Meta
Journal des traducteurs
Translators' Journal

The Good Guys and the Bad Guys: The Behavior of Lenient and Demanding Translation Evaluators

Tomás Conde

Volume 57, numéro 3, septembre 2012

URI : https://id.erudit.org/iderudit/1017090ar
DOI : https://doi.org/10.7202/1017090ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Conde, T. (2012). The Good Guys and the Bad Guys: The Behavior of Lenient and Demanding Translation Evaluators. *Meta*, *57*(3), 763–786. https://doi.org/10.7202/1017090ar

Résumé de l'article

Le comportement d'évaluateurs qualifiés d'exigeants ou d'indulgents fait l'objet du présent article. On possède très peu de données sur le processus d'évaluation, plus particulièrement sur la façon dont différents types d'évaluateurs procèdent. Les 88 sujets de cette étude ont été dits exigeants ou indulgents à partir de la moyenne des jugements de valeur qu'ils ont portés sur 48 textes traduits. Leur profil a été déterminé en fonction d'une série de paramètres et de catégories établis à partir de l'analyse des textes évalués. Les évaluateurs indulgents sont davantage intervenus sur le texte, se préoccupaient de la qualité du produit, présentaient un rendement régulier, semblaient plus sûrs d'eux et s'étaient probablement plus impliqués dans l'évaluation qui leur avait été confiée. Les évaluateurs exigeants sont moins intervenus sur le texte, retournaient en général des commentaires, ne sont intervenus que dans certains passages, exprimaient moins de certitude, et semblaient plus conscients du caractère expérimental de la recherche. Tandis que les évaluateurs exigeants paraissent mieux convenir à un contexte professionnel et aux formations avancées, les évaluateurs indulgents semblent être plus appropriés pour la recherche et l'enseignement aux débutants. Le présent travail pourrait ainsi ouvrir la voie à des recherches portant sur les profils d'évaluateurs.

The Good Guys and the Bad Guys: The Behavior of Lenient and Demanding Translation Evaluators

TOMÁS CONDE
Universidad del País Vasco, Vitoria-Gasteiz, Spain
tomas.conde@ehu.es

RÉSUMÉ

Le comportement d'évaluateurs qualifiés d'exigeants ou d'indulgents fait l'objet du présent article. On possède très peu de données sur le processus d'évaluation, plus particulièrement sur la façon dont différents types d'évaluateurs procèdent. Les 88 sujets de cette étude ont été dits exigeants ou indulgents à partir de la moyenne des jugements de valeur qu'ils ont portés sur 48 textes traduits. Leur profil a été déterminé en fonction d'une série de paramètres et de catégories établis à partir de l'analyse des textes évalués. Les évaluateurs indulgents sont davantage intervenus sur le texte, se préoccupaient de la qualité du produit, présentaient un rendement régulier, semblaient plus sûrs d'eux et s'étaient probablement plus impliqués dans l'évaluation qui leur avait été confiée. Les évaluateurs exigeants sont moins intervenus sur le texte, retournaient en général des commentaires, ne sont intervenus que dans certains passages, exprimaient moins de certitude, et semblaient plus conscients du caractère expérimental de la recherche. Tandis que les évaluateurs exigeants paraissent mieux convenir à un contexte professionnel et aux formations avancées, les évaluateurs indulgents semblent être plus appropriés pour la recherche et l'enseignement aux débutants. Le présent travail pourrait ainsi ouvrir la voie à des recherches portant sur les profils d'évaluateurs.

ABSTRACT

The behavior of demanding and lenient evaluators is analyzed and discussed. Little is known about the process of translation evaluation, specifically on how different types of evaluators perform. The 88 subjects of this study were classified as demanding or lenient on the basis of the average quality judgments they made on 48 translated texts. Their profiles were outlined according to a series of parameters and categories starting from the observation of their products, i.e., the evaluated texts. Lenient evaluators carried out more actions on the text, were fairly product-oriented, showed a fairly steady performance, seemed to be more confident, and were probably more committed to the evaluation assignment they were given in this research. Demanding evaluators intervened less, were usually feedback-oriented, preferred to carry out actions in certain segments and text parts, expressed less certainty, and were possibly more aware of the particular circumstances surrounding the experiment. While demanding evaluators appear better suited for professional environments and advanced level teaching, lenient evaluators seem more suited to research and teaching at initial stages. The present work might pave the way for further research into evaluative profiles.

MOTS-CLÉS/KEYWORDS

évaluation, évaluation de la traduction, niveau de la demande, études empiriques, processus de recherche
evaluation, translation assessment, level of demand, empirical studies, process research

1. Introduction

The everyday world is full of both demanding and indulgent people. In the realm of pedagogy, the distinction between demanding and lenient subjects gives rise to interesting controversies, such as whether a high level of demand is associated with a higher quality of teaching (Downey 2010[1]; Kvanvig 2008[2]), whether the threat of low grades can be disadvantageous for student motivation (Gross Davis 1999), or even over the – apparently arbitrary – disparity among different teachers' expectations (Hoffman 2008). Moreover, teachers' level of demand is equally taken into consideration among students, and can sometimes affect their decision whether to enroll in certain courses (Downey 2010).

In our field – Translation Studies – the distinction between more or less demanding evaluators poses an immediate challenge from the very moment a value is allocated to a translation. As in everyday life, the reasons for high or low levels of demand can only be inferred from the evaluators' behavior. Fortunately, however, we have real data on such behavior, since evaluators tend to leave traces of their work on the object of their evaluations, i.e., on the translations themselves. It therefore seems paradoxical that, given the availability of this material, to date no comprehensive analyses have been carried out on their performance.

The impact of the level of demand on the *way* translations are assessed is clearly of interest in both teaching and professional environments. Teachers often assume the role of potential addressees of the translation, an abstraction that sometimes leads them to be overly-demanding, emphasized by personal preferences (Conde 2009: 101-102). Such excessive severity by teachers contrasts with the level of demand present in many professional environments, where poor quality is not directly related to the presence of translation errors, but a "mismatch of assumption and goals between the people requesting a translation and the people supplying it" (Muzii 2006: 17), in other words, a correct translation in the profession is a translation where "the total errors are within the desired threshold in a quality index" (Muzii 2006: 24). Some teachers' decision to raise levels of demand with regard to the market is often justified because it ensures that students achieve a minimum quality to meet professional standards and helps improve the average quality of products (Conde 2010: 253), but from a social constructivist perspective, it has some drawbacks since learning concerns not just the trainee, but the whole teaching environment (Pym 2004).

Within the classroom, the other end on the scale of demand levels might be represented by fellow students who, when called upon to evaluate their peers' translations, show understandable empathy and tend to be overly generous. Perhaps because of this sense of solidarity, Haiyan (2006) advises that peer assessment should avoid assigning numerical grades.

The degree of severity shown by teachers is sometimes a source of conflict because of the inevitable comparisons made between teachers (Conde 2009: 103) – especially those who share subjects – and generally based on numbers of students that pass or fail, average grades, and so on.

Another aspect that may affect evaluators' level of demand is the text type (Conde 2009: 170). Future research contrasting evaluators' performance on different textual genres would therefore be highly desirable, but the complexity of the source texts must first be defined.

