QUALITY RESEARCH REVISITED

Franz Pöchhacker
Center for Translation Studies, University of Vienna

1. Introduction

Considering the broad range of topics and the great diversity of research approaches in the field of interpreting studies, research on quality in interpreting stands out as an impressively rich and cohesive area of study. One line of investigation in particular – survey research on interpreters' and users' quality expectations and preferences – has been around for about twenty years and could be said to form a distinct research model, or 'paradigm' (in the narrower sense often used in various sciences). As such it is productive in various ways: it embodies a set of underlying theoretical assumptions and thus supplies the necessary conceptual framework for empirical research. Crucially perhaps, it also consolidates a set of methodological choices, thereby facilitating repeated application (replication). This in turn helps extend the base of empirical data from which conclusions may be drawn. As an accepted standard of sorts, the research model offers a working method that can readily be adopted also by less experienced investigators.

At the same time, and on a different level, a research model's prominence may also expose it to closer scrutiny within the scientific community. Careful (re)examination of its conceptual and methodological choices will put the research model to the test and either confirm or question its validity. Either way, such methodological criticism serves to consolidate and refine research practices and results. It is this hopeful assumption that lies at the heart of the present paper, which revisits and critiques some studies on interpreters' and users' quality expectations and preferences. Most of the revisiting will be done in rather practical methodological terms, with an emphasis on statistical procedures for the analysis of survey data. Aside from this re-analysis component, the paper also doubles as a review of some recent research, with special emphasis on methodological issues and on the gatekeeping function of the editorial process leading to quality publications. In either dimension, my discussion will pivot on a recent paper by Delia Chiaro and Giuseppe Nocella, of the University of Bologna, which both raises important methodological doubts about previous studies and prompts some concerns about research published in our field.

2. A reliable springboard

Like any piece of serious research, the present contribution should begin by reviewing the state of the art. Given the breadth of the topic, however, the scope of such a review must be strictly limited. It would be impossible here to summarize the expansive literature on quality in interpreting, as reflected, for instance, in the bibliography by Shlesinger (2000) and in the two proceedings volumes of the international conference on the topic convened in 2001 by Ángela Collados Aís of the University of Granada (Collados Aís *et al.* 2003a, 2003b). The same is true of survey research on interpreting quality, which has been the subject of several review papers (e.g. Kurz 2001a, 2003; Pöchhacker 2001). Indeed, I will (have to) narrow my focus to one particular line of investigation, namely questionnaire-based surveys on the quality criteria and expectations of conference interpreters and users of simultaneous interpreting (SI) – QE research, for short.

QE research was pioneered in the 1980s by colleagues at the University of Vienna, Hildegund Bühler (1986) and Ingrid Kurz (1989). Their work proved seminal to most subsequent efforts, including the user expectation study commissioned by AIIC (Moser 1996) and the 'matched-guise' experiments by Collados Aís (1998, 2002) and Garzone (2003). Most recently, an innovative survey using the World Wide Web (Chiaro and Nocella 2004) has again shone the spotlight on these 'classic' studies, albeit in a rather exposing way. Before reporting their empirical study, Chiaro and Nocella (2004) offer a review of methodological issues in quality-oriented research, including a rather harsh critique of Bühler (1986) and Kurz (1989, 1993). Their paper can therefore serve as a convenient peg both for a more detailed account of the studies in question and for addressing some basic methodological problems.

2.1. The interpreters' perspective

Chiaro and Nocella (2004) depart from the observation that "there appears to be little harmony concerning which perspective to take when undertaking research" (279). Framing their choice as one between the perspectives of the interpreter and the user, they opt for the former to provide "a helpful starting point" and hope for their findings to serve as "a reliable springboard for further research" (279).

Though Chiaro and Nocella supply no further rationale for adopting the interpreters' perspective, it is obvious from their research design that it was actually Bühler (1986) who provided the springboard for their survey: "The criteria used in this investigation are the same as those used by Bühler (...)" (Chiaro and Nocella 2004: 283).

As described very briefly by Chiaro and Nocella (2004: 282), "the well-known study conducted by Bühler (1986)" was based on a list of sixteen "linguistic" (performance-related) and "extra-linguistic" (interpreter-related) criteria which Bühler suggested AIIC members might consider more or less important when sponsoring candidates for membership. Bühler's all too sparse description, in an endnote, of her sample of 41 interpreters who received and returned the questionnaire "at the Council Meeting and the International Symposium [...] convened by AIIC in Brussels in January 1984" (1986: 233-234) does not draw any critical remarks; rather, it is her results that lead Chiaro and Nocella to conclude that "something was faulty in the research design of the study" (2004: 283). According to Chiaro and Nocella (2004: 282), "interpreters valued most of the items as important or highly important, thus highlighting their difficulty in assigning an order of importance". This assessment, according to which "the interpreters were incapable of discriminating and were giving equal importance to all the criteria" (283), invites a look at Bühler's actual findings. Figure 1 was drawn up on the basis of the percentage values published as an annex to Bühler's paper (1986: 235).

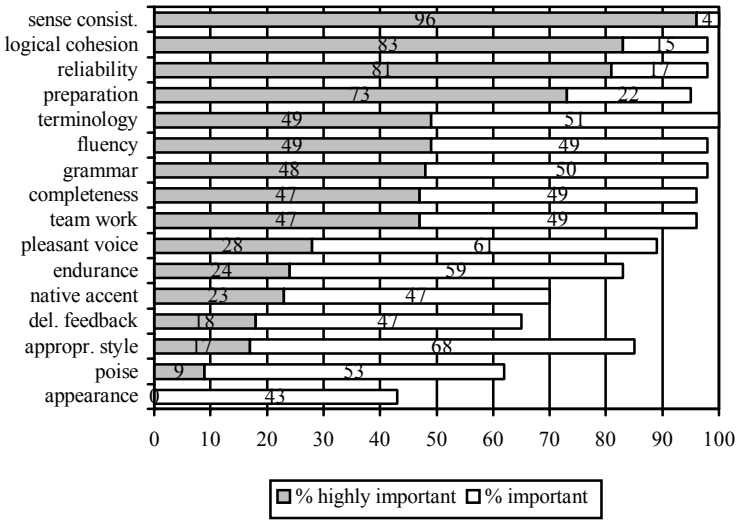

Figure 1. Quality criteria rated as "(highly) important" by 47 AIIC members (Bühler 1986)

Ordered according to the percentage of respondents who gave a rating of "highly important", the sixteen criteria displayed in Figure 1 reflect a rather clear-cut differentiation, from the top-rated demand for "sense consistency with original message" to the least important criterion, the interpreter's "pleasant

appearance", which a majority of respondents considered "less important" (43%) or "irrelevant (13%). While it is true that all other criteria received a rating of at least "important" from a clear majority of respondents, exclusive use of the two highest ratings was made for only two criteria – "sense consistency with original message" and "use of correct terminology".

