LOOKING QUALITY RESEARCH IN THE EYE: ARE WE BEHOLDING THE SPLINTER AND IGNORING THE BEAM?

Delia Chiaro
University of Bologna at Forlì

Giuseppe Nocella
University of Bologna at Cesena

1. Introduction

The aim of Chiaro and Nocella's article (2004) which has been much quoted by Franz Pöchhacker in this issue, was not to boast their competence in slick and sophisticated statistical techniques. Neither was it to be excessively harsh on researchers who were, after all, pioneers in bringing survey techniques to Interpreting Studies (IS) in the first place. Chiaro and Nocella's aim, nonetheless, was to underscore a certain lackadaisical attitude rampant in several attempts at questionnaire based quality research (QBQR) in this field. However, as Pöchhacker casts doubts on the reliability of their study, the authors cannot do otherwise but jump to their own defence[1]. In fact, much as Pöchhacker's re-elaboration of existing data does him honour (this issue: 150-154), as we intend to demonstrate in this essay, it cannot, and indeed does not disguise the existing general lack of methodological expertise and rigour present in many attempts at QBQR in IS. Furthermore, while grateful to Pöchhacker for having pointed out a series of shortcomings in their work, Chiaro and Nocella wish to accept total responsibility for each and every weakness, rather than take refuge behind the shield of poor refereeing. Presumably all attempts at research have their strengths and weaknesses. What is important is that the latter do not outnumber the former, otherwise our incessant quest for knowledge could well go awry.

Moreover, the present authors would like to highlight the fact that they are flattered to see that their infinitesimal contribution to the field has triggered off an animated response by such an eminent scholar. In fact, Chiaro and Nocella

1 The authors are grateful to the editorial board of *The Interpreters' Newsletter* for having given them the opportunity to respond and go into print in the same issue in which Franz Pöchhacker's article appears. The editorial decision to allow two lesser known researchers to respond so openly to such a renowned scholar is evidence of transparency and a true credit to the journal. Furthermore, they would also like to express their appreciation of Pöchhacker's sense of fair play and sportsmanship for having given them prior access to his critique and consequently the opportunity to elaborate the present reply.

believe that IS could benefit from some lively, albeit constructive discussion, a common practice in other scientific discourse communities but, until now, rather lacking in this one. In other words, with the present discussion, the authors welcome the opportunity to defend their work and intend, good heartedly, not only to stick to their guns, but also (hopefully) to trigger off a wider debate.

Taking Pöchhacker's re-visiting of QBQR in this issue as a starting point, we too will follow the same path and respond to his critique while simultaneously providing our own (over)view, where relevant, of other, similar existing research**.** In addition, similarly to Pöchhacker, most of our revisiting will also be carried out in practical methodological terms, with special emphasis on research hypotheses underlying previous studies, the nature of research design and finally statistical procedures for the analysis of survey data. We will also (re)visit Pöchhacker's detailed reanalyses of the work of his colleagues Bühler and Kurz and naturally bear out his critique of our own methods, results and conclusions. The above argumentation will be arranged in two major sections, the first regarding a detailed discussion and defence of what we consider to be a series of unjust criticisms of our work brought to light by Pöchhacker. In a separate section we will unearth a number of significant flaws in the works of others that Pöchhacker appears to have overlooked after which, we will attempt to demonstrate beyond reasonable doubt that our claim that "research undertaken so far (in QBQR) is surprisingly lacking in methodological rigour" (Chiaro and Nocella 2004: 278) is anything but inaccurate. However, we feel obliged to underscore the fact that we are not taking issue with the worth of research in IS *tout court.* We are not disputing the wealth of existing descriptive and experimental work in the field. Our criticism was originally, and still is, limited to QBQR alone.

Nonetheless, before embarking on this enterprise, the present authors would like to begin by seriously challenging the suggestion that survey based research offers "a working method that can readily be adopted also by less experienced investigators" (Pöchhacker this issue: 143).

2. A working method for less experienced investigators?

It would appear that after the well known work of Bühler (1986) and Kurz (1989), survey work in IS has become trendy and *à la mode* as more and more researchers jump onto the questionnaire bandwagon (e.g. Meak 1990; Vuorikoski 1993; Mack and Cataruzza 1995; Moser 1996; etc.). Yet those who think that developing a questionnaire is simply a matter of sitting at a desk and thinking up a list of questions are mistaken. Questionnaire development is a demanding and challenging process which requires time and energy spent first and foremost in preliminary qualitative research methods. These consist of

preparatory processes such as setting up and conducting in depth interviews and/or focus groups or adopting projective techniques such as association, completion, construction or expressive techniques (Malhotra 1996) which provide essential input for setting up a survey. A glance at the extensiveness of the literature on interview techniques alone can provide us with a fair idea as to how far such pre-survey qualitative methods have developed, while closer examination reveals how complex such practices actually are (Malhotra 1996; Tull and Hawkins 1993). And even if the principal investigator wears two hats and is also an experienced practitioner, as often appears to be the case in IS, this should not exempt them from this preliminary phase. In a certain sense this stage is even more important when the investigator is a practitioner because a researcher-cum-practitioner by default may well be inclined to increase the "observer's paradox" (Labov 1972)[2] as such a researcher will be even more lacking in the psychic distance required for unbiased study.

Let us now turn to what we shall crudely define the second stage in questionnaire design. Once researchers have obtained sufficient input from an adequate number of external informants to enable them to outline a questionnaire, they will need to know exactly what, as well as how, information is to be collected from the population under examination. This may sound trite and obvious, yet poor judgment at this stage may lead to results that are not relevant to the purpose of the study, or else that are incomplete. Questions require choosing the appropriate measurement scales, formatting and careful wording, as well as proper sequencing and layout (Aaker *et al*. 1995: 291); tasks which are easier said than done. Less than careful framing of questions can lead to distorted results. The following anecdote should illustrate the point we are trying to make:

> Two priests, a Dominican and a Jesuit, are talking about whether it is a sin to smoke and pray at the same time. After failing to reach a conclusion, each goes off to consult his respective superior. The next week they meet again:
> "Well, what did your superior say?" asks the Dominican.
> "He said it was all right", the Jesuit responds.
> "That's funny", replies the Dominican, "my superior said it was a sin."
> "What did you ask him?" inquires the Jesuit. "I asked him if it was all right to smoke while praying"', says the Dominican.
> "Oh," says the Jesuit, "I asked my superior if it was all right to pray while smoking!" (Dillon *et al*. 1994)

2 We would like to point out that we are not using the term in its strictest Labovian meaning but in its broader sense to embrace all types of biases which can occur owing to the relationship between researcher and informant.

If a mistake occurs in a measurement scale, problems are bound to arise. Several drafts as well as extensive piloting are essential before arriving at a final version. Once satisfied with the instrument, aspects such as deciding upon a method of administration (e.g. face to face, telephone, mail etc.), selecting a random[3] sample, choosing *a priori* the statistical technique to test the research hypothesis, elaborating raw data and in the final stages, interpreting results are not aspects to be taken lightly. We believe that many of the shortcomings inherent to many such studies in IS have been due to this very underestimation of what designing a survey instrument actually entails. In fact, as far as we know, there are no Translation and Interpreting faculties which offer foundation courses in empirical research methods at either undergraduate or postgraduate level, so it is understandable that interpreters often lack in necessary know how. If the single investigator is unable to see beyond data collection they may well be walking up a blind alley. Thus investigations involving researchers with different types of expertise and the adoption of an interdisciplinary attitude to IS research can only be fruitful, as long as the single researchers do not work independently and are involved in every single stage of the study. In other words, a statistician brought in *a posteriori* is unhelpful. A statistician (or better, a researcher trained in methodology and statistical analysis) at this point will indeed be capable of elaborating existing data, but his or her cooperation would have been more productive at the stage of research design. What we are trying to say is that researchers should already have in mind the kind of statistical tests they want to carry out on resulting data in order to test the initial research objective before carrying out the survey. "Here's my data see what you can do with it" is out of order in serious empirical research.

Last but not least, one of Pöchhacker's many objections to our work is 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 1997)" (this issue: 162). With all due respect to the two studies which Pöchhacker suggests should have been included in our discussion and which apparently slipped the mind of the journal's referees, we would like to state that we preferred to refer the reader to authors specialized in qualitative and quantitative research methods (i.e. Aaker *et al*. 1995; Hair *et al*. 1995 and Schiffman *et al*. 1981) rather than scholars of interpreting. This was not to belittle the two renowned scholars in question but simply because surely it is IS which is drawing from well established methodologies of the Social Sciences rather than vice-versa. Now what we were suggesting from the start was that IS should look more closely at the rules of qualitative and quantitative research methods which were born and bred outside this discipline. Interesting as both

3 "A random sample allows a known probability that each elementary unit will be chosen." (Lapin 1990: 104)

articles may be, they appear to remain, however, within the somewhat self-referential boundaries of IS.