Another difficulty is that evaluation is considered an eminently subjective activity (Bowker 2000: 183; Li 2006: 84; Varela and Postigo 2005; Gouadec 1980: 116); many stakeholders therefore endeavor to use evaluation scales, parameters and criteria that are as objective as possible. Other researchers accept subjectivity as an intrinsic part of any judgment (Fox 1999: 5; Hönig 1997: 14; Muñoz 2007: 262), and attempt to integrate it into the evaluative process. In any event, a better understanding of the regularities concerning evaluation would be useful for the discipline. Certainly, studying these regularities means analyzing the evaluation process, an activity that – it must be said – has not enjoyed much attention in Translation Studies: many authors have contributed their experiences to propose models for evaluating translations (Nord 1991; Lauscher 2000; House 2001; Adab 2002, amongst others), but few have stopped to investigate what ultimately happens during this evaluation process. An effort to trace the actual behavior of different types of evaluators through rigorous research (Conde 2008: 93) would be particularly helpful in approaching such an uncharted area of activity.

But, of course, there have been valuable exceptions to the aforementioned lack of empirical research on the process of translation evaluation. Christopher Waddington, in his doctoral thesis (1999) and subsequent work (2001), showed that holistic models were at least as effective for measuring translation quality as the (most widespread) analytical models, although the former could be implemented more rapidly. In his thesis, Conde (2009) compared the evaluation carried out by professional translators, translation teachers, translation students and potential addressees, and confirmed Waddington's results (holistic evaluators reached the same conclusions as analytical evaluators). In an earlier study (Muñoz and Conde 2007), evaluators' performance was for the first time described according to their level of demand. The authors found that lenient evaluators had a more homogeneous behavior, performed many actions on the texts and seemed to focus their evaluation on improving the text; whereas demanding evaluators made far fewer actions and tended to focus their evaluation on learning feedback (Muñoz and Conde 2007: 437). However, the study examined the effects of serial translation evaluation (evaluating several translations from the same original), and included only ten subjects, so although it provided some interesting results, they could not be extrapolated to all situations. The Muñoz and Conde study may nonetheless be considered as the predecessor to the present paper, which aims to complete the description of evaluator profiles according to their level of demand, but starting from the process and results of all the subjects who participated in Conde's doctoral thesis (2009) which constitutes a rich data set for analysis.

The next section (2) describes the materials and methods employed; the results are presented and discussed in section 3, and finally the conclusion (section 4) includes a description of the two profiles of evaluators (demanding and lenient), a summary of the strengths and weaknesses of this work and some suggestions for future research.

2. Materials and Methods

This section details the circumstances in which the experiment was carried out, and describes the profile of the subjects who participated in the study, the evaluation task,

and the parameters used. The main hypothesis is that *there are significant differences in the way lenient and demanding evaluators work, which will be evident in the evaluated product.*

2.1. Subjects and task

Marks made on the translation during the evaluation conducted by a total of 88 subjects were analyzed in order to test the hypothesis. The subjects had different backgrounds and experience with respect to translation and its evaluation:

- 25 translation students, on an advanced Translatology course;
- 13 professional translators, both in-house and freelancers;
- 10 translation faculty, at Spanish and Mexican universities;
- 40 potential addressees, students on English for Specific Purposes courses within degrees related to the translations' subject matter.

The preferred method of contact with the groups was via email, but university students also attended meetings held to explain the procedure and encourage participation. In all cases, documentation for the evaluation task always included a sheet with the few instructions (see below) needed to carry out the evaluation. Previous studies (e.g., Conde 2009) have analyzed the different ways the four groups of evaluators behaved. In the present study, however, these groups are only of interest to classify subjects – as lenient or demanding evaluators – within each population group, since the circumstances under which each group carried out the task were not always identical.

The evaluation task consisted in assessing 48 Spanish translations of four English originals, following the aforementioned sheet, and afterwards filling in a final questionnaire. Two of the originals were political texts for a wide readership (DP1 and DP3), and the other two dealt with industrial painting techniques (CT2 and CT4). Sets of different topics were alternated to prompt evaluators to think of them as separate tasks; subjects performed the evaluation in the order defined by the code numbers: DP1, CT2, DP3, and CT4. Each set included 12 randomly ordered Spanish translations done by translation students in their third year. The subjects received the evaluation task and were given generous deadlines to complete it; this was a pertinent issue, especially for professional translators and translation teachers, who were expected to be more reluctant to participate in the project (which eventually proved to be the case, as evidenced by the number of subjects collaborating in each group).

Instructions were intentionally general, since the concept to be observed was the evaluation process itself. Subjects could work on screen or on printed copies of the documents. Thus, evaluators were given only three instructions:

1) Follow the set order;
2) Work on each set of 12 translations in one sitting;
3) Classify the quality of translations as very bad, bad, good or very good.

Apart from these indications, the evaluators were instructed just to "assess, review or correct [the translations], according to their personal beliefs, intuition and knowledge." When the evaluators submitted their work, data were stored and managed through an MS Access database, and then exported to several files for statistical

processing using the SPSS v15 (initially) and v17 (for further analysis) software packages. Figures were originally designed – and later edited – by MS Excel and Word. Together with the tables included in this article, they represent the results corresponding to the parameters that are described in detail in the next section.

2.2. Parameters and categories

As previously mentioned, examining the subjects' performance enabled us to define the parameters that were used to outline the profile of the two groups (lenient and demanding). The choice of parameters was based on where, how and why something had been noted, as the reasons given by the evaluators (and their consistent behavior) pointed out. Some of these parameters were related either to actions or to phenomena; others were considered as cross-cutting, since they are particularly informative when contrasted with other parameters. Table 1 summarizes the parameters and categories included in this study; a full explanation of these concepts is provided below.

TABLE 1

Parameters and categories

Actions	*Number of actions*	
	Types of changes: feedback-oriented (linkups, highlights, classifications); product-oriented (additions, suppressions, substitutions)	
	Comments	*Number of comments*
		Location: at the beginning, at the end, in the margin, in the text, separately
		Contribution: correction, solution, alternative, assessment, vocative, procedural, null
		Source: personal, external, null
		Certitude: certain, uncertain, null
Phenomena	*Scope:* sentence, paragraph, text, set, task	
	Nature: normalized (typos, punctuation, format, spelling, proper nouns, terminology, concordance, cohesion, syntax, weights and measures); non-normalized (appropriateness, clarity, usage, divergent interpretations, omission, perspective); others (combined, unknown)	
	Reaction: negative, positive, very negative, neutral	
	Saliency: zero, very low level, low level, medium level or high level of coincidence	
Crosscutting	*Quality judgment:* very bad, bad, good, very good	
	Order and segmentation: sets (DP1, CT2, DP3, CT4); stretches (I, II, III); sections (initial, central, final); poles (title, ending); typography (outstanding, regular)	

The operational definition of an action is "any mark introduced by the evaluator in the text or file." The first parameter, then, is the number of actions carried out by each type (lenient or demanding) of evaluator. Actions may be simple or complex (consisting of a change and a comment). Changes are modifications of the body text itself, whereas comments refer to information that is attached to the text but does not belong to the body text. The number of changes is, logically, equivalent to that of actions, so the analysis focused on their nature rather than on their number. Accordingly, there are several types of changes, classified into two groups:

1) Feedback-oriented: these changes provide information for the author of the translation and include the following subcategories: linkups (mark the text to introduce comments), highlights (shaded text), and classifications (systematic highlighting following a code);
2) Product-oriented: these changes directly improve the text (adding, suppressing or substituting ideas that are missing, unnecessary, or that were poorly expressed by the translator).