It may also be noted that among the nine top-ranking criteria in Figure 1 (at least 47% "highly important") there are three interpreter-related ("extra-linguistic") qualities: "reliability", "thorough preparation of conference documents" and "ability to work in a team". This is of interest here because subsequent QE surveys – up to the study by Chiaro and Nocella – largely neglected Bühler's extra-linguistic criteria, so that comparisons have been possible only for her output-related ("linguistic") criteria.

2.2. Interpreters vs. users

The shift from conference interpreters' criteria for sponsoring AIIC candidates – and, presumably, for a "first class interpretation" (cf. Bühler 1986, note 2) – to the expectations of end-users was brought about by Ingrid Kurz, who questioned Bühler's (1986: 233) assumption that her criteria "reflect the requirements of the user as well as [the] fellow interpreter". Narrowing down the list of criteria to the first eight items in Bühler's questionnaire, Kurz (1989) introduced a comparative view on quality expectations, most famously presented in her 1993 paper on "expectations of different user groups" in *The Interpreters' Newsletter* (reprinted in *The Interpreting Studies Reader*).

While there is no need here to say more about Kurz' (1993) widely noted findings, the ostensible methodological weaknesses of her work, as pointed out by Chiaro and Nocella, require closer examination. Chiaro and Nocella (2004: 282) observe that "Kurz' samples were very small and uneven" and even speak of "discouragingly poor returns". Given the actual number of respondents (124), this critique is hardly justified. One might point out, for instance, that the sample size of the AIIC survey (Moser 1996), in which 94 interpreters conducted questionnaire-based interviews at 84 different meetings with a total of 201 conference participants, by no means dwarfs what was achieved single-handedly by Kurz in three conferences. Her sample, made up of participants in a medical conference (47), a meeting of engineers on quality control (29) and a Council of Europe meeting on equivalences (48), also compares well with the work of Vuorikoski (1993) and Mack and Cattaruzza (1995), who had 177 and 75 questionnaires, respectively, completed at five meetings with SI.

Elsewhere in their paper, Chiaro and Nocella (2004: 284) speak of "the 10-15% rate of questionnaire returns that is normal for traditional surveys". Assuming that not all participants would make use of the SI services offered,

Kurz' four dozen questionnaires each from two of her meetings could easily amount to a 15% response rate in a conference with some 400 participants. Admittedly, though, this conjecture may well err on either side, and it is indeed regrettable that no information on the number of questionnaires distributed is available. A laudable model in this regard is provided by Mack and Cattaruzza (1995: 40), whose return rate, incidentally, was three times higher (roughly 80% to 90%) in meetings where the survey had been announced to the participants than in meetings without such announcement (roughly 25%). Again, it is not known for Kurz' surveys how the questionnaires were brought to the attention of the conference participants.

While these methodological shortcomings go unmentioned, Chiaro and Nocella level a different, rather curious charge against Kurz' (1993) work, namely that her questionnaire was "administered in three very different moments in time and in different contexts, thus weakening the rigour of the experiment" (2004: 282). Though one may well ask for more detailed information on the meetings concerned, it is hard to see how the aim of studying different user groups could be achieved without surveying participants in different meetings, as was indeed done purposely also in the AIIC survey (Moser 1996).

In the abstract of their paper, Chiaro and Nocella (2004: 278) note that "research undertaken so far is surprisingly lacking in methodological rigour". In the text, at the outset of their review of methodological issues, they similarly state that "attempts at more scientific research in interpreting often appear to be based on rather uncertain methodological principles" (279). Aside from the shortcomings mentioned above, the most serious criticism brought against the studies by Bühler (1986) and Kurz (1993) would seem to concern their statistical analysis of the data. According to Chiaro and Nocella (2004: 283), "a substantial shortcoming of this particular study is that the mean was used as the descriptive statistic for analysing and discussing data and drawing conclusions when dealing with ordinal data". And here they have a point. Though Chiaro and Nocella voice this criticism, erroneously, with reference to Bühler's (1986) study (cf. Fig. 1) and are more benign toward Kurz' statistical analysis, the latter does indeed suffer from the infelicitous choice of using the arithmetic means to describe her ordinal data. Having asked her respondents, as Bühler did, to rate the individual quality criteria on a four-category scale ("highly important" – "important" – less important" – "irrelevant"), Kurz (1993) should have described her results, as Bühler did, in terms of the percentages for the various ratings. Essentially, the intervals between the four items making up the scale cannot be assumed to be the same, so metric conversion is, strictly speaking, not permissible. But even if Kurz had used a four-point metric scale, e.g. ranging from "least important" to "most important", with numbered values in-between,

statisticians would be wary of using the arithmetic mean to describe the data because too much of the variability and actual distribution of the data between "1" and "4" may be lost to an average value in the middle. Whereas some would accept such calculations for a five-point metric scale, many authors suggest that rating scales analyzed in terms of means should consist of at least seven points (cf. also Gile 1983: 241).

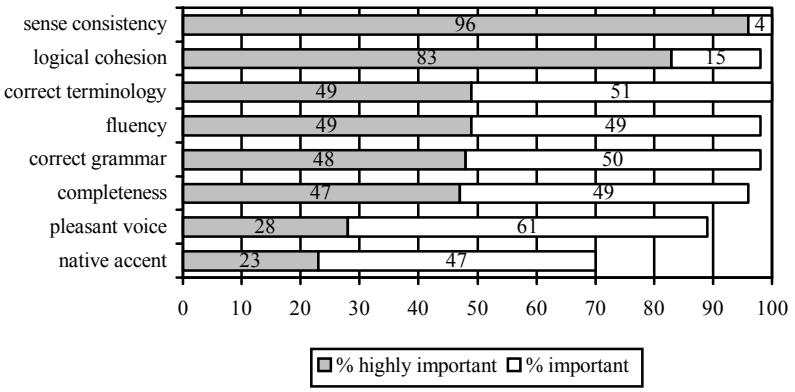

Figure 2a. Eight criteria as rated by 47 AIIC members (based on Bühler 1986)

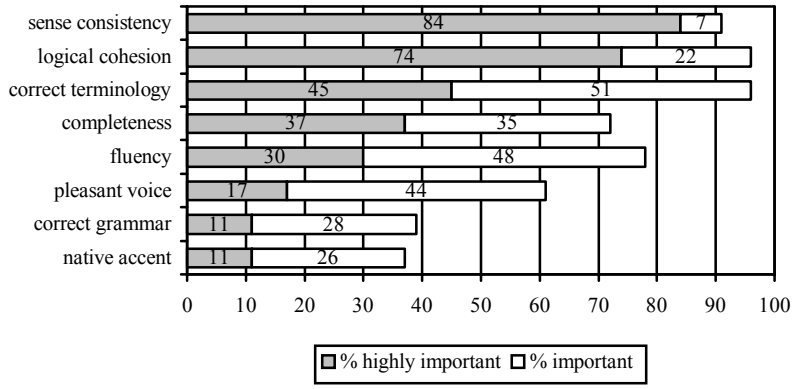

Figure 2b. Eight criteria as rated by 47 medical conference participants (based on Kurz 1989)

It should be noted, however, that a description in percentages was in fact offered in Kurz (1989), where the values for ratings of "highly important" and

"important" by AIIC interpreters and medical conference participants (47 each) were juxtaposed in a table. Using 'valid percent' of responses, i.e. percentages adjusted for the 2 missing values in Bühler's and the 9 missing responses in Kurz's data,[1] the results can be visualized as shown in Figures 2a and 2b.

