so that cross evaluations could be made while preserving anonymity. Furthermore, in evaluating the team essays and not individuals, the potential negative effects of giving low grades due to personal rivalries, or high grades in reciprocity, vanishes. But the most important aspect of this procedure is that all students developed a theme for their own collective essay throughout the semester, on the same topic that they later had to evaluate, so they were particularly familiar with the subject, even to the point of becoming more knowledgeable and up-to-date on the subject than the lecturer of the course. Thus the co-evaluation is also an overview of the subject worked during the semester, but now seen from another perspective, from the critical reading and reflection that demands from them to make their own valid judgments about the different aspects that are subjected to the evaluation using the headings. According to Bloom et al. (1956) taxonomy, and its revision (Anderson et al., 2001), to make judgments (evaluation) is one of the most important cognitive skills, so it must be induced during the formative process.

Thanks to computer resources and online media it is possible, in a virtual classroom, to promote co-evaluation easily and expeditiously, which also creates an environment that helps students develop their digital skills. This way there can be a detailed record of every action in databases that allow the whole process to be transparent, which in itself is of great importance for social legitimacy and the credibility of the evaluation. Electronic formats can be used that are easy to answer and correspond to the digital environment that is now so familiar to students.

Example - Contribution: There are original contributions by the authors

○ Strongly agree ○ Agree ● Disagree ○ Strongly disagree

Argument to support your view: _____

Important aspects for collaboration can be evaluated and thus induce students to work collectively:

Example - Respect: Offensive adjectives were used on arguments exposed

○ Strongly agree ● Agree ○ Disagree ○ Strongly disagree

Argument to support your view: _____

Working through collaboration and the co-evaluation are central to the course objectives, as they are intended to develop two formative principles in students: a sense of responsibility towards the commitments that ought to be made, as well as respect based on being appreciative and considerate towards peers as well as to induce the willingness to reach agreements through consensus.

Hypothesis:

If the evaluation is systematized through teams of peers (diverse views), under well-defined specific criteria (headings), and through a procedure that preserves anonymity (impartiality), well supported qualitative evaluations can be obtained on complex tasks (the preparation of essays) from the judgments of students themselves.

Scenario:

The research was conducted with students from two undergraduate groups who were enrolled in the fifth semester of psychology during the 2010-1 school year, at UNAM (FES Iztacala). The course was designed under a mixed modality that included both classroom seminars in which we reviewed all of the course curriculum, and activities that take place in a virtual classroom online (we used a Moodle platform) where essays and evaluations were conducted through collaboration.

In order to develop teamwork skills, essays were done by teams (4 to 5 members each) using a collective-writing platform (Wiki) embedded in the virtual classroom, where there were also forums and chat rooms to support the deliberation on the essay development.

A list of course curriculum topics was provided to students in the virtual classroom and they were asked to give them ordinal numbers according to their own preference, so they could be assigned to a team that would write an essay about one of their thematic preferences. At least three teams (A, B, C, ...) were integrated for each subject so that team A could mark essay B, team B essay C and C essay A, so no teams were evaluating each other's essay, thus avoiding possible bias by complicity.

Students electronically marked the essays in the virtual classroom, where all the headings were visible, thus considering the 7 properties for the evaluation. These are Organization: The order or structure for presenting the ideas is adequate; Clarity: It is well written, the text is easily understood; Sufficiency: It touched on the subject's most relevant points, Treatment of the case: it properly addressed and analyzed a case study according to the framework developed; Contribution: There are original contributions by the authors; Conclusions: There is a good work of synthesis and highlighting the contributions of the work, supported by the case; Bibliography: The references are up-to-date and relevant to the subject. The evaluation of each category was based on a Likert scale with a supported response, which expresses the degree of agreement or disagreement with the statement provided, writing the reason underlying such evaluation.

Example - Contribution: There are original contributions by the authors

● Strongly agree ○ Agree ○ Disagree ○ Strongly disagree

Argument to support your view: _____

Finally, at the end of the course students were invited to evaluate different attributes of the course experience using a Likert scale, as described above. One of the attributes is for them to express their views on the experience of co-evaluation, for which they were asked 8 specific questions.

Example - Favoritism: Some classmates show favoritism while marking

○ Strongly agree ● Agree ○ Disagree ○ Strongly disagree

Results

The analysis of the meta-evaluation (evaluation of the evaluation) reported in this research was done under the principle of correspondence and consistency of the evaluations performed by students, based on two instances and conditions. The first was to assess the consistency of scores settled by the students. The second was an assessment of scores based on the arguments put forward as support, departing from the estimates made independently by three professors of the course (Judge A, Judge B and Judge C).