The total number of comments was examined, together with four aspects, namely location, contribution, source and certitude.

Comments may be introduced at the beginning or end of each text, in the margin or embedded in the text itself. Further, some comments are attached to the text, i.e., on a separate sheet. Comments can have the following functions: as corrections when the evaluator reports the existence of a specific problem; as solutions, when they provide solutions to the detected problems; as alternatives, when they offer several equally valid solutions or explain that the version proposed is as good as the one chosen by the translator of the text; as assessments, when they simply define the quality of a specific fragment; as vocatives, when they appeal to the researcher and, finally, as procedurals, when they provide information about the evaluation system.

Most sources are personal, because evaluators were asked to use their own discretion in carrying out the task. However, sources are external when evaluators explicitly refer to other persons, institutions or (grammar) rules. This extra information – since it was not required – may be indicative of a higher or lower self-confidence vis-à-vis the evaluation performed. This parameter is therefore related to another aspect of comments: certitude. In principle, all comments are certain (otherwise, subjects would not introduce them) except when evaluators express uncertainty, insecurity, doubt or irony; for example, when they include questions (and question marks), or state clearly that they are "not sure." The variables of contribution, source and certitude include an extra category ("null") that accounts for the few instances in which other classifications are impossible, for example when evaluators introduced comments (with the corresponding word processor tool) that, for some reason, were left blank.

In order to move away from popular approaches to evaluating translations which focus on *mistakes* (evaluators do not only mark mistakes), we defined phenomenon as "what motivates or may motivate an evaluator to act on a particular text fragment" (thus, every action presupposes – at least – a phenomenon). Other parameters were related to this operational concept, namely: scope, nature, reaction and saliency.

Scope involves the portion of text affected by a phenomenon. There are five categories: sentence (including phenomena that refer to shorter fragments), paragraph (that is, two or more consecutive sentences in a block of text), text (complete), set (complete), and task (complete, two sets for the potential addressees, and four for the other groups).

As regards the nature, a distinction was made between phenomena in which there is a regulatory body that sanctions a correct option (normalized) and the rest (non-normalized). The subcategories within each group are:

- Normalized: typos, punctuation, format, spelling, proper nouns, terminology, concordance, cohesion, syntax, and weights and measures;
- Non-normalized: appropriateness, clarity, usage, divergent interpretations, omission and perspective.

In addition, two other subcategories that belong to neither normalized nor non-normalized phenomena were considered: combined phenomena, where several phenomena could be ascribed to the same action; and unknown cases, when the phenomenon referred to by the evaluators when performing a specific action was not clear. Even though the classification is not homogeneous nor totally sharp, it "responds to the nature of the phenomena pretty well, does not demand a strong heuristic effort" (Muñoz and Conde 2007: 429) and brings about a considerable reduction of unknown phenomena.

Certain actions (e.g., comments or classifications) enable the researcher to infer the evaluator's reaction to the phenomena detected. Most, of course, are negative reactions, since evaluators tend to mark errors; however they may also be positive (when they identify good choices of translation), very negative (when they emphasize the gravity of certain errors), and neutral reactions (when the action undertaken is not directly related to a positive or negative reaction, or when there is a combination of both).

Finally, phenomena are more or less salient depending on the number of subjects that single them out. Accordingly, there were three main categories: those phenomena noted by only one evaluator (zero coincidence), those marked by over 25 evaluators (high level of coincidence) and the remaining phenomena (intermediate coincidence). Within the latter category, three sub-levels were distinguished: very low level (noted by 2 or 3 evaluators), low level (between 4 and 12), and medium level of coincidence (between 13 and 25). For a more detailed explanation of these levels, see Conde (2009: 321).

Further, as mentioned above, some parameters cross over the others. Such is the case for quality judgment, i.e., the average grade assigned to the translations by each subject. The four quality levels were given a numerical value (1 for very bad, 2 for bad, 3 for good, and 4 for very good translations) to allow for the statistical contrast with the rest of the parameters. As well as quality judgment, order[3] was considered to be a cross-cutting parameter. It was analyzed across sets (DP1, CT2, DP3, and CT4), stretches and sections. Stretches were considered to observe order effects within the sets:

1) Stretch I includes translations 1-4 of each set;
2) Stretch II, 5-8;
3) Stretch III, 9-12.

To examine the order effects within translations, texts were divided into three sections:

1) Initial: first third of text;
2) Central: second third;
3) Final: Last third.

Each translation was split into three sections by counting the total number of words and then dividing by three; however, some adjustments were needed (Conde

2009: 269) to avoid truncated sentences within the sections. The first (title) and the last (ending) sentences of the translations were also taken into account; these were termed *poles*. Some analyses also distinguished between outstanding and regular segments, depending on whether they appeared on typographically emphasized fragments or not, respectively.

3. Results and discussion

According to the manner in which subjects undertook the task, they were considered to be either concise or detailed evaluators. The concise evaluators carried out very few actions in the texts, as they apparently based their assessments on holistic approaches. Their actions, if any, followed the instructions regarding issuing of judgments; subjects usually did so by writing comments at the beginning or at the end of each translation or set. The detailed evaluators, on the other hand, performed many actions, probably because their approaches were based on error analysis.

The most important difference between the two types of evaluators is that concise evaluators show the outcome of their evaluation (their quality judgments), but provide little information about the process, while the process carried out by detailed evaluators can be inferred from their work. Therefore, the analysis of the process, and the contrast with the outcome, is based on the work of detailed evaluators (Conde 2009: 342).

3.1. Quality judgments

In all groups, most subjects – according to the average quality judgment issued – were detailed (61% of the lenient evaluators and 64% of the demanding evaluators). Thus, the approach evaluators take does not appear to affect their level of demand (Conde 2009: 441).

To analyze the evaluators' behavior, *demand* was defined as "the sum of conscious and unconscious expectations an evaluator seems to think that a translation should meet." Once atypical subjects – any observation that is statistically distant of the rest of the data – had been suppressed, demanding evaluators were defined as those who, within their population group (potential addressees, translation students, translation teachers and professional translators), issued below average quality judgments. This process resulted in 47 lenient and 36 demanding evaluators (the remaining five of the total of 88 subjects did not issue quality judgments). When concise subjects were suppressed, 29 lenient evaluators and 23 demanding evaluators remained. Hence, the data (reported below) reflects the evaluation of these 52 subjects.

First, Table 2 reports the average quality judgment issued by both groups (lenient and demanding evaluators) for the whole task, each set and each stretch. The average judgment of each translation from lenient subjects was 0.54 points higher than that of demanding evaluators. The biggest differences in sets were observed on the last set (0.59), and the smallest differences, in the first of the specialized sets (0.11). In addition, stretch II (which comprises translations 5-8 within each set) revealed the greatest differences: 0.64 points.