Before beginning our discussion proper, we would like to raise one more small issue. Over and over again we read that investigating quality in interpreting is not an easy task due to the huge number of variables involved not only in the process itself, but also in conditions which regard operators, users and even the contractors of the service (e.g. Shlesinger 1997; Garzone 2003 etc.). The general idea which comes across to the reader is that dealing with the enormous heterogeneity of circumstances in and around interpreting verges on the insurmountable. This may well be true and we certainly do not wish to claim that quality research in IS is unproblematic. But is not apparent insuperability typical of scientific enquiry? Was Watson and Crick's model easy to identify? And what of the excogitation of a formula that shows that distance and time are not absolute? And discovering penicillin? The list of seemingly intractable problems is endless. But is it not this very complexity that is what makes research fascinating and irresistible?

> A researcher is a detective or a spy who is out to discover or uncover something that is in some way, unnoticed, hidden, secret or problematic. Researchers, like detectives, find that their sources sometimes lie, sometimes offer conflicting stories, and sometimes behave in baffling ways. That is why research is so exciting … (Berger 1991: 7).

2.1. A harsh critique or calling a spade a spade?

Pöchhacker accuses Chiaro and Nocella of offering a "rather harsh critique" of the work of his Viennese colleagues Bühler (1986) and Kurz (1989). We hereby express regret for our lack of tact and for having couched our criticism harshly. Our exact words were: "Unfortunately, a substantial shortcoming of this particular study (Bühler) is that the mean was used as the descriptive statistic for analyzing and discussing data[4]. Percentage, mode or median would have described the data more correctly." (Chiaro and Nocella 2004: 283) and, with regard to Kurz, our claim was that "percentage would have given a better comparison" (282). Admittedly, each turn of phrase could be seen as being rather heavy handed and inconsiderate. We could have perhaps been less direct and softened matters slightly by using words to the effect of: "Let us see what would have happened if the median/percentages had been used instead?" or possibly relegating the entire issue to a couple of footnotes. But would this have really changed anything if the analyses did not have a clear direction? More

4 For a detailed discussion on the concept of measurement see (Aaker *et al*. 1995: 56 and Tull 1993: 309).

seriously however, Pöchhacker criticises the present authors for "their erroneous criticism of Bühler's analysis", well, for the sake of diplomacy, much as we have tried to fault our analysis, mathematics is not an opinion and we will show that it is not in the least "erroneous" (4.1).

And if we were "rather harsh", Pöchhacker's critique is hardly tender. The use of subtle irony (or biting wit?) in the title of the essay[5], or indeed in the heading of the section entitled "Into print?" is not exactly gentle either. A question mark can be every bit as cutting, if not more so, than a word. And here we are talking in terms of our academic credibility. Are we certain that the words spent on Kurz and Bühler deserve such a scathing attack?

Our extensive, hands on experience in questionnaire based surveys (admittedly in other fields of research) led us both (foolhardily it would appear) to try our hand at applying our expertise to IS and also to feel (erroneously it now seems) that we had something to contribute to other less experienced researchers trying their hand at such surveys. Any impression of overconfidence surfacing from our study was quite unintentional, and by the same token, we wish to assert that, despite our experience, we are perfectly aware of how very little we do know and how much there is for us still to learn. However, we do feel that Pöchhacker is actually implying that as our study was less than perfect we should not have criticised others. And in a sense he is right. In an ideal world casting stones should be restricted to those without sin. Now we dared cast stones despite being less than immaculate ourselves. But the point is that our offences were venial rather than mortal and that most of the accusations for which we have been charged are fallacious. We sincerely believe that Pöchhacker's critique is disproportionate and that the faults in our work are in no way connected with methodological mishandling and will demonstrate that, in contrast, other studies quoted by Pöchhacker contain major inadequacies.

However, having said that, it would be a true pity if all the thought and energy which have gone into both Pöchhacker's critical assessment of Chiaro and Nocella's essay and this present retaliation were to degenerate into a lengthy scuffle of *quid pro quo*.[6] Rather it would be desirable that both

---

5 The present authors would like to point out that 'Revisiting and reanalyzing the work of Bühler and Kurz and replying to the work of Chiaro and Nocella on quality research' would have been a more fitting title to the essay to which they are responding. Nevertheless, as a scholar of Humour Studies, Delia Chiaro cannot help but relish in the clever and, admittedly, successful inherent paronomasia coined by Pöchhacker for his title.

6 In line with Pöchhacker's anecdotal style it is also perhaps worth mentioning that ample correspondence via e-mail as well as a lengthy and affable telephone conversation between the two parties involved in this discussion had occurred prior

Pöchhacker's critique and the present defence should serve to shed light upon what we still believe to be a shadowy area in IS, ie research design and implementation in QBQR and thus promote ample and, above all, constructive discussion.

3. Beholding the splinters: interpreters on the Web

Having been accused of several deficiencies, we now intend to tackle each and every one throughout the course of this comeback. Although these faults appear in a somewhat jumbled order in Pöchhacker's critique, we have tried to disentangle them and present them, together with our rebuttal, in a logical order so as to facilitate both the reader and the force of our argument.

3.1. Conceptual frameworks and operational definitions

3.1.1. Key terms

The liberal use of the word 'perception' is dangerous. Rather like cigarette smokers, consumers of the term should be made aware that its use may well present several hazards. In fact, Chiaro and Nocella dared to adopt the term liberally without defining it in operational terms and, as a result, have not only been accused of "imprecise use of key words" (this issue: 162) but, perhaps more significantly, also appear to have been severely misunderstood.

In psychology and in the cognitive sciences the word 'perception' refers to the concept of acquiring, interpreting, selecting and organizing sensory information:

> The sense organs provide our brain with a steady flow of information about our environment and the brain's task is then to take this raw material and use it to help us make sense of that environment through the process of *perception*. And the brain does its job so smoothly and well that we're not even aware of what it does. (Statt 1997: 46)

Now we have been, quite appropriately, criticized for our unclear use of the term in our study. And this is one criticism which we openly acknowledge. However, our use of the term 'perception' was quite deliberate. We were in no way confounding 'perception' with interpreters' 'generic expectations' as suggested by Pöchhacker (this issue: 158), such confusion would indeed have "fallen short of the art" (Pöchhacker this issue: 158). Besides, why should we

---

to going into print. It would not be unfair to say that communication concluded in a reciprocal decision to remain united in our diversity.

confuse perception with expectations? Is the study of expectations in IS compulsory? Why cannot perception be taken as a starting point instead? While aware of the fact that there is a strong tradition of investigations into expectations in IS (e.g. Kurz 1993, Moser 1996 etc.), expectations were not what Chiaro and Nocella were investigating at all, yet Pöchhacker seems to imply that wanting to look at interpreting from a different angle is not viable. Or rather that what we were really studying were expectations. Well, let us put the records straight and underscore that we were not seeking to access respondents' awareness or judgment of performance. What we were trying to establish was interpreters' *consciousness of mental selections which they constantly make* (our emphasis). To put it another way, we were plainly asking respondents to consider and attempt to untangle a complex mental process and express their awareness in terms of how they weighted a set of essential criteria against each other in their effort to transform incoming sensory information into verbal output in a different language. If this was erroneously confused with the expectations of a final product we trust that we have now clarified our position and again are obliged to the Editors of *The Interpreters' Newsletter* for having given us the opportunity to make amends. Incidentally, is it not also the case that 'expectations' are more relevant when one is interviewing end-users, less so when the subjects are interpreters themselves? Surely, regardless of all, 'perception' seems a more appropriate term to refer to self-monitoring by an interpreter? Over and above this, our essay contains a perceptual map (290) which displays interpreters' mental image of the various criteria. Without wishing to be tautological, a perceptual map represents perception and not expectations. How can this have been construed as confusion on our behalf?