Subjects

We worked with two groups (a and b) of the undergraduate degree program in psychology, one with 36 students (a) and the other with 27 (b). Of the 63 students enrolled 12 dropped out (19%), the remaining 51 were integrated into 14 teams to develop the thematic essays, ensuring that teams in both groups had more or less the same number of students.

Rigor of students evaluation

The first analysis was to assess the consistency of the evaluations made by students regarding their colleagues' essays. The first observation was to determine whether the scores were differentiated (showing variance), showing no "ceiling" effect of giving only extremely high scores, or conversely, the "floor" effect of giving only low scores. Only three students had showed no variance assigning all items the same score. Three others did a self evaluation of their own essays, not those prepared by their peers, and were therefore discarded and not considered in this analysis.

The average score for all essays was 49.9 on a scale from 1 to 100, meaning that on average they are not high nor low scores, but they range on an intermediate level. The essay that received the highest score was 57.7 points and the lowest 40.2, which does not show a great difference (17.5 points). Average variance was 0.43, reflecting there was a heterogeneity of scores, acknowledging differences in the quality of components of the essays evaluated. This dispels the prejudice that if students are in charge, good grades are given away, or on the contrary, that they will give low grades due to potential rivalries. No student was observed showing this type of behavior, and it may be said that scores were assigned rigorously.

Reliability is a substantive condition in any evaluation procedure. It refers to the extent an evaluation can be trusted and how accurate the diagnosis is. For example, if there are two judges evaluating the same test, it could be expected for the two to make the same diagnosis, it would be reliable if they give the same score, or at least a very similar one. A method often used is to have three judges, to see if they all coincide in their assessment, or at least two out of the three, therefore

having teams of students evaluating offers very favorable conditions for assessing reliability.

To assess the level of reliability of scores awarded by different students, comparisons on their assessments were made. Out of the 322 (46 times 7) scores by the 46 students under the 7 headings (1- Organization, 2- Clarity, 3- Sufficiency, 4- Treatment of the case, 5- Contribution, 6- Conclusions, 7- Bibliography), 203 coincided, i.e. there was an agreement in 63% of cases. As an indicator of internal consistency of scores, we obtained the value of Cronbach's Alpha, which was 0668. Although this value is acceptable, it is not very high, yet there were 14 different essays, hence correlations were obtained for each team that evaluated the same essay, correlating their scores and average scores in each of the 7 headings, assuming that the average corresponds to the best estimate of the group consensus. The average of all correlations was 0.59, the student who produced the best correlation was 0.97, 19 students (37%) obtained very good correlations of above 0.70, 17 (33%) were between 0.50 and 0.70, the remaining 15 students (29%) had relatively low correlations (between 0.50 and 0.20).

These estimates show that 70% of evaluations by students are highly consistent, were made responsibly, with rigor and therefore they are considered valid and reliable.

Assessment of student evaluations

To make a careful estimate of the rigor of assessments made by students, three judges (the course teachers) independently reviewed one by one the arguments offered by students to support the scores they awarded. An assessment was made on whether students stuck to the explicit criteria contained in the headings, using a Likert scale of 4 values (0- very bad, 1- bad, 2- good and 3- very good).

When reviewing the arguments in support of the evaluations, as already indicated, it was found that three students had self-evaluated, not realizing that they were supposed to evaluate their peers essays and not their own, so they were discarded in this analysis, consequently the number of evaluations amounted to 48 students in 7 categories, which makes for a total of 336 judgments, large enough to assess the quality of evaluations by students.

The purpose of this analysis is to show whether teachers could recognize the quality of evaluations by students, by coinciding (reliability) in recognizing which of the students were good, average or poor evaluators, that stuck to the criteria presented under headings for each item.

The level of agreement between teachers' assessments was estimated from the degree of correspondence, as measured based on the number of agreements (matches in the ratings), by the correlations between the average generated by the 3 judges (teachers) and the scores by each teacher, as well as the heterogeneity of the scores measured by the variance, and finally, according to the differences between the mean scores given by each teacher based on an analysis of variance (corresponding to the null hypothesis: there are differences in the judgments of each evaluator).

Number of agreements:

Based on traditional psychometric procedures for estimating reliability departing from three judges, the number of agreements, when three or two of them gave the same rating, were counted. Of the 336 arguments to support the scores given by students (48 students - by 7 items = 336), the three judges gave the same rating on 32 occasions (9.5%), at least two of them agreed 224 times (66.7%) and there were disagreements among the three in 80 cases (23.8%), which allows us to note that in most cases (76.2%) at least two teachers agreed to give the same rating. Of the 256 cases (32 + 224) in which at least two judges agreed with the same rating, Judge A agreed in 69.5% of cases (178 times), Judge B coincided in 60.9% of cases (156), and Judge C in 82% (210).