TABLE 2
Quality judgment

	Lenient	Demanding	*Difference*
Total	2.72	2.18	*0.54*
Sets			
DP1	2.53	2.14	*0.39*
CT2	2.76	2.65	*0.11*
DP3	2.68	2.44	*0.24*
CT4	2.65	2.06	*0.59*
Stretches			
I	2.53	2.18	*0.35*
II	2.76	2.12	*0.64*
III	2.78	2.26	*0.52*

These data support our description of the subjects' profiles. For example, lenient evaluators are especially tough on the first set, but subsequently their level of demand decreases, particularly in the first of the technical sets. They are also tougher in stretch I, and then their judgments become more balanced in the translations comprising II and III. Demanding evaluators are also tough in DP1, but especially in CT4: after relaxing their level of demand in CT2 (perhaps because of the topic), they seem to regain confidence – and level of demand – in the last set. They issued a lower average judgment on stretch II, and it increased again in III (even more sharply than lenient evaluators).

3.2. Actions

As mentioned in 2.2, an action is any change made by the evaluator in the text or any comment made. This section presents and discusses the results concerning: the number of actions, types of changes, number and types of comments according to their location, contribution, source and certitude.

3.2.1. Number of actions

The number of actions carried out by evaluators was the first parameter to be analyzed. Table 3 shows averages for lenient and demanding evaluators. In general, subjects performing more actions are also less demanding. Differences between sets are small (following the decrease between DP1 and CT2) in both groups. Additionally, both groups tend to carry out a similar number of actions on technical sets. There is a gradual decrease in both groups by stretch, although the gap in the number of actions between II and III is higher among the demanding group.

TABLE 3
Number of actions

	Lenient	Demanding
Total	1022.59	849.13
Sets		
DP1	492.42	361.34

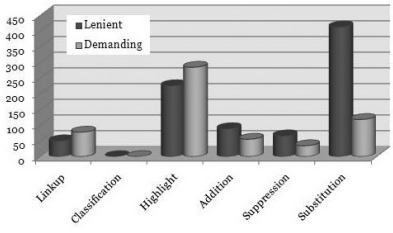
CT2	237.73	207.14
DP3	223.02	148.31
CT4	172.32	158.81
Stretches		
I	378.93	311.81
II	330.32	273.45
III	313.33	195.70
Sections		
Initial	365.02	293.14
Central	316.85	261.80
Final	340.71	255.04
Poles		
Title	38.31	33.60
Ending	43.45	45.08
Typography		
Outstanding	98.91	77.17
Regular	923.71	706.58

Within each translation, the activity of demanding evaluators decreases moderately but gradually. However, lenient evaluators perform fewer actions in the central section and the number increases again toward the end (but not to the level of the initial section). Two distinct patterns of behavior therefore seem to emerge: the gradual fall in the number of actions within translations may indicate that the purpose of demanding evaluators' actions is mainly to build an opinion on the text quality (which is not necessarily low), whereas lenient evaluators might dissociate actions from the judgments they issue on the texts. On the other hand, both types of evaluators prefer to act on the endings (as compared to titles), but the difference between these two categories is highest among demanding evaluators. This could be due simply to a tendency within this group to render quality judgments at the end. No differences are apparent in the number of actions carried out by lenient and demanding evaluators on (typographically) outstanding segments: in both groups, actions in these segments account for around 10% of total actions.

3.2.2. Changes

Every action involves a change in the text, as the analysis was based on the marks subjects made when carrying out their evaluations. Changes were either product-oriented (if they directly improved the translation, by means of fragment additions, suppressions or substitutions) or feedback-oriented (if they provided information independent of the text contents, in the form of linkups, highlights and classifications). Figure 1 shows the types of changes in lenient and demanding subjects. A quite similar use of linkups, additions and suppressions can be seen in both groups. Once atypical values were removed, neither of the groups used classifications. Lenient evaluators substitute much more (average 412.76) than demanding evaluators (118.03), whereas the latter introduce more highlights (284.22) than the former (225.59).

FIGURE 1
Changes



Lenient subjects prefer text substitutions, indicating that they might be more product-oriented: they want to leave the text ready for use, and their extra effort to find the best solution seems to increase their empathy toward the evaluated subjects, since they assign them higher quality judgments. Another possible explanation is that substitutions modify the original draft and improve the quality of the text in the eyes of evaluators, so they could possibly have a distorted perception of the quality in subsequent partial or final assessments. In contrast, demanding evaluators perform more highlights, and also (although the difference is insignificant) of linkups, i.e., their evaluation is fairly feedback-oriented: they emphasize mistakes straight away, which could lead them to issue lower judgments.

Table 4 shows the evolution in the number of product- and feedback-oriented changes among lenient and demanding evaluators, through sets, stretches and sections. Lenient subjects perform many more product-oriented changes, with an abrupt decrease between DP1 and CT2 and another slighter fall between DP3 and CT4. Demanding evaluators carry out fewer product-oriented changes, but in essence they behave similarly. As for feedback-oriented changes, lenient subjects show a steady decline, whereas demanding evaluators introduce fewer changes in DP3.

TABLE 4
Changes per segmentation

	Product		Feedback	
	Lenient	**Demanding**	**Lenient**	**Demanding**
Sets				
DP1	326.87	69.88	149.48	132.40
CT2	103.51	9.85	125.41	152.13
DP3	104.43	13.58	87.90	73.45
CT4	73.92	6.68	77.11	90.53
Stretches				
I	212.79	120.72	166.14	174.74
II	181.02	57.38	149.33	145.76
III	173.77	72.78	139.56	155.00
Sections				
Initial	197.67	100.99	167.38	175.02
Central	185.06	58.47	131.79	136.89
Final	184.86	58.61	155.85	140.58

There are minor differences across stretches. Both types of changes made by lenient subjects tend to decrease, whereas demanding evaluators are less active in II. As for sections, demanding subjects again make gradually fewer product- and feedback-oriented changes. Lenient subjects, however, are more balanced in their product-oriented changes, but their activity decreases in the central section.

3.2.3. Comments

Apart from the changes, actions could include comments. This section provides an overview of the number of comments made by the lenient and the demanding evaluators. Table 5 shows the average for both groups, which is higher among the lenient group. Across sets, the evaluators' level of demand has little effect on the evolution of comments throughout the task. Within sets, the decrease in actions is almost parallel in the two groups; however, lenient subjects make more comments in the initial section. Both groups increase their activity in the final section, but the rise is more significant among the demanding subjects. Perhaps the comments in the final section affect the overall assessment of the demanding evaluators, since they are still fresh in their memory when they make their final evaluation. As for the poles, both groups prefer to make comments on the ending rather than the title; the difference is, however, larger among the demanding subjects. Finally, the percentage of comments on typographically outstanding segments is considerably higher within the group of demanding evaluators. That is, the subjects who pay more attention to the outstanding segments (compared to the regular ones) are also more demanding.