Furthermore, we are also accused of not having distinguished between research on "generic expectations […] direct assessment, or judgement […]" (Pöchhacker this issue: 158). Needless to say, this omission was not because we didn't know the difference or because we had deliberately decided to ignore the issue. Yet, operational definitions of these terms are nowhere to be found in any of the QBQR we have examined. Moser, for example, freely uses the term perception (1996: 148, 159) imprecisely when in effect what he was investigating were "judgements, needs and expectations" (145). We are criticised for using it. He gets away with it scot-free. Again, Mack and Cataruzza suddenly introduce the term with no further definition too (1995: 45) and more recently Garzone (2003: 23-24) also adopts it freely. Are we to be the first to be accused of a lack of operational definitions? Since Moser-Mercer introduced the concept of "optimum quality" (1996: 44), the issue of attempting to define the concept of quality itself any further appears to have slipped almost everyone's mind until quite recently (Kurz 2001: 395 and 2003: 17-18). Again, the concept of multi-dimensional models of quality begin to be mentioned

(Garzone 2003: 23) while the only serious attempt at modeling the multifaceted issue of quality in interpreting has been produced by Gile (2003: 110).

Now, let us turn to the term 'experiment', a word which, according to Pöchhacker, we have used improperly in reference to Kurz's survey (Chiaro and Nocella 2004: 282). For the sake of argument, let us accept that we did use the term 'experiment' inappropriately. By the same token, Pöchhacker sets off our work against the controlled laboratory studies of Gourevich and Mateeff (1989)[7]; Collados Aís (1998, 2002); and Garzone (2003) in his defence of sound QBQR. Is this because he ignores the difference between an experiment and a survey? Or does Pöchhacker wish to widen the present dispute to colleagues adopting different methods by paying them homage? We repeat, we were/are only criticizing QBQR. Collados Aís and Garzone have carried out laboratory style research with which we have no bones to pick. And yes, we are aware of the difference between an experiment and a survey; the former is:

> A controlled situation in which the experimenter systematically changes the values of one or more variables [the independent variable(s)] to measure the impact of these changes on one or more other variables [the dependent variable(s)]. (Tull 1993: G-6)

while the latter refers to the "systematic collection of information directly from respondents" (Tull 1993: 61). Furthermore,

> The important distinction between the survey and the experiment is that the survey takes the world as it comes, without trying to alter it, whereas the experiment systematically alters some aspects of the world in order to see what changes follow. (Simon 1969: 229)

Or would Pöchhacker prefer us to adopt Vuorikoski's vague definition of experimentation as something through which "… it is possible to arrive at clear causal inferences" (1993: 318)?

Moreover, lexical networks are created within texts by the writer and false or close synonymy are simply textual strategies of reiteration (for ample discussion see Halliday and Hasan 1976: 278-279 and Hoey 1983). For the purpose of textual cohesion special synonymy with words which are not 'normally' synonymous are often created – if such a thing as 'normal' or absolute synonymy exists. Of course, this not only applies to our specific use of terminology but also to the other researchers who we have quoted above as a counter-argument (see Moser, Mack and Cattaruzza and Garzone's use of the term 'perception' above). A similar argument can just as easily be constructed

7 We do not have access to this paper and are thus relying on Pöchhacker's description of the study (Pöchhacker this issue: 160).

for the criticism of our use of the term "intonation" as a synonym of "fluency of delivery" (Pöchhacker *this issue*: 159 note 6).

3.1.2. Conceptual frameworks

Although Pöchhacker has understood that "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)' (this issue: 158), a more accurate reading reveals that our distinction was, in effect, threefold and that we had we split QBQR into "analyses of the product" (Approach number 1), field work on end-users (Approach number 2) and field work on interpreters (Approach number 3) (280)[8]. Furthermore, we are also accused of not having considered Vuorikoski's (1993) "fourfold distinction" created "specifically for the purpose of research on interpreting quality" (Pöchhacker *this issue*: 158). Pöchhacker is quite right, we do indeed ignore this "fourfold distinction". However, the reason we have done so is simply because we were unable to locate this distinction. The only mention of anything remotely "fourfold" in Vuorikoski's essay are the multi research methods she sets out to discuss. Therefore, the comparison Pöchhacker makes between our study and Vuorikoski's is quite vain. Vuorikoski's "fourfold distinction" regards the application of diverse research methods simultaneously. Our threefold distinction regards the ways in which quality research had been carried out so far in IS. In fact we state that

> …attempts at empirical research carried out so far on quality interpreting reflect these three perspectives [supplier of service, client and service itself] and have thus been based on a) analyses of product; b) field work based upon…end user perception and c) … interpreter perception…of interpretations in general. (Chiaro and Nocella 2004: 280).

We know full well of the existence of multimodal research methods but to the best of our knowledge, no attempts have as yet been made to adopt them in QBQR.

We are next charged with not providing sufficient rationale for choosing to examine interpreters rather than end-users. The pros and cons of one or the other have been argued at length in the field (Kurz 1993, Moser 1996 etc.) and we were (mistakenly it seems) convinced that we had argued our case adequately by explaining that interpreting is a service which is used by clients who presumably require assistance in understanding a language with which they are

8 Having argued about the meaning of perception (3.1.1.) it seems clear that Approaches 2 and 3 are diverse.

not familiar. This lack of knowledge of the source language renders end user quality judgement of the service of interpreting difficult as clients would be unable to judge a basic characteristic such as fidelity to the original (Chiaro and Nocella 2004: 281-282). Judging the quality of an interpretation is quite different from that of judging a regular marketable good. (We suggest that those convinced by our argument skip the rest of this section and move on to 3.2). A housewife asked to judge the quality of a pot of jam, for example, has a range of tangible and highly perceptible characteristics upon which to base her evaluation. The colour of the jam, how much it costs, it's shelf life, nutritional information on the label, packaging and, last but not least, it's flavour. In fact, we state that "Interpreting is a service and according to Economics a service is an intangible and non-transferable economic good and thus quite distinct from a physical commodity. Therefore the special nature of interpreting makes its evaluation difficult for people who consume the service but know very little about it" (281). Of course, a conference delegate can judge a variety of features connected to an interpreter's voice quality, he or she can judge clarity and coherence of speech as well as their command of the language. But a genuine delegate is likely to be hard put to be able to judge the fidelity of an interpreted speech with the original. Thus our choice of respondents naturally fell on interpreters, and with two seminal studies to rely on, using Bühler as a springboard seemed a natural choice.

3.2. The survey instrument

3.2.1. Design

Pöchhacker's first incursion regarding our survey instrument concerns the fact that we did not discuss the reasons for not adopting Bühler's criteria *tout court*. In fact, we adapted 7 criteria and we included a new one, namely "absence of stress" which twenty years ago may not have been an issue for Bühler's interpreters. And here Pöchhacker has a point so we shall immediately make amends. The input of the experts who helped us construct our instrument together with our own common sense led us to accept that Bühler's criteria "pleasant appearance" and "poise" could perhaps be cut as they possibly do not contribute to the quality of an interpretation. As for Bühler's inclusion of "positive feedback from delegates", this was considered to be a criterion which is not part of the interpreter's self-perception and therefore jars with the truly linguistic and extra-linguistic criteria which we had decided to examine. Again the concept of "reliability" was excluded for similar reasons. Our sample of interpreters were asked *verbatim* to "rank (a list of) factors contributing to the quality of interpreting". The concept of reliability was felt to be in a

hyperonymous relationship with the other factors. If an interpretation is of good quality it follows that it can be considered reliable precisely because it is made up of a positive relationship between the criteria listed. Finally, we changed Bühler's "completeness of interpretation" to "completeness of information" and "thorough preparation of conference documents" to "preparation of conference documents" upon the advice of our informants.

We also accept Pöchhacker's criticism of our somewhat cavalier description of how we constructed our instrument basing it on "several interviews" and "endless brainstorming sessions" with interpreters (283). With neither of us being a practitioner we had to look outwards and seek professionals and academics for help in devising our instrument. Few of the well known studies in QBQR appear to have bothered with any preliminary research or if they did, they certainly do not mention it in their work. Moser (1996) is the only scholar to describe a preparatory phase of his survey but we are sure that there can be no disagreement that he is as offhand as we are in his description of this stage.