As discussed in detail by Kurz (1989), conference participants (MDs) generally tended to give lower ratings than the AIIC members in Bühler's study. A noteworthy exception is "use of correct terminology", which was rated "important" by 51% of interpreters and users alike and for which the interpreters' ratings of "highly important" were only slightly higher (49% vs. 45%). It is also evident that the two criteria given the least importance, "pleasant voice" and "native accent", have distinctly lower percentage ratings among the SI users at the medical conference.

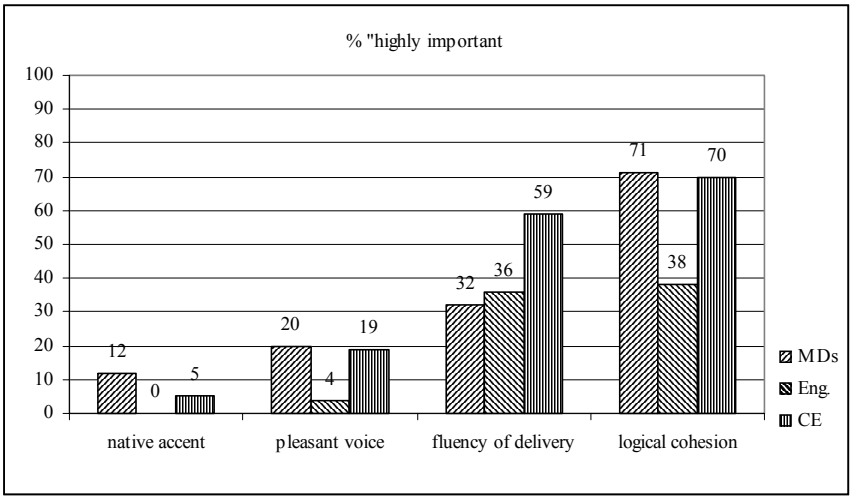

Figure 3a. Quality criteria ratings (% "highly important") by three user groups (cf. Kurz 1993)

While it is thus quite feasible to compare the findings of Bühler (1986) and Kurz (1989) in terms of percentages, the matter is more difficult in the case of Kurz (1993), which requires a comparative analysis of at least three sets of findings.[2] Tables or charts of percentages describing Kurz' user expectation

1 Special thanks are due to Ingrid Kurz, who kindly provided me with her original survey materials for statistical reanalysis.

2 It was for this purpose that Kurz (personal communication) enlisted the help of someone with training in statistics – and was supplied with mean values for comparative analysis of her data (see in particular Kurz 1993: 16-17).

dataset are likely to yield a rather complicated picture. Figures 3 is an attempt to describe the ratings of "highly important" for the three different user groups: medical conference participants (MDs), engineers (Eng.) and Council of Europe delegates (CE). To facilitate comparison with the charts published in Kurz (1993: 17), the eight criteria are shown in two charts (Figs. 3a and 3b).

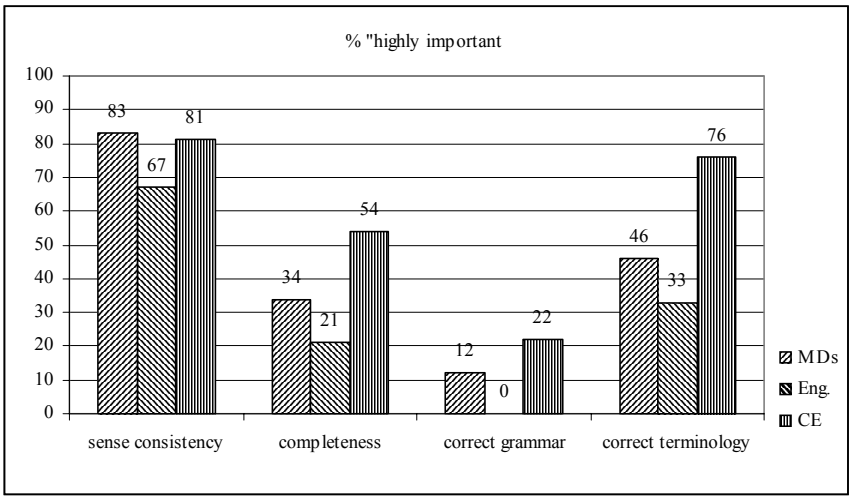

Figure 3b. Quality criteria ratings (% "highly important") by three user groups (cf. Kurz 1993)

Such charts, which offer a rather detailed but cumbersome description, could be drawn up for all four response options. Ideally, however, our statistical analysis should not stop at mere description but should help us understand what the various differences in the data mean – if they mean anything at all. In other, statistical words, we would ask whether these differences are significant, that is, based on some principled relationship in the data, or whether they are equally likely to result from pure chance. Although I cannot claim any special statistical expertise, I will attempt such an analysis in the section below, using some widely available analytical tools.[3] While my main goal here is to illustrate a few basic methodological options in processing ordinal data, the analysis will also serve to test some of the longest-standing findings in QE research for their statistical significance.

3 The statistics software SPSS for Windows (version 12.0) was used to process the data and perform the various calculations and tests.

3. Significance

3.1. Crosstabulation

In examining Kurz's (1993) ordinal data for significant relationships between the three user groups, the most elementary option would be crosstabulation. This involves the cross-classification of two categorical variables – in our case, a given criterion's degree of importance (an ordinal variable) and the nominal variable of 'user group'. The four response options ("highly important", "important", "less important", "irrelevant") and the three user groups (MDs, Eng., CE) result in a three-by-four data matrix for each of the eight criteria. It is on the basis of such contingency tables that various measures of association can be calculated. Chief among them is the chi-square test, a nonparametric test that compares observed frequencies to their expected values.

Unfortunately, the sample of 124 respondents is not quite large enough to ensure an adequate number of expected values in all twelve cells of the three-by-four table. For each criterion the distribution yields at least two cells (20% of cells and more) for which the expected frequency in the chi-square test is smaller than five, which renders any interpretation of the test invalid.

For a chi-square test to be viable for the given data set, the values should have a more balanced distribution. This can be achieved by collapsing some categories containing low-frequency values. When this is done by recoding "less important" and "irrelevant" into a single value ("not important"), crosstabulation yields better results. Though there are still too many cells with low-frequency values in the tables for four of the criteria, Pearson's chi-square test indicates a significant relationship in two cases, namely "completeness of interpretation" (Table 1) and "correct grammatical usage" (Table 2).