TABLE 5
Number of comments per segmentation

	Lenient	Demanding
Total	276.62	224.38
Sets		
DP1	99.01	87.55
CT2	74.29	69.82
DP3	37.27	37.71
CT4	37.16	38.16
Stretches		
I	116.13	99.49
II	90.18	72.45
III	64.96	49.00
Sections		
Initial	114.87	69.77
Central	77.01	62.80
Final	84.67	80.76
Poles		
Title	21.09	15.30
Ending	29.01	35.28
Typography		
Outstanding	34.94	42.16
Regular	225.32	174.10

Table 6 shows the number of comments made by lenient and demanding evaluators, with respect to the parameters of location, contribution, source and certitude. The first, location, refers to where the comment was written. Lenient evaluators introduce many comments in the margin and hardly any in the text. They make comments at the end rather than at the beginning. The comments written on a separate sheet are virtually anecdotal. The average demanding evaluator makes more comments in the text than in the margin, and more at the end than at the beginning. Furthermore, the demanding subjects do not make use of comments in a separate file. Visually, comments in the text make documents look messier than comments in the margin (which are easier to ignore). This may have affected the evaluators' level of demand: they may associate a clean text with a text without errors, and a more cluttered text with a text full of errors. A cognitive explanation for this is that subjects, knowing the body text would be left unmodified, may be more emboldened to make comments in the margin (even while considering them less important). In contrast, those who prefer comments in the text, knowing that their comments would make the text messy, would likely decline to do so when the comments are not really necessary and, consequently, the comments they do make become decisive in their overall quality judgment. Obviously, this is based on simple intuition and needs to be tested by further empirical work.

TABLE 6

Number of comments by location, contribution, source and certitude

	Lenient	Demanding
Location		
On a separate sheet	0.26	0.00
At the beginning	12.82	4.68
At the end	28.16	34.83
In the margin	91.67	52.15
In the text	8.07	62.59
Contribution		
Vocative	0.00	0.00
Alternative	3.26	0.94
Correction	108.76	79.12
Null	0.00	0.00
Procedural	0.04	0.44
Solution	35.32	9.20
Assessment	33.84	39.16
Source		
External	1.79	1.86
Personal	295.57	219.52
Null	0.00	0.00
Certitude		
Certain	278.16	198.62
Uncertain	9.80	19.99
Null	0.00	0.00

The second aspect – contribution – is useful for classifying comments on the basis of the information entered. Lenient evaluators essentially make corrections, solutions and assessments; in their comments, they rarely introduce alternatives or discuss the procedure they follow. In turn, demanding subjects also make many corrections, assessments and solutions. Their average for alternative and procedural comments is also low. Once atypical subjects were suppressed, neither group introduced vocative or null comments. Therefore, the most obvious difference between the two groups is that the demanding subjects prefer assessment to solution comments, whereas within the lenient subject group these two types of comments are balanced. As a consequence, demanding evaluators offer fewer solutions, which could be due to a lower commitment to the evaluated subject's learning, as they do not provide answers to the questions they find. Another explanation could be that the demanding evaluators want students to find solutions themselves, as part of a pedagogical strategy.

The last two aspects of comments referred to the source (defined as the authority responsible for the information contained therein) and the certainty with which they are expressed. Regarding the comment source, there are no major differences between lenient and demanding evaluators, although the relative weight of the external references is slightly higher within the latter (0.85%) than within the former group (0.61%). Having suppressed the atypical values, groups do not introduce null comments in either the source or the certainty category. As far as this latter aspect is concerned, the demanding evaluators make about three times as many uncertainty comments as the lenient group. One possible explanation is that demanding evaluators feel more responsible for the truthfulness and correctness of their comments: they decide to introduce a piece of information about which – they sometimes admit – they are not completely sure. This also explains why demanding subjects refer more to external sources, although, as mentioned above, the differences in this case were insignificant.

3.3. Phenomena

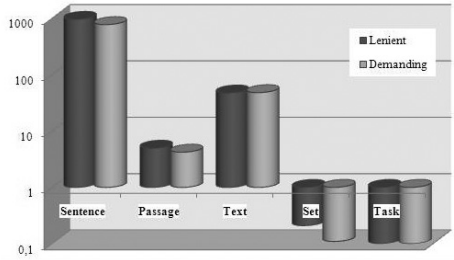
Parameters other than those concerning actions were taken into account. These variables depend on the phenomena – defined in 2.2 as motivators of the actions carried out by the evaluators – and are discussed below in the following order: scope, nature, reaction and saliency.

3.3.1. Scope

Actions are performed on specific fragments throughout the files, but can refer to various distinct portions of text. Figure 2 illustrates the number of phenomena identified with respect to each category by both lenient and demanding evaluators. A base-10 logarithmic scale is used for easier interpretation of the data.

FIGURE 2

Scope



After atypical values were suppressed, no phenomena were registered in the scope of the task. Lenient evaluators identified an average of 957.86 phenomena in the scope of the sentence, as opposed to the 771.57 for demanding subjects. The other categories show few differences, although demanding evaluators tend to act more in the scope of the text (47.92) than lenient ones (47.34), a pattern that was repeated in the scopes of the passage and the set. We may conclude that, apart from the difference in phenomena at sentence level – an obvious result, taking into account the general results described for actions (3.2) – lenient and demanding evaluators behave similarly.

Table 7 shows phenomena above sentence level to see whether these (and those in the scope of the sentence) evolve similarly across sets, stretches and subsections. While among lenient subjects the phenomena at sentence level fell gradually across sets, demanding subjects present a decrease between DP and DP3, and a subsequent slight increase in CT4. Phenomena at sentence level, mainly related to details, are more numerous in CT4 than in DP3 within the demanding evaluators group. As for phenomena above sentence level, lenient evaluators are particularly active in DP3, perhaps because they noticed or more strongly emphasized the fact that most translations in DP3 are incomplete (the original text was longer and most students were unable to finish the translation).

TABLE 7

Scope per segmentation

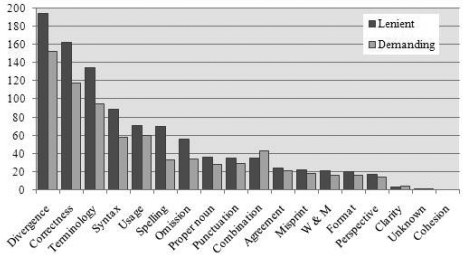
	Sentence		Above sentence	
	Lenient	**Demanding**	**Lenient**	**Demanding**
Sets				
DP1	405.14	250.85	13.47	12.95
CT2	191.48	154.90	12.10	11.71
DP3	167.73	111.73	14.83	12.60
CT4	149.13	119.32	8.82	11.12
Stretches				
I	355.98	284.36	19.99	24.36
II	309.72	249.61	17.79	17.87
III	292.17	168.59	16.91	20.57
Sections				
Initial	342.26	286.98	19.86	23.82
Central	307.34	247.73	3.11	3.67
Final	308.26	216.67	32.44	34.90

In both groups of evaluators, the number of phenomena at sentence level identified in successive stretches decreased. Phenomena above sentence level across stretches also fell among lenient subjects, but demanding subjects identify more phenomena in III than in II. Typically, phenomena above the sentence level accumulate in III, where quality judgments on the whole set and task are usually introduced. However, lenient subjects identify fewer phenomena above sentence level in this stretch, perhaps because – as they perform many more actions than the demanding evaluators – at this point (stretch III), they are more fatigued and the decrease in their activity is more evident. Regarding sections, the evolution of the two groups is very similar in terms of phenomena above the sentence level: their activity dramatically increases in the final section, where judgments on the text, the set and the entire task are usually included. Demanding evaluators identify increasingly fewer phenomena at sentence level across sections, whereas lenient evaluators show similar numbers in central and final sections, perhaps because of their greater commitment to address all erroneous aspects of the translations, in other words, their pedagogical approach.