Furthermore, we also acknowledge the fact that our questionnaire did not contain a request for information regarding respondents' specialized fields of expertise. Neither was information solicited regarding working language combinations. But why regard the lack of such information as methodological deficiencies? The exclusion of queries to elicit such data were choices which we intentionally made and not slips of the mind.[9] Firstly we had to keep the questionnaire as brief as possible as, in the days before the advent of widespread broadband connections, we did not know how much time people could spend on line, thus we opted for essentiality. More simply, interpreters may simply not want to waste their time filling out endless questions. Furthermore, the aim of the study was to provide a (reliable) springboard for further study. Thus, what we were searching for was broad-spectrum data. In other words, what we were interested in was obtaining a general idea of what the average conference interpreter perceived as being important and less important in his or her choices. In fact, our study was devised to be a starting point which might act as a spur for more particular, fine tuned studies. If, generally speaking, *n* conference interpreters reported that they perceive criteria *x* to be important when working to and from languages *a*, *b*, *c* or *d*, (to put it another way in and out from *any*

9 The exclusion of such data from our work was deliberate, however, let us imagine that we had simply skipped this variable for a number of reasons which could range from sloppiness and forgetfulness to sheer ignorance. Let us remind readers that the seminal works of Bühler and Kurz contain no socio-demographic variables at all, while one of Moser's is based on guesswork (for a discussion see section 4.5.). Also, one might wonder where one should stop when it comes to assembling socio-demographic information, surely you can always think up another variable that might be potentially relevant and that had not been taken into account!

non-specified language), it would then be interesting to see how the same test stands to trial when applied to specific language combinations. As Pöchhacker suggests "interpreting styles may differ from one sociocultural context to another"(this issue: 158) – well let's go out there and support this claim. Or else reject it. Who knows, perhaps interpreters' perception of choices they make may even prove to be universal. After all, surely scientific experimentation starts from the general to the particular rather than vice-versa?

As for challenging the language of the questionnaire's administration, the use of English was again a conscious choice. Unlike other surveys in which we have been engaged where the issue of language was indeed a concern, here we were looking for all-encompassing generalized data. Furthermore, we are quite certain that we were not erring in an excessive credence in the linguistic colonialism of the English language by assuming that the hypothetical average interpreter would be likely to have a working knowledge of English.

And in response to one of Pöchhacker's most critical charges, to wit, the fact that the majority of our respondents claimed that they did not work into their native tongue, again, "baffling" as this may sound, this is how the sample responded and, like it or not, the information needs to be taken at face value. Should we have excluded these findings just because he is not happy about them? Or should we have manipulated our data and claimed the contrary? We would also like to take issue with Pöchhacker's charge that the question which led us to the above claim was due to a "poorly worded questionnaire item" (Pöchhacker this issue: 159 note 7). In fact, Pöchhacker is basing this claim on an early draft of the instrument which we had sent him, and not to the final pluri-piloted version in which the wording had been improved and which we were unable to send him. However, if Pöchhacker is unhappy with our results and their subsequent interpretation, we suggest he rerun the test on a different sample.

3.2.2. Distribution

Next, we are accused of having given "an all too sparse description of their sampling procedure" (this issue: 159). Way back in 2000 when we conducted the survey, Web based questionnaires were indeed a novelty and today, the way we sampled at the time makes us both smile at our naïve techniques. What we did, which would be highly irregular today (as well as being almost impossible with the number of fire-walls and anti-spamming programs which have been widely installed in computers), was to spam an invitation to visit the site containing our questionnaire to a number of mailing lists of conference interpreters world wide. These lists were collected by networking and included a list of EU interpreters, and national associations across the world. At this point

Pöchhacker could easily argue that our sample is unreliable because it was restricted to Internet users and that we only invited about 1000 interpreters to participate. True, there surely are more than a 1000 interpreters in the world and of course we are aware that we did not contact every single one of them. But the point is that we were sampling and not contacting the entire population of interpreters. We are well aware that the 1000 interpreters we contacted had not been selected according to the table of random numbers. In other words, we cannot be sure that every interpreter with an e-mail address received our invitation to participate and that others did not receive the information twice. Nevertheless, we would like to call attention to our good faith by highlighting that neither of us are practitioners and between the pair of us we only knew about a score of interpreters at the time. This means that we were unable to use personal networking to create our sample, so at least Pöchhacker should give us our due and allow us to go down in IS history as being the first QBQR researchers who did not depend on a self selected, albeit a convenient, sample.

3.3. The mathematics behind the scores

One comment of Pöchhacker's which the authors (partially) agree with is the lack of accessibility of the statistical analysis. Or rather, for an IS readership the statistics may well be inaccessible whereas in fields such as psychology, economics and marketing research there would no need to explain the mathematics behind well-known techniques unless data is being modelled introducing innovative elements. The "sum of the scores"[10] (Chiaro and Nocella 2004: 288) is a descriptive statistic, so if there is a need to explain it we should clarify every descriptive statistic from mean to mode and from standard deviation to range and so on.

Finally, Pöchhacker introduces "A finer point, which deserves comment only in the context of aspirations to maximum methodological rigor" and criticizes our "use of unequal scales for the visualization of comparable percentages, as in the autors' Figure 2 (Chiaro and Nocella 2004: 287)" (this issue: 163) We really do not understand this comment. How can the scales be unequal if all the data summarized in Figure 2, labelled "Distribution of the degree of importance given to each linguistic criterion",[11] was obtained from the same rank scale. Instead of presenting our data in one crowded graph which may have been confusing, we simply split the data into three different line graphs to allow

10 The sum or total of the values, across all the cases with non-missing values.

11 A typo which Pöchhacker did not spot is the plural form CRITERI**A** which appears instead of singular CRITER**ION** above figure 2.

readers to follow the distribution of the 9 criteria under investigation more easily.

4. Ignoring the beams

What follows is a brief overview of the QBQR quoted by Pöchhacker in this issue. We wish, however, to begin by reiterating that most of the contributions Pöchhacker mentions and sets off against our own work present a series of gross methodological deficiencies. Secondly, we would like to declare that we are somewhat uncomfortable with having to draw attention to these studies, but having had our own work publicly scrutinized for what are patently much lesser faults, we cannot but support our claims that "research undertaken so far is surprisingly lacking in methodological rigour" (Chiaro and Nocella 2004: 278) by pointing to these examples. In fact, if we had originally spoken of "uncertain methodological principles" (279), a sense of delicacy had led us to go no further and remain somewhat vague. Now, while we are aware that our lack of humility in criticizing others has led to the disparagement of our own work, what still remains a mystery is why Pöchhacker should consider the splinters in our eyes and yet demonstrably overlook the beams in those of others. Thus the necessity to safeguard our own faces internationally now leads us to bring these beams to light. This will be done following a chronological order and restricting the review to the field of surveys pertaining to quality alone.

Interestingly, most of the surveys which Pöchhacker plays off against Chiaro and Nocella's Web survey reveal a strikingly similar series of faults which principally regard three areas, namely the sampling frames, the measurement scales adopted and the choice of statistical test. We will briefly tackle all three with reference not only to the work of Bühler and Kurz, but also to that of Vuorikoski, Mack and Cataruzza, Moser and finally Pöchhacker.

4.1. Bühler

Pöchhacker appears to be puzzled by the fact that the present authors did not take issue with regard to the rather small sample of 41 interpreters who returned questionnaires in Bühler's well-known study. We really see no cause for bewilderment simply because this study as most of the others regarding QBQR is simply descriptive in nature i.e. there is no use of any technique of inferential statistics. As a result, in absence of any use of probability theory, there is no need to argue about sample size[12], Bühler was not inferring from the sample to

12 Most text books on general statistics and methods in social research tackle the issue of sample size (i.e. budget constraint, sampling error, interval estimation, etc.).

the population but simply commenting percentages on the criteria investigated. On the other hand, what Pöchhacker should really be asking is why Bühler decided to use a sample size of 41 as this cannot be deduced from her article. Why indeed 41? Budget constraints? Time constraints? Rule of thumb? Or was sample size determined according to statistical theory considering factors such as reliability, confidence, tolerable error and precision?

However, what we would like to highlight once more in Bühler's explorative work, the importance of which is still relevant in IS today, regards the way in which these criteria were assessed. In order to test the importance of these criteria Bühler adopted the following measurement scale:

Highly important, Important, Less important, Irrelevant

The fact that most of her respondents only chose the two highest points of the scale: "highly important" and "important" was what led us to question the validity of the instrument. This is also confirmed in Pöchhacker's graphic effort (bar chart this issue: 145, 148) to reanalyse Bühler's data. Instead of showing what he claims to be a "clear cut differentiation" (this issue: 145) it shows a clearly skewed sampling distribution with a tail on the right for almost all the criteria investigated. So in the light of this observation we asked ourselves, was the measurement scale adopted the most appropriate? Had the questionnaire been properly piloted? Furthermore, why use a scale which is unbalanced towards the importance of linguistic and extra-linguistic criteria with no mid-point of neutrality and no escape route for those who did not know what to answer?[13] How can we verify whether most of these items are really so important or highly important to interpreters?