Considering the agreed scale (0- very bad, 1- bad, 2- good and 3- very good), none of the 48 students got rated as "very bad" based on the opinion of the 3 judges, 5 (10.4%) were considered "bad", 39 (81.2%) "good" and four (8.3%) were rated as "very good". That is almost 90% (89.6%) of students were considered good to very good evaluators, with discrepancies in the remaining 10% of cases.

The average of the evaluations done by the three judges on the 7 items was 1.99, which according to the scale corresponds to "good." In the range of variation on the average score, out of the three judges, Judge C (- 0.15) was the closest, and Judge A (0.53) was the farthest. Whereas the plausible values are between 0 and 3, meaning that this deviation (0.53) with respect to the average was of only 17.6%.

Correlations

We correlated the marks given by each of the judges with the average of the 3 marks given by each student (48) in each of the 7 items. All correlations were positive and most above 0.70, implying that the correlation was very high. Judge A obtained an average correlation of 0.78, the lowest was 0.57 in the category of Conclusions and the highest 0.87 in the Bibliography. Judge B had an average correlation of 0.81, the lowest was 0.73 in the category of Conclusions and the highest of 0.87 in Sufficiency. Judge C's average was 0.69, the highest being 0.87 in Treatment of the case and the lowest 0.38 under the heading of Clarity. This correlation indicates that this judge, in this area, was far from the average evaluation, but this was the only correlation that obtained a value under 0.50. All correlations are statistically significant at p. <0.01. In this analysis, Judge B was the closest to the average scores variance and therefore the best.

Variance

It may be argued that a judge that does not generate variance in his or her views is a poor judge, as everything is the same and therefore she or he does not discern. The lower the variance the more the judge gave homogeneous scores, and conversely, the higher the variance the more heterogeneous. Variance was calculated for each of the three judges in the 7 categories, and the average variance for each of them. Judge A had an average variance of 0.51, in the case of Judge B it was 0.50 and Judge C's 0.51, i.e. the heterogeneity of scores was almost the same for all three cases.

Variance must be assessed systemically taking into account other indicators of weight as was done in this study (number of agreements, correlations, size of discrepancies, deviations from the mean), and given that homogeneity of the scores does not necessarily indicate that the judge has no ability to discriminate, because it would be feasible, although improbable, that the performance was very good or very poor in all areas and therefore a good judge may assign a consistent score which does not necessarily mean that the judge has evaluated poorly.

Difference between means

Finally, another way to assess the level of agreement between two judges is a test to estimate whether the difference between the mean and the variance in the scores is statistically significant with respect to the other judge, so that if there are no significant differences it can be said that the scores are equivalent. The difference between the scores of one judge and another are well within a range of variance that is not statistically significant, so it could be said that it would not matter which judge did the evaluation because the result would be virtually the same. This was estimated running an analysis of variance factors, with post hoc Scheffe multiple comparisons, with a significance level of $p < 0.05$ (Silva, 2004).

For this analysis we used the SPSS processor, comparing each of the 7 items, Judge A's means with Judge B's and C's, and the same with Judges B and C. Of the 14 possible comparisons for Judge A (7 items with B and 7 with C), only in one case were there no significant differences with Judge C. Judge B and C did not differ significantly in 6 of 7 items, i.e. in 85.7% of cases their judgments were equivalent. In this analysis, Judge C was the best, given that in the 7 items his or her average grades were statistically equal to at least one of the two other judges.

This analysis is rigorous because it assumes that the judgment of the assessment is given on one same scale, resulting in an excessively objective view, while there are indeed elements of variance. Each judge did a separate assessment of the evaluation, e.g. in one case Judge B and C had virtually no difference coinciding in 86% of cases, however Judge A agreed with Judge C in only one case. These scores mean that Judge A has a value of 2.51, Judge B

of 1.61 and Judge C of 1.84, it can be said that Judge B was more rigorous in his or her assessments, this is closer to Judge C's, whereas Judge A seemed to consider that students evaluations were overall better. It is like using two rulers to measure the same height, one in centimeters and the other in inches, there will always be fewer inches than centimeters, but the evaluations could be completely equivalent. Another way to interpret this is: one judge thinks that 6 inches is high enough and another thinks that 8 inches is sufficient. This inter-subjectivity is a source of uncontrolled variance in the assessment of complex tasks. What could be expected is that there is at least ordinal correspondence: an agreement that these are poor, these are average, these are good evaluators, as is reflected in the correlations.