3.3.2. Nature

This is probably the most subjective parameter, since it depends on the researcher's interpretation: he or she must discern the reasons underlying the evaluator's action, that is, the nature of the phenomenon. Figure 3 shows the average of phenomena of different nature in lenient and demanding subjects, ranked from the highest to the lowest frequency in lenient evaluators, who in general identified more phenomena. Figure 3 shows that demanding and lenient evaluators follow the same patterns of frequency except in four instances: the phenomena of usage (within the group of demanding evaluators, above syntax), omission (above spelling), punctuation (above proper noun) and, particularly, combination, above the phenomena of spelling and, therefore, the sixth most important phenomenon for demanding evaluators.

FIGURE 3
Nature



Lenient subjects introduce combined actions to discuss several aspects at once, but attach more importance to other phenomena: spelling, omission, proper nouns and punctuation. The higher frequency of combined phenomena among demanding evaluators may be due to their greater sensitivity to feedback learning, which leads them to draw attention to general aspects of the text beyond the specific errors or phenomena.

Table 8 shows normalized and non-normalized phenomena across sets, stretches and sections, distinguishing between lenient and demanding evaluators. The behavior of the two groups with regard to normalized phenomena is similar, except for the initial decline, which is greater among the lenient subjects. Additionally, lenient evaluators identify many more non-normalized phenomena in the technical sets than in the popularizing sets; hence, their evaluation process seems to be affected by the texts' subject matter – which does not lead to lower quality judgments, but rather the opposite, *cf.* 3.1. In CT4, the upturn of normalized phenomena for both types of evaluators seems to disprove the hypothesis of fatigue *across sets*. Rather, this behavior could be explained as an adjustment in the order of priorities, where non-normalized phenomena would be less important in technical texts.

TABLE 8

Nature per segmentation

	Normalized		Non-normalized	
	Lenient	**Demanding**	**Lenient**	**Demanding**
Sets				
DP1	215.81	124.28	209.31	141.43
CT2	96.06	75.46	95.58	80.18
DP3	65.17	45.44	104.75	72.35
CT4	75.31	57.83	70.58	53.70
Stretches				
I	167.67	131.32	196.33	158.28
II	146.47	112.50	170.11	119.28
III	148.61	111.18	147.70	90.82
Sections				
Initial	167.23	130.98	170.81	147.17
Central	143.13	109.66	167.40	128.52
Final	152.39	114.36	166.38	115.96

Broadly speaking, both groups evolve similarly across stretches. Most notable is that the evaluators' level of demand affects neither the gradual reduction of non-normalized phenomena, nor the more stable behavior with respect to normalized phenomena. The former result may be explained by the tradition of considering non-normalized phenomena as less necessary. In fact, they are *non-binary* errors (Pym 1991: 281), since they are more complex and more difficult to correct and justify. Thus, the accumulated fatigue in the evaluation of the set would be reflected in the evaluators' behavior at the expense of non-normalized phenomena.

Among demanding evaluators, the evolution across sections seems to support the previous assumption, since the number of normalized phenomena picks up in the final section, whereas that of the non-normalized phenomena continues to fall. In other words, within each translation, lenient evaluators continue to mark the phenomena they identify (whether normalized or non-normalized), whereas demanding subjects limit their efforts, especially as regards phenomena that perhaps they consider less important, i.e., non-normalized phenomena.

3.3.3. Reaction

Phenomena might provoke reactions of various kinds, depending on whether the evaluators consider them errors (more or less serious), good decisions or a mixture of both. Figure 4 illustrates the phenomena identified by lenient and demanding evaluators, classified according to the type of reaction they provoked. A base-10 logarithmic scale is used for easier interpretation of the data. Once the atypical values were suppressed, the lenient subjects have more negative (average 962.84) and positive reactions (20.95) than the demanding subjects (688.03 and 10.93, respectively). In contrast, the demanding subjects reported more very negative (15.45) and neutral reactions (6.33) than their lenient colleagues (4.75 and 2.53, respectively). As lenient evaluators intervene more in the text (and all evaluators tend to focus on errors), they have many negative reactions. On the other hand – and in keeping with the average quality judgments – they register more positive reactions. Among demanding evaluators, the higher number of very negative reactions (compared to the lenient subjects) was expected. But they also have more neutral reactions, which may be due to the tendency among the demanding subjects to carry out shallower evaluations, usually including positive and negative comments in the same actions.

FIGURE 4
Reaction

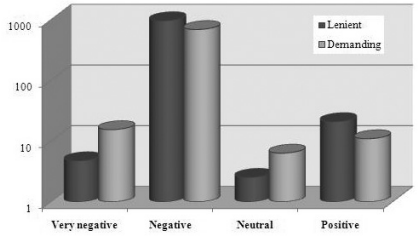


Table 9 distinguishes between negative and other reactions, within both the lenient and demanding evaluator groups. The results concerning negative reactions show a sharper decline from DP1 to CT2 within the group of lenient evaluators, and a slight rise from DP3 to CT4 within the demanding evaluator group. Non-negative reactions produced quite different results between the groups. Lenient evaluators have more negative reactions in CT2 than in DP1, after which negative reactions gradually decrease. Demanding evaluators have more negative reactions in the technical than in the popularizing texts. It is also noteworthy that, for both categories of reactions, the variations in demanding evaluators are more moderate than in lenient evaluators, which may be due, once again, to a lower effect of fatigue with respect to their lenient evaluator colleagues, who intervene more.

TABLE 9

Reaction per segmentation

	Negative		Other	
	Lenient	**Demanding**	**Lenient**	**Demanding**
Sets				
DP1	411.55	218.04	7.67	8.76
CT2	191.14	151.59	12.06	9.42
DP3	170.03	110.52	8.32	8.46
CT4	145.98	114.66	5.97	10.32
Stretches				
I	359.44	275.18	10.81	20.46
II	309.91	219.31	12.50	18.43
III	293.52	177.69	13.31	15.99
Sections				
Initial	342.87	255.83	18.20	19.54
Central	304.74	232.62	4.09	12.25
Final	315.27	216.34	18.24	23.09

Demanding evaluators' reactions fell gradually across stretches in both categories, whereas negative reactions among lenient evaluators decreased and their other reactions increased. Perhaps demanding subjects do not deem it necessary to comment or explain the serious errors (very negative reactions) or the suitability of some solutions (positive) each time they appear, but only on the first occasions. A possible explanation for this behavior would be that these evaluators feel they have already done their duty by marking the mistakes the first times, and they know their evaluations will not be read by students but by researchers.

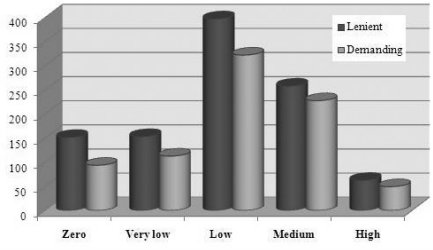
Neutral reactions are not expected to fall: quality judgments in the scope of the set and the task are usually entered toward the end of the sets. These phenomena often include a summary of the good and bad aspects of the text or texts and, consequently, have been considered as neutral reactions.