		User group			Total
		MDs	Eng.	CE	
not important	Count	13	6	5	24
	% of group	28.3%	20.7%	10.4%	19.5%
important	Count	16	17	16	49
	% of group	34.8%	58.6%	33.3%	39.8%
highly important	Count	17	6	27	50
	% of group	37.0%	20.7%	56.3%	40.7%
Total	Count	46	29	48	123
	% of group	100%	100%	100%	100%

Pearson chi-square = 13.103; $p = .011$
(0 cells with expected frequency < 5; min. exp. = 5.66)

Table 1. Crosstabulation for "completeness of interpretation"

As can be seen from the percentages in Table 1, the participants in the Council of Europe meeting attributed significantly more importance to "completeness" (56% "highly important" vs. 10% "not important") than either medical doctors (37% vs. 28%) or engineers (21% vs. 21%). According to Pearson's chi-square test, this difference is clearly significant at the 95% confidence level ($p < .05$) and even approaches significance at a probability level of 99% ($p < .01$).

		User group			Total
		MDs	Eng.	CE	
not important	Count	28	22	19	69
	% of group	60.9%	78.6%	40.4%	57.0%
important	Count	13	6	19	38
	% of group	28.3%	21.4%	40.4%	31.4%
highly important	Count	5	0	9	14
	% of group	10.9%	0%	19.1%	11.6%
Total	Count	46	28	47	121
	% of group	100%	100%	100%	100%

Pearson chi-square = 12.512; $p = .014$
(1 cell (11.1%) with expected frequency < 5; min. exp. = 3.24)

Table 2. Crosstabulation for "correct grammatical usage"

As regards users' differential appreciation of "correct grammatical usage", the significant relationship confirmed by Pearson's chi-square test clearly holds between the ratings of engineers and Council of Europe delegates. Whereas the former assign particularly little importance to grammatical correctness (79% "not important"), a majority of CE delegates consider it "important" (40%) or even "highly important" (19%). Again, the difference is highly significant ($p = .014$).

3.2. Other nonparametric tests

Aside from the chi-square test, there are other nonparametric tests for identifying significant relationships among different sets of rank-ordered data. The most appropriate procedure here is the Kruskal-Wallis *H*-test, applied to multiple independent samples for determining whether the values of a particular variable differ between two or more groups. The Kruskal-Wallis test, which involves comparisons of rank orders, can be viewed as the nonparametric

equivalent of the one-way analysis of variance (ANOVA) commonly used to determine whether the means of various groups are significantly different.[4]

	Chi-square	df	Asymptotic significance
1. native accent (n=123)	.595	2	.743
2. pleasant voice (n=121)	.987	2	.610
3. fluency of delivery (n=113)	12.468	2	.002
4. logical cohesion of utterance (n=118)	10.798	2	.005
5. sense consistency with original message (n=120)	1.843	2	.398
6. completeness of interpretation (n=123)	9.558	2	.008
7. correct grammatical usage (n=121)	11.766	2	.003
8. use of correct terminology (n=124)	19.122	2	.000

Table 3. Results of Kruskal-Wallis test for quality ratings by user group

The results of the Kruskal-Wallis test for the ordinal data under study (Table 3) indicate group-related differences significant at the 99% confidence level for five of the eight criteria (cf. note 4). For "native accent" and "pleasant voice" as well as "sense consistency with original message", quality expectations are not significantly different among the three user groups. For the remaining criteria, paired tests are required to identify the nature and location of the differences between groups. This can be done using the Mann-Whitney *U*-

4 If the means used in Kurz (1993) were accepted as a valid descriptive statistic, the test used to identify significant differences among the three user groups would be an analysis of variance. Its results, calculated for illustration, indicate significant relationships in four of the eight criteria: fluency, $F(2,110) = 7.037$, $p = .001$; logical cohesion, $F(2,115) = 3.79$, $p = .025$; completeness, $F(2,120) = 5.056$, $p = .008$; and correct terminology, $F(2,121) = 9.958$, $p = .000$. (The values for correct grammar fail the preliminary test for homogeneity of variances and must therefore be excluded from the interpretation.) Upon further examination in paired post-hoc tests (e.g. Bonferroni), particularly clear-cut differences are found for completeness and correct terminology, where the mean ratings of Council of Europe delegates differ significantly from each of the other groups (cf. Table 4).

test, which tests for significant differences between two independent samples. The Mann-Whitney test results for the three possible comparisons (MDs vs. Eng., MDs vs. CE, Eng. vs. CE) suggest that the medical doctors have the least to do with the overall between-group differences: Only one criterion in comparison with engineers shows a significant relationship (logical cohesion, $p = .002$), and four criteria are significantly different in relation to Council of Europe delegates (fluency, $p = .001$; completeness, $p = .022$; correct grammar, $p = .046$; correct terminology, $p = .002$). It is the comparison between the latter and the engineers that yields significant differences for all five of the criteria identified as significant by group in the Kruskal-Wallis test (Table 3). For illustration, detailed results are shown in Table 4.

	Group	N	Mean Rank	Sum of Ranks	Mann-Whitney U	Asymptotic Significance (2-tailed)
fluency of delivery	Eng.	26	29.29	761.5	410.5	.037
	CE	43	38.45	1653.5		
logical cohesion of utterance	Eng.	28	29.70	831.5	425.5	.013
	CE	44	40.83	1796.5		
completeness of interpretation	Eng.	29	30.10	873.0	438.0	.003
	CE	48	44.38	2130.0		
correct grammatical usage	Eng.	28	27.84	779.5	373.5	.001
	CE	47	44.05	2070.5		
use of correct terminology	Eng.	29	27.09	785.5	350.5	.000
	CE	48	46.20	2217.5		

Table 4. Results of Mann-Whitney *U*-test for differences between groups "Eng." and "CE"

3.3. Significance and meaning

This (re)analysis of Kurz' user surveys has focused on the statistical options and tools for describing the data and examining them for significant associations between them. It has highlighted in particular the importance of choosing the appropriate procedures in accordance with the nature of the data and the assumptions holding for various analytical tools. While a thorough understanding of statistics would be highly desirable for anyone carrying out such analyses, it is suggested here by way of demonstration that PC-based statistics software has become accessible enough to be used, with proper guidance, also by the 'semi-skilled' analyst.

However, as much as some statistical know-how can and should well be expected of interpreting reseachers today, the above exercise in significance testing should not obscure the fact that analyzing empirical data, whether from survey research, fieldwork or experiments, is not a question of mathematical skills but, essentially, a matter of meaningful interpretation, of making sense of the relationships indicated by the data. In other words, a statistical significance test does not explain anything but merely points reliably to what needs to be explained. Such (possible) explanations of their survey findings are amply discussed in the papers by Bühler (1986) and Kurz (1989, 1993), and there is neither need nor space in this methodology-oriented paper to revisit this – crucial – part of QE research. Two comments may be in order, though, since they relate to fundamental issues of research methodology (see also section 4.2 below).