Students' opinion on the co-evaluation

Finally we analyze the 8 questions in which students were asked their views on the co-evaluation experience. This assessment was done openly and voluntarily after completion of the course, the number of students who entered the virtual classroom was low (N = 25), however this information allows us to learn about the prevailing views on the co-evaluation.

The questions and answers were as follows. First question: Students were asked if they considered that some of their peers marked with favoritism. The majority (40%) declined answering choosing the "I do not know", which is perhaps the best answer because they certainly could not know, 32% said no and 24% said yes. Second question: whether they thought it was desirable for students to contribute to the co-evaluation. 68% agreed, 16% disagreed and the rest abstained (18%). Third question: whether they thought the way the co-evaluation was carried out was appropriate. 52% said yes, 36% no and the rest abstained. Fourth question: Was it convenient to do the co-evaluation? 64% agreed, 24% disagreed, the rest abstained. Fifth question: Was it objective? The majority (52%) felt that it was, 12% did not think it was and the remaining 36% answered "I do not know" or declined to respond. Sixth question: Did some peers act in bad faith in the process? 52% said no, 12% said yes and the remaining 36% abstained. This question is complementary to the first, it is interesting to note that students are more inclined to think someone acted in bad faith than to mark with favoritism. Seventh question: Participants were asked if students under the supervision of teachers should be responsible for the evaluation. Here opinions were divided, 44% agreed and the same percentage disagreed. Then the eighth question: whether students should always be solely responsible for the evaluation. In this question, 88% disagreed and only 12% said yes, it is noteworthy that there were no abstentions here.

It is interesting to note based on these opinions that students are uncertain about the arbitration by their peers, and would clearly prefer the evaluation to be supervised by teachers, which seems a relevant indication.

At the end of the last questionnaire students were invited to openly express their comments, critical remarks or suggestions about the course. In

this section, only two students referred to the co-evaluation, one expressed he was wary of being marked based on his participation in the development of a collaborative essay instead of his own personal work. Another student felt the co-evaluation should be combined with the teacher's evaluation. In fact this is so, as the co-evaluation of the essays represents a score which is weighted with many other elements, such as participation and personal contributions, compliance with the course readings, participation in seminars, attendance and punctuality among other performance indicators.

Conclusions

There are three issues of great relevance for higher education in Mexico. One is the high number of school-age youth that do not study, whether because of dropping out and/or being rejected due to lack of space at public universities. Part of the problem is that the school model is expensive, consuming personnel (teachers), facilities are insufficient and easily overwhelmed by ever-increasing demand. There is also evidence of failure in keeping pace with the rapid changes that are currently taking place. Hence, it is necessary to find new ways to offer more viable alternatives, such as distributed education, involving learners more actively in their own educational processes, transforming the role of the teacher, while ensuring that decision making that affects students is made to be more horizontal.

Another problem is the quality of the educational process, given it is predominantly focused on a traditional model, of a teacher who instructs, where attention is convergent in what the teacher presents and does not promote students' entrepreneurial spirit, who are instead asked to wait for instructions, passively and react only to respond to instructions received. Students are not taught to be proactive, to self-regulate themselves, developing their own initiatives, creativity and originality. Why is it then not considered essential to encourage the pursuit of divergent teaching approaches, focusing on the students while promoting their initiatives, critical thinking and creative capabilities.

The third challenge is to promote students' ethical development, shaping civic life, developing social ethical values of respect for others, respect for the principle of reciprocity, shared responsibility and commitments undertaken in collaboration for joint projects, as well as responsible co-evaluation of academic performance by peers through elaborating on critical, thoughtful, fair and well supported judgments.

Hence, this work is intended as a contribution for rethinking the educational process, placing students at the center of this learning process, based on three main approaches: distributed learning, divergent teaching and co-evaluation, all integrated into a mixed strategy combining in-classroom and virtual learning. The results of this investigation suggest that the co-evaluation conducted by students was carried out responsibly, fair marks were given, that were neither high nor low, which shows they are consistent and therefore reliable.

The assessment carried out by three of the course's professors mostly coincides with the arguments stated by students to justify the scores they awarded to their peers, implying that the criteria of the headings were clear. Also the average score was very close for the three judges, correlations were quite high, levels of variance were almost the same, leading to the conclusion that about 90% of students can be considered good evaluators, who are rigorous and reliable to undertake such a complex task as that of evaluating an essay, which involves the assessment of complex cognitive processes through writing.