By sections, demanding evaluators' reactions decrease steadily; moreover, their negative reactions are similar to those of the lenient subjects, with one exception: the decline in the central section is not as sharp as among lenient evaluators, perhaps because they are more selective with their actions, so that those made in the text are less dispensable, especially those with an extreme weighting (positive or very negative) or totally neutral (which refer to several factors simultaneously; thus, removing them eliminates a greater amount of information). This does not happen with only negative reactions, which are possibly presupposed.

3.3.4. Saliency

The last parameter related to phenomena is saliency, defined as the extent to which evaluators react to phenomena. Figure 5 shows the average of the five original levels (zero coincidence, very low, low, medium or high coincidence)[4], as identified by the lenient and the demanding evaluators.

FIGURE 5
Saliency



The lenient evaluators identify more phenomena from all categories than the demanding ones. Proportionally, demanding evaluators have more medium-coincident and fewer non-coincident phenomena. The latter result is consistent with what has been suggested so far: lenient evaluators intervene more, thus reacting to a greater number of individual events (identified by only one evaluator), probably because they are driven more by personal preferences than their demanding evaluator colleagues but also because the more changes and comments are made, the more likely it is that some of them will be unique.

Table 10 groups phenomena into three levels (zero coincidence, intermediate coincidence and high coincidence), and relate them to the evaluators' level of demand.

TABLE 10
Saliency per segmentation

	Zero		Intermediate		High	
	Lenient	**Demanding**	**Lenient**	**Demanding**	**Lenient**	**Demanding**
Sets						
DP1	52.07	32.51	356.12	217.21	44.46	28.21
CT2	28.39	25.27	170.52	144.34	12.23	9.28
DP3	34.63	21.15	146.21	103.98	3.73	2.99
CT4	22.95	11.86	131.08	114.17	7.24	6.27
Stretches						
I	57.42	42.30	304.22	254.53	17.82	13.42
II	46.51	29.84	258.47	214	25.39	19.64
III	47.66	30.94	245.88	169.08	19.80	16.21
Sections						
Initial	50.51	40.00	288.77	234.36	25.77	20.91
Central	52.36	39.93	242.32	201.92	22.17	17.48
Final	48.71	26.52	277.47	228.24	14.53	11.75

Across sets, both groups show a similar evolution with respect to the intermediate and high phenomena, with a sharper decrease after DP1 among the lenient subjects. These evaluators also identify more singular phenomena in the popularizing than in the technical texts, whereas demanding evaluators' actions on phenomena of no coincidence decrease throughout the task. That is, personal preferences, which may be the reason for most singular phenomena identified by the lenient subjects, are especially present in the popularizing sets, where, perhaps, they feel more confident.

By stretches, the two types of evaluators present a similar distribution of phenomena: in zero coincidence (more instances in I, then III and finally in II), intermediate coincidence (decrease across stretches), and high coincidence phenomena (more instances in II, then in III and finally in I). Regarding sections, the two types of evaluators show similar figures for the zero coincidence phenomena in the initial and central sections, as well as a decrease in the final section. Again, the most striking difference between groups is the reduction of singular phenomena identified by demanding evaluators, who apparently already have a clear idea of the quality of the text and stop marking phenomena they do not consider essential. Both groups show a similar distribution with respect to the phenomena of intermediate coincidence: a decrease between the initial and final sections, with a sharp fall in the central section compared to the other two. Finally, the two types of evaluators' actions on phenomena of high coincidence fall across sections.

4. Conclusions

Little is known about the evaluation of translations. There are many studies on how to evaluate, but a lack of research describing the evaluators' actual behavior. This seems striking in such a personal, subjective area, in a field which calls out for empirical research to explore an activity which is as widespread as it is unknown. This paper deals with probably one of the least studied aspects: the potential impact of the level of demand on evaluators' behavior. The assumption is that the outcome of the evaluation helps to outline different types of behavior among lenient and demanding evaluators. The results discussed in the previous section support the hypothesis, thus contributing to uncover specific aspects of translation evaluation.

Following the introduction, the paper summarized the circumstances in which the experiment was conducted[5]: the procedure, the participants, the evaluation task and the software used. The main results were then introduced and discussed, thus enabling the description of the two evaluator profiles; these are described in the following paragraphs.

Lenient evaluators prefer analytical to holistic evaluation. They intervene more than demanding evaluators, especially in non-specialized texts and in the final sections of texts. Their quality judgments are initially low (first sets and stretches), and then moderate. Lenient evaluators are fairly product-oriented, since they introduce a large number of inclusions, suppressions and substitutions, and they usually write their comments in the margin (where they offer many alternatives). They evaluate in a steady way across task sets and text sections (across stretches, however, they seem to be affected by fatigue due to their particularly thorough behavior, in general), especially on non-normalized phenomena, but at all levels of saliency or coincidence. The lenient evaluators appear to be more confident than their demanding evaluator colleagues, as they mark more singular phenomena (zero coincidence), especially in the sets where they presumably feel more comfortable with the subject matter (DP1 and DP3). Their number of non-negative reactions increases per set stretch; on the whole, they have more negative and positive reactions than demanding evaluators. In short, the orientation of the lenient evaluators seems to be to assess the whole text, introduce information to improve the translations, and underscore positive aspects. They also show a greater commitment to the research, since they were willing to

assess each translation as if it really were intended for their producers (translators), even though they knew their work would be evaluated by a researcher. This behavior contrasts with that of their demanding evaluator colleagues, whose characteristics are summarized below.

Demanding evaluators intervene less (but increase their level of demand) in stretch III, and more in the technical sets, in the second of which their judgments are especially low. They tend to be feedback-oriented: they use many highlights, classifications and comments. These refer (slightly) more to external sources; and subjects prefer to introduce them in the text and in the final sections, which could affect their level of demand for several reasons: comments are more noticeable, they make the text messier, and are probably fresher in the evaluators' minds when they are making the decision to assign the quality judgment. The demanding evaluators also tend to write comments in text endings and (typographically) outstanding segments; and they express more uncertainty than lenient evaluators, possibly because they are more self-demanding or, perhaps, because they feel less confident about their corrections. They may also feel under greater scrutiny than their lenient evaluator colleagues, or more aware of the *experimental research* situation in which they are taking part, a factor that could also explain their reluctance to mark repeated phenomena. They identify more combined phenomena and their negative reactions decrease across stretches. Further, they express more neutral and very negative reactions. The demanding evaluators ultimately seem to aim to formulate an idea on the quality of the text, so they stop marking phenomena when they have reached an opinion. Nonetheless, they may lack self-confidence, which they could be trying to offset by turning to external sources and reacting very negatively when dealing with errors they do feel sure about.

All in all, both groups show particular characteristics, which might make them better suited to different environments and aims. While demanding evaluators do not contribute (as much) to improving the final product, they seem better suited for evaluating in professional environments, since their behavior is less time-consuming. Lenient evaluators, who are more informative and thorough, appear to be more suited to research. In teaching environments, demanding evaluators might come closer to constructivist approaches whereas lenient evaluators would better fit roles where they are expected to transmit information, rather than prompting trainees to find out solutions on their own. Hence, lenient evaluators might be better suited to initial stages, and demanding evaluators, to advanced training.