In other words, we asked ourselves whether interpreters should evaluate each item independently or whether they should play off the items against each other. Therefore, the question of how to measure the importance of these criteria led us to consider sets of non-comparative scales (e.g. continuous rating scales, itemized rating scales such as Likert scales, semantic differential scales and staple scales) and comparative scales (e.g. paired comparisons, graded paired comparisons, constant sum scales and rank order scales) in order to decide whether to modify the scale used or to choose a new measurement scale.[14] With the intention of testing whether interpreters could discriminate in terms of importance, a rank order scale seemed to be the most appropriate because interpreters could compare these criteria in one fell swoop according to their

13 An example of such a scale could be: 'Highly important', 'Important', 'Neither important nor unimportant', 'Unimportant', 'Irrelevant', 'I don't know'.

14 For a detailed discussion on measurement scales see Aaker *et al*. (1995), Maholtra (1996), Tull and Hawkins (1993).

level of importance. What is more, as the cognitive exertion involved in choosing from 16 factors would have been extremely high, we split the criteria under investigation into two separate sets: linguistic criteria and extra-linguistic criteria. Hopefully, this should have somewhat eased respondents' efforts at selection. If there are any shortcomings in the chosen scale, they could be linked to the fact that the process of selection was controlled by an algorithm in *Java script* which did not allow interpreters to give the same level of importance to two or more factors. However, having noted in the initial piloting stages of the project that nobody took issue with this characteristic it remained unchanged throughout. At the end of the sampling a total of three interpreters complained about this restriction in choice.

For the sake of argument regarding our criticism of the mathematics employed by Bühler and challenged by Pöchhacker, we must remind readers the objects were measured on an ordinal scale which was also unbalanced, thus

> Because we don't know the amount of difference between objects, the permissible arithmetic operations are limited to statistics such as the median or mode (but not the mean). Our emphasis. (Aaker *et al*. 1995: 257)

Finally, we note in passing that, Bühler's survey contains no information about the socio-economic characteristics and professional experience of the interpreters who took part in the survey.

4.2. Kurz

In order to defend the work of Kurz, Pöchhacker compares her sample with Moser's (1996) arguing in favour of the greater validity of the former sample of 124 end users which had been collected at only three conferences while Moser had to gather data at 84 different conferences in different parts of the world in order to collect a final sample of 201 end users. Clearly, being based on 84 extremely diverse conferences, Moser's sample could[15] surely have been more representative than Kurz's sample. But this is not the point. Let us examine the precise date in which data was gathered at Kurz's conference on general medicine. According to the original publication of this study, Kurz gathered her data in 1988 (Kurz 1993) while according to the reprinted version it was apparently collected in 1989 (Pöchhacker 2001). After the ruthless critique of the editorial process of the journal *Meta* with regard to Chiaro and Nocella, how could such a significant detail have escaped Pöchhacker's notice? Are we to

15 We are adopting a tentative conditional form because when we discuss the study by Moser we will illustrate why his sample is equally unrepresentative.

assume that the date was erroneous in the 1993 article? If so, why did not Pöchhacker add a footnote to clarify the point? Was 1989 the correct date of the study or was it a typo? Or is this to remain a mystery? However, as we referred to Kurz's original study (1993) we wondered how much time had lapsed between her three data sets. It is worth remembering that Kurz compares the data gathered from Bühler's 47 interpreters in 1986 with data from her own three samples collected between 1988 and 1989. Could it be that there were almost two years between Kurz's three sub-samples and almost four years between Bühler's study and the Council of Europe meeting? Now the point is whether sample size is so important with such a large time gap in sequential sampling? Sequential sampling is a technique adopted when taking decisions (usually in business and marketing) which depend upon laws of probability. If sampling had been deliberately sequential in nature we would need to know whether Bühler's interpreters and Kurz's end users inhabited an immutable world or a dynamic one. Surely four years must have brought a minimum of technological and scientific progress to the world of interpreting. Now if progress has zero impact on the world then a comparison between samples collected at different points in time may be plausible. However, as occurs in most human activity, technological and scientific progress travel at breakneck speed, thus a comparison is almost bound to present problems. Why? Progress (better working conditions in booths, use of PCs, more sophisticated technology, more competent interpreters, more fastidious end users etc.) could, on the one hand, have an impact on the average performance of interpreters and, on the other, on end users' power of assessment and thus the comparison of expectations of the different groups will be problematic, unless, of course, this is accounted for methodologically.

As for testing, Kurz appears to have adopted Bühler's measuring scales even though this is not clearly mentioned apart from a vague reference to evaluating "[…] the quality of interpretation on a four point scale" (15). It is clear that the issues we raised regarding Bühler's measurement scales apply here too. Furthermore, Kurz claims that she wishes to test the hypothesis that "different groups of end users have different expectations and needs" (15) and yet presents a set of descriptive data which remain untested. Furthermore, the same information in Table 1 (16) is repeated in Figs. 1, 2 and 3 (17) in the form of bar charts. Therefore, are we, like Pöchhacker to assume that peers were not consulted and/or that refereeing was slack?

Finally, Kurz does not attempt to compare the different groups (124 users and 47 interpreters). Instead of grouping her end users together as a single set and comparing them to the 47 professionals, she seems to lose her thread and goes on only to compare the three sub-sets to each other. Furthermore, the CACL group (6 experts) from Bühler's study are merged into the 47 interpreters

yet remain unmentioned. Again there is no mention of socio-demographic characteristics of her samples. Would this not have influenced opinions?

4.3. Vuorikoski

Vuorikoski makes an attempt to import multi-method research to quality research in IS. Unfortunately, her efforts at innovation fall short as she brings neither methodological innovation nor any empirical contribution to the field. Although she mentions a variety of methods available, her own survey does not reflect the spirit of multi-method research which she so strongly advocates. In fact, it is quite unclear where exactly the "eclectic" (1993: 318) dimension in her study is. The author, in fact, claims that

> … the small size of typical fieldwork research was compensated for with survey techniques. By covering five different seminars, each having about 100 participants, there would be more ground for generalizing the results. The size of the seminars was closer to that of fieldwork, and consequently no statistical sampling method was necessary: the seminars were considered to be theoretically relevant populations as such, and large enough for statistical analysis when treated as one population. (Vuorikoski 1993: 318)

This is, of course, a clearly contradictory statement regarding sample size. Is Vuorikoski saying that both samples are large, or is she saying that they both are small, or is one large and the other small? It would appear that the author swings back and forth from population to sample making sweeping statements yet with no mention of theory when she should have quoted some law of probability theory in support of her argument.

As for the survey itself, Vuorikoski declares that respondents were asked to give a phone number so as to allow for follow up phone interviews. Here too we find a contradiction as the author claims that "Telephone interviewing was selected as an alternative to the more traditional face-to-face interview" (323). Now rather than an alternative which allows the comparison of two independent samples and would have thus given force to a multi-method approach, what we appear to have here is a paired sub-sample of the same respondents being interviewed before and after the conference. In terms of number of people interviewed telephonically and what they were asked, these elements remain unknown. Hardly multi-method. Finally, Vuorikoski entitles a chapter "The eleven statements in the questionnaire" (321). Overlooking the fact that some of these "statements" turn out to actually be "questions", in questionnaires statements are usually measured on 5 or 7 point Likert scales. And here Vuorikoski is finally innovative as she adopts a two point scale which includes

the two options "agree" and "disagree" with no mid-point. However, to do her justice, she did include an "I don't know" escape route.

Over and above all these problems, the study is purely descriptive, research hypotheses are vague and no attempt is made to test them or even to really argue in favour of the much quoted multi-research methods she so fervently upholds. In the light of this discussion, is Pöchhacker still certain that we should have quoted this study? And again, we are forced to ask ourselves, were the referees caught out sleeping on the job just for us?

4.4. Mack and Cattaruzza

With regard to the descriptive survey carried out by Mack and Cattaruzza (1995) we do not wish to discuss sampling, measurement scales and statistical tests as we have with the others. The reason for this is to be found in the conclusions of their work in which they claim that

> Since this survey was conducted and elaborated using non-professional statistical means, no attempt was made to generalize its results nor achieve full comparability with previous studies, as this would require more sophisticated methods (47).