Challenges remain in distributed learning, because not all students share the same competences nor commit themselves to the task in the same measure, which creates disruptive inequalities. Arrangements must be devised to induce and better regulate the collaborative work, where shared responsibilities and work distribution are equitable. One significant problem arises when due to lack of compliance with commitments by one or more students, the collective work falls either short or has to be compensated by the work of another student. This results in substantial and justified disagreements, making the evaluation of the collective essay difficult. The mechanism used to mitigate this problem has been to follow up on individual performance in the virtual classroom while weighing this based on several indicators, albeit new ways must be found to address this problem.

References

Ahumada, P. (2005). *Hacia una evaluación auténtica del aprendizaje*. México, Paidós.

Anderson, L. W. y Krathwohl, D. R. (Editors); Airasian, P. W.; Cruikshank, K. A.; Mayer, R. E.; Pintrich, P. R.; Raths, J. and Wittrock, M. C. (Contributors). (2001). *A Taxonomy for Learning, Teching, and Assessing A Revision of Bloom's: Taxonomy of Educational Objectives*. New York. Addison Wesley Longman.

Baker, E. (2009). Consideraciones de validez prioritaria para la evaluación formativa y de rendición de cuentas. *Revista de Educación*. 384, pp. 91-109.

Bartolomé, Antonio (2004). Blended Learning. Conceptos básicos. *Revista de Medios y Educación*, 23, pp. 7-20.

Bloom, B. S. (ed.), M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. (1956) *Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*. New York: David McKay. Bearison, D. J. and Dorval, B. (2002). *Collaborative cognition*. Westport: Albex.

Biggs, J. (2003). *Teaching for Quality Learning al University*. Open University Press. McGraw Hill.

Castañeda, S. (2006). *Evaluación del aprendizaje en el nivel universitario*. México, UNAM.

Castillo, S. y Cabrerizo, J. (2007). *Evaluación educativa y promoción escolar*. Madrid, Pearson Prentice Hall.

Canales, A y Gilio, M. (2008). La actividad docente en el nivel superior: ¿diferir el desafío? In Rueda. M. (Coord.) *La evaluación de los profesores como recurso para mejorar su práctica*. México, UNAM, Plaza y Valdés Editores.

Díaz Barriga, F. (2006). *Enseñanza situada: vínculo entre la escuela y la vida*. México: McGraw Hill.

Díaz Barriga, F. y Hernández, G. (2002). *Estrategias docentes para un aprendizaje significativo; una interpretación constructivista*. México, McGraw Hill.

Estes, Ch. (2004). Promoting Student-Centered Learning in Experiential Education. *Journal of Experiential Education*, 27(2), pp. 141-161.

García-Aretio L., Ruiz M. y Domínguez D. (2007). *De la educación a distancia a la educación virtual*. Barcelona, España, Ariel

Laurillard, D. (2009). The pedagogical challenges to collaborative technologies. *International Journal of Computer-Supported Collaborative Learning*. 4 (1), pp. 5-20

Ligorio, M. B. & Ritella, G. (2010). The collaborative construction of chronotopes during computer-supported collaborative professional tasks. *International Journal of Computer-Supported Collaborative Learning*. 5 (4), pp. 433-452

Mihaela Sgera F y Stoicescu D. (2011). Using blended learning as a tool to strengthen teaching competences. *Procedia Computer Science*. 3, pp. 1527–1531.

Olson, D. R. (1998). *El mundo sobre el papel. El impacto de la escritura y la lectura en la estructura del conocimiento*. Barcelona, Gedisa.

Quesada, R. (2003). *Cómo planear la enseñanza estratégica*. México, Limusa.

Silva, A. (2004). *Métodos Cuantitativos en Psicología, Un enfoque Metodológico*. México, Trillas.

Stegmann, K., Weinberger, A. y Fischer, F. (2007). Facilitating argumentative knowledge construction with computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*. 2 (4), pp. 421–447

Tirado, S.F. (2007). "La formación ciudadana y la ética social". En Vidales, I y Maggi, R. (Compiladores) *La democracia en la escuela. Un sueño posible*. México, Colegio de Estudios Científicos y Tecnológicos del Estado de Nuevo León.

Tirado F., Bustos A. y Miranda A. (2007). Enseñanza divergente, diferenciada y distribuida a partir de un aula virtual. Memorias: *IX Congreso Nacional de Investigación Educativa*. Consejo Mexicano de Investigación Educativa, Facultad de Educación, Universidad Autónoma de Yucatán, Mérida, Yucatán.

Whittaker, T. y Bonakdarian, E. (2011). Face-to-face experiences for online students: effective, efficient, and engaging hybrid classes. *The Journal of Computing Sciences in Colleges*. Volume 26 (4), pp.140-148.