This work may be considered innovative in that the related literature contains few descriptive antecedents on the actual behavior of translation evaluators. Perhaps because of this singularity, some scientific criteria could be no doubt improved in future experiments on the evaluation of translations. In particular, ecological validity would be enhanced if the evaluators' performance could be measured with real textual genres and commissions, but in such instances the comparison between different population groups would be practically impossible. Moreover, the experiment would have been scientifically more economic (Neunzig 1999: 10-15), in terms of effort (for both the evaluators and the researcher), if it had used shorter originals, since the evaluation of the subjects generated a tremendous amount of data. However, some good decisions were also taken. The evaluators' behavior could be quantified due to the emergence of a number of variables and categories; an effort was also made

to check the reliability of the experiment through indexes that counterbalanced the differences in size of the original texts and their corresponding translations. Other criteria taken into account in the design and implementation of the experiment were those of applicability (Neunzig 1999: 10-15), objectivity, replicability and generalizability (Orozco 2001: 99-100).

In sum, more studies on evaluative profiles are needed. Research could also examine how the inclusion of an intermediate level of evaluators (between *lenient* and *demanding*) affects these results. But perhaps the most interesting further research line would be to observe the purpose and actual process of evaluation in translation companies and agencies, so that the methods and criteria used in the professional world could be transferred to teaching environments (especially in advanced courses). This is an enormous challenge, but it would be worth the effort, given the confusion surrounding the activity and the lack of agreement among evaluators, but mostly because evaluation is an unavoidable necessity in all translation environments. And also because quality, a crucial concept for the profession, is intimately linked to the idea we have of what it means to evaluate.

NOTES

1. Downey, Maureen (Updated last: 3 March 2010): *Students use online reviews to find easy graders on faculty*. Visited 6 August 2010, <http://blogs.ajc.com/get-schooled-blog/2010/03/03/students-use-online-reviews-to-find-easy-graders-on-faculty/>.
2. Kvanvig, Jon (Updated last: 11 July 2008): *Evaluating Faculty Quality, Randomly*. Visited 1 August 2010, <http://el-prod.baylor.edu/certain_doubts/?p=844#comments>.
3. For a fuller explanation of the expected impact and actual significance of order on the evaluation, see Muñoz and Conde (2007).
4. As outlined in section 2, phenomena of zero coincidence were marked by only one evaluator; very low coincidence, by 2 or 3 evaluators; low coincidence, by between 4 and 12 evaluators; medium coincidence, by between 13 and 25 and high coincidence, by over 25 evaluators.
5. This article is based on a Ph.D. thesis submitted at the University of Granada, Faculty of Translation and Interpreting (Conde 2009). The focus of the thesis was on the process and result of translation evaluation.

REFERENCES

Adab, Beverly (2002): The Translation of Advertising: A Framework for Evaluation. *Babel*. 47(2):133-157.

Bowker, Lynne (2000): A Corpus-Based Approach to Evaluating Student Translations. *The Translator*. 6(2):183-210.

Conde, Tomás (2008): La evaluación de traducciones, a examen. *In*: Mariela Fernández and Ricardo Muñoz, eds. *Aproximaciones cognitivas al estudio de la traducción*. Comares: Granada, 67-100.

Conde, Tomás (2009): *Proceso y resultado de la evaluación de traducciones*. Doctoral thesis, unpublished. Granada: Universidad de Granada.

Conde, Tomás (2010): Propuesta para la evaluación de estudiantes de traducción. *Sendebar*. 20:245-269.

Fox, Olivia (1999): *The evaluation of inter- and intra-rater reliability in the application of uniform and diverse correction criteria: a case study*. Internal paper, unpublished. Barcelona: Universitat Autònoma de Barcelona.

Gouadec, Daniel (1980): Paramètres de l'évaluation des traductions. *Meta*. 36(2):96-116.

Gross Davis, Barbara (Updated last: 1 September 1999): *Motivating students*. Visited on 6 August 2010, <http://orgs.bloomu.edu/tale/documents/Davis_Motivating_Students.pdf>.

Haiyan, Li (2006): Cultivating Translator Competence: Teaching & Testing. *Translation Journal*. 10(3). Visited 9 August 2010, <http://accurapid.com/journal/37testing.htm>.

Hoffman, Alexander (Updated last: 24 June 2008): Review of *Measuring Up: What Educational Testing Really Tells Us*. Visited 6 August 2010, <http://www.amazon.com/review/R3PU9U3ELR9YNO>.

Hönig, Hans G. (1997): Positions, Power and Practice: Functionalist Approaches and Translation Quality Assessment. *Current Issues in Language and Society*. 4(1):6-34.

House, Juliane (2001): Translation Quality Assessment: Linguistic Description versus Social Evaluation. *Meta*. 46(2):243-257.

Lauscher, Susanne (2000): Translation Quality Assessment. Where Can Theory and Practice Meet? *The Translator*. 6(2):149-168.

Li, Defeng (2006): Making Translation Testing More Teaching-oriented: A Case Study of Translation Testing in China. *Meta*. 51(1):72-88.

Muñoz, Ricardo (2007): Evaluación, corrección, revisión y edición. *In:* Juan Cuartero and Martina Emsel, eds. *Brücken: Übersetzen und interkulturelle Kommunikation. Festschrift für Gerd Wotjak zum 65. Geburtstag*. Frankfurt: Peter Lang, 255-268.

Muñoz, Ricardo and Conde, Tomás (2007): Effects of Serial Translation Evaluation. *In*: Peter A. Schmidt and Heike E. Jüngst, eds. *Translationsqualität*. Frankfurt: Peter Lang, 428-444.

Muzii, Luigi (2006): Quality Assessment and Economic Sustainability of Translation. *Rivista internazionale di tecnica della traduzione*. 9:15-38.

Neunzig, Wilhelm (1999): *Sobre la investigación empírica en traductología – cuestiones epistémicas y metodológicas*. Doctoral dissertation, unpublished. Barcelona: Universitat Autònoma de Barcelona.

Nord, Christiane (1991): *Translation as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

Orozco, Mariana (2001): Métodos científicos en traducción escrita: ¿Qué nos ofrece el método científico? *Sendebar*. 12:95-115.

Pym, Anthony (1991): Translation Error Analysis and the Interface with Language Teaching. *In*: Cay Dollerup and Anne Loddegaard, eds. *Teaching Translation and Interpreting Training Talent and Experience*. Amsterdam/Philadelphia: John Benjamins, 279-288.

Pym, Anthony (2004): Propositions on cross-cultural communication and translation. *Target*. 16(1):1-28.

Varela, María-José and Postigo, Encarnación (2005): La evaluación en los estudios de traducción. *The Translation Journal*. 9(1). Visited 9 August 2010, <http://accurapid.com/journal/31evaluacion.htm>.

Waddington, Christopher (1999): *Estudio comparativo de diferentes métodos de evaluación de traducción general (inglés-español)*. Doctoral thesis. Madrid: Universidad Pontificia de Comillas.

Waddington, Christopher (2001): Different Methods of Evaluating Student Translations: The Question of Validity. *Meta*. 46(2):311-325.