The awareness and unassuming nature of the two researchers admission of their lack in methodological know how makes criticism of their shortcomings absolutely unnecessary. Surprisingly too, of all the studies in QBQR, it is the study with fewest methodological weaknesses.

4.5. Moser

Once more, taking Pöchhacker's comments as a starting point, we would like to specify that rather than defend the size of Kurz's sample in terms of having "by no means" being dwarfed by that of Moser (this issue: 146), we should instead ask ourselves how the giant (i.e. Moser's sample 1996) was produced. In fact, even if Moser's sample size is considerable, an impressive 201 respondents was pretty remarkable for the field of interpreting at that time, the way in which participants were interviewed at 84 different meetings clearly shows that the sample was self selected. How do we know this? First and foremost because an average of 2.4 interviews took place at each conference and of these 1.2 involved speakers as opposed to delegates. Surely speakers and end users cannot be considered as the same thing unless speakers are considered a particular

segment of end users?[16] Furthermore, out of scores of participants how were the respondents who were not speakers actually selected? Did each of the participants have an equal chance of being chosen? What we are saying is that sampling is not simply a question of size but, also importantly about how a sample is selected.

Furthermore, also in the case of Moser, the nature of the study is just descriptive and explorative. In fact, there is no application of any statistical test even if in this case the author in the central concerns of the survey states that he wants to investigate the "hypothesis … that different user groups would have different expectations of interpretation" (1996: 146). Yet how this hypothesis is going to be tested and the relative results are left to the imagination of the readers; similarly where the author sees the "positive correlation" (157) between increasing conference-going experience and the fact that users want the interpretation to match the original also remains statistically unexplained. Naturally, considering the self selected sample one could argue about the parameters of distribution involved in the statistical test chosen, but since this was not the case we can only leave the answer to the reader's imagination.

Now, in order to understand the way in which Moser measured his items we have been forced to draw on both the work published by AIIC (1995) and a different version of the same study published in *Interpreting* (1996). We have had to look at both articles because interestingly, the same study published in the journal omits a great deal of background information present in the initial study. Measurement scales, for example, are not stated in the 1996 article. So, let us pick on a couple of examples to examine how Moser in his survey measured some items relating to end users' needs and expectations. regarding "completeness of rendition", "clarity of expression" and "correct terminology" for which the following scale was used:

Very important	Fairly Important	Fairly unimportant	Unimportant	Ambiguous	I don't know

Firstly, we can see that the scale seems to be lacking in a central point (ie "neither important nor unimportant"), unless of course the reader is supposed to assume that the item "ambiguous" is filling the gap. Now, if our first assumption is true, that is that the scale is lacking in a central point, then it follows that the scale is incomplete. If, on the other hand, "ambiguous" is a deliberate choice in the scale for the central point, then there is obviously a problem of wording in communicating the points of the scale. Wording, as pointed out previously (see 2), is a very important aspect of setting up measurement scales. Moreover,

16 Of course we are well aware that at some conferences speakers are there also as delegates.

Moser only comments on criteria which respondents judged as being "very important" (162 and 163) according to conference type. In other words, he is giving readers a somewhat incomplete picture of user expectation by concealing other information.

Now let us take another example from the AIIC publication (1995: C1, C2). After having asked respondents to indicate the importance of three criteria ("completeness of rendition", "clarity of expression" and "correct terminology") according to the scale mentioned above, they were then asked, under the label "other", to identify criteria not specified in the preceding questions.[17] Now the issue of the word "other" followed by a list begs the following question: why after such extensive preliminary research were none of these criteria identified and inserted in the questionnaire to be measured on the same scale reported above? Could it be that the observer's paradox has reared its head? In other words, what we are trying to say is that if the target population were end users it should have clearly been tested on a group of end users before final administration. This is not clear from the final version of the questionnaire in German neither in terms of numbers, nor in terms of the people interviewed (interpreters again or end users?). However, over and above this, the list of assorted criteria detected by informants leads the reader, in any case, up a blind alley as their degree of importance is not measured with the same scale as the first three criteria. In other words, these criteria are incomparable with the first three. An example of a more gross error is finding the item "correct terminology" under "other" when the respondent had already given an opinion on that criterion in the previous question.

For the sake of argument, let us consider one more example of a scale adopted in the same study (1995: C4, C5):

Very irritating	Fairly irritating	Not really irritating	Unimportant	Don't know	Ambiguous

In this case the researcher is trying to measure end users' degree of irritation of particular behaviour of interpreters. The question which arises here is why include "importance" in a scale which is trying to measure irritation? And again we find the baffling item "ambiguous" occurring once more. Last but not least, we have yet another unbalanced scale, with no central point. How come?

17 The choice of criteria listed by Moser under a stark "Other" are "synchronicity, emotional congruence, pleasant voice, correct terminology, focus on essentials, technical knowledge, faithfulness to the original, faithfulness to the meaning of the original (*sic*.), clarity of expression, neutrality towards the speaker, lively, animated delivery, translation of jokes and asides, native sounding accent, stop when a mistake is made, other." (1995: C2)

So how did the editorial process work here?[18] Do high standards of refereeing apply only to the work and Chiaro and Nocella published in *Meta* or are these criteria universal? We will however return to this issue in a dedicated paragraph (5).

Curiously, at a certain point Moser's study introduces the concept of attitudes towards providers of the service in order "to shed additional light on the study" (p. 159) and sets out to ask end users what they consider to be particularly interesting and particularly difficult about the interpreting profession.[19] Well, how attitudes are used in this context is not clear despite the fact that already at the time in which this survey was conducted attitude models such as the Fishbein model (theory of reasoned action) and the Ajzen model (theory of planned behaviour) to try to measure attitudes were already well established (Ajzen 1991; Solomon 2004). However, the important conclusion at which Moser arrives using the term attitude (in the broadest possible sense?) is that it is linked to "the broad educational and cultural background for which they (interpreters) are envied" (160).

Finally, also in this survey, information about respondents' education and professions which could have played an important role in testing the unproven hypothesis were not solicited. And unusually, instead of directly asking respondents (end users?) how old they were, ages were supplied by interviewers (interpreters) who "were asked to estimate the age of persons interviewed" (1996: 151) thus bringing to mind vets who estimate the age of horses by examining their teeth. Why were interviewers not asked to guess other socio-demographic data too? Presumably because apart from evaluating a person's sex, the rest is quite difficult. Again, it would appear that *Meta* is not the only journal to suffer from lax refereeing.

4.6. Pöchhacker

According to Mark Twain there are three kinds of lies: "lies, damned lies and statistics." And it is undeniably a truism that statisticians can manoeuvre numbers at their will. And this is precisely what Pöchhacker attempts to do by offering readers fresh analyses of his colleagues' data.

---

18 We have chosen just a couple of the numerous faults in Moser's work simply for the sake of argument.

19 Moser's question "What do you find particularly interesting about the profession, and what particularly difficult?" (1996: 159) is actually two questions in one and thus would require rewording.

4.6.1. Kurz's calculations according to Pöchhacker

Pöchhacker occupies more or less a third of his essay re-elaborating the data of Ingrid Kurz. His re-elaboration of the percentages in Bühler and Kurz's and comments on figures 2a, 2b, 3a e 3b have already been amply discussed in paragraphs 4.1 and 4.2 above. However, before we offer our own interpretation of this recent amplification of the data, we wish to make a short premise. Without a shadow of a doubt the energy which Pöchhacker has exerted into the re-elaboration of his colleagues' data is, to say the least, admirable. Nevertheless, his efforts recall the period between the two World Wars when the first social scientists were lacking in a compass (a research hypothesis) to guide them through their studies. In fact, they would start off by gathering data willy-nilly and subsequently observing what emerged. The only guide they had at the time was their personal capacity to elaborate data with the means of sound techniques (Guidicini 1996). In other words, our predecessors possessed neither computers nor sophisticated, click-of-the-mouse software. Fortunately, their somewhat careless manner of conducting research was soon to be replaced by one which started off by forming a hypothesis, gathering and elaborating specific data and subsequently either confirming or rejecting the initial assumptions aided with new technology which was to come to their rescue.

In his discussion of the use of statistical tests on Kurz's data Pöchhacker seems to have lost his compass as he appears to oscillate between testing group differences on continuous variables, testing relationships among discrete variables and testing both together. Does he see the nature of the variable? Can the same variable be both discrete and continuous at the same time?[20] Pöchhacker is analyzing categorical data from an unbalanced ordinal scale.[21] First he applies chi squared testing to check whether there are any relationships between two or more categorical variables and then on the same data he explores differences amongst the groups treating the variables as though they were continuous.