One is prompted by the rather striking findings for the role of terminological correctness. "Use of correct terminology" ranked high in Bühler's (cf. Figs. 1 and 2a) as well as Kurz's (1989) findings (Fig. 2b), and was also given special attention by Mack and Cattaruzza (1995), who even found correct terminology to be the top-rated criterion (cf. also Kopczyński 1994). Bühler, herself an expert in the area of terminology, had argued that "[o]ne has to use correct terminology if one aspires to render the message faithfully" (1986: 232). Acknowledging this reasoning, Kurz (1989: 144) also suggested that "the strong emphasis on correct terminology observed here may well be a specific feature of medical (and other highly technical) conferences". When she put this assumption to the test in her subsequent surveys, the prominent role of correct terminology was undiminished but showed a clear peak among Council of Europe delegates (cf. Fig. 3b). Kurz (1993) sought to explain this finding with reference to the institution-specific terminology of international organizations. Judging from the program of the CE conference in question, however, one should also consider an alternative explanation, as suggested also by Mack and Cattaruzza (1995: 46-47). The conference, held in Vienna and Budapest under the auspices of the Council of Europe, was devoted to equivalences in education, that is, the comparability and recognition of certificates and degrees granted by institutions of secondary and higher education in Europe. On the face of it, interpreters at that meeting would have grappled with the rendition of concepts linked to different sociocultural traditions and institutions – a daunting translational task in any case, which was probably not made any easier by the organizers' request, in the preliminary conference program, that speakers limit their oral presentations to five minutes. In this light, it is quite conceivable that the thematic context of the meeting made terminology a prized asset to the proceedings, and that the CE delegates' high expectations for terminological correctness were a function of the conference topic, if not the actual interpreting

services received. In her conclusions, Kurz (1993: 20) makes explicit reference to "the importance of situationality and communicative context" for her comparative study as such; based on the information available, it appears that this awareness should extend also to the situational and thematic context in which her QE survey data were collected.

This methodological issue in data collection, which bears on the interpretation of the survey findings, is connected to another point that may deserve further consideration, namely the language used to collect responses. Kurz used a bilingual (English/German) questionnaire (see Kurz 1996: 57) in the first two of her surveys (MDs and engineers) and an English-only version in the CE meeting. One might therefore ask whether the language in which respondents (MDs and engineers) filled in the questionnaire could have influenced the results. Crosstabulation of the (three-category) ratings by language indeed reveals such an effect for the criterion of completeness, which received significantly higher ratings from the 39 respondents using the English version than from the 36 German-language users (Pearson's chi-square; p = .005). When analyzed by conference (MDs vs. Eng.), this effect appears to obtain irrespective of user group (Mann-Whitney *U*-test; MDs: p = .034, Eng.: p = .018). As for a possible explanation of this finding, it may again be of a methodological nature. Bühler's English term "completeness of interpretation" was rendered in German as "vollständige Wiedergabe des Originals" (complete rendition of the original). One might speculate whether the greater redundancy of the German version, which foregrounds "rendition" rather than completeness ("Vollständigkeit"), led German-language users to give lower ratings to this criterion, not least because it followed immediately upon "sense consistency with original message", another "a priori" feature of interpreting. Additional support for this hypothesis might be seen in the fact that the CE delegates, who received only the English version of the questionnaire, gave significantly higher ratings to completeness than the other two groups (see Table 1).

4. The way forward

The re-examination of previous QE research findings undertaken in the previous section essentially suggests that progress in interpreting studies, especially with regard to research methodology, may come not only from the introduction of novel techniques but also from a more detailed, critical engagement with previous work. This applies in particular to the recent contribution by Chiaro and Nocella (2004), whose criticism of previous QE research prompted the discussion offered in the preceding sections, and whose own research will be reviewed and used as a starting point for additional methodological reflections in the sections to follow.

4.1. Interpreters on the Web

With a keen awareness of methodological limitations in previous QE research, apparently inspired by Gile's (1994) critical view of research skills in interpreting studies, Chiaro and Nocella report an innovative study in which "great care was taken (...) not to fall into the traps that previous studies had failed to avoid." (2004: 283). With Bühler's (1986) criteria as their starting point, the authors drafted a questionnaire which included quality criteria as well as background variables (age, place of birth, qualifications, experience). Rather than a rating of individual criteria on a scale with several response options, the survey instrument designed by Chiaro and Nocella (2004) called for a ranking of the criteria in descending order of importance, i.e. from the most important to the least important item in the list. The questionnaire was administered through the World Wide Web by sending out 1,000 invitations by e-mail "to interpreters belonging to several professional associations" (284). A total of "286 conference interpreters across five continents" responded to the web-based survey (279).

The sample was 29% male and 71% female, with a mean age of 45 years and an average of 16 years of experience. 44% of respondents had their birthplace in Western Europe and had a degree in interpreting. Chiaro and Nocella also report that the interpreters in the sample are mostly freelancers and that, rather strikingly, "most respondents do not interpret into their mother tongue" (285).

To facilitate the ranking task, the list of quality criteria was offered to the respondents in two groups, "linguistic" and "extra-linguistic", the first of which comprised the first nine items in Bühler's questionnaire (i.e. the eight used by Kurz plus "appropriate style"). Displaying the percentages for the various ranks (first to ninth) for three sets of three criteria, Chiaro and Nocella (2004: 287) find the following pattern of relative importance: "consistency with original", "completeness of information" and "logical cohesion" as the three most important factors, followed by "fluency of delivery", "correct grammatical usage" and "correct terminology", with "appropriate style", "pleasant voice" and "native accent" ranking lowest. These findings are further explored by multidimensional scaling, a statistical technique for plotting the similarity structure found in the data in a two- or three-dimensional conceptual space. The three most important and the three least important criteria are found to cluster at opposite ends of a "discriminating quality" dimension, while grammar and terminology occupy a middle ground and "fluency of delivery" appears in a unique intermediate position.

As regards the set of extra-linguistic criteria, the authors do not find a neat pattern, except for the two top-rated items, "concentration" and "preparation of

conference documents". Results are given as summary scores (from 1932 to 1024), the calculation of which is left unexplained in the paper.[5]

4.2. Methodological issues

There is no doubt that Chiaro and Nocella have tread new ground by harnessing the Internet for QE research among interpreters, and their innovative study deserves praise and recognition. Their use of advanced statistical methods for data analysis is likewise apt to encourage the use of more sophisticated analytical techniques in future studies. And yet, in light of the authors' aspirations to methodological soundness and their somewhat heavy-handed criticism of previous studies, one cannot but question some aspects of research design and presentation that would have demanded more attention.

The first of these weaknesses concerns the authors' conceptual framework as reflected in their use of basic terms. Aside from their liberal use of the term 'experiment' in referring to Kurz' surveys, Chiaro and Nocella base their review section on a two-fold distinction between product analysis and "field work (based upon the results of questionnaire surveys)" (2004: 280). While there are indeed many ways of distinguishing various types of approach, it is not clear how the authors' categorization improves on earlier proposals, such as the four-fold distinction made by Vuorikoski (1993) specifically for the purpose of research on interpreting quality. More critically, though, Chiaro and Nocella use the term "perception" as the principal keyword in their work (and its title), obscuring the fundamental distinction between QE research on generic expectations (as pioneered by Bühler and Kurz) and the direct assessment, or judgment, of an actually perceived interpreting performance, as introduced by Gile (1990) and combined with QE research by Mack and Cattaruzza (1995). This distinction is crucial to the work of Collados Aís (1998, 2002) and Garzone (2003), which has taken user-oriented studies of interpreting quality to a new level. Confounding preferences and perception could therefore be said to fall short of the state of the art.

Another methodological uncertainty concerns the authors' survey instrument, with regard to both design and distribution. Though Chiaro and Nocella (2004: 283) state that their criteria "are the same as those used by Bühler", they actually use 17 rather than 16 criteria, several of which are not the same as those in Bühler's (1986) questionnaire. While a critical appraisal and, if

5 The scores become clear from the questionnaire which the authors kindly provided to me after receipt of a first draft of this paper: Respondents were instructed to give "8 to the most important and 1 to the least important"; the scores were thus calculated by multiplying the rank values by the number of respective responses.

necessary, appropriate modification of previous instruments would certainly be welcome, Chiaro and Nocella do not offer any discussion of this part of their work. There is mention of "several interviews" and "endless brainstorming sessions" with interpreters as the basis for devising the questionnaire (2004: 283), but no explanation why two of Bühler's linguistic criteria were apparently rephrased and five new ones substituted for items in the extralinguistic category.[6] At any rate, it would have been desirable to reproduce the relatively short (one-page) questionnaire in an annex to the paper.

Most consequentially perhaps for a paper boasting an innovative approach to QE research, Chiaro and Nocella (2004) give an all too sparse description of their sampling procedure (see section 4.1). It would be interesting to know *which* professional associations were targeted for the survey and, if AIIC was among them, how individual interpreters were selected from the membership list (which in the case of AIIC includes more than 2,600 entries. It is thus not even clear whether the survey was addressed to conference interpreters only: The indication of workload in terms of "hours per month" (with the minimum reported as 0 and the maximum as 200 hours = about 30 days per month), and the baffling finding that "most respondents do not work into their mother tongue" (285) raises some serious doubts which could easily have been dispelled by asking respondents to indicate their professional affiliation and domain of work.[7]

Another methodological issue in survey research of such a comprehensive scope is the language and cultural context of survey administration. With one third of respondents originating from (though not necessarily residing in) South and Central America and Eastern Europe, one cannot be sure that the questionnaire was equally accessible to all recipients (unless they were included in the sample for having English among their working languages). Moreover, there is some evidence in the literature that preferred interpreting styles may differ from one sociocultural context to another (e.g. Ločmele 2001); Chiaro and

---

6 A number of critical comments are on record regarding the criteria used in QE research, beginning with detailed reflections on possible misunderstandings by Bühler (1986) herself and the immediate "Comment" by Seleskovitch (1986). The fact that Chiaro and Nocella (2004: 290) use "intonation" as a synonym of "fluency of delivery" highlights the problem of definition and the need for terminological clarity.

7 As it happens, the clue can be found in the poorly worded questionnaire item (cf. note 5): "Do you interpret mostly esclusively [sic] towards your mother tongue? (Yes/No)". Nevertheless, further information on respondents' professional domain could also be expected from the last item in the questionnaire ("Is your interpreting: Mostly consecutive / Both consecutive and simultaneous / Mostly simultaneous"), the results for which are not reported.

Nocella do not examine their findings for such differences, or do not report any such attempts in their paper. Even if the interpreting profession in various parts of the world were homogeneous enough to render such linguistic and cultural effects negligible, translation scholars conducting surveys across cultural boundaries should probably be the first to demonstrate an awareness of this delicate methodological issue (see, e.g., Harkness *et al.* 2003).

Contextual effects ought to be considered also in a more concrete sense, as illustrated in connection with particular user expectations in Kurz' (1993) surveys (see section 3.3 above). At least since the comprehensive survey commissioned by AIIC (Moser 1996), QE researchers have been aware that users' (and possibly interpreters') quality criteria may differ depending on the type of conference (large vs. small, technical vs. general). Studies on quality requirements for interpreting in media settings (e.g. Elsagir 2001, Kurz 2001b) are another case in point. Asking interpreters to give an opinion regardless of meeting type (cf. Gile 1989, Pöchhacker 1995) therefore precludes a more differentiated view of quality among the respondents.

The way respondents were asked to give their opinion deserves special attention also in a more technical sense. Asking interpreters to rank rather than rate the individual criteria is of course perfectly valid, and represents an innovative aspect of the study. However, there is some evidence in the literature (e.g. Bradburn and Sudman 1979) that ranking more than five to six items may be an overly difficult task for reliable performance. (As explained by Chiaro and Nocella, ranking Bühler's first nine criteria requires 36 mental comparisons.) In light of the authors' interpretation of Bühler's findings, that respondents had "difficulty in assigning an order of importance" (Chiaro and Nocella 2004: 282), their forced-choice approach for a list of nine items therefore seems less than ideal for bringing out subtle distinctions. In future studies it may be preferable – and more user-friendly – to design the questionnaire as a combination of rating scales and rankings, e.g. with a list of criteria to be rated on a multi-point scale followed by a request to rank the three or five most-important ones in the list.

Another option is the paired-choice approach adopted by Gourevich and Mateeff (1989), who asked 50 experienced interpreters to state a preference for one of each pair of criteria offered to them on 28 test cards (which reflected all possible combinations of eight criteria, including completeness, correctness, usefulness, smoothness, calmness and pleasantness). Though the mathematics of their scaling analysis are daunting, the findings suggest that, despite disagreement among the experts concerning the importance of various characteristics of SI, "correctness" and "usefulness" outweigh prosodic characteristics on the scale of relative importance.

The study by Gourevich and Mateeff (1989) offers an interesting parallel to the work of Chiaro and Nocella. Admittedly, the latter could not easily have

been aware of that paper, published as it was in a rare journal and language.[8] Still, the comparative discussion, or lack thereof, of the survey findings is yet another broadly methodological issue to note. Since Bühler's (1986) pioneering survey constituted their basis and point of departure, Chiaro and Nocella (2004) could be expected to draw some explicit comparative conclusions. Instead, the authors vaguely state that "contrary to common belief, results highlight that interpreters do not consider all the criteria in question as being of more or less equal importance" (291). Leaving aside the rather crude interpretation of Bühler's findings (cf. Fig. 1), the conclusion drawn by Chiaro and Nocella is circular, since the design of the web-based questionnaire did not allow respondents to assign equal importance to any two or more items.