Moreover, during the application of the chi squared test, Pöchhacker quite rightly observes that more cells have the expected frequency which is smaller than five and begins by admitting that " the expected frequency in the chi-square test is smaller than five, which renders any interpretation of the test invalid" (*this issue*) and that the sample is not big enough, and then that the data should have had a more balanced distribution. But, in order to resolve the problem he collapses the 2 categories of the scale adopted by Bühler and Kurz (i.e. "less

20 Obviously it is possible to transform a continuous variable into a discrete one but the reverse would be more complex.

21 For a detailed discussion on statistical tests on categorical data see Agresti 2002.

important" and "irrelevant") into a single category labelled "not important". Is it really plausible to collapse "irrelevant" and "less important" into a single category? In other words, surely "very important", "important" and "less important" have more in common semantically with each other than "less important" and "irrelevant"? Would it not have been more reasonable to collapse all the categories which measured importance so that the new dichotomic variable would have been acceptable? Pöchhacker would then at least have had all the dominions of "importance" in one category and "irrelevance" in the other. One last point, Pöchhacker includes an explanation of an elementary concept such as cross tabulation yet does not elucidate chi square distribution, the significance of probability *p* or acceptance or rejection of the null hypothesis. Surely if anything needed clarification it would be the latter concepts and not the former.

At the beginning of the sub paragraph "Other non parametric tests (*this issue*?)" Pöchhacker states that "Aside from the chi-square test, there are other nonparametric tests for identifying significant relationships among different sets of rank-ordered data" and he uses the Kruskal-Wallis H and the Mann-Witney U tests. Here again, it is unclear whether he is looking for relationships or differences. And in applying these tests, if respondents had originally been asked to express their opinion on a single unbalanced item scale for the various criteria, how did he obtain his rank ordered data? Is he still using mean scores? If so, once more is the variable continuous or discrete? Moreover he does not explain that the use of asymptotic significance for the exact test may not be a good measure of significance if the variables are poorly distributed, which seems the case with these data sets. It would appear that Pöchhacker is simply looking for anything significant in the dataset without clarifying how he is manipulating his data. Moreover to explain differences among the 3 groups on the criteria which resulted as being significant, he runs the Mann Witney U test taking into consideration the three combinations of the three independent samples. Well, in this case we would like to point out that by following this path Pöchhacker is falling into the so called "familywise" or "experimentwise" error rate (Field 2000)[22]. In other words, is he aware that the probability of making at least one type I error is increasing from 0.05 to 0.143? If the Bühler group was included, and we do not understand why in his re-analysis this group has been omitted, this probability would have jumped to 0.185.

But why, we wonder, twenty years on does Pöchhacker want to show significance at all costs? Undeniably, more than one statistical test can be

22 We apologize but space does not allow us to explain testing hypotheses. However, the topic is treated in almost all text books on general statistics. As regards the figure assuming the independence of the samples we apply the independent law of probability: in the case of 3 groups $(0.95)3 \approx 0.857$ and $(1\text{-}0.857) \approx 0.143$.

carried out on both Bühler and Kurz's data, on our data, or on anyone else's data come to that, but is there any point? If a researcher originally sets out to either accept or reject a particular hypothesis, why demonstrate that they could have done something different?

Our discussion of the beams seems to have highlighted the difficulty both of choosing and applying a statistical test in the studies examined so far. We would now like to dedicate a few words to this issue as succinctly as possible for obvious reasons of space.

4.6.2. The choice of a statistical test: an overview

In the light of our previous discussion, it now seems evident that the choice of a statistical test cannot be dictated simply by the significance or lack of significance deducted from *p* values produced by any "PC-based statistics software […] accessible enough to be used, with proper guidance, also by the 'semi-skilled' analyst" (Pöchhacker this issue: 154). On the contrary, the choice of a statistical test should be made in function of three general conditions at the same time: the research question, the nature of the data and the plan or design of the research.

The research question should veer in a clear direction. From the start, the researcher should know whether the aim is to find differences or correlations between or among the variables which are object of the study. Once the researcher has decided which direction the study will take, inferential statistics offer numerous tests which test the hypothesis underlying the research question: univariate, bivariate and multivariate techniques[23]. Naturally, the choice of a test also depends upon the nature of the data. Does the data consist of discrete or continuous variables? And what are the forces of the measurement scales upon which the variables were measured: nominal, ordinal, interval or ratio-scaled? As a result, descriptive statistics and statistical tests must be also gauged in function of the metric or non metric nature of the variable and of the force of the measurement scale. The research design used to generate the data also affects the choice of a statistical test. So, decisions regarding the independence of the samples, number of groups, number of variables and variable control must be taken *a priori*. Moreover, when a technique is used, the assumptions regarding that technique have to be satisfied before applying the technique. So if one wishes to apply ANOVA for example, to explore differences among groups, the assumptions of independent random samples, normality and equal variances of

23 There is really no room to explain even briefly the statistical techniques included in these three groups, however explanations can be found in basic and more higher level text books in general and advanced statistics.

all populations must be assessed. So, the analyst should explore the dataset in order to understand whether these conditions have been satisfied. But if the necessary assumptions are violated what should be done? Well, it depends on which assumptions have been violated. If normality or equal variances are involved then transformation to symmetry[24] could be applied (Ryan 1985) to approach a Gaussian distribution[25]. However, if one or more samples differ in a significant way from the population of interest then it could be very difficult to draw any conclusions from the dataset. An explorative analysis becomes more stringent and of paramount importance when more variables are involved, i.e. when multivariate techniques are used (Hair *et al*. 1995). In this case, the relationship among variables, the analysis of missing data, the detection of outliers through graphical output (e.g. stem and leaf diagrams or box plots) or statistics (e.g. Mahalanobis D^2) and the verifications of the assumptions such as normality, homoscedasticity[26] and linearity are something which cannot be solved just by a few clicks of a mouse.

Finally, while we are perfectly aware that the advent of user-friendly statistical software has facilitated the application of statistical tests and the mathematical calculations behind them, we firmly believe that these very packages require a sound knowledge of the field of statistics. We cannot possibly agree that

> […] analyzing empirical data […] is not a question of mathematical skills but, essentially, a matter of meaningful interpretation, of making sense of the relationships indicated by the data (Pöchhacker this issue: 155).

This does not do the field of research methodology justice. Without understanding what he or she is doing in terms of statistics, the researcher not only runs the risk of misapplication of tests but also of a poor interpretation of results. Not only, but whether a statistical significance test "does not explain anything but merely points reliably to what needs to be explained" (Pöchhacker this issue: 155) is highly debatable too.

---

24 Of all transformations made on data in practice, the three most popular are the square root (moderate), the logarithm (strong) and negative reciprocal (very strong).

25 When assumptions are violated one could also think of applying non parametric tests which are less stringent in matching assumptions; in the case of ANOVA one could use the Kruskal-Wallis test.

26 Homoscedasticity is an assumption related primarily to dependence relationships between variables. It refers to the assumption that dependent variable(s) exhibit equal levels of variance and homogeneity of variance across the range of predictor variable(s). (Hair *et al*. 1995: 67).

5. To print or not to print

It is clear from his use of ironic punctuation ("Into print": 161) that Pöchhacker did not consider our study worthy of publication. For reasons of delicacy we would rather avoid the hearsay and the chitchat surrounding our article's journey from Italy to Canada.

Moreover, in defence of *Meta* and the referees of our article, we wish once more to take total responsibility of all shortcomings which are totally our own and not imputable to the journal[27]. From our point of view, our article was sent to Canada in mid-2001, refereed about six months later, corrected, accepted and finally published in the summer of 2004. However, perhaps it would have been more correct if Pöchhacker's article had appeared in *Meta* rather than *The Interpreters' Newsletter* seeing that it is the former journal which is under attack. Unfortunately, when Pöchhacker thoughtfully sent us his paper, he had already sent it to *The Newsletter* thus we too had no option but to respond in the same journal. But thinking more precisely on the matter, perhaps a Special Issue on quality is exactly where this discussion should take place. However

> Aristotle argues that there are three kinds of rhetorical proof; that is three ways in which a speaker can persuade an audience of his position – *ethos*, *pathos* and *logos*. *Ethos* is ethical proof, the convincing character of the speaker (…) *Pathos* is an appeal to the emotions of the audience (…) *Logos* is logical proof, or argument, the kind of proof that appeals to reason (Root 1987: 16-18).