The various problems noted for the authors' handling of the literature and of their own findings, and the methodological issues raised by the design and presentation of their study, bear strongly on the broader theme of research standards in interpreting studies, as addressed most consistently by Daniel Gile (e.g. 1994, 1999). Research training, international and interdisciplinary cooperation, and joint supervision of theses have been suggested as measures to improve the quality of research done in interpreting studies. Such initiatives notwithstanding, a crucial aspect of quality assurance in our field, as in any other scholarly/scientific discipline, is a screening procedure prior to publication. With the article by Chiaro and Nocella (2004) as a case in point, this issue will be discussed in the following, final section of this paper.

4.3. Into print?

For a research paper to be published in an edited volume or academic journal, it has to meet certain requirements with respect to both substance and presentation. One or more editors will usually be responsible for making sure that this is the case. For scientific journals in particular, the editorial process relies heavily on a peer review system, in which colleagues with appropriate expertise examine the manuscript for its contribution to the state of the art, making sure that the research reported is theoretically and methodologically sound. A highly informative description of this process is offered by Gile and Hansen (2004) with reference to the proceedings volume of the 2001 EST Congress in Copenhagen. The following remarks on the paper 'under review' will have to be more anecdotal, but should serve to highlight some of the issues nevertheless.

---

8 Knowledge of that study came to me through Ingrid Kurz, whose cooperation in this endeavor is again gratefully acknowledged.

The research reported by Delia Chiaro and Giuseppe Nocella in volume 47 (2004) of the Canada-based translators' journal *Meta* was conducted in the fall of 2000, prior to the International Conference on Interpreting at Forlì, where the survey and preliminary findings were presented by Giuseppe Nocella. Nocella subsequently submitted his paper for publication in the proceedings which were to be edited by the conference organizers, Giuliana Garzone and Maurizio Viezzi. Instead of the editors' original plan to publish two volumes with a leading international publisher in translation studies, only one book was eventually published in John Benjamins' Translation Library series (Garzone and Viezzi 2002). A second volume was published locally in the same year (Garzone *et al.* 2002). Neither volume contains the paper by Nocella, with whom the present author had exchanged manuscripts by e-mail at the time of submission for the proceedings. Instead, an extended version co-authored by Delia Chiaro appeared in *Meta* two years after the publication of the Forlì Conference proceedings volume(s).

It is difficult to establish to what extent and at what stage in this process the author(s) received feedback from any editorial screening or peer reviewing. A comparison between Nocella's original paper and the joint version, mainly enlarged by the critical review of previous studies, suggests that this was not the case for the shortcomings noted here.

Aside from the fact that peer reviewers might have suggested that Chiaro and Nocella include some key references in their discussion of methodological issues (e.g. Moser-Mercer 1996, Shlesinger *et al*. 1997), referee reports by colleagues with a background in QE research would most probably have pointed out the authors' imprecise use of key terms (e.g. perception); their erroneous criticism of Bühler's analysis; the ambiguity surrounding the criteria in their questionnaire; the missing information on the sampling procedure; and the highly unlikely finding that most interpreters would not work into their mother tongue. Assuming the necessary degree of motivation (cf. Gile and Hansen 2004: 301) and active editorial interest in the reviewer(s), the authors might also have received feedback and recommendations on making their text more focused, particularly in the introductory and concluding sections, and making their statistical analysis more accessible to a wider readership.

Moreover, formal defects of the paper, though not as consequential as issues of research design and interpretation, should not be ignored. A keen reviewer or editor might have noticed, for instance, that the three subheadings in section 2 are on different levels (2.1, 2.1.1, 2.1.2) and thus at odds with the authors' conception of three different methodological perspectives (product analysis, user surveys, interpreter surveys) to which the subheadings refer. (In Nocella's original manuscript, the headings were numbered 2.1, 2.1.2 and 2.1.3, indicating some, albeit unsuccessful, editorial intervention or revision.) A finer point,

which deserves comment only in the context of aspirations to maximum methodological rigor, is the use of unequal scales for the visualization of comparable percentages, as in the authors' Figure 2 (Chiaro and Nocella 2004: 287). More blatantly, in contrast, the consistent misspelling of 'Kopczyński' as "Kopezynski" (282, 293) and other infelicities in the bibliography (Bassnett misspelled; entry for Kopczyński truncated; Kurz 1989 listed as 1988; no data for Tommola's 1995 volume) suggests that the editorial process in this case proved less than fully effective in ensuring optimum standards for the quality of published research.

5. Conclusion

As illustrated by the present review paper on methodological issues in QE research, the field of interpreting studies reflects an evolution toward higher scientific standards at the same time as leaving ample room for improvement with regard to both analytical rigor and editorial procedure. The aspiration to greater methodological sophistication underlying the paper by Chiaro and Nocella (2004) thus deserves special acknowledgment. The authors point to a number of issues in previous research which deserve more critical attention, and their paper is greatly appreciated as a starting point for this endeavor. Unwittingly, however, Chiaro and Nocella, in their commendably innovative study, also provide material for a critical discussion of methodological rigor in quality research. While offering a convincing demonstration of the power of the Internet and advanced statistical analyses in QE research, the authors give insufficient consideration to various aspects of design and presentation for the paper to meet their own stringent requirements for high-quality research. The fact that these weaknesses were not corrected in the course of the – rather extended – editorial process suggests that quality assurance in the academic publishing process in translation studies is not as systematic and reliable as it could and should be.

Apart from constructive criticism sought from fellow researchers before submission, the peer review system for scholarly manuscripts is mostly anonymous, and its content and effect remain hidden to the research community at large. That a critique of published papers should be offered here is therefore rather delicate. In the case of Ingrid Kurz, a colleague at the University of Vienna as well as in professional interpreting practice, such published scrutiny and comment might be considered awkward, were it not for her active cooperation to allow a reassessment and elaboration of her data. As regards the work of Delia Chiaro and Giuseppe Nocella, this public feedback ex post facto is offered in support of their welcome ambition to raise the methodological standards of research in this field. Understandably, these colleagues would

rather not see their published work become an object of methodological criticism. However, while we certainly owe respect and appreciation to fellow members of our scientific community, we also owe it to the next generation of researchers, in search of guidance and inspiration for their work, to refine our research models and methodological standards as much as our skills and resources will permit. This paper, and the present issue of *The Interpreters' Newsletter*, will hopefully serve to further promote quality research in our field and help the discipline of interpreting studies earn the academic recognition it deserves.

References

Bradburn N.M. and Sudman S. (1979) *Improving Interview Method and Questionnaire Design*, San Francisco, Jossey-Bass.

Bühler H. (1986) "Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters", *Multilingua* 5 (4), pp. 231-235.

Chiaro D. and Nocella G. (2004) "Interpreters' perception of linguistic and non-linguistic factors affecting quality: A survey through the World Wide Web", *Meta* 49 (2), pp. 278-293.

Collados Aís Á. (1998) *La evaluación de la calidad en interpretación simultánea: La importancia de la comunicación no verbal*, Granada, Comares.

Collados Aís Á. (2002) "Quality assessment in simultaneous interpreting: The importance of nonverbal communication", in *The Interpreting Studies Reader*. Ed. by F. Pöchhacker and M. Shlesinger, London/New York, Routledge, pp. 327-336.

Collados Aís Á., Fernández Sánchez M.M. and Gile D. (ed.) (2003a) *La evaluación de la calidad en interpretación: Investigación*, Granada, Comares.

Collados Aís Á., Fernández Sánchez M.M., Pradas Macías E.M., Sánchez Adam C. and Stévaux E. (eds) (2003b) *La evaluación de la calidad en interpretación: Docencia y profesión*, Granada, Comares.

Elsagir I.M. (2001) "Anforderungen an Dolmetschleistungen im Fernsehen aus Zuschauersicht: Eine Fallstudie", in *Dolmetschen: Theorie, Praxis, Didaktik*. Ed. by S. Kalina, S. Buhl and H. Gerzymisch-Arbogast, St. Ingbert, Röhrig Universitätsverlag, pp. 107-123.

Garzone G. and Viezzi M. (eds) (2002) *Interpreting in the 21st Century: Challenges and Opportunities*, Amsterdam/Philadelphia, John Benjamins.

Garzone G., Mead P. and Viezzi M. (eds) (2002) *Perspectives on Interpreting*, Bologna, CLUEB.

Garzone G. (2003) "Reliability of quality criteria evaluation in survey research", in *La evaluación de la calidad en interpretación: Investigación*. Ed. by Á. Collados Aís, M.M. Fernández Sánchez and D. Gile, Granada, Comares, pp. 23-30.

Gile D. (1983) "Aspects méthodologiques de l'évaluation de la qualité du travail en interprétation simultanée", *Meta* 28 (3), pp. 236-243.

Gile D. (1989) "Le flux d'information dans les réunions interlinguistiques et l'interprétation de conférence: premières observations", *Meta* 34 (4), pp. 649-660.

Gile D. (1990) "L'évaluation de la qualité de l'interprétation par les délégués: une étude de cas", *The Interpreters' Newsletter* 3, pp. 66-71.

Gile D. (1994) "Methodological aspects of interpretation and translation research", in *Bridging the Gap: Empirical Research in Simultaneous Interpretation*. Ed. by S. Lambert and B. Moser-Mercer, Amsterdam/Philadelphia, John Benjamins, pp. 39-56.

Gile D. (1999) "Use and misuse of the literature in interpreting research", *The Interpreters' Newsletter* 9, pp. 29-43.

Gile D. and Hansen G. (2004) "The editorial process through the looking glass", in *Claims, Changes and Challenges in Translation Studies*. Ed. by G. Hansen, K. Malmkjær and D. Gile, Amsterdam/Philadelphia, John Benjamins, pp. 297-306.

Gourevich A. and Mateeff S. (1989) "Study of characteristics of simultaneous interpretation by the method of paired comparisons" (in Bulgarian), *СЪПОСТАВИТЕЛНО ЕЗИКОЗНАНИЕ / Сопоставительное языкознание / Contrastive Linguistics* 14 (1), pp. 32-38.

Harkness J.A., van de Vijver F.J.R. and Mohler P.P. (eds.) (2003) *Cross-Cultural Survey Methods*, Hoboken NJ, John Wiley.

Kopczyński A. (1994) "Quality in conference interpreting: Some pragmatic problems", in *Translation Studies: An Interdiscipline*. Ed. by M. Snell-Hornby, F. Pöchhacker and K. Kaindl, Amsterdam/ Philadelphia, John Benjamins, pp. 189-198.

Kurz I. (1989) "Conference interpreting – user expectations", in *Coming of Age: Proceedings of the 30th Annual Conference of the American Translators Association*, Medford NJ: Learned Information, pp. 143-148.

Kurz I. (1993) "Conference interpretation: Expectations of different user groups", *The Interpreters' Newsletter* 5, pp. 13-21.

Kurz I. (1996) *Simultandolmetschen als Gegenstand der interdisziplinären Forschung*, Wien, WUV-Universitätsverlag.

Kurz I. (2001a) "Conference interpreting: Quality in the ears of the user", *Meta* 46 (2), pp. 394-409.

Kurz I. (2001b) "Mediendolmetschen und Videokonferenzen", in *Dolmetschen: Theorie, Praxis, Didaktik*. Ed. by S. Kalina, S. Buhl and H. Gerzymisch-Arbogast, St. Ingbert, Röhrig Universitätsverlag, pp. 89-106.

Kurz I. (2003) "Quality from the user perspective", in *La evaluación de la calidad en interpretación: Investigación*. Ed. by Á. Collados Aís, M.M. Fernández Sánchez and D. Gile, Granada, Comares, pp. 3-22.

Ločmele G. (2001) "Interpreting norms in Latvia", in *Dolmetschen: Beiträge aus Forschung, Lehre und Praxis*. Ed. by A.F. Kelletat, Frankfurt, Peter Lang, pp. 179-184.

Mack G. and Cattaruzza L. (1995) "User surveys in SI: A means of learning about quality and/or raising some reasonable doubts", in *Topics in Interpreting Research*. Ed. by J. Tommola, Turku, University of Turku, Centre for Translation and Interpreting, pp. 37-49.

Moser P. (1996) "Expectations of users of conference interpretation", *Interpreting* 1 (2), pp. 145-178.

Moser-Mercer B. (1996) "Quality in interpreting: Some methodological issues", *The Interpreters' Newsletter* 7, pp. 43-55.

Pöchhacker F. (1995) "Simultaneous interpreting: A functionalist perspective", *Hermes. Journal of Linguistics* 14, pp. 31-53.

Pöchhacker F. (2001) "Quality assessment in conference and community interpreting", *Meta* 46 (2), pp. 410-425.

Seleskovitch D. (1986) "Comment: Who should assess an interpreter's performance?", *Multilingua* 5 (4), p. 236.

Shlesinger M. *et al*. (1997) "Quality in simultaneous interpreting", in *Conference Interpreting: Current Trends in Research*. Ed. by Y. Gambier, D. Gile and C. Taylor, Amsterdam/Philadelphia, John Benjamins, pp. 123-131.

Shlesinger M. (2000) "Evaluation issues in interpreting. A bibliography", *The Translator* 6 (2), pp. 363-366.

Vuorikoski A.-R. (1993) "Simultaneous interpretation – user experience and expectations", in *Translation – the Vital Link. Proceedings. XIIIth World Congress of FIT*, vol. 1. Ed. by C. Picken, London, Institute of Translation and Interpreting, pp. 317-327.