And we have attempted to defend ourselves from Pöchhacker's accusations taking the philosopher's advice by blatantly appealing above all to the reader's understanding of our competence in research methodology, as well as to his or her emotions and reason. We hope to have clarified above all our use of the term perception (3.1.1.); that our sampling frame was accurate (3.2.2.) and demonstrated that our criticism of Bühler was all but erroneous (4.1.). In doing so, we have been forced, albeit unwillingly, to be harsh on others.

However, what emerges from the present discussion is that over the years, macroscopic faults in the refereeing process in this field have been common across the board. Admittedly our argument was perhaps circular. This is one criticism which Pöchhacker makes that we feel we must accept, but we still wish to claim that our study was methodologically sound in design,

---

27 Above all, we humbly apologize for our misspelling of Kopczyński, for depriving Susan Bassnet's surname of an 's' and the "infelicities in the bibliography." However, IS must be in an embarrassingly poor state if, in an attempt to punish two authors who have (apparently) stepped out of line, Pöchhacker feels he must include typos in a critique of methodology.

administration and data elaboration. Indeed, we could have extended our article with more detailed information, rationale and discussion. But could not the same be said of the other works mentioned? The methodological faults we have found in others are substantial and incomparable to the display of nitpicking displayed by our plaintiff. Circular? Maybe. But what have been the conclusions that other researchers adopting QBQR have reached so far? Have they been so insightful? But again, whatever their findings and conclusions at least they had a go, unlike our complainant who simply sits and looks and then comments from high with a critical eye. Indeed, one wonders whether such an eye is really critical. For us the word misguided seems more fitting. What is more, as two researchers looking in from the periphery, the argumentation put forward by Pöchhacker makes the field of QBQR in IS appear somewhat self referential to say the least.

And we certainly could have done without the author glibly offering the quasi-total demolition of our work "… in support of their welcome ambition to raise the methodological standards of research in this field" (Pöchhacker this issue: 163). Are we supposed to feel honoured by this insult to our intelligence? Yet still not satisfied, Pöchhacker turns the dagger in the wound by stating that "Understandably, these colleagues would rather not see their published work become an object of methodological criticism." Well, Pöchhacker certainly notches up full marks in insight and sensitivity on that score, yet at the end of the day, what we find most objectionable, is not so much the criticism itself, but the rather patronizing tone in which it is couched. Criticize us by all means, but superciliousness we can do without. As far as survey research is concerened, Pöchhacker is still in an early stage of infancy.

6. Beams of light?

Last, but certainly not least, despite our criticisms of the studies mentioned, we would like to emphasize our respect for all those researchers who have tried their hand at field work. Our harshness towards these people has been dictated by the need to demonstrate that if we "Unwittingly … provide material for a case study of methodological rigor in quality research" (Pöchhacker this issue: 163) others provide just as much, if not more so and presumably, until this moment, just as inadvertently. Nevertheless, however faulty and elementary their instruments, only people who have actually rolled up their sleeves and personally tried to obtain answers from complete strangers will have experienced the blood, sweat and tears behind each single return. Which naturally makes the whys and wherefores of less than rigorous sampling understandable. Of course, it is simpler to announce a questionnaire to a roomful of delegates than to stop them one by one thus chancing a higher risk of refusal. But then we must be aware that the sample is no longer random.

Similarly, asking friends and colleagues to take part in studies is equally open to criticism. It was precisely this type of nonchalant way of surveying that made us want to contribute with our five (Euro)cents.

We would like nonetheless to express our discomfort with the present response. This time we are fully aware of our heavy handedness regarding the work of others. But if originally we had been vague, here we have had to argue our case as clearly as possible and hopefully readers and, above all, the researchers involved will understand that we had no option. From our point of view, Pöchhacker's critique of our work was short sighted and in places erroneous.

Re-reading the QBQR in IS in preparation for this reply, the lack of knowledge in the tools and methods of the social sciences is self evident. Measurement scales, sampling frames, statistics and statistical tests are constantly defective and studies are strikingly self-referential. If the field of IS aims to "earn the academic recognition it deserves" (Pöchhacker this issue: 164) at excellence in research design and applications, then it should be open to the views and criticism of outsiders who are free of the institutional shackles of unassailable individuals within that group. We hope to have shown that no one is exempt from developing clay feet. Having said that, if the field wishes to remain self-referential, then so be it.

However, we wish to conclude on a positive note. The present essay is a display of academic argument in which we have criticized rather old studies. Perhaps the time has come to let sleeping dogs lie. And perhaps it is also time that translation and interpreting faculties began introducing courses in research design and statistics so that students wishing to embark upon the fascinating field of research are well equipped to do so with a working knowledge of how to go about it. Perhaps now is the moment to learn from disciplines which have been working in social research for decades. In fact, in more recent publications, it is highly uplifting to find that IS scholars are beginning to look tentatively outside IS towards the social sciences for insights into quality research (e.g. Kurz 2003). Surely, if there is something to be learnt from the successful marketing of a good or a service it is the collaboration of experts with diverse expertise who together construct high quality products and facilities. If it is truly excellence which interpreters desire for themselves and their clients in the real world, then the path of interdisciplinary research of practitioners and objective outsiders is surely a good one. If, on the other hand, the issue of quality is to be restricted to the philosophical argument and mutual back patting of a few, then let it remain trapped and stagnant in its ivory towers.

References

Aaker D.A. *et al*. (1995) *Marketing Research*, New York, John Wiley.

Agresti A. (2002) *Categorical Data Analysis*, New Jersey, John Wiley.

Ajzen I. (1991) *Attitudes, Personality and Behaviour*, Milton Keynes, Open University Press.

Berger A. (1991) *Media Research Techniques*, Newbury Park, Sage.

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

Berenson M. L. and Levine D.M. (1986) *Basic Business Statistics: Concept and Applications*, New Jersey, Prentice-Hall.

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.

Dillon W.R, Madden T.J and Firtle N.H. (1994) *Marketing Research in a Marketing Environment*, Boston, MA, Irwin.

Field A. (2000) *Discovering Statistics using SPSS for windows*, London, Sage.

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. (2003) "Quality assessment in conference interpreting: methodological issues" in *La evaluación de la calidad en interpretación: Investigación*. Ed. by Á. Collados Aís, M.M. Fernández Sánchez and D. Gile, Granada, Comares, pp 109-124.

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.

Guidicini P. (1996): "La ricerca sociologica nella sua evoluzione storica" in: *Nuovo manuale della ricerca sociologica* Ed. by P. Guidicini, Milano, Franco Angeli, pp 11-37.

Hair J.F., Anderson R.E., Tatham R.L., Black W.C. (1995) *Multivariata data analysis with readings*, Fourth edition, New Jersey, Prentice-Hall.

Halliday M.A.K. and Hasan R. (1976) *Cohesion in English*, London, Longman.

Hoey M. (1983) *On the surface of Discourse*, London, Allen and Unwin.

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

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.

Labov W. (1972) *Language in the Inner City*, Philadelphia, Pennsylvania University Press.

Lapin L.L. (1990) *Statistics for Modern Business Decisions*, Fifth Edition, San Diego, Harcourt Brace Jovanovich.

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.

Malhotra N.K. (1996) *Marketing Research: an applied orientation*, New Jersey, Prentice Hall.

Meak L. (1990) "Interprétation simultanée et congrès medical : attentes et commentaries", *The Interpreters' Newsletter* 3, pp. 8-13.

Moser P. (1995) *Survey on Expectations of users of conference interpretation*, Final report, Vienna, AIIC.

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

Root R.L. Jr. (1987) *The rhetorics of popular culture: Advertising, advocacy, and entertainment*, New York, Greenwood.

Ryan B.F., Joyner B.L., Ryan T.A. (1985) *Minitab handbook*, Boston, PWS Kent.

Schiffman S.S. *et al*. (1981) *Introduction to multidimensional scaling: Theory, methods, and applications*, Orlando, FL, Academic Press.

Shlesinger M. (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. 123-132.

Simon J.L. (1969) *Basic Research Methods in Social Science: the art of empirical investigation*, New York, Random House.

Solomon M.R. (2004) *Consumer behaviour*, New Jersey, Prentice Hall.

Statt D.A. (1997) *Understanding the Consumer: a psychological approach*, Basingstoke, Palgrave.

Tull D.S and Hawkins D.I. (1993) *Marketing Research: measurement and method*, New York, Macmillan.